Rough Set Extensions for Feature Selection

Neil S. Mac Parthaláin

Department of Computer Science
Aberystwyth University

September 2009

PhD Thesis

Abstract

Rough set theory (RST) was proposed as a mathematical tool to deal with the analysis of imprecise, uncertain or incomplete information or knowledge. It is of fundamental importance to artificial intelligence particularly in the areas of knowledge discovery, machine learning, decision support systems, and inductive reasoning. At the heart of RST is the idea of only employing the information contained within the data, thus unlike many other methods, probability distribution information or assignments are not required. RST relies on the concept of indiscernibility to group equivalent elements and generate knowledge granules. These granules are then used to build a structure to approximate a given concept. This framework has unsurprisingly proven successful for the application to the task of feature selection.

Feature selection (FS) is a term given to the problem of selecting input attributes which are most predictive of a given outcome. Unlike other dimensionality reduction methods, feature selection algorithms preserve the original semantics of the features following reduction. This has been applied to tasks which involve datasets that contain huge numbers of features (in the order of tens of thousands), which would be impossible to process otherwise. Recent examples of such problems include text processing and web content classification. FS techniques have also been applied to small and medium-sized datasets in order to discover the most information-rich features. The application of rough sets for FS has resulted in many efficient algorithms. However, due to the granularity of the approximations generated by the rough set approach there is often a resulting level of uncertainty. This uncertainty in information is usually ignored for FS (by nature of the very fact that it is 'uncertain').

In this thesis, a number of methods are proposed which attempt to use the uncertain information to improve the performance of rough sets and extensions thereof for the task of FS. These approaches are applied to two application domain problems where the reduction of features is of high importance; mammographic image analysis and complex systems monitoring. The utility of the approaches is demonstrated and compared empirically with several other dimensionality reduction techniques. In several experimental evaluation sections, the approaches are shown to equal or improve classification accuracy when compared to results obtained from unreduced data.

Based on the new fuzzy-rough approaches, further developments and techniques are also presented in this thesis. The first of these is the application of a nearest neighbour classifier for the classification of real-valued data. This technique is evaluated within the mammographic imaging application. Also, a novel unsupervised feature selection approach is proposed which reduces features by eliminating those which are considered redundant. Both the fuzzy-rough classifier mentioned above, and UFRFS are employed and evaluated for the complex systems monitoring application.

Datganiad

Nid yw sylwedd y gwaith hwn wedi cael ei dderbyn o'r blaen ar gyfer unrhyw radd, ac nid yw'n cael ei gyflwyno ar yr unpryd wrth ymgeisio am unrhyw radd.

Llofnod .....(ymgeisydd)

Dyddiad .....

Gosodiad 1

Canlyniad fy ymchwiliadau i yw'r gwaith hwn, oni nodir yn wahanol. Lle mae **gwasanethau cywiro*** wedi ei defnyddio, nodir maint a natur y cywiriad yn eglur mewn troednodyn.
Cynabyddir ffynonellau eraill mewn troednodiadau sy'n rhoi cyfeiriadau eglur. Atodir llyfryddiaeth

Llofnod .....(ymgeisydd)

Dyddiad .....

Gosodiad 2

Yr wyf drwy hyn yn rhoi caniatâd i'm gwaith, os yw yn cael ei dderbyn, fod ar gael i'w lun-gopïo ac ar gyfer benthyciadau rhwng-llyfrgellol ac i'r teitl a'r crynodeb fod ar gael i gyrff allanol.

Llofnod .....(ymgeisydd)

Dyddiad .....

Acknowledgements

I would like to express my gratitude and thanks to my supervisor, Prof. Qiang Shen, for his encouragement, effort, and guidance throughout the project. I must also sincerely thank Dr. Richard Jensen, my second supervisor, particularly for his advice and involvement in this research.

Additionally, I am grateful to:

- Dr. Reyer Zwiggelaar for his help with the mammographic image analysis application and his donation of the data used therein.
- Angharad Elias, a fuoedd yna i fi wastad - diolch.
- Mo Mham, a cabhraigh liom i slí amháin, no slí eile - Go raibh míle maith agat.
- Laura and Alex for their help with the proof-reading.
- Ac i unrhyw un arall bo fi wedi anghofio sôn amdan / Also to anyone else whom I have neglected to mention.

Contents

1 Introduction **10**
- 1.1 Feature Selection . . . 12
- 1.2 Limitations of Current Approaches . . . 14
- 1.3 Distance Measure Assisted Rough Set Feature Selection and Extensions . . . 15
  - 1.3.1 Additional Developments . . . 16
  - 1.3.2 Applications . . . 17
- 1.4 Thesis Structure . . . 18

2 Background **20**
- 2.1 Dimensionality Reduction . . . 20
  - 2.1.1 Transformation-based Approaches . . . 21
  - 2.1.2 Selection-based Approaches . . . 24
- 2.2 Feature Selection Models . . . 30
- 2.3 Rough Sets for Approximate Modelling . . . 35
  - 2.3.1 Basic Concepts and Theoretical Background . . . 37
  - 2.3.2 Rough Set Dependency and Other Measures . . . 39
  - 2.3.3 Minimal Reducts and Reduct Discovery . . . 43
- 2.4 Rough Set Extensions . . . 45
  - 2.4.1 Tolerance Rough Sets . . . 45
  - 2.4.2 Variable Precision Rough Sets . . . 47
  - 2.4.3 Dominance-based Rough Sets . . . 48
  - 2.4.4 Vaguely Quantified Rough Sets . . . 50
  - 2.4.5 Other Rough Set Extensions . . . 52
- 2.5 Combining Rough Sets with Other Techniques . . . 53
  - 2.5.1 Rough Set Hybridisation . . . 53
  - 2.5.2 Fuzzy-Rough Sets . . . 54
- 2.6 Applications . . . 57
  - 2.6.1 Classification . . . 57
  - 2.6.2 Clustering . . . 61
  - 2.6.3 Feature Selection . . . 64
- 2.7 Summary . . . 68

3 Exploring the Boundary Region: Rough Sets **70**
- 3.1 Distance Metric and Mean Positive Region . . . 71
- 3.2 Distance Measure-based Selection Algorithm . . . 73
- 3.3 Computational Complexity . . . 74
- 3.4 A Worked Example . . . 75

3.5 Experimental Evaluation - Comparison with Other Approaches . . . 80
3.5.1 Classifier Learners . . . 81
3.5.2 Comparison with RSAR . . . 81
3.5.3 Comparison with PCA . . . 83
3.5.4 Comparison with FRFS . . . 84
3.5.5 Comparison with TRSM . . . 86
3.5.6 Hausdorff Metric Implementation . . . 87
3.6 Summary . . . 88

4 Exploring the Boundary Region: Tolerance Rough Sets and Fuzzy-Rough Sets 90
4.1 Tolerance-based Feature Selection . . . 91
4.2 Distance Metric-Assisted Tolerance Rough Set Feature Selection . . . 92
4.2.1 Distance Metric-based ToleranceQuickReduct . . . 93
4.2.2 Worked Example . . . 95
4.3 Experimentation . . . 97
4.3.1 Experimental Setup . . . 98
4.3.2 Comparison of Classification Accuracy . . . 98
4.3.3 Subset Sizes . . . 100
4.3.4 Comparison with Randomly Selected Subsets . . . 102
4.3.5 Hausdorff Metric Implementation . . . 102
4.3.6 Comparison with Existing FS Methods . . . 103
4.4 The Fuzzy-Rough Set Boundary Region . . . 109
4.4.1 Fuzzy-Rough Feature Selection (FRFS) . . . 110
4.4.2 Fuzzy Boundary Region-based FS . . . 111
4.4.3 Integration of Fuzzy Entropy . . . 115
4.4.4 Fuzzy-Rough Reduction with Fuzzy Entropy . . . 115
4.4.5 Fuzzy-Rough Reduction with Fuzzy Gain Ratio . . . 118
4.5 Experimentation . . . 118
4.5.1 Experimental Setup . . . 118
4.5.2 Experimental Results . . . 120
4.6 Summary . . . 121

5 Association Learning 122
5.1 Classifier Learning . . . 122
5.1.1 Nearest Neighbours Classification . . . 123
5.1.2 Fuzzy Nearest Neighbours Classification . . . 124
5.2 Fuzzy-Rough Set Theory . . . 125
5.2.1 Fuzzy-Rough Ownership kNN . . . 127
5.2.2 Fuzzy-Rough Nearest Neighbours . . . 128
5.2.3 Worked Example . . . 129
5.3 Unsupervised Feature Selection . . . 131
5.3.1 Unsupervised Fuzzy-Rough Feature Selection . . . 132
5.3.2 Fuzzy Dependency . . . 132
5.3.3 Algorithm . . . 133
5.3.4 Worked Example . . . 133
5.4 Experimentation . . . 135
5.4.1 Experimental Setup . . . 135

5.4.2 Experimental Results . . . 136
5.5 Summary . . . 136

6 Application to Mammographic Image Analysis 138
6.1 System Overview . . . 140
6.1.1 Dimensionality Reduction . . . 142
6.1.2 Nearest Neighbour Classification . . . 143
6.1.3 Fuzzy-Rough Ownership kNN . . . 144
6.1.4 Vaguely Quantified Rough Sets (VQRS) . . . 144
6.2 Fuzzy-Rough Nearest Neighbours . . . 145
6.2.1 FRNN Algorithm . . . 146
6.3 Experimentation . . . 146
6.3.1 Mammographic Risk Analysis Data . . . 147
6.3.2 Experimental Setup . . . 148
6.3.3 Unreduced Data . . . 148
6.3.4 Reduced Data . . . 149
6.3.5 Investigation I: Unreduced Data for All Classifiers . . . 151
6.3.6 Investigation II: Comparison with Current State-of-the-art . 153
6.4 Summary . . . 157

7 Application to Plant Monitoring 161
7.1 Rule Induction . . . 161
7.2 Unsupervised Fuzzy-Rough Feature Selection . . . 163
7.2.1 Fuzzy Dependency . . . 163
7.2.2 Classification and Rule Induction . . . 164
7.3 A Realistic Application . . . 165
7.4 Experimental Results . . . 166
7.4.1 Classification of Unreduced and Reduced Features . . . 166
7.4.2 Rule Induction . . . 167
7.4.3 Comparison with Other FS Techniques . . . 168
7.5 Summary . . . 170

8 Conclusion 171
8.1 Distance Measure Assisted Rough Set Feature Selection . . . 172
8.2 Unsupervised Fuzzy-Rough Feature Selection . . . 173
8.3 Mammographic Image Analysis . . . 173
8.4 Industrial Plant Monitoring . . . 174
8.5 Future Work . . . 174
8.5.1 Short-term Developments . . . 174
8.5.2 Long-term Developments . . . 176

A Publications Arising from the Work in this Thesis 179
A.1 Published or Accepted for Publication . . . 179
A.2 Currently under Review . . . 180

B Glossary of Terms 181

List of Figures

1.1 Process of manual knowledge discovery . . . 11
1.2 A few of the many real-world applications for feature selection . . . 13

2.1 Dimensionality reduction approaches . . . 22
2.2 2-dimension normal point cloud with corresponding principal components . . . 23
2.3 Feature selection . . . 26
2.4 The feature selection problem . . . 26
2.5 Subset generation techniques . . . 28
2.6 FS models . . . 31
2.7 Filter FS . . . 31
2.8 Generalised filter algorithm . . . 32
2.9 Wrapper FS . . . 32
2.10 Generalised wrapper algorithm . . . 33
2.11 Hybrid FS . . . 33
2.12 Generalised hybrid algorithm . . . 34
2.13 Basic rough set concepts . . . 38
2.14 The QuickReduct algorithm . . . 43
2.15 A taxonomy of rough set extensions . . . 45
2.16 Tolerance rough set model . . . 46
2.17 Document clustering using tolerance rough sets - stage 1 . . . 63
2.18 Document clustering using tolerance rough sets - stage 2 . . . 64
2.19 Feature selection for gene expression data . . . 67

3.1 Objects of the lower approximation and boundary region . . . 72
3.2 The rough-set distance metric-based algorithm . . . 74
3.3 RSAR and DMRSAR runtimes for 50-350 attributes . . . 76
3.4 RSAR and DMRSAR runtimes for 500-8000 objects . . . 77

4.1 The DMTQuickReduct algorithm . . . 94
4.2 DM-TRS vs. randomly selected subsets . . . 102
4.3 Performance: JRip . . . 120
4.4 Performance: PART . . . 121

5.1 Nearest neighbours (NN) classification . . . 124
5.2 The fuzzy kNN algorithm . . . 125
5.3 The fuzzy-rough ownership NN algorithm . . . 128
5.4 The FRNN algorithm . . . 129
5.5 The UFRQuickReduct algorithm . . . 133

6.1 Example mammograms where breast tissue density increases from L-R Corresponding to BIRADS class I(far left) to class IV (far right) 139
6.2 Mammographic density classification 141
6.3 Unified fuzzy-rough framework for mammographic data analysis . . 142
6.4 Experimental setup 147
6.5 Classification accuracy: Unreduced MIAS data for the four methods and different values of k – 10CV 149
6.6 Classification accuracy: Unreduced DDSM data for the four methods and different values of k – 10CV 150
6.7 Classification accuracy: DMTRS reduced MIAS data for the four methods and different values of k – 10-fold CV 151
6.8 Classification accuracy: DMTRS reduced MIAS data for the four methods and different values of k – 10-fold CV 152
6.9 Classification accuracy: DMTRS reduced DDSM data for the four methods and different values of k – 10-fold CV 153
6.10 Classification accuracy: DMTRS reduced DDSM data for the four methods and different values of k – 10-fold CV 154
6.11 Classification accuracy: FRFS reduced MIAS data for the four methods and different values of k – 10-fold CV 155
6.12 Classification accuracy: FRFS reduced DDSM data for the four methods and different values of k – 10-fold CV 156
6.13 Classification accuracy: FRFS reduced MIAS data for each of the four methods and different values of k – LOOCV 157
6.14 Classification accuracy: FRFS reduced DDSM data for each of the four methods and different values of k – LOOCV 158
6.15 Classification accuracy: DMTRS (τ=0.97) reduced MIAS data for the four methods and different values of k – LOOCV 159
6.16 Classification accuracy: DMTRS (τ=0.98) reduced MIAS data for the four methods and different values of k – LOOCV 159
6.17 Classification accuracy: DMTRS (τ=0.98) reduced DDSM data for the four methods and different values of k – LOOCV 160
6.18 Classification accuracy: DMTRS (τ=0.99) reduced DDSM data for the four methods and different values of k – LOOCV 160

7.1 Modular decomposition of the implemented system 164
7.2 Water treatment plant overview 166
7.3 Example rules generated from the 3-class data: unreduced and reduced 169

List of Tables

2.1 Example dataset 39

3.1 Example dataset: crisp attributes 76
3.2 Average classification accuracy – crisp data 82
3.3 Comparison of reduct size, dependency value, & run times – crisp Data 83
3.4 Subset size and classification accuracy results for PCA 84
3.5 Classification accuracy of unreduced, DMRSAR reduced, and FRFS reduced, data using JRIP, PART, and J48 classifiers 84
3.6 Comparison of subset size, dependency value, & run times – FRFS . 85
3.7 Comparison of subset size for each tolerance threshold value 86
3.8 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.90) 87
3.9 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.95) 87
3.10 DMRSAR – Hausdorff metric implementation subset size and run-times 88

4.1 Example dataset 95
4.2 Classification accuracy using QSBA 99
4.3 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.80) 99
4.4 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.85) 99
4.5 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.90) 100
4.6 Classification accuracy using JRIP, PART, and J48 classifiers ($\tau =$ 0.95) 100
4.7 Comparison of subset size for each tolerance threshold value 101
4.8 DMRSAR – Hausdorff metric implementation ($\tau =$0.90) 103
4.9 PCA & DM-TRS – Comparison of classification accuracy 104
4.10 CFS Subset size and classification accuracy 105
4.11 Average subset size and classification accuracy for DM-TRS 106
4.12 Closest comparable subset size and classification accuracy for DM-TRS 106
4.13 Subset size and classification accuracy results for consistency-based FS 108
4.14 Subset size and classification accuracy results for ReliefF 108

4.15 Subset size and classification accuracy results for consistency-based FS . . . 109
4.16 Example dataset . . . 113
4.17 Reduct size and time taken . . . 119
4.18 Resulting classification accuracies JRip (%) . . . 119
4.19 Resulting classification accuracies PART (%) . . . 119

5.1 Example training data . . . 129
5.2 Example test data . . . 129
5.3 Example dataset . . . 134
5.4 Fuzzy similarity relations . . . 134
5.5 Subset sizes for UFRFS . . . 136
5.6 Unreduced, supervised FRFS, and UFRFS Classification accuracies (%) . . . 137

6.1 Reduct sizes for the MIAS dataset following the application of DMTRS and FRFS . . . 150
6.2 Reduct sizes for the DDSM dataset following the application of DMTRS and FRFS . . . 150
6.3 MIAS - Average classification accuracy, and standard deviation results using 10-fold CV . . . 153
6.4 DDSM - Average classification accuracy, and standard deviation results using 10-fold CV . . . 154
6.5 DDSM - Average classification accuracy, and standard deviation results using LOOCV . . . 155
6.6 MIAS - Average classification accuracy, and standard deviation results using LOOCV . . . 156

7.1 Subset sizes obtained using UFRFS . . . 167
7.2 Classification accuracy results: reduced and unreduced data . . . 167
7.3 Maximum rule arity for unreduced and UFRFS reduced data . . . 167
7.4 Number of rules generated for unreduced and UFRFS reduced data 168
7.5 Subset sizes returned by consistency-based, CFS, and C4.5 wrapper, feature selection methods . . . 169
7.6 Subset size and classification accuracy results for consistency-based FS . . . 170

Chapter 1

Introduction

"I have travelled the length and breadth of this country and talked with the best people, and I can assure you that data processing is a fad that won't last out the year." – the editor in charge of business books for Prentice Hall, 1957.

In almost every field imaginable, data is now collected and collated at a staggering pace. In fact, as the capability to process data increases, so too does the means to gather and record it. This has led to the storage and maintenance of huge amounts of data, of which a small percentage (in spite of today's advances in computing technology) will ever be used to any advantage. There is therefore, a pressing need for the development of approaches and automated tools to assist humans in extracting useful information (knowledge) from these rapidly expanding mountains of data. Such approaches and tools are the subject of the ever-growing field of knowledge discovery from data.

The idea behind the search for useful patterns in data has had many names in the past, including data mining, knowledge extraction, information discovery, information harvesting, and data pattern discovery. At a fundamental level however, knowledge discovery from data is concerned with the development of methods and techniques which 'make sense of data'. The basic issue which the knowledge discovery process is attempting to address is that of mapping low-level data, which is typically too expansive for human comprehension into other forms which may be more compact, more humanly interpretable (e.g. a descriptive model of the process that created the data), or more useful (e.g. a predictive model for estimating unseen cases). At the heart of the process is the application of specific data-mining methods for pattern recognition and extraction.

The traditional manual process of converting data into knowledge [59] as illustrated in Fig. 1.1, relies on human analysis and interpretation. For example, in market research, it is common for commercial organisations to periodically employ an expert to analyse data relating to current consumer trends and pur-

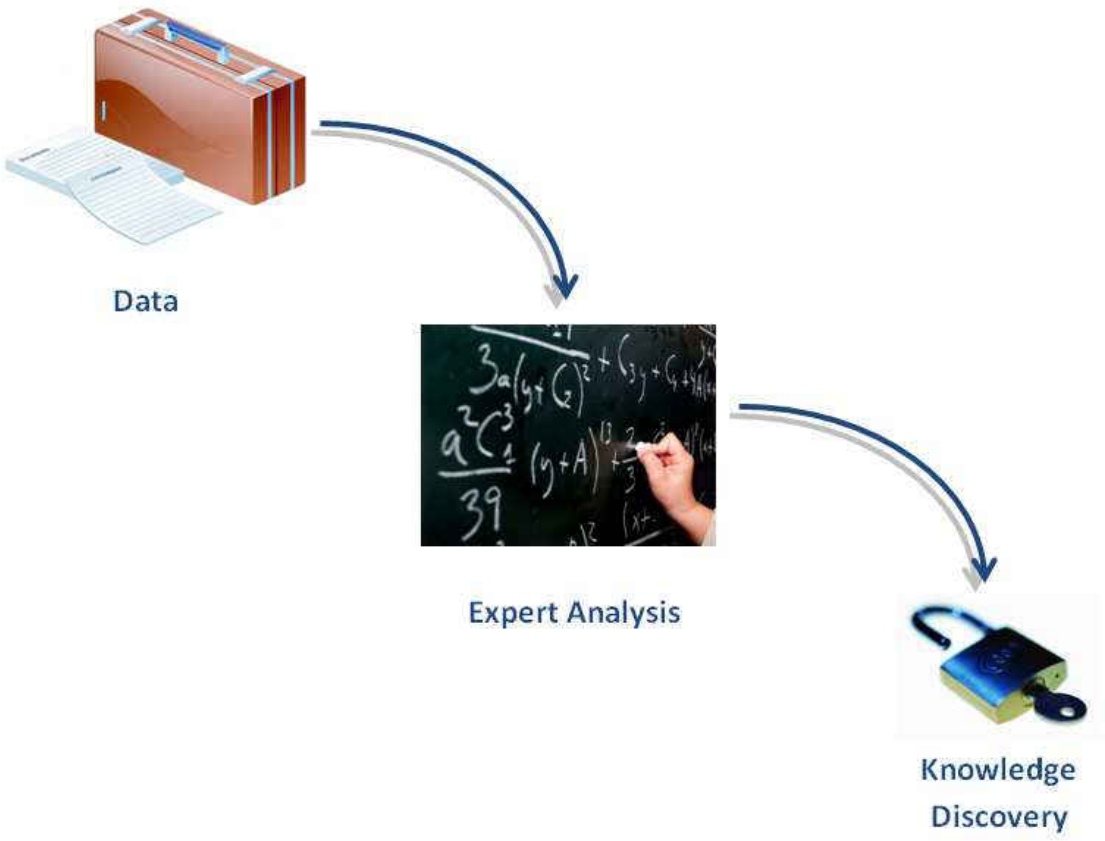

Figure 1.1: Process of manual knowledge discovery

chasing habits. The expert(s) then compile an assessment detailing the analysis and present it to the organisation. This assessment forms the basis for future marketing strategies, and decision-making with regard to target audiences, etc. In another, completely different field, criminal forensic scientists sift through huge amounts of scientific evidence, carefully analysing and cataloging objects, material fragments, fingerprints, etc. before generating plausible scenarios or events. Whether in science, crime detection, finance, machine performance, or any other field, this classical approach to data analysis relies fundamentally on the fact that at least one expert must have a detailed and intricate understanding of the data.

For many other applications, this form of manual manipulation of data is time-consuming, expensive, and of-course, highly subjective. Indeed, as the volume of data grows dramatically, this type of manual data analysis is becoming completely impractical in some domains, if not impossible in most. Data is increasing in size both in terms of the number of data objects, and the number of features or attributes which describe each data object. Datasets which contain in the order of 10^9 data objects are now commonplace [131], as are those which contain 10^3 or greater features [32]. Clearly, these types of problems are beyond the scope of the human being, and hence, such analysis requires automation.

As discussed previously, not only is there often a large number of objects in todays data, but there may also be a large number of features or attributes. Therefore the dimensionality of the problem may be high. High-dimensional datasets create problems in terms of the size of the search space, and it can be shown that the addition of extra dimensions (features) causes an exponential increase in the complexity of the problem - termed *computational explosion*. This is further exacerbated at the point of data collection, as it is often naively assumed that *more features = more knowledge*, thus increasing the likelihood of having enough information to distinguish between classes. Unfortunately, this is not the case if the size of the training dataset does not also increase remarkably with each additional feature that is included. This problem is commonly known as the *curse of dimensionality* [10]. It is a problem that often frustrates the effective application of machine learning techniques for knowledge discovery. Additionally, it increases the possibility that an induction algorithm will find spurious patterns that are invalid, due to high levels of noise. Solutions to this problem include methods which reduce the overall dimensionality of the data. Such approaches are known as dimensionality reduction techniques [118], and can be classified into one of two categories: those which transform the underlying meaning (or semantics) of the data, and those that do not - known as semantics preserving. The methods known as feature selection (FS) are those which fall into the latter category [98]. These methods select a subset of the original features using a suitable evaluation function, and are particularly useful for knowledge discovery as they preserve the human interpretability of the original data and resulting discovered knowledge.

1.1 Feature Selection

Feature selection is common in machine learning, where it may also be termed *feature subset selection*, *variable selection*, or *attribute reduction*. Fundamentally, it can be considered as the process of selecting the input attributes of a dataset that most closely define a particular outcome. FS attempts to focus selectively on relevant features, whilst simultaneously attempting to ignore the (possibly misleading) contribution of irrelevant features. From a computational complexity point of view, it is beneficial to have a minimal set of features involved in the classification phase, and as noted previously, many learning algorithms scale up rapidly with the inclusion of additional features. In addition to the improvement in classifier performance, the costs associated with collecting large amounts of (feature) measurements can also be reduced by ensuring a minimal feature set.

It should be noted that it is not possible for even the most advanced learning algorithms to compensate for poor FS techniques which select irrelevant or redun-

dant features. This highlights the importance of performing efficient and robust FS in the first instance.

Feature selection is often employed in areas where the dimensionality of the original data is such that it is impossible for humans to comprehend, but where it is imperative for the reduced data to retain the underlying meaning of the reduced features (e.g. rule induction). The diagram in Fig. 1.2 represents just a few of the many real-world applications for FS.

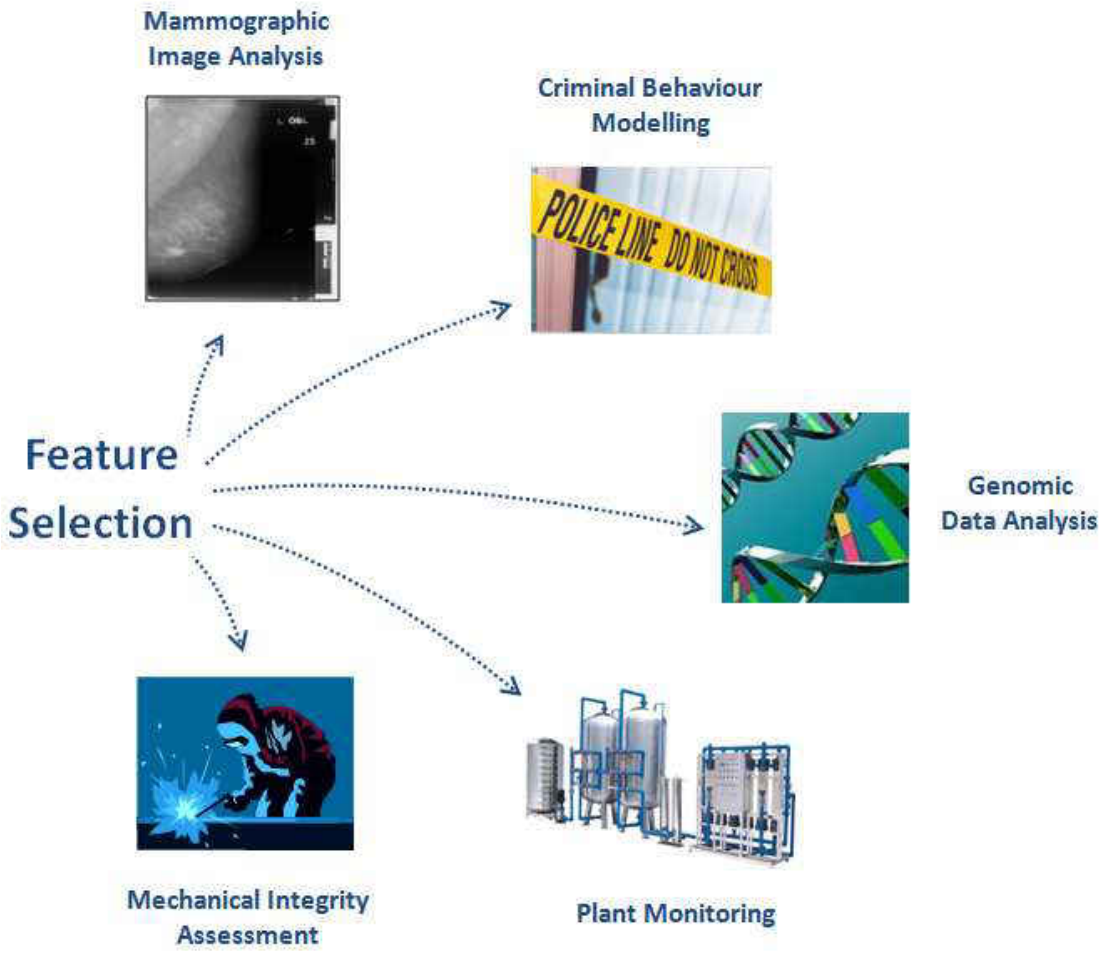

Figure 1.2: A few of the many real-world applications for feature selection

Feature selection algorithms are commonly used to improve classification performance for image analysis systems [167], [199]. The motivation for this stems from a phenomenon which is commonly observed during the training phase of classification known as 'peaking'. This occurs when the number of features is increased, the classification rate of the classifier begins to decrease after a 'peak'. In mammographic data analysis, for instance, the accuracy of radiologists in identifying tissue density lies between 60% and 84% [16]. With the application of FS algorithms, automated recognition systems can produce classification accuracies above 90% [141].

Automation of the inspection of welded metal joints has led to images from which large amounts of features must be extracted, slowing the analysis considerably. Attempts at classification of such high dimensional data also results in poor performance. In [88], the authors employ FS for a dataset which contains radiographic image data for welded pipe seams. They succeed in reducing the dimensionality of the data by over 80% whilst simultaneously increasing the classification accuracy, thus improving both the efficiency and the accuracy of defect detection.

Another area where FS has been employed is that of crime detection. In [245] the authors present methods for modelling and predicting criminal behaviour based on historical data. Most of this data is based on movements of criminals at or around the time a crime was committed, and includes such information as geographic location data. Feature selection algorithms are then used to identify important patterns hidden in the data (i.e. proximity to main roads, personal care expenditure per household, etc.), which allows more accurate modelling of the location of crime incidents and patterns.

Rough set theory (RST) [172] has proven popular for data dimensionality reduction [32], [154], [262], and has attracted much interest from researchers. This is borne out by the wide variety of application areas, such as classification [63], clustering [76], fault diagnosis [201], plant monitoring [32], etc. The popularity of rough sets for feature selection tasks in particular relates mainly to its simplicity - both conceptually and computationally. In addition to this, only the facts in the data are analysed, and no subjective thresholding parameters, expert advice, or domain knowledge is required. Employing RST as a preprocessor, a subset (or reduct) of useful original features can be selected. This type of application of RST for feature selection is usually performed as preprocessing step prior to the use of an induction or learning algorithm.

A rough set [172] is the approximation of a vague concept by a pair of precise concepts which are known as upper and lower approximations. The lower approximation is a definition of the domain objects which are known to belong to the concept of interest with full certainty, whilst the upper approximation is the set of all those objects which *definitely* or *possibly* belong to the concept of interest. The difference between the upper and lower approximation is an area known as the boundary region or region of uncertainty.

1.2 Limitations of Current Approaches

Current rough set and rough set extensions have demonstrated much success when applied in the area of FS [98]. Within the current rough set framework however,

most existing approaches only consider the certain information or the information of the lower approximation concept [70], [120], [121], [154], [203], [221], [262] described previously. These approaches have adopted this strategy, as the certainty that is embodied in the lower approximation is associated with greater importance in scientific analysis. However, although often overlooked, there is also additional information to be gathered from the boundary region or region of uncertainty. This information by nature of the fact that it is uncertain, is often ignored as it is assumed that it will not be able to offer any further advantage for the approximation task at hand. Within the current framework of RST, this may be true, or at least this information may only be able to offer some additional assistance if considered in the context of attempting to minimise it, thereby maximising the certain information. However, there is also the possibility in reality, that there are objects in the boundary region which only differ from those in the lower approximation as a result of noise.

Although no direct attempts have been made in the literature to qualify the uncertain information of the boundary region, some approaches have been proposed which consider the rough set upper approximation [47], [87]. The problem with such approaches however is that the upper approximation is examined as a whole; they do not consider the possibility that only *some* of the information in the upper approximation may be relevant. That is to say; no attempt is made to separate the upper approximation, into the lower approximation (certain information) and boundary region (uncertain information) and deal with the data in this manner.

It is important to emphasise that the uncertain information of the boundary region of rough sets is only uncertain within the context of the RST framework. In other words it is not possible to determine from the granular information structure of rough sets whether this information relates to the concept which the theory is attempting to approximate. This leads to the question - can this information be used within the RST framework to assist in improving the performance of rough sets and thus the approximation capability?

1.3 Distance Measure Assisted Rough Set Feature Selection and Extensions

The fundamental concept for the new approaches proposed in this thesis, is that of attempting to extract knowledge from the rough set boundary region. This is done by proposing a framework whereby a significance value (generated by a set of distance metric values) is assigned to the boundary region information for the

set under consideration. This information, along with the certain information of the lower approximation is then used as an approximation metric. This approach performs well but is restricted to the discrete or crisp-valued data domain, so in order to consider data from real-valued domains a discretisation step is employed. This is not adequate however and can result in information loss, as the degrees of membership of values to discretised values are not considered.

To address the shortcoming, new approaches have been proposed based on both tolerance rough sets [209] and fuzzy-rough sets [52]. These approaches allow a greater degree of flexibility when compared to the strict, rigid requirements of crisp rough sets that can deal only with full or zero membership. Tolerance rough sets allow the consideration of real-valued data by assigning a tolerance threshold and similarity relation such that 'similar' data objects are allowed to be considered equivalent, thus easing the hard equality of rough sets, and introducing an element of fuzziness to the crisp rough set model. Fuzzy-rough sets [52] encapsulate the related but distinct concepts of vagueness for fuzzy sets [254], and indiscernibility for rough sets [172], both of which occur as a result of uncertainty in knowledge. A fuzzy-rough set is defined by two fuzzy sets, fuzzy lower and upper approximations, obtained by extending the corresponding crisp rough set notions.

The main contribution of this thesis relates to the use of the information contained in the boundary regions of rough sets, tolerance rough sets and fuzzy-rough sets. Three different approaches are presented:

- An approach based on classical rough set theory, and is shown to improve the performance of the basic underlying model
- Another which is based on tolerance rough sets and can handle real-valued data.
- Also, a further development which is based on fuzzy-rough sets which attempts to exploit the information gain value of the boundary region.

1.3.1 Additional Developments

Although the main contribution of this thesis is the exploitation of the boundary region information for the task of feature selection, a number of other developments have also been proposed. These have come about due to various reasons, and include the application of a nearest neighbour (NN) classifier to mammographic data analysis, and an unsupervised feature selection algorithm, both approaches are based on fuzzy-rough sets.

The NN classifier takes advantage of the fuzzy upper and lower approximation concepts as an indicator to predict test object classes. This approach is data-

driven and does not require a k parameter value (although one can be specified if required). The approach takes advantage of the complementary nature of fuzzy sets and rough sets. Compared with the popular fuzzy nearest neighbour algorithm, and a number of other fuzzy classifiers, it performs well and is successful in classifying noisy data such as that obtained from images.

The unsupervised fuzzy-rough feature selection (UFRFS) algorithm compares the conditional features of a dataset with one another and removes those that are correlated and hence redundant. Fuzzy-rough sets are used to determine the level of dependency of a feature for elimination on subsets of selected features. UFRFS is data-driven and uses only the information contained in the data itself - no subjective thresholding parameters are required. Additionally it has the ability to handle real-valued data, and experimental evaluation demonstrates that it also selects useful feature subsets.

1.3.2 Applications

The new methods proposed in this thesis can be applied to any of the domains discussed previously where feature selection or classification have been employed. In this thesis however, two important domains of interest were chosen; mammographic risk analysis [167], and complex systems monitoring [204]. These illustrate the potential utility of the approaches detailed in this work.

Knowledge discovery from images often requires the maximisation of all of the information contained within a given image. This means that large numbers of features are often extracted initially. These features typically contain high levels of redundancy, irrelevance, and noise. However, given that it is not known *a-priori* which features are most valuable and which are not, this is a necessary step. In the work described here the tolerance rough set-based FS method is employed to identify the most valuable features such that the process of extracting large amounts of features can be avoided. The selected features are then fed back into the extraction phase ensuring that only those features need to be identified in future. Use of fewer features means that any algorithms employed in both the training and testing phases of the classifier will not only potentially be more accurate as there are fewer noisy features present, but also execute in quicker time. This helps to reduce the demands on expert radiologists' time in examining mammographic images. Most importantly however it can result in more accurate breast abnormality risk assessments. The new fuzzy-rough NN classifier is also used to classify both the unreduced data and reduced data in this work, which is more applicable to real-valued data than previous approaches.

In systems monitoring, it is important to reduce the number of features in-

volved for a number of reasons. First and foremost, is the cost associated with feature measurement, as each measurement not only requires additional sensors and monitoring equipment but also specialist data-logging of all the measurements. Also, the monitoring process can be simplified if fewer variables are involved. Finally, it is often observed that the accuracy of the monitoring system can be significantly improved by a reduction in measurement variables [204] as there is a lower level of noise due to noisy or irrelevant variables. UFRFS is applied a water treatment plant dataset [158] to demonstrate how this fuzzy-rough method can be used within the systems monitoring domain.

1.4 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2:** *Background.* This chapter presents a systematic overview of both dimensionality reduction and current rough set techniques as well as their extensions. It begins with basic concepts of dimensionality reduction with a particular focus on feature selection and its various models and some algorithms. Then, rough set theory is explored in detail with an examination of the underlying mathematical model. This is followed by a detailed description of each of the rough set extensions and hybridisations. The work in this chapter has been published in [145]

- **Chapter 3:** *Exploring the Boundary Region of Rough Sets.* In this chapter, the theoretical developments of a new feature selection method are presented. This novel approach uses the information contained in the rough set boundary region (or region of uncertainty) to improve the performance of the rough set model. The operation of the approach and its benefits are demonstrated through the use of some simple examples. To evaluate the technique, comparative investigations are carried out with the current leading techniques. This chapter and parts thereof have been published initially in [148], with a further and more in-depth version in [149] .

- **Chapter 4:** *Exploring the Boundary Regions of Tolerance Rough Sets and Fuzzy-rough Sets.* This chapter builds on the work presented in Chapter 3. It further expands on the initial ideas of exploiting the information contained in the rough set boundary region for feature selection to tolerance rough sets and fuzzy rough sets. The developments for tolerance rough sets have been published in [144], and those for fuzzy-rough sets in [142].

- **Chapter 5:** *Association Learning.* By exploiting fuzzy-rough sets, this chapter presents further development of techniques for association learning. In particular a novel classifier learning technique for application to the problem of image analysis is described. A simple example is also presented which demonstrates the approach fully. Also a new unsupervised fuzzy-rough feature selection (UFRFS) technique is presented and is applied to 8 benchmark datasets. The techniques described in this chapter are published in [138], [141], and further work is currently under review for journal publication [146].

- **Chapter 6:** *Application to Mammographic Risk Analysis.* Mammographic risk analysis from images is an important area of research as it provides an important indicator for the likelihood of a woman developing breast cancer, which is the leading cause of death of women in their 40's in the EU and US. Like many areas which deal with image data, there are large amounts of redundancy and noise in the data. With the use of FS, these extraneous features can be removed. Additionally, with the aid of an accurate classifier learner such as that described in Chapter 4, a unified approach to mammographic risk analysis is formulated which can increase the accuracy of risk analysis and thus reduce the potential for misdiagnoses. This unified technique and application is currently under review for journal publication [141].

- **Chapter 7:** *Application to Plant Monitoring.* Complex application problems, such as reliable monitoring and diagnosis of industrial plant equipment, usually present large numbers of features, many of which are redundant for the task at hand. Employing UFRFS, these correlated features can be removed. This not only makes resultant rulesets generated from such data much more concise and readable, but can reduce the equipment and monitoring cost involved in measuring redundant features. The monitoring system is applied to water treatment data, and results in similar classification accuracies than those of the full feature set. An application oriented journal article summarising the work described in this chapter has been submitted for consideration for publication [147].

- **Chapter 8:** *Conclusion.* A summary of the key findings from the research is presented, together with a discussion of topics which form the basis for future work.

Chapter 2

Background

"If variable elimination has not been sorted out after two decades of work assisted by high-speed computing, then perhaps the time has come to move on to other problems." R.L. Plackett, discussion with Miller in [137]

This chapter is concerned firstly with the basic concepts of dimensionality reduction with a particular focus on feature selection. Broadly speaking, there are two different types of dimensionality reduction techniques; transformation-based reduction, and selection-based reduction. The transformation-based methods transform the data features or attributes, whereas the selection-based methods (such as feature selection) do not. An overview of both types of approach is presented here, followed by a more in-depth discussion of feature selection. All of the common feature selection models are examined in detail with a simple pseudocode example for each one. This first initial examination of dimensionality reduction is followed by a view of rough set theory and its recent extensions, which is also presented in [145]. The preliminary concepts and theoretical foundation of rough set theory are covered in detail. Various rough set extensions (both past and recent) such as tolerance rough sets, variable precision rough sets, dominance-based rough sets, vaguely quantified rough sets, and others are also explored thoroughly. The hybridisation of rough set theory with other techniques is discussed later, with particular emphasis on fuzzy-rough set theory as this is a useful technique which takes advantage of the complementary nature of fuzzy and rough sets. A range of both theoretical and real-world example applications with regard to rough set theory, and the above mentioned extensions are demonstrated also.

2.1 Dimensionality Reduction

The inclusion of a dimensionality reduction (DR) step in a variety of problem-solving systems [24] may be proposed for a number of different reasons. For many

real-world application problems, data is processed in the form of a collection of real-valued object vectors e.g. text classification [247], bookmark categorisation [93], mining of medical data [29], mammographic image analysis [2], etc. If such data is of high dimensionality, it can be beneficial to employ a DR step. Indeed, it is often necessary where the dimensionality of the data prior to reduction may be prohibitively large. The central idea behind DR therefore is the reduction of the data to a size which is computationally tractable, without information loss. Hence, a DR step is usually included as an integral part of a data preprocessing system.

There are often cases where high-dimensional phenomena are governed by significantly fewer, simple features [57]. Here, the process of dimensionality reduction acts as a tool for modelling these phenomena, thus improving clarity. Additionally, a significant amount of redundant or misleading information is also present; this requires removal prior to any further processing. For instance, the problem of deriving classification rules from large datasets often benefits from a data reduction preprocessing step. Not only does this reduce the time required to perform induction, but the resulting rules are more comprehensible and this can potentially improve the classification accuracy [98], [130], [131].

Many DR techniques destroy the underlying meaning (the semantics) behind the features present in a dataset [48] - this is an undesirable property for many applications. This is particularly true where the understanding of the data processing method and that of the resulting processed data is as important as the accuracy of the resultant lower dimensional dataset in use. For example, in medical imaging it may be important to be able to identify particular areas of the image or data which are of greatest interest both prior to, and following reduction [18].

It is important at this point to emphasise that DR can be divided into two categories; transformation-based approaches, and selection-based approaches - see Fig. 2.1. The former, is a set of approaches which perform dimensionality reduction but in doing-so irreversibly transform the descriptive dataset features. The latter approaches however, preserve the original meaning or semantics of the data through the removal of redundant, noisy, or irrelevant features - i.e. the set of survival features is a subset of the original unreduced features.

2.1.1 Transformation-based Approaches

The decision of which DR approach to use is often governed by whether or not subsequent applications or systems will need to be able to refer to the underlying data. For instance, if an application or system requires the use of the meaning

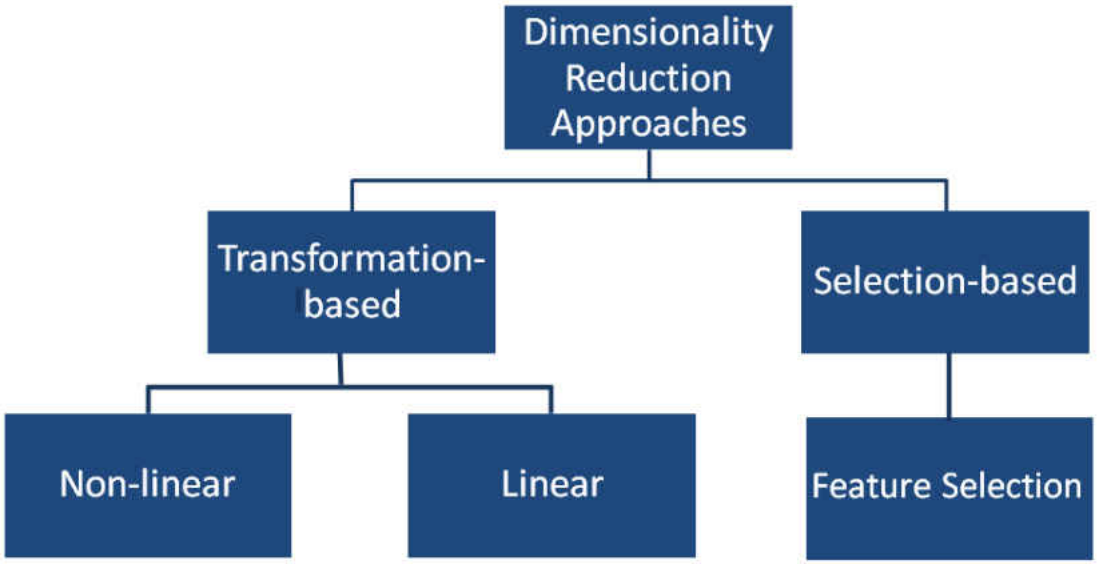

Figure 2.1: Dimensionality reduction approaches

of the original feature set, then a feature selection approach is chosen which will ensure this. However, conversely if an application requires a visualisation of relationships within the data, then a DR approach that transforms the data into a small number of dimensions whilst emphasising those relationships may be more suitable.

Basically, transformation-based DR approaches can be further classified into two distinct groups (see Fig. 2.1): linear, and non-linear. Linear methods have enjoyed much popularity, and include such approaches as Principal Component Analysis (PCA) [48], and Multidimensional scaling (MDS) [226]. PCA is perhaps the most popular amongst the linear techniques due mainly to its relative simplicity and the number of computationally efficient algorithms which are available. The approach transforms the original data features with a (usually) reduced number of uncorrelated features. These features are termed principal components. At the heart of the approach is the assumption that large feature variance is indicative of useful information, and conversely that small variance is considered less useful. Fig. 2.2 demonstrates this, where the principal components of a two-dimensional normal point cloud are shown. The second principal component (PC2) indicates the direction of maximum variance of the data, as it is most dispersed along this new axis in the example. Data is transformed such that transformed features with small variance are allowed to be removed. In order to do this, first the eigenvectors of the covariance matrix of data points or objects must be found. Then, a transformation matrix is constructed from the ordered eigenvectors, and finally the original data is transformed by matrix multiplication.

One of the problems with PCA is that the number of variables to be discarded

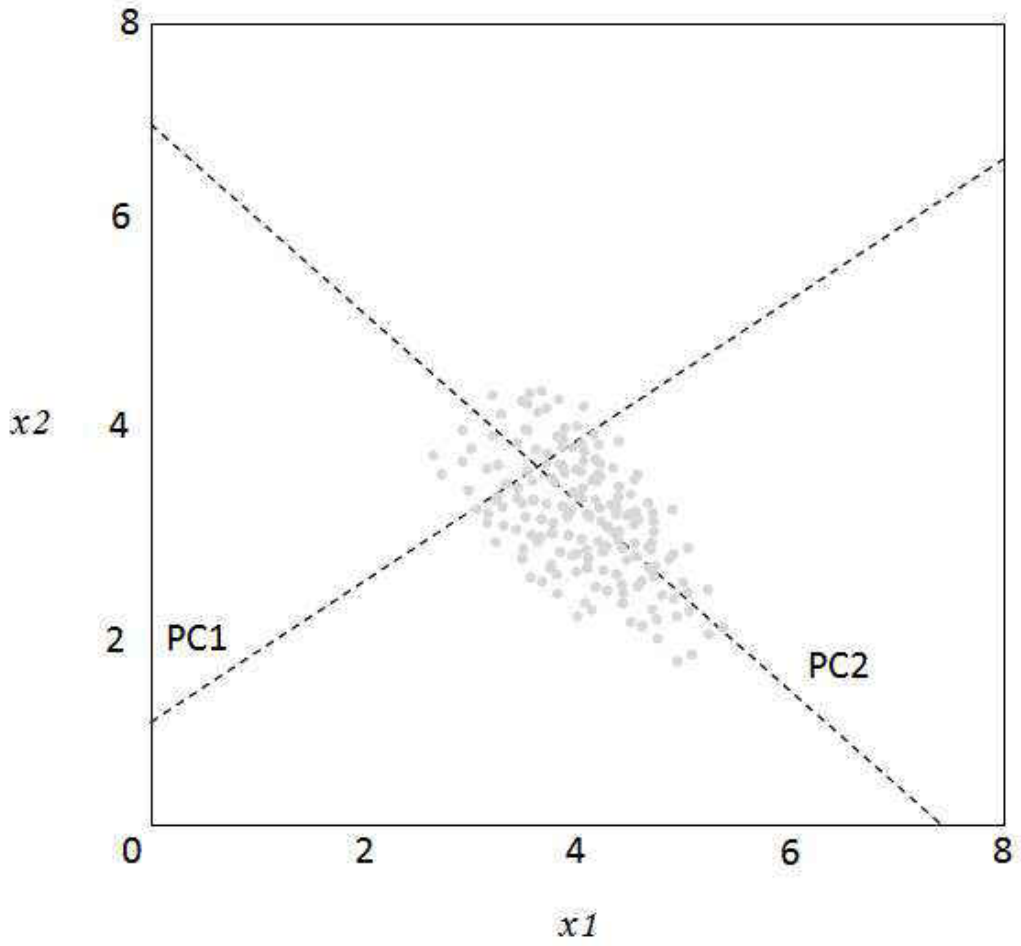

Figure 2.2: 2-dimension normal point cloud with corresponding principal components

is not known *a-priori*, introducing a potential source of error. A subjective estimation must be made therefore, as to how many transformed features should be retained. There are other problems also, including the inability of PCA to operate on nominal data. This relates to the matrix calculations which are obviously inapplicable in this case. Additionally, PCA is not effective at handling data which is correlated in a non-linear fashion. PCA is applied to a number of benchmark datasets in Sections 3.5.3 and 4.3.6.1, and compared with other approaches.

Multidimensional scaling (MDS) [136], [226], [252], refers to a set of techniques that use the proximity of objects as input and display its structure as a geometric construct. The proximity values are a measure of the similarity of objects. The resulting transformation to fewer dimensions aims to preserve the original proximity values as far as possible.

The previously mentioned inability of methods such as PCA and MDS to deal with non-linear data led initially to unsuccessful extensions based on PCA [21], [60], [112]. These extensions however still suffered from the problems originally

exhibited by PCA, and this provided new impetus for the development of techniques which had the ability to handle non-linear data effectively. Isomap [222] is one such technique which is based on MDS. In this approach embeddings are optimized to preserve geodesic distances between pairs of data objects. These are estimated by calculating the shortest paths through large sublattices of data. The algorithm can discover non-linear degrees of freedom as the geodesic distances act as the true low-dimensional geometry of the manifold. This approach unlike PCA however requires the specification of a parameter known as the 'neighbourhood value' such that the edges of the manifold can be determined correctly prior to reduction. Locally Linear Embedding (LLE) [191] is another of the non-linear DR techniques. It uses an eigenvector method for the problem of non-linear DR. It works by computing low dimensional 'neighbourhood-preserving' constructs or embeddings of high dimensional data. This is achieved through the exploitation of the local symmetries of linear constructs. In more informal terms, non-linear structures are modelled by piece-wise linear steps. Again, as with Isomap, LLE suffers from the same 'neighbourhood value' specification problem described previously. A detailed and more comprehensive review of other non-linear approaches can be found in [118].

2.1.2 Selection-based Approaches

The DR techniques examined in the previous section irreversibly transform data, thus destroying the semantics or underlying meaning of the features. Feature selection (FS) however, is a DR technique which obtains a minimal feature subset from a problem domain whilst retaining a suitably high accuracy in representing the original features. FS is necessary in many real-world problem domains due to the level of noisy, irrelevant or misleading features which can lead to to the discovery of spurious or irrelevant patterns in the data. Through the removal of such factors, the performance of techniques which learn from data can be greatly improved. A number of detailed reviews of feature selection techniques can be found in [40], [130], and [131].

Feature selection is a commonly used approach in machine learning (may also be known as feature subset selection, variable selection, or attribute reduction) and can be considered as the process of selecting the input attributes of a dataset that most closely define a particular outcome. FS enables the selective focus on relevant attributes whilst ignoring the (possibly misleading) contribution of irrelevant attributes. From a computational efficiency standpoint, it is advantageous to have a minimal set of features involved in the classification process as many learning algorithms scale up quickly ($O(n^2)$ or worse) with the addition of features.

FS has countless application domains including (but not limited to) image recognition/retrieval [217], complex plant monitoring [204], text categorisation [247], computer network intrusion detection [50] [119], genomic analysis [244], and data mining [40], [39].

Amongst the advantages of FS are:

- Facilitation of data visualisation through the reduction of the data to fewer dimensions. This makes trends within the data more easily identifiable and can be very important where few features may have an influence on data outcomes.
- Reduction in measurement and storage requirements. In domains where features correspond to particular measurements (for instance, a water treatment plant [204]), fewer features are highly desirable due to the expense and time-cost of taking such measurements.
- Reduction of training and utilisation times. With smaller datasets, the runtimes of learning algorithms improve significantly, for both training and classification phases.
- Improvements in prediction performance. Classifier accuracy can be increased as a result of feature selection, through the removal of noisy or misleading features.

Additionally, for those methods which extract knowledge from data (e.g. rule induction) other benefits of FS include an improvement in the readability of discovered knowledge. When induction algorithms are applied to reduced data, the resulting rules are more compact. A good feature selection method will remove unnecessary attributes which affect both rule comprehension and rule prediction performance.

There have been many attempts in the literature to categorise FS methods using a review type approach [40], [50], [67], [130], [155]. All of these reviews overlap to a greater or lesser degree when describing the FS process, which can broken down into three discrete steps (as illustrated in Fig. 2.3): generation of subsets, evaluation, and criteria (or a criterion) that will halt the FS process.

It is possible therefore to broadly define each approach to FS on the basis of the following characteristics:

- Search strategy
- Generation of subsets
- Evaluation measure

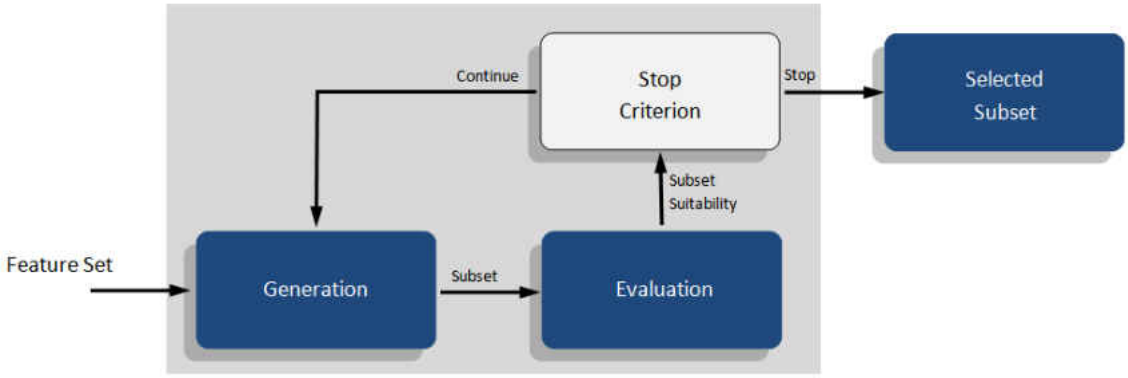

Figure 2.3: Feature selection

Note that the search strategy and the subset generation steps are closely related (they are integrated for some approaches) and in some FS algorithms may impact on each other. FS therefore can be viewed as a search in a feature space for a solution which maximises some predefined evaluation measure. Let Y = the feature set to be examined, and a cardinality $|Y| = n$. For any problem this can be defined as in Fig. 2.4:

FEATURE SELECTION
J, an evaluation measure to be optimised in the manner $J : Y' \subseteq Y \rightarrow R$;
Y, the feature space.
Y', a subset of Y
The problem of subset selection can now be viewed in the following ways:

- Set a desired value for J called J_{des}. Search for $Y' \subseteq Y$ such that $|Y'| < |Y|$, and $J(Y') \geq J_{des}$
- Let $|Y'| = m < n$. Search for $Y' \subset Y$ such that $J(Y')$ is maximum.
- Search for a subset such that $|Y'|$ is minimised whilst simultaneously aiming to maximise $J(Y')$

Figure 2.4: The feature selection problem

2.1.2.1 Search Strategy

The search strategy of a FS process is the manner in which the feature space is traversed in an attempt to locate valid feature subsets. Each state in the search space can be thought of as having a weighting $ws_1, ..., ws_n$ of possible features of Y, with $|Y| = n$. For a binary example, $ws_i \in \{0, 1\}$, or discrete example, $ws_i \in$

$[0, 1]$. Exponential Search includes approaches where the complexity is $O(2^n)$. An exhaustive search will always guarantee an optimal solution. However a search which returns an optimal solution does not necessarily need to be exhaustive. If the evaluation measure is monotonic, a branch-and-bound search [159] is optimal. An evaluation measure J is monotonic if for any two subsets S_1, S_2, and $S1 \subseteq S2$ then $J(S1) \geq J(S2)$.

Sequential search selects a single successor amongst all the successors to the current state. This is performed in an iterative manner. Obviously the number of steps for this type of search is limited to $O(n)$ - it could not otherwise be referred to as a sequential search. The complexity can be determined by including the number of evaluated subsets in each change of state - (k). With this in mind the complexity can be viewed as $O(n^{k+1})$. Sequential search methods therefore cannot guarantee optimality, as the optimal subset may be found in an area of the search space that is not examined.

The use of random selection of features is a strategy which ensures that the search moves through a number of states and does not become trapped in a single area of the search space. This may mean however that the search will return several sub-optimal solutions.

2.1.2.2 Generation of Subsets

As with search techniques, a number of different approaches can be adopted when deciding how to generate subsets for evaluation. Forward selection for example, is carried out by adding a feature or a number of features once per iteration to the current subset candidate from a subset of those features that have not already been considered for addition. For each iteration a feature (or features) that results in an increase in J is added to the current subset candidate. Typical complexity of forward selection is $O(n)$.
Let $Y' = \emptyset$, one forward iteration will be:

$$Y' := Y' \cup \{y_i \in Y \setminus Y' | J(Y' \cup y_i) \text{ is greater } \}$$

A typical stopping criterion in this case might be $|Y'| = n'$ (where the value n' is predetermined). Alternatively thresholds such as - *if the value of J has not increased after x iterations then stop* - can be imposed. Despite the efficiency of this of selection technique, it has one major weakness - the combined behaviour of features is not considered. Take the features f_1, f_2, suppose that $J(\{f_1, f_2\}) \gg J(\{f_2\}), J(\{f_2\})$, then neither of the individual features f_1, or f_2 will be selected despite their obvious combined value.

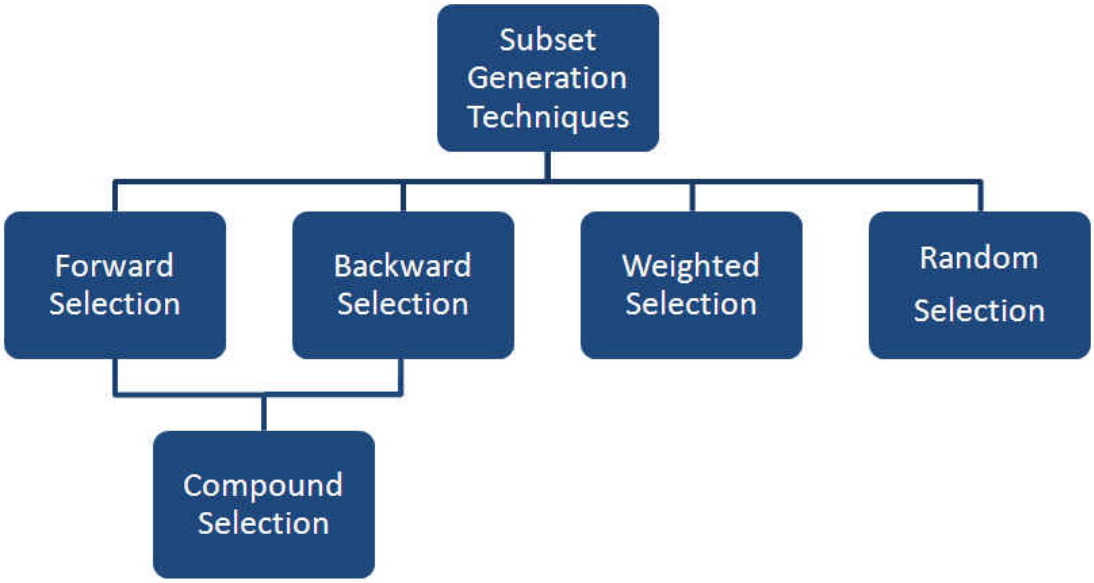

Figure 2.5: Subset generation techniques

Backward selection works in a manner that is essentially contrary to that of forward selection. All of the features of the original feature subset form the initial subset candidate for consideration. Features are then removed from the current subset candidate from a set of features that have not already been considered for removal. For each iteration a feature is removed that results in an increase in J. The complexity of backward selection is again $O(n)$ although software implementations have shown that it requires higher computational overhead [106], [107], [117] to obtain similar performance to that of forward selection. Backward selection has also shown a tendency to result in subsets which are suboptimal in comparison to those obtained using forward selection [67].

Random selection as mentioned previously selects features at random and again tries to optimise J for each iteration of subset generation. The selection process may be restricted by some measure which is used to determine the feature subset from which features are randomly selected. Examples include random-start, hill climbing and simulated annealing [50]. An alternative approach [19] is to generate the subset which will be selected from in a completely random manner thus any current subset does not grow or shrink from any previous subset - i.e a new subset is selected for each iteration. This is known as the Las Vegas algorithm. The use of random techniques allows the search to escape local optima.

Weighted selection [115] is a method in which all of the features in the feature space are present in the solution to a certain degree (weighting). A successor space will have a different weighting. Searches for optimal subsets are performed on this basis in an iterative manner.

The selection of features using both backward and forward selection on the same feature space is known as compound selection. Through applying several forward steps (f) and backward steps (b) separately on the same subset, the value of J is assessed. If the magnitude of J increases to a greater value in the forward method then forward selection is continued until an optimal subset is obtained and vice-versa. If the value of J appears to be equal then the next step (forward/backward) is chosen at random.

2.1.2.3 Evaluation Measures

Optimality of subsets is subjective, and a subset that is selected as optimal using one particular evaluation function may not be equivalent to that of a subset selected by another evaluation function. There are a number of different evaluation functions employed by the various FS approaches. These include, interclass distance measures, probability of error measures, information measures e.g. [40], [110], dependence measures [172], consistency measures [84], and classifier error rate [184] measures.

Interclass Distance or Euclidean measures are based on the assumption that objects of different classes are distant in the feature space. It is sufficient therefore to define a metric (D) which can be used to differentiate between classes and use it as a measure:

$$D(\omega_i, \omega_j) = \frac{1}{N_i N_j} \sum_{k1}^{N_i} \sum_{k_2=k_1+1}^{N_i} d(x_{(i,k1)}, x_{(j,k2)}) \tag{2.1}$$

$$J = \sum_{i=1}^{m} P(\omega_i) \sum_{j=i+1}^{m} P(\omega_j) D(\omega_i, \omega_j) \tag{2.2}$$

Where $x_{(i,j)}$, the object j of class ω_i, and N_i is the number of objects of class ω_i. The most common distance measures (d) utilised for this measure are Euclidean.

When considering the accuracy of a classifier, the ability to classify instances generated by the same probability distribution correctly is the objective. One of the most obvious ways to approach this is to use the Bayesian probability error measure. Minimising the probability of error P will result in a classifier that is obviously more accurate as the probability of error is reduced. The suitability of this measure for J is therefore clear.

Let $\vec{x} \in \mathbb{R}^{\kappa}$ represent the unlabeled objects, and $\Omega = \{\omega_1, ...\omega_n\}$ a set of classes such that: $c : \mathbb{R}^{\kappa} \rightarrow \Omega$. Probability such as that defined in [48] can then be applied:

$$P = \int [1 - maxP(\omega_i|\vec{x})]\, p(\vec{x}) d\,\vec{x} \tag{2.3}$$

Where $p(\vec{x}) = \sum_{i=1}^{m},\ p(\vec{x}|\omega_i),\ P(\omega i)$ is the unconditional probability distribution of the objects, and $P(\omega_i,\ vecx)$ is the *a posteriori* probability of ωi being of class $\vec{x}$.

Some classifiers such as those which use the single-nearest-neighbour type of approach are directly related to probability of error [48].

Information measures are typically concerned with the information gain (entropy measure) of a feature [40]. The information gain value from a feature f can be defined as the difference between the prior uncertainty and the expected posterior uncertainty of y. Feature y is more desirable than feature z if the information gain using y is greater than that using z.

Dependence or correlation is a measure of how closely two features are associated. It can be used in deduction such that if the value of a feature is known, it is possible to deduce the value of another. This can be used to easily eliminate highly correlated features.

Inconsistency in Y' and S can be defined as two instances in S that are equal when considering only the features in Y' and that belong to different classes. The aim is therefore to find the minimum subset of features that will lead to no inconsistencies[3]. The *inconsistency index*(or *inconsistency count* [130]) of an object ($A \in S$) can be defined as:

$$INDEX_{Y'}(A) = Y'(A) - maxY'_k(A) \tag{2.4}$$

Where Y' is the number of objects in S equal to A using only the features in Y' and,$Y'_k(A)$ is the number of objects in S of class k equal to A again using only the features in Y'. The *inconsistency rate* for a feature subset can therefore be defined as:

$$INRATE(Y') = \frac{\sum_{A \in S} INDEX_{Y'}(A)}{\mid S \mid} \tag{2.5}$$

Consistency is the measure used in the FS algorithm FOCUS [3]. Also the author in [98] has noted that this is identical to the rough set dependency evaluation measure as used in [32].

2.2 Feature Selection Models

There are a number of frameworks or models which can be employed for the FS task, and broadly these can be divided into four types: filter, wrapper, hybrid,

and embedded methods.

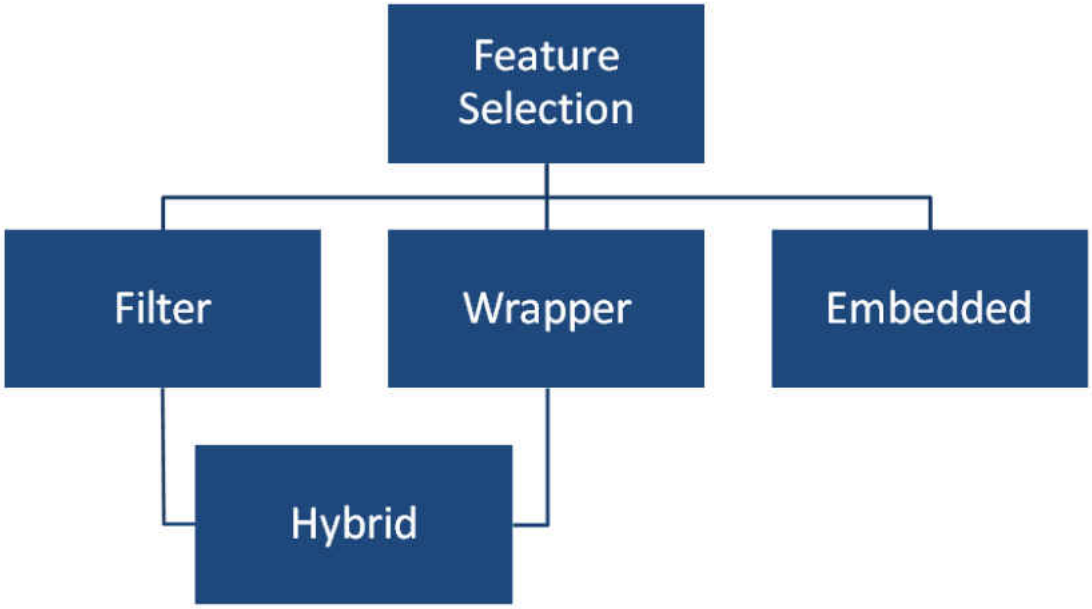

Figure 2.6: FS models

Algorithms that perform FS in isolation of a learning algorithm, are termed filter approaches. Essentially irrelevant attributes are *filtered* out prior to performing induction. Filters are useful for most domains as they are not integrated with any induction algorithm.

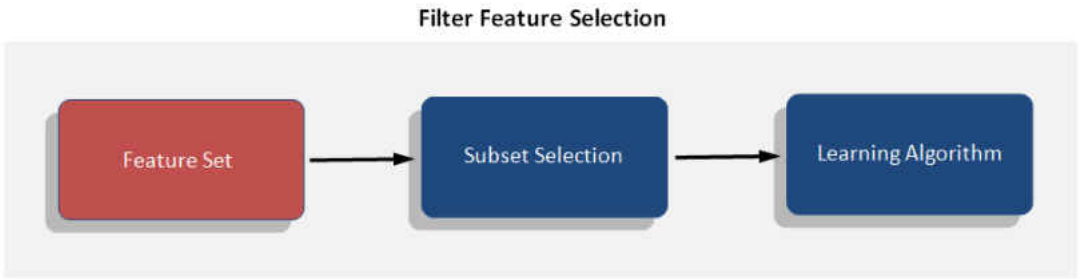

Figure 2.7: Filter FS

Filter approaches are generally employed where redundancy/irrelevance removal is the aim due to their execution speed. Fig. 2.8 shows a generalised algorithm for the filter approach.

For a given dataset D, the algorithm searches from a given subset S_0 (initially empty when using forward selection), using a predefined search strategy to traverse the feature space. Each generated subset is then evaluated by an independent measure J and compared with the current best subset. If as a result of the evaluation the generated subset offers an increase in the value of J it becomes the new current best subset. The search continues until a predefined stopping

FILTER FS (D, S_0, δ, S).
D, a training dataset with N features;
S_0, initially empty subset.
δ, stopping criterion.
S_{opt}, an optimal subset.

(1) **start**
(2) $S_{opt} \leftarrow S_0$;
(3) $\gamma_{best} = \text{eval}(S_0, D, J)$
(4) **do**
(5) $S = generate(D)$;
(6) $\gamma = eval(S, D, J)$;
(7) **if** $(\gamma > \gamma_{best})$
(8) $\gamma_{best} \leftarrow \gamma$
(9) $S_{opt} \leftarrow S$
(10) **until** (δ is reached)
(11) **return** S_{opt}

Figure 2.8: Generalised filter algorithm

criterion δ has been reached. The algorithm finally outputs the last 'best current subset' S_{opt}.

Wrapper methods [25], [55], [106] in contrast to filter approaches are often used in conjunction with a learning or data mining algorithm, where the learning algorithm forms part of the validation process.

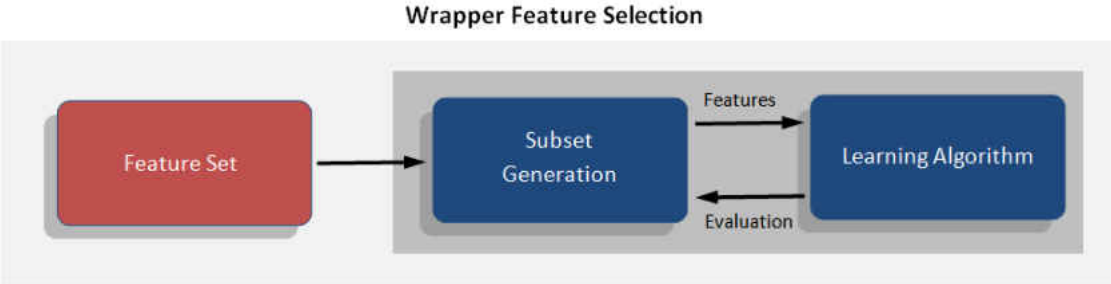

Figure 2.9: Wrapper FS

The wrapper approach is generally seen as being inferior [130] to other approaches because of the computational overhead that is required for the examination of each subset by the learning algorithm. However, this is not necessarily true as the application domain should to be taken into account, and if the application is data mining or related then a wrapper approach will return better results.

The generalised wrapper algorithm is similar to the filter approach apart from the fact that a learning algorithm (LA) is employed in place of a measure (J) as used in the filter approach.

WRAPPER FS (D, S_0, δ, S).
D, a training dataset with N features;
S_0, initially empty subset.
δ, stopping criterion.
S_{opt}, an optimal subset.

(1) **start**
(2) $S_{opt} \leftarrow S_0$;
(3) $\gamma_{best} = \text{eval}(S_0, D, LA)$
(4) **do**
(5) $S = generate(D)$;
(6) $\gamma = eval(S, D, LA)$;
(7) **if** $(\gamma > \gamma_{best})$
(8) $\gamma_{best} \leftarrow \gamma$
(9) $S'_{opt} \leftarrow S$
(10) **until** (δ is reached)
(11) **return** S_{opt}

Figure 2.10: Generalised wrapper algorithm

Hybrid methods [38], [160], [244] are those which try to take advantage of both previous models (filter and wrapper).

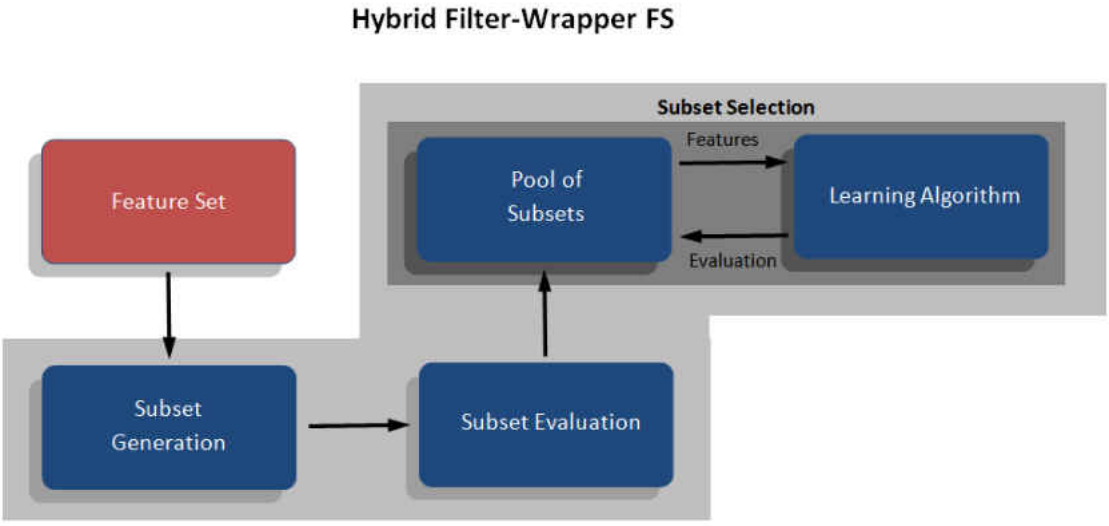

Figure 2.11: Hybrid FS

The ideology behind hybrid approaches is to make use of both a measure, and a learning algorithm to evaluate feature subsets. The measure is used to decide which subsets are the 'best' for a given cardinality. The learning algorithm is then used to select the final 'best' overall feature subset from a pool of feature subsets of different cardinalities.

The search starts from a defined subset S_0 and traverses the feature space

HYBRID FILTER-WRAPPER FS(D, S_0, δ, S).
D, a training dataset with N features;
S_0, initially empty subset.
S_{opt}, an optimal subset.
F_j, a feature chosen using j.

```
(1)  start
(2)  S_opt ← S_0;
(3)  c_0 ←| S_0 |;
(4)  γ_best ← eval(S_0, D, J)
(5)  θ_best ←eval(S_0, D, LA)
(6)    for c = c_0 + 1 to N begin
(7)     for i = 0 to N – c begin
(8)      S ← S_opt ∪ {F_j}
(9)      γ ←eval(S, D, J)
(10)      if (γ¿γ_best)
(11)         γ_best ← γ
(12)         S'_opt ← S
(13)     end
(14)     θ ←eval(S'_opt, D, A)
(15)     if (θ is better than θ_best)
(16)         S_opt ← S'_opt
(17)         θ_best ← θ
(18)     else
(19)         break
(20)         return S_opt
(21)   end
(22)   return S_opt
(23) end
```

Figure 2.12: Generalised hybrid algorithm

in order to find the best subset for each level of cardinality. For each iteration a 'best' subset with a cardinality c, is searched for through all possible subsets with cardinality of $c+1$ by adding a feature from the remaining features. Each newly generated feature subset is evaluated by measure M and then compared to current optimal candidate S'_{opt}. If S is determined to be 'better' it becomes the 'best' candidate $S \leftarrow S_{opt}$ at level $c+1$. Before performing the next iteration, the learning algorithm LA is applied to S'_{opt} at cardinality level $c+1$ and the goodness or 'quality' of the learned result θ is compared to that of the best feature subset at cardinality level c. If S'_{opt} is better, the algorithm continues to search for the 'best' subset at the next level. Otherwise it stops and will return the current best subset candidate. The 'quality' of the results from a learning algorithm ensures that the search will terminate automatically.

Finally, there is the embedded method approach. This approach simply means that embedded within the learning algorithm is an implicit or explicit FS mechanism e.g. [184]. Decision trees are an example of the embedded approach, indeed some decision tree algorithms as well as allowing the use of their own internal embedded FS process allow other FS algorithms to be 'plugged-in' [48].

2.3 Rough Sets for Approximate Modelling

The ability to deal effectively with insufficient or imperfect knowledge is a central motivating factor in much of the research in the field of computational intelligence. In the areas of machine learning, data-mining, pattern recognition, and intelligent control, the ability to handle such knowledge is of primary importance both in terms of theoretical advancement and practical applications. The work in the area of rough set theory (RST) [172], [175] offers perhaps one of the most distinct and recent approaches in this respect.

Such is the world-wide nature of the attention that RST has attracted since its inception [108], that much research and development has been carried out not only in applying the theory to many and varied problem domains, but also to extending it theoretically. This has resulted in a significant breadth and depth of work in the area. Rough set theory [172] has been used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information [172], [175], [182], and [205]. Over the past ten years, RST has become a topic of great interest to researchers and has been applied to many domains. Indeed, since its inception, this theory has been successfully utilised to devise mathematically sound and often, computationally efficient techniques for addressing problems such as knowledge discovery from data, data reduction, data significance evaluation, decision rule generation, and data-driven inference interpretation [171]. Given a dataset with discretised attribute values, it is possible to find a subset (termed reduct) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss. Rough set theory possesses many features in common (to a certain extent) with the Dempster-Shafer theory of evidence [197] and fuzzy set theory [254]. It works by making use of the granular structure of the data only. This is a major difference when compared with Dempster-Shafer theory and fuzzy set theory, which require probability assignments and membership values, respectively. The use of only the data and its granularity ensures that no other assumptions are made about the data. This approach has led to some researchers suggesting that this is a disadvantage rather than an advantage, of rough set theory [108] as other numerical and contextual aspects

are effectively ignored. However, in disregarding such supplemental information, model assumptions can be minimised.

Formally, a rough set is the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations, which are a classification of the domain of interest into disjoint categories. The lower approximation is a description of the domain objects which are known with certainty to belong to the concept of interest, whereas the upper approximation is a description of the objects which possibly belong to the concept. The approximations are constructed with regard to a particular subset of attributes or features.

One of the primary drawbacks of RST lies in its inability to deal with real world data. Due mainly to the granular approach that RST uses to handle data, and the strict structure of equivalence imposed, it does not allow any flexibility when dealing with measurement noise, or imperfection that are prevalent in real-world data. However, most datasets contain real-valued features and so it becomes necessary to perform a discretisation step prior to employing RST for knowledge discovery. Take for instance a weather forecasting system which records a number of meteorological attributes, one in particular might be *average rainfall*, in reality this is a continuous and real-valued measurement. However, in order to apply RST to such a problem, this attribute must be discretised with a set of labels such as *light*, *medium*, and *heavy*. This imposes subjective human judgment on what is otherwise an objective measurement.

The deficiency of RST in handling real-valued data has led over the years to the development of a number of extensions which aim to address this problem. There are two areas of RST which have been considerably exploited in order to achieve this; modification of the equivalence relation, and manipulation of the subset operator. These are the primary operations of RST and it is unsurprising therefore that a number of extensions have been proposed with regard to these areas. The tolerance rough set model (TRSM) [208], is a typical example of an attempt to address this problem through the modification of the equivalence relation. Variable precision rough sets (VPRS) [263] allow the relaxation of the subset operator of traditional RST. This approach was originally formulated in order to analyse and identify data patterns which represent statistical trends.

In addition to the use of alternative equivalence relations and modification of the subset operator, there is also a third aspect of RST which has been exploited, that of the use of the information contained in the boundary region, or region of uncertainty between the lower and upper approximations, [84], [148]. This information although uncertain can be useful in maximising the performance of RST without changing the underlying model or modifying the subset operators.

As well as directly extending RST, it has also been hybridised with other soft

computing methods such as fuzzy sets [254], genetic algorithms (GAs), neural networks, and statistical methods such as principal component analysis (PCA) [48] etc. Such hybridisation has highlighted the value of employing RST, as its use often results in methods which outperform such methods individually. In particular, the hybridisation of RST with fuzzy set theory [254] to form fuzzy-rough set theory [52] is perhaps the most important of all. Fuzzy-rough set theory [52] attempts to take advantage of the complementary nature of fuzzy sets and rough sets. The significance of this work is reflected in the level of research carried out in this area and also to the number of applications of fuzzy-rough set theory.

2.3.1 Basic Concepts and Theoretical Background

In this section, the basic notions, definitions, and operations of rough set theory are described. The upper and lower approximation concepts, as well as how these can be used to minimise data are also explored. A small example is used to demonstrate all of the concepts described and show the individual steps involved in employing RST. Heuristics for discovering reducts, and search techniques are also discussed.

Central to rough set theory is the concept of indiscernibility. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse) and $\mathbb{A}$ is a non-empty finite set of attributes such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. For any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 \mid \forall a \in P,\ a(x) = a(y)\} \tag{2.6}$$

The partition of $\mathbb{U}$, generated by $IND(P)$ is denoted $\mathbb{U}/IND(P)$ and can be defined as follows:

$$\mathbb{U}/IND(P) = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \tag{2.7}$$

where,

$$\mathbb{U}/IND(\{a\}) = \{\{x|a(x) = b,\ x \in \mathbb{U}\}|b \in V_a\} \tag{2.8}$$

and,

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \tag{2.9}$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$.

Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of X:

$$\underline{P}X = \{x|\ [x]_P \subseteq X\} \tag{2.10}$$

$$\overline{P}X = \{x|\ [x]_P \cap X \neq \emptyset\} \tag{2.11}$$

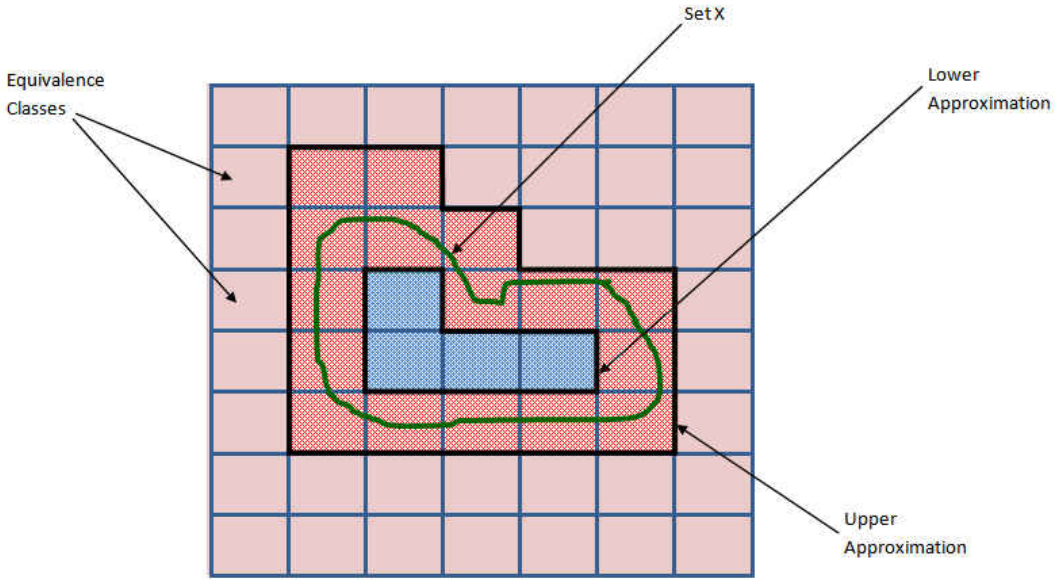

Figure 2.13: Basic rough set concepts

Let P and Q be equivalence relations over $\mathbb{U}$, then the positive, negative and boundary regions are defined by:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{2.12}$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{2.13}$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{2.14}$$

The positive region contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}/Q$ using the information in attributes P. The boundary region, $BND_P(Q)$, is the set of objects that can possibly, but not certainly, be classified in this way. The negative region, $NEG_P(Q)$, is the set of objects that cannot be classified to classes of $\mathbb{U}/Q$.

2.3.1.1 Example

To illustrate the above concepts a short example in the form of an information system is employed. There are four conditional attributes: a, b, c, and d, and a single decisional attribute, e.

$x \in U$	a	b	c	d	$\rightarrow$	e
1	M	L	N	N		H
2	L	M	M	M		F
3	M	M	L	N		F
4	M	L	N	L		G
5	N	N	L	M		G
6	N	M	M	M		F
7	L	M	M	L		G

Table 2.1: Example dataset

Using the indiscernibilty concept, the data in Table 2.1 can be partitioned according to the outcome. V_a is the set of values that attribute a may take (in this case L, M, or N). In a decision system $A = \{\mathbb{C} \cup \mathbb{D}\}$ where $\mathbb{C}$ denotes the set of condition attributes and $\mathbb{D}$ denotes the set of decision attribute(s). There are associated equivalence relations with any $P \subseteq \mathbb{A}$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 \mid \forall a \in P,\ a(x) = a(y)\} \tag{2.15}$$

For the data in Table 2.1 - the partition of $\mathbb{U}$ by the attribute a would be:

$$\mathbb{U}/IND(\{a\}) = \{\{1, 3, 4\}, \{2, 7\}, \{5, 6\}\} \tag{2.16}$$

And for the same table using attributes $\{b, c\}$

$$\mathbb{U}/IND(\{b, c\}) = \{\{1, 4\}, \{2, 6, 7\}, \{3\}, \{5\}\} \tag{2.17}$$

This relates to the partition or grouping of the attributes where: $a = L$ (objects 1 3 and 4), where $a = M$ (objects 2, and 7) and, where: $a = N$ (objects 5 and 6). The equivalence classes of the P-indiscernibility relation are denoted by $[x]_p$. Let $\mathbb{X} \subseteq \mathbb{U}$. X can be approximated using only the information within P by formulating lower and upper approximations of X as described previously.

2.3.2 Rough Set Dependency and Other Measures

An important aspect of data analysis is the discovery of dependencies between attributes. From an intuitive point-of-view an attribute or a set of attributes Q can depend on a set of attributes P, denoted $P \Rightarrow Q$ if all values of attribute(s)

in Q are determined uniquely by values of attribute(s) from P. Another way of describing this is that, Q depends totally on P if a functional dependency exists between the values of Q and P.

Referring to the example in the previous section, the rough set dependency for the set of attributes Q on a set of attributes P can be shown. For $P, Q \subset \mathbb{A}$, it can be said that Q depends on P in a degree k (where $k \in [0,1]$) denoted $P \Rightarrow_k Q$, if:

$$k = \gamma_P(Q) = \frac{\mid POS_P(Q) \mid}{\mid \mathbb{U} \mid} \tag{2.18}$$

where

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}(X) \tag{2.19}$$

is the positive region of the partition of the universe with respect to P (i.e. the set of all elements that can be classified uniquely into sets of the partition $\mathbb{U}/Q$ in terms of P).

If k=1, Q is completely dependent on P, if $k < 1$ Q is partially dependent (to a degree - k) on P and obviously, if $k = 0$, Q is completely non-dependent on P. Calculation of the relevant dependencies of each attribute (or group of attributes) allows the significance of that attribute (or group) to be realised.

Taking the data from the example decision table (Table 2.1), the degree of dependency of attribute $\{e\}$ upon the attributes $\{b, c\}$ is:

$$\gamma_{\{b,c\}}(\{e\}) = \frac{\mid \{POS_{\{b,c\}}(\{e\}) \mid}{\mid \mathbb{U} \mid}$$

$$= \frac{\mid \{3,5\} \mid}{\mid \{1,2,3,4,5,6,7,\} \mid} = \frac{2}{7}$$

For the application of feature selection, the minimisation of attributes can be realised through the comparison of equivalence relations generated by sets of attributes ($\{b, c\}$ for the purpose of the previous example). Attributes are removed such that the minimised set provides an equivalent predictive characteristic as the initial decision features. This minimised set is termed a reduct and can be defined as a subset R of the conditional attribute set *cond* such that $\gamma_R(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})$.

Other measures have also been used to discover rough set reducts, for instance in [69], a feature selection method which is based on an alternative dependency measure is presented. This technique was proposed in order to avoid the expensive calculation of discernibility functions or positive regions. The authors replace the

traditional rough set dependency measure with the relative dependency measure, defined as follows for an attribute subset P:

$$\kappa_P(\mathbb{D}) = \frac{|\ \mathbb{U}/IND(P)\ |}{|\ \mathbb{U}/IND(P \cup \mathbb{D})\ |} \tag{2.20}$$

The authors then demonstrate that R is a reduct if and only if $\kappa_R(D) = \kappa_C(D)$ and that $\forall X \subset R,\ \kappa_X(D) \neq \kappa_C(D)$.

In addition, the entropy measure has been used in [94] to discover smaller reducts than the rough set dependency measure alone. In this approach, although entropy is used in the search for reducts, rough set dependency is still used as a termination criterion.

2.3.2.1 An Example Feature Selection Algorithm - Rough Set Attribute Reduction

To demonstrate how the concepts described in the last few sections can be applied, a feature selection algorithm which exploits all of the measures discussed previously is discussed here. This approach has been used extensively, and successfully [32], [202], [203] etc.

At the heart of the RSAR approach is the concept of indiscernibility [172]. Let $I = (\mathbb{U}, \mathbb{S})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse) and $\mathbb{S}$ is a non-empty finite set of attributes so that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{S}$. V_a is the set of values that a can take. For any $P \subseteq \mathbb{S}$, there exists an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P, a(x) = a(y)\} \tag{2.21}$$

The partition generated by $IND(P)$ is denoted $\mathbb{U}/IND(P)$ and is calculated as follows:

$$\mathbb{U}/IND(P) = \otimes\{\mathbb{U}/IND(\{a\}) : a \in P\} \tag{2.22}$$

where,

$$S \otimes T = \{X \cap Y : \forall X \in S, \forall Y \in T, X \cap Y \neq \emptyset\} \tag{2.23}$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained in P by constructing the P-*lower* and P-*upper* approximations of X:

$$\underline{P}X = \{x|[x]_p \subseteq X\} \tag{2.24}$$

$$\overline{P}X = \{x|[x]_p \cap X \neq \emptyset\} \tag{2.25}$$

Let P and Q be equivalence relations over $\mathbb{U}$, then the concepts of the positive, negative and boundary regions can be defined:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{2.26}$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{2.27}$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{2.28}$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of a set of attributes Q on a set of attributes P. This can be achieved as follows: For P,$Q \subseteq S$, it can be said that Q depends on P in a degree k ($0 \leq k \leq 1$), thus the higher the value of k the more dependent Q is upon P. This is denoted ($P \Rightarrow_k Q$) if:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \tag{2.29}$$

The reduction of attributes can be achieved through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features, assuming of course that the dataset is consistent. A reduct of set $\mathbb{N}$ is a minimal set of attributes $B \subseteq A$ such that $IND_N(B) = IND_N(A)$. In other words, a reduct is a minimal set of attributes from A that preserves the partitioning of the universe and hence the ability to perform classifications as the whole attribute set A does.

The QuickReduct algorithm [32] shown in Fig. 2.14 searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until the search produces its maximum possible dependency value for that dataset ($\gamma_c(\mathbb{D})$). Note that this type of search does not guarantee a minimal subset and may only discover a local minimum.

QuickReduct($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

(1) $R \leftarrow \{\}$
(2) **do**
(3) $T \leftarrow R$
(4) $\forall x \in (\mathbb{C} - R)$
(5) **if** $\gamma_{R \cup \{x\}}(\mathbb{D}) > \gamma_T(\mathbb{D})$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) **until** $\gamma_R(\mathbb{D}) == \gamma_{\mathbb{C}}(\mathbb{D})$
(9) **return** R

Figure 2.14: The QuickReduct algorithm

2.3.3 Minimal Reducts and Reduct Discovery

A method for reducing data, demonstrated in Section 2.3.1.1, identifies equivalence classes using the available attributes [172]. If only those attributes that preserve the indiscernibility relation are retained as demonstrated in Section 2.3.2.1, any remaining attributes are redundant since their omission will not affect classification. There are usually many such subsets of attributes, however those which are minimal are termed minimal reducts. A minimal reduct is therefore a minimal set of attributes that preserves the partitioning of the universe and hence the ability to perform the same classification as the complete dataset. In practical terms this means that no attributes can be removed from the subset without affecting the dependency measure. Let $\mathbf{R}$ be the set of all reducts then minimal reducts $\mathbf{R_{min}} \subseteq \mathbf{R}$ can be defined as:

$$\mathbf{R_{min}} = \{X : X \in \mathbf{R}, \forall Y \in \mathbf{R}, \mid X \mid \leq \mid Y \mid\} \tag{2.30}$$

The search for minimal reducts is however non-trivial [206], [221] and it can be demonstrated that the number of reducts for a given information system with n attributes can be as much as:

$$\begin{pmatrix} m \\ \lfloor m/2 \rfloor \end{pmatrix} \tag{2.31}$$

The intersection of all the sets in $\mathbf{R}$ is termed the *core*. This set contains the attributes which cannot be eliminated without the introduction of contradictions in the data.

Many rough set approaches for dealing with data opt for search techniques

which tend to balance the need for the discovery of minimal reducts with the computational overhead involved in searching for such reducts. The greedy hill-climbing search [32] is such an example, and although it will not guarantee minimality it is relatively efficient in terms of time/space complexity - $(n^2 + n)/2$ for a data dimensionality of n. Other search techniques which also do not guarantee minimality but which have been employed for the rough set methodology include backward-elimination (similar to hill-climbing) [40], compound selection [155], and stochastic selection [19]. However, where the discovery of minimal reducts is necessary, this approach may not be acceptable, and this has frustrated efforts to apply the rough set methodology to application domains which involve large numbers of features and relatively few objects [108] such as gene expression data.

There are various search techniques and heuristics which can be used to alleviate this problem however. Genetic algorithms (GAs) are an obvious candidate for this type of problem, and indeed the work in [93], and [238] employ such techniques to search for minimal reducts. Although such techniques cannot guarantee minimality, they do offer an alternative which will avoid local minima. Problems may arise when employing GAs for situations where the number of data attributes is high, as the amount of time taken to discover reducts may increase considerably.

Another approach similar to GA is particle-swarm-optimisation (PSO) [233], which does not require operations such as crossover and mutation, but primitive and simple mathematical operators, and is also efficient in terms of time/space complexity. Again, PSO will not guarantee minimality of any reducts discovered but like GAs allows the search to escape local minima. Other techniques similar to GA and PSO include ant-colony-optimisation (ACO) [93], [95], [102] and simulated annealing [93]. The approach in [262] also offers an interesting insight into possible heuristics for finding minimal reducts.

The only way in which to ensure minimality is to conduct a complete search of all possible reducts. An exhaustive search is an example of a complete search, but it does not necessarily follow that a complete search must be exhaustive. A branch-and-bound search [159] is typical of a complete search that is non-exhaustive, others include Boolean propositional satisfiability (SAT) [43]. In [98] the authors use a SAT solver algorithm [43] to perform a complete search for rough set reducts. The SAT algorithm can be used to perform a complete search of the feature space and thus discover minimal reducts. This technique is both computationally efficient and can guarantee the minimality of any discovered reduct. One of the principal drawbacks of SAT however, is that it can only be applied to discrete data domains.

2.4 Rough Set Extensions

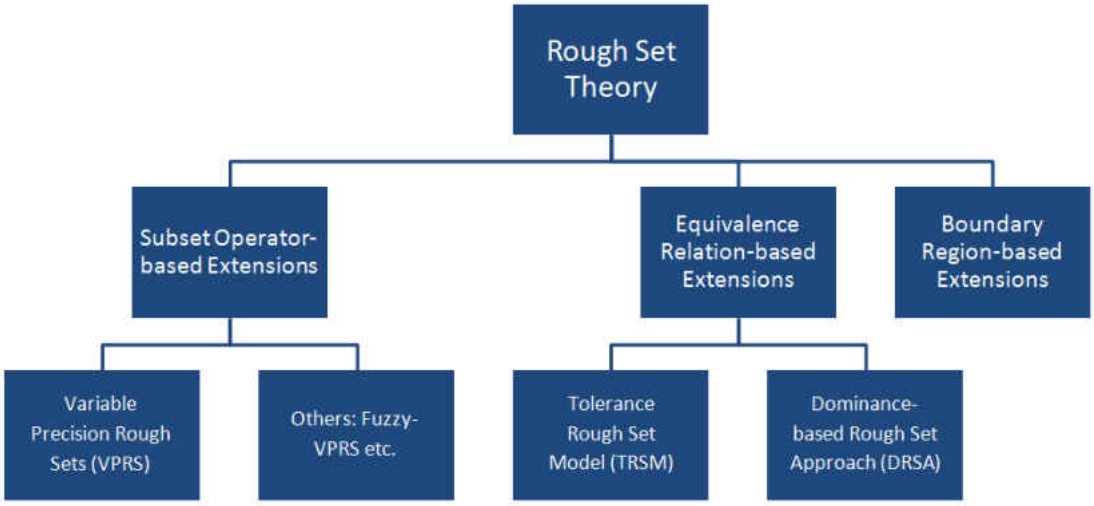

Figure 2.15: A taxonomy of rough set extensions

The conceptual simplicity of the rough set approach is undoubtedly one of the main reasons for its success. The two areas which are most often exploited in order to extend the approach are the equivalence relation, and the subset operator. These aspects are therefore the subject of a number of extensions. In addition to these extensions, there is also a third aspect of RST which has been exploited, that of the use of the information contained in the boundary region, or region of uncertainty. The illustration in Fig. 2.15 shows the main RST extensions in relation to the aspects of the theory they extend. The approaches are discussed here with reference to their underlying concepts as well as their respective merits and drawbacks.

2.4.1 Tolerance Rough Sets

The tolerance rough set model (TRSM) [208] can be useful for application to real-valued data. TRSM employs a similarity relation to minimise data as opposed to the indiscernibility relation used in classical rough-sets. This allows a relaxation in the way equivalence classes are considered. The effect of employing this relaxation, means that the granularity of the rough equivalence classes has been blurred slightly. Fig. 2.16 attempts to demonstrate this, and also how tolerance classes are allowed to 'overlap', whereas traditional RST lacks the ability to consider equivalence classes in this manner. This flexibility enables a change to occur in the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one so-called tolerance class which is TRSM equivalent of a rough set equivalence class.

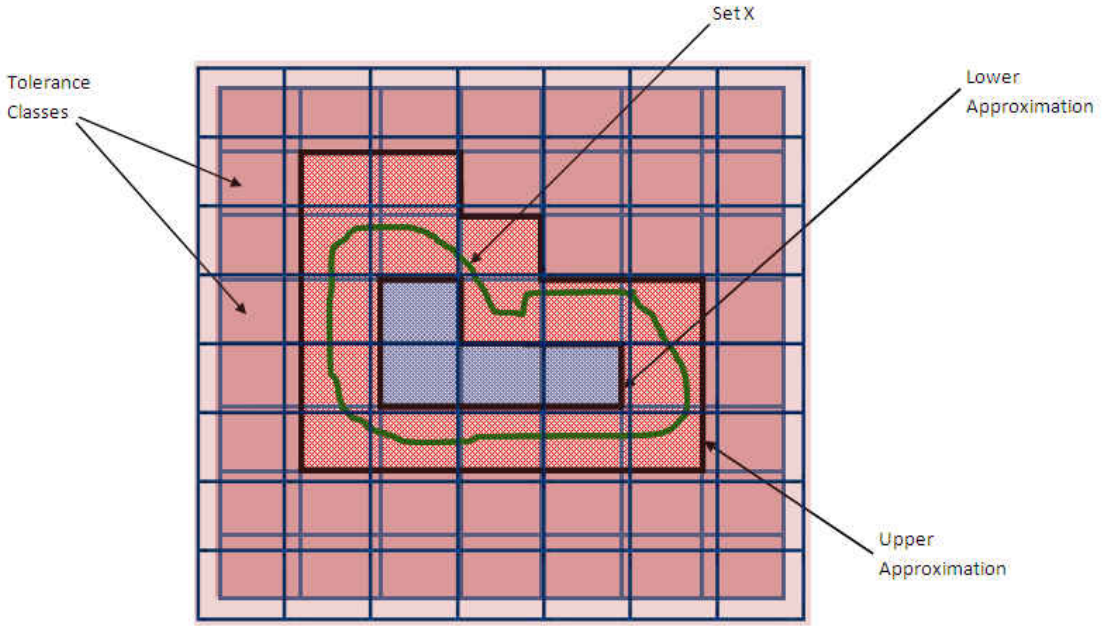

Figure 2.16: Tolerance rough set model

The tolerance threshold (τ) is a global similarity threshold and determines the required level of similarity for inclusion within a tolerance class. The specification of this threshold however is a departure from the traditional rough set approach, which relies only upon the information contained in the data. The framework also allows for the specific case of traditional rough sets by defining a suitable similarity measure (e.g. complete equality of features and the use of equation (15)) and threshold ($\tau = 1$). Further similarity relations are summarised in [163], but are not included here. From this, the tolerance classes that are generated by a given similarity relation for an object x are defined as:

$$SIM_{P,\tau}(x) = \{y \in U \mid (x, y) \in SIM_{P,\tau}\} \tag{2.32}$$

Lower and upper approximations are defined in a similar way to those of traditional rough set theory:

$$\underline{P_\tau} X = \{x \mid SIM_{P,\tau}(x) \subseteq X\} \tag{2.33}$$

$$\overline{P_\tau} X = \{x \mid SIM_{P,\tau}(x) \cap X \neq \emptyset\} \tag{2.34}$$

The tuple $\langle \underline{P_\tau} X, \overline{P_\tau} X \rangle$ is known as a tolerance rough set [208]. Using this, the positive region and dependency functions can be defined as follows:

$$POS_{P,\tau}(Q) = \bigcup_{X \in U/Q} \underline{P_\tau} X \tag{2.35}$$

$$\gamma_{P,\tau}(Q) = \frac{\mid POS_{P,\tau}(Q) \mid}{\mid U \mid} \tag{2.36}$$

These definitions are analogous to the traditional rough set concepts and can be applied in the same way as demonstrated in Section 2.3.1.1.

2.4.2 Variable Precision Rough Sets

The variable precision rough sets (VPRS) approach [263] extends rough set theory by relaxing the subset operator. It was originally proposed in order to analyse and identify data patterns which represent statistical trends rather than those which are functional. At the heart of VPRS, is the idea of allowing objects to be classified with an error smaller than a given predefined level or threshold. The introduction of this threshold means that unlike the traditional rough set approach, VPRS requires additional information other than that contained within the data.

Let $X, Y \subseteq \mathbb{U}$, the relative classification error is defined by:

$$c(X, Y) = 1 - \frac{\mid X \cap Y \mid}{\mid X \mid} \tag{2.37}$$

Note that $c(X, Y) = 0$ if and only if $X \subseteq Y$. A degree of inclusion can therefore be achieved by allowing a certain level of error, β, in classification:

$$X \subseteq_\beta Y \iff c(X, Y) \leq \beta,\ 0 \leq \beta \leq 0.5 \tag{2.38}$$

Thus by replacing $\subseteq$, with the operator $\subseteq_\beta$, the β-upper and β-lower approximations can be formulated:

$$\underline{R_\beta} X = \{x|\ [x]_R \subseteq_\beta X\} \tag{2.39}$$

$$\overline{R}_\beta X = \{x|\ c([x]_R, X) < 1 - \beta\} \tag{2.40}$$

Note that when $\beta = 0$, $\underline{R_\beta} X = \underline{R} X$.

Using this extension, the positive, negative and boundary regions can now also be defined:

$$POS_{R\beta}(Q) = \bigcup_{X \in U/Q} \underline{R_\beta} X \tag{2.41}$$

$$NEG_{R\beta}(Q) = U - \bigcup_{X \in Q} \overline{R}_\beta X \tag{2.42}$$

$$BND_{R\beta}(Q) = \bigcup_{X \in Q} \overline{R}_\beta X - \bigcup_{X \in Q} \underline{R}_\beta X \tag{2.43}$$

A more comprehensive investigation of reducts for the VPRS approach may be found in [11], [12], and [113]. No general comparative studies appear to have been carried out with regard to comparing the rough set and the VPRS methods, although in [223], the authors compare feature selection methods based on both RST and VPRS.

As indicated previously, the VPRS approach requires the specification of an additional parameter (β). This parameter can be approximated by repeated experimentation. However, problems may arise if searching for true reducts, as the VPRS approach incorporates an element of inaccuracy in determining the number of classifiable objects.

2.4.3 Dominance-based Rough Sets

The Dominance-based Rough Set Approach (DRSA) [62] is an extension of RST for multi-criteria decision analysis. In contrast to traditional RST, DRSA employs a dominance relation instead of an equivalence relation. This allows DRSA to deal with the inconsistencies which are typical of criteria and preference-ordered decision classes.

The ordering of data describing decision situations is naturally related to preferences of considered condition and decision attributes. Traditional RST does not have the ability to deal with ordinal data in the same way that DRSA does. This is because DRSA employs a dominance relation in place of the traditional rough set equivalence relation.

In DRSA, data is represented in decision table form. Let $S = \langle \mathbb{U}, Q, V, f \rangle$, where $\mathbb{U}$ is a non-empty set of finite objects, Q is a finite set of criteria, $V = \bigcup_{q \in Q} V_q$ where V_q is the set of values that the criterion q can take, and $f : \mathbb{U} \times Q \rightarrow V$ is an information function such that $f(x, q) \in V_q$ for every $(x, q) \in \mathbb{U} \times Q$. The set Q consists of condition criteria $\mathbb{C}$, and the decision class $\mathbb{D}$. Note that $f(x, q)$ is the evaluation of object x, on criterion $q \in \mathbb{C}$, while $f(x, d)$ is the decision class assignment for that object.

In order for DRSA to operate effectively on preordered data, the approach employs an 'preferencing' or 'outranking' relation. A typical example is: $\succeq_q; x \succeq_q y$ which means that x is preferential to or 'outranks' y with respect to q. The values that q can take is a subset of real numbers - $\mathbb{R}$, such that $V_q \subseteq \mathbb{R}$, and the

preference relation is a simple order between real numbers $\geq$ such that $x \succeq_q y \iff f(x,q) \geq f(y,q)$ holds. This relation is straightforward for simple maximisation criterion, e.g. an exam result - 'the higher, the better'. For criteria where the opposite is true, e.g. student failure-rate ('the less, the better'), the relation can be satisfied by negated values of V_q. Let $P \subseteq \mathbb{C}$, it can be said that x dominates y, denoted by xD_py, if x is 'better' than y for every criterion from P, $x \succeq_q y$, $\forall q \in P$. For each $P \subseteq \mathbb{C}$, the dominance relation D_P is reflexive and transitive. Given that $P \subseteq \mathbb{C}$ and $x \in \mathbb{U}$,

$$D_P^+(x) = \{y \in \mathbb{U} \colon yD_px\} \tag{2.44}$$

$$D_P^-(x) = \{y \in \mathbb{U} \colon xD_py\} \tag{2.45}$$

These are termed the P-dominating set and P-dominated set respectively.

As the DRSA deals with ordinal data and objects, the manipulation of the data is carried out with respect to the ranking of decision classes. Let $T = \{1, \ldots, n\}$. The domain values of decision criterion, V_d consists of n elements (it is assumed that $V_d = T$) and induces a partition of $\mathbb{U}$ into n classes $Dc = \{Dc_t, t \in T\}$, where $Dc_t = \{x \in \mathbb{U} \colon f(x,d) = t\}$. Each object $x \in \mathbb{U}$ is assigned to only one decision class $Dc_t, t \in T$. All of the classes are preference-ordered according to an increasing order of class indices, i.e. $\forall r, s \in T \mid r \geq s$, objects from Dc_r are preferential to the objects from Dc_s. Thus the upward and downward unions of classes, can be defined respectively, as:

$$Dc_t^{\geq} = \bigcup_{s \geq t} Dc_s \qquad Dc_t^{\leq} \bigcup_{s \leq t} Dc_s \qquad t \in T \tag{2.46}$$

In DRSA, the knowledge being approximated is a collection of upward and downward unions of decision classes. The knowledge granules employed for approximation in DRSA are the P-dominating and P-dominated sets, these are analogous to the equivalence classes of traditional RST. The P-lower and the P-upper approximation of $Dc_t^{\geq}, t \in T$ are denoted $\underline{P}(Dc_t^{\geq})$ and $\overline{P}(Dc_t^{\geq})$, respectively, and can be defined as follows:

$$\underline{P}(Dc_t^{\geq}) = \{x \in U \colon D_P^+(x) \subseteq Dc_t^{\geq}\} \tag{2.47}$$

$$\overline{P}(Dc_t^{\geq}) = \{x \in U \colon D_P^-(x) \cap Dc_t^{\geq} \neq \emptyset\} \tag{2.48}$$

Similarly, the P-lower and the P-upper approximation of $Dc_t^{\leq}$, denoted $\underline{P}(Dc_t^{\leq})$ and $\overline{P}(Dc_t^{\leq})$, respectively, can be defined thus:

$$\underline{P}(Dc_t^{\leq}) = \{x \in U : D_P^-(x) \subseteq Dc_t^{\leq}\} \tag{2.49}$$

$$\overline{P}(Dc_t^{\leq}) = \{x \in U : D_P^+(x) \cap Dc_t^{\leq} \neq \emptyset\} \tag{2.50}$$

As with traditional RST, the boundary regions of $Dc_t^{\geq}$ and $Dc_t^{\leq}$ can also be defined:

$$BND_P(Dc_t^{\geq}) = \overline{P}(Dc_t^{\geq}) - \underline{P}(Dc_t^{\geq}) \tag{2.51}$$

$$BND_P(Dc_t^{\leq}) = \overline{P}(Dc_t^{\leq}) - \underline{P}(Dc_t^{\leq}) \tag{2.52}$$

2.4.4 Vaguely Quantified Rough Sets

In traditional RST, an object is a member of the upper approximation of a set if it is related to one of the elements in the set, while the lower approximation only retains those objects related to all the elements in the set. This is a result of the use of an existential quantifier in the definition of the upper approximation, and the use of a universal quantifier for the lower approximation. For real-world data which includes noise to a greater or lesser degree, this approach will inevitably suffer from classification errors and inconsistency. The associated definition of the upper approximation may be too general (thus resulting in very large sets), while the definition of lower approximation might be too rigid (resulting in an empty set in the extreme case). Fuzzy rough set theory (which is covered in the next section), exhibits similar behaviour where the quantifiers $\exists$ and $\forall$ are replaced by the *sup* and inf operations [36]. These operators however, can be equally as susceptible to the effects of noise as their crisp counterparts.

As demonstrated previously in Section 2.4.2, thresholds are introduced in VPRS to deal with these problems for the crisp case. In general, given $0 < l < u < 1$, an element y is added to the lower approximation of a set A if at least (100 x u) % of the elements related to y are in A. Likewise, y belongs to the upper approximation of A if more than (100 x l) % of the elements related to y. This can be interpreted as a generalisation of the rough set model using crisp quantifiers *at least (*100 *x u) %*, and *more than (*100 *x l) %* to replace the universal quantifier which demands rigid (at least 100%) membership for an element to be included in the lower approximation, and the existential quantifier which demands membership which is non-zero (greater than 0%) for an element to be included in the upper approximation.

In perhaps what is one of the most recent extensions of rough sets, the authors

of [36] introduce vague quantifiers like 'most' and 'some' to the rough set model. As a result of this, an element y now belongs to the lower approximation of A if most of the elements related to y are included in A. Similarly, an element belongs to the upper approximation of A if some of the elements related to y are included in A. Also, the vague quantifiers are modeled mathematically in terms of the notion of fuzzy quantifiers in [254], so not only does the VQRS model inherit the flexibility of VPRS for dealing with classification errors mentioned previously, but also that of fuzzy sets for the expression of partial constraint satisfaction - by distinguishing between varying levels of membership of both the upper and lower approximations.

The definitions used for the upper and lower approximations in VPRS can be relaxed through the use of vague quantifiers, to express that y belongs to the upper approximation of the set X to the extent that some elements of y's equivalence class (Ry) are in the set A, and y belongs to the lower approximation of A to the extent that most elements of Ry are in X. In VQRS, it is implicitly assumed that the approximations are fuzzy sets, i.e. mapped from X to $[0,1]$, that evaluate the degree to which the associated condition is fulfilled. The concept of a fuzzy quantifier in [254] is employed, i.e. a $[0,1] \rightarrow [0,1]$ mapping Q. The set Q is said to be regularly increasing, if it is increasing *and* it satisfies the boundary conditions $Q(0) = 0$ and $Q(1) = 1$. Examples of fuzzy quantifiers can be generated by means of the following parameterised formula, for $0 \leq \alpha < \beta \leq 1$, and $x \in [0,1]$,

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \tag{2.53}$$

For instance, $Q_{(0.1,0.6)}$ and $Q_{(0.2,1)}$ may be used to reflect the vague quantifiers *some* and *most* respectively from natural language.

The VQRS upper and lower approximations can be defined once the quantifier pair (Q_l, Q_u) has been fixed such that:

$$\mu_{\underline{R_P}X}^{Q_u}(y) = Q_u\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{2.54}$$

$$\mu_{\overline{R_P}X}^{Q_l}(y) = Q_l\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{2.55}$$

In other words, an element y belongs to the lower approximation of X if most of the elements related to y are included in X. Likewise, an element belongs to the upper approximation of X if some of the elements related to y are included in

X. Notice that when X and R_P are a crisp set and a crisp equivalence relation respectively, the approximations may still be non-crisp because of the use of vague quantifiers.

2.4.5 Other Rough Set Extensions

As mentioned previously, perhaps one of the most appealing aspects of traditional RST lies in its simplicity. It is based on straightforward set operations, and is computationally efficient. Examining the concepts described earlier in Section 2.3.1, the most obvious areas for further exploration and extension are the equivalence relation and the subset operator, both of which are extended by the VPRS/VQRS[263], [36] and TRSM/DRSA [208], [62] approaches respectively. One possible avenue for further exploration which has not been examined previously lies in a variable precision tolerance rough set approach. Although this would involve the specification of two parameters, it could take advantage of the benefits offered by both models: the ability to deal with real-valued data from TRSM and the ability to handle noise from the VPRS approach.

There is also one further aspect of RST that is often overlooked however; the upper approximation concept and its potential contribution to improving the performance of the rough set model. Work in this area has included an approach which generates reducts which preserve the rough upper approximation [87], as well as an approach which considers the upper approximation and proposes a feature selection algorithm based on a rough upper approximation measure [47].

Other techniques such as those presented in [84], [148], and [144] consider the positive and boundary regions as conceptually different entities, and attempt to use the boundary region information for both feature selection and classification.

In particular in [84], the authors employ a consistency measure for feature selection in order to determine the classification of objects in the rough set boundary region and use this information to search for reducts. The approach uses a greedy-type search to select attributes which result in the greatest increase in the consistency value. Problems may arise however, if the data on which the approach is operating is inconsistent, in these cases a stopping threshold must be specified to avoid overfitting.

The approach in [148] however treats the data in the same way as that of traditional crisp RST. The central idea of this approach is that from an intuitive point-of-view objects in the boundary region of a given concept are more likely to belong to that concept if those objects are *close* to the objects of the positive region. Thus, a distance measure is employed to determine the *'closeness'* or proximity of boundary region objects to those objects in the positive region. This

proximity information is then used in feature selection as a measure to determine the 'goodness' or value of potential reducts.

An approach which examines the boundary region of tolerance rough sets (and thus can also handle real-valued data) based on [148] has also been proposed [144]. Also in [162] the authors discuss what they term 'approximate reducts', based on exploiting the rough set boundary. However the work does not outline their application.

Another interesting idea which is explored in [210] and [211] is the re-definition of the upper and lower approximation concepts of RST. The definitions propose the use of fuzzy similarity, and tolerance, as opposed to indiscernibility, although otherwise the framework remains unchanged from that of traditional RST. Similar treatment is also given by the authors in [257] to VPRS to extend the β-upper and β-lower approximations, however only similarity is explored in this case.

2.5 Combining Rough Sets with Other Techniques

The combination of RST with other soft computing techniques to form hybrid systems has highlighted the value of employing RST as a part of a wider framework for improving the overall performance of such systems. Such hybrids include the combination of RST with neural networks, genetic algorithms (GAs), evolutionary algorithms, and fuzzy sets. Very significantly, there is the hybridisation of rough sets and fuzzy sets to form fuzzy-rough set theory.

2.5.1 Rough Set Hybridisation

It has been demonstrated that RST can be very effective for preprocessing data input for neural networks [91]. More recent work [133] compared the rule extraction capabilities of both rough sets and neural networks and hybrid methods with ID3 [184]. The work of [246] further reinforces the utility of employing RST either as a neural network's preprocessor or as a combined inference mechanism for medical diagnosis and tested on a hepatitis disease dataset. Another approach for medical image classification is reported in [199] that uses RST as a dimensionality reduction step prior to the application of a neural networks based classifier. Further detail with regard to the use of rough sets and hybrid methods for medical applications can be found in [170].

In [123], a hybrid rough set and neural networks approach for rule induction is presented. This technique is applied to relatively large data sets in order to generate more concise and accurate rules than either neural networks or rough sets alone. A feature selection algorithm is proposed and rules are generated from

a decision table based on the rough set discernibility matrix. Reducts and rules are obtained using RST with neural networks employed to remove noisy data. Other rough set/neural network hybrid approaches are also to be found in [91], [151], [218], and [230]. Additionally, it has been demonstrated that rough sets can help to generate new models of neurons in [124], and [125].

A review of the hybridisation of RST with genetic algorithms (GAs) is documented in [35]. Prior to this, the first hybridisation based on lower and upper bounds of numeric ranges was proposed as a rough-genetic algorithm in [126]. Others include: genetic encoding in order to generate rough set representations of clusters [129], and a hybrid decision support system for cancer detection [152]. Genetic programming has also been allied with rough sets for bankruptcy classification [135].

RST has also been hybridised with classical statistical methods such as Principal Component Analysis (PCA) [220], Bayesian methods [219], or wavelets [229]. Such integration has resulted in classifiers of better quality than those constructed through the use of RST alone [22].

In terms of hybridising rough set extensions, a number of approaches have been proposed such as fuzzy-rough VPRS [190], dominance based rough sets and VPRS [81]. An interesting idea that has not yet been explored is a VPRS and TRSM hybrid. This would allow the flexibility to deal with real-valued data inherited from the TRSM approach and the noise-tolerance of the VPRS method. This would mean the specification of two parameters however, which would involve significant experimentation in order to establish ideal values for a given set of data.

2.5.2 Fuzzy-Rough Sets

Fuzzy set theory (FST) was first proposed nearly forty four years ago [254] and RST will celebrate its twenty-eighth anniversary this year [172]. FST and RST complement one another [52] and much advantage has been taken of this fact. This is reflected in the breadth and depth of research which has been undertaken in this particular hybridisation of rough sets.

Note that fuzzy-rough sets should not be confused with existing approaches that directly combine the use of RST for dimensionality reduction and that of FST for knowledge modeling e.g [203]. Whilst successful in real-world applications, the underlying ideas of such work are straightforward and hence are omitted from the discussions below.

There have been two main lines of thought in the hybridisation of fuzzy and rough sets [128], the constructive approach and the axiomatic approach. A general

framework for the study of fuzzy-rough sets from both of these viewpoints is presented in [251]. For the constructive approach, generalised lower and upper approximations are defined based on fuzzy relations. Initially, these were fuzzy similarity/equivalence relations [52] but have since been extended to arbitrary fuzzy relations [251]. The axiomatic approach is primarily for the study of the mathematical properties of fuzzy-rough sets [240].

In [52], the authors define the fuzzy P-lower and P-upper approximations as follows:

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\} \qquad \forall i \tag{2.56}$$

$$\mu_{\overline{P}X}(F_i) = \sup_x \max\{\mu_{F_i}(x), \mu_X(x)\} \qquad \forall i \tag{2.57}$$

where F_i is a fuzzy equivalence class and X is the (fuzzy) concept to be approximated. The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is known as a fuzzy-rough set. Also in literature are definitions for rough-fuzzy sets [51], [214], which can be seen as a particular case of fuzzy-rough sets. A rough-fuzzy set is a generalisation of a rough set, derived from the approximation of a fuzzy set in a crisp approximation space. In [249] it is argued that, in order to remain consistent, the approximation of a crisp set in a fuzzy approximation space should be called a fuzzy-rough set, and the approximation of a fuzzy set in a crisp approximation space should be called a rough-fuzzy set, thus ensuring that both models are complementary. In this framework, the approximation of a fuzzy set in a fuzzy approximation space is considered to be a more general model, unifying both theories. However, most researchers consider the traditional definition of fuzzy-rough sets in [52] as standard. The specific use of *min* and *max* operators in the above definitions is expanded in [186], where a wide range of fuzzy-rough sets are constructed, with each member represented by a particular implicator and t-norm. The properties of three typical implicators are investigated. Further investigations in this area can also be found in [44], [224], [241], [251].

In [17], [157], an axiomatic approach is taken, but restricted to fuzzy T-similarity relations (and hence fuzzy T-rough sets). The properties of generalised fuzzy-rough sets are investigated in [239], and a pair of dual generalised fuzzy approximation operators are defined based on arbitrary fuzzy relations. The approach presented in [150] introduces definitions for generalised fuzzy lower and upper approximation operators determined by a residual implication. Assumptions are found that allow a given fuzzy set-theoretic operator to represent a lower or upper approximation from a fuzzy relation. Different types of fuzzy relations produce different classes of fuzzy-rough set algebras.

The work in [187] generalises the fuzzy-rough set concept through the use of residuated lattices. An arbitrary residuated lattice is used as a basic algebraic structure, and several classes of lattice-valued fuzzy-rough sets (a fuzzy-rough hybridisation of L-fuzzy sets)and their properties are investigated. In [28], a complete completely distributive (CCD) lattice is selected as the foundation for defining lower and upper approximations in an attempt to provide a unified framework for rough set generalisations. It is demonstrated that the existing fuzzy-rough sets are special cases of the approximations on a CCD lattice for T-similarity relations. The relationships between fuzzy-rough set models and fuzzy topologies on a finite universe have been investigated. The first such research was reported in [17], where it was proved that the lower and upper approximation operators were fuzzy interior and closure operators respectively for fuzzy T-similarity relations. The work carried out in [251] investigated this for arbitrary fuzzy relations. In [183], [242] it was shown that a pair of dual fuzzy rough approximation operators can induce a topological space if and only if the fuzzy relation is reflexive and transitive. The fuzzy interior (closure) operator, is also examined.

In addition to the previous approaches to fuzzy-rough hybridisation, other generalisations are possible. One of the first attempts at hybridising the two theories is reported in [243], where rough sets are expressed by a fuzzy membership function to represent the negative, boundary and positive regions. All objects in the positive region have a membership of one and those belonging to the boundary region have a membership of 0.5; whilst those of the negative region have a membership of 0 as they do not belong to the set of interest. Thus, in adopting this approach a rough set can be defined using FST. This also means that the rough set operators of union and intersection are modified accordingly. In [177] the author attempts to address the problem where the fuzzy set representation of a rough set may be too-precise, such that a concept is described exactly once its membership function has been defined. The solution to this is to employ an approximation of a family of fuzzy sets which the author terms a *shadowed set*. Shadowed sets do not use exact membership values but instead use truth values and a zone of uncertainty. A similar approach to that of [243] is applied where elements may belong to a set with certainty (membership value 1), possibility (unit interval), or not belong (membership value 0). These ideas of course correspond to the rough set positive, boundary and negative regions respectively.

Another approach is reported in [31] where the rough set lower approximation is employed, and elements are allowed to belong to this with certainty, however the boundary region or uncertain region is fuzzified and membership values of elements are expressed in terms of a fuzzy membership function. The authors of [190] apply a fuzzy-rough sets extension to the VPRS model described in Section

2.4.2 in an attempt to capitalise on the advantages of both rough sets and fuzzy sets within the VPRS framework. However, the VQRS approach of [36] as detailed in Section 2.4.4 also takes advantage of these in a single approach as it employs fuzzy quantifiers and extends the VPRS approach simultaneously.

The interest in the hybridisation of fuzzy sets with rough sets is borne out by the level of publication in this area. The marriage of these approaches has resulted in methods which take advantage of the ability of rough sets to model vagueness and that of fuzzy sets to model uncertainty. In this sense both approaches are complimentary, furthermore when hybridised as described in this section no tunable parameters are required and only the data is used. There is much scope for further research in relation to the development of fuzzy-rough sets. In particular there is much interest in the area of type-2 fuzzy sets [255] at the present moment. However, a hybridisation with rough sets has not been proposed as yet. Additionally, there are a number of aspects in respect of fuzzy measures with application to fuzzy-rough sets which remain unexplored, and these may offer some new and interesting research areas.

2.6 Applications

In this section a number of theoretical and real world application areas of RST, rough set extensions, and fuzzy-rough set theory are examined. Note that these examples are for representative purposes and do not serve to demonstrate the whole spectrum of possible applicable areas. The sheer number of applications and amount of work that has been published in the area means that it would be impossible to cover all areas in sufficient depth. Therefore, in this chapter three important areas of machine learning have been chosen for close examination; classification, clustering, and feature selection. A review of each of these areas is documented in the following sections. In each section a further subsection is devoted to an example real-world application.

2.6.1 Classification

Classification concerns any problem in which a decision is taken or a forecast is made on the basis of available knowledge or information. A classification algorithm allows repeated forecasts to be made with regard to accumulated knowledge for new situations. Such algorithms can then be applied in order to classify previously unseen objects. Each new object can be assigned to a predefined set of classes, based on the observed values of suitably chosen attributes or features.

It is interesting to note that, despite the level of interest in rough set classi-

fication which is borne out by the number of publications in the area, no comprehensive survey of rough classification has been published to date. Perhaps this is due in part to the fact that RST is often married with other approaches when applied to the classification problem. Nevertheless, a number of RST-based classifiers have been proposed. The first application of RST to the classification problem is demonstrated in [173]. The authors of [176], [207], and [212] discuss the fundamentals of rough set rule induction for classification, but no algorithms are proposed.

The earliest RST-based classification algorithm is described in [174]. Later examples were proposed in [9], and [46], although the latter focused on database mining. Much use has been made of rough classifiers which were integrated into the learning from examples based on rough sets (LERS) framework [63], [64]. In these methods, descriptions of concepts are constructed through the calculation of all reducts for a given dataset, by means of the decision rules. In [8], it is argued that these methods are not appropriate for classifying unseen data, thus a number of rough set classification methods are proposed which address this problem. Additionally, some new methods for rule induction from reducts, as well as ways of dealing with real-valued data discretisation are also described (also within the LERS framework). Similar aspects are also examined in [65] and [66]. Other research such as [215] also concentrates on addressing some of the shortcomings of the use of rough sets for rule induction as an aid to classification.

Rough set extensions have also been employed for classification. In [264], the author discusses the use of VPRS for building decision tables from data models. Others which also employ VPRS include [61] and [259] for email spam filtering, and general classification [256]. In [231] the authors have combined VPRS with fuzzy clustering techniques to discover rules in process planning. In the same way that VPRS has been applied to the classification task, so too has the tolerance rough set model (TRSM), and a number of papers have been published in this area. Applications include handwriting classification [104], web document classification [250], geographical land classification [253]. Although a relatively new approach, VQRS has also been applied to the classification of mammographic data (see Section 2.6 for further detail) [141]. The DRSA has also been employed for rule induction [200], and classification [111] albeit with application to ordinal data.

Initial attempts to use fuzzy-rough sets for classification were presented in [193], which adopted a nearest neighbour (NN) type classifier approach. This approach attempted to handle both the fuzzy uncertainty due to overlapping classes and the rough uncertainty caused by lack of informative features. A fuzzy-rough ownership function (a value which is influenced by all training objects) was employed in an effort to capture both of the aforementioned aspects. Additionally,

this also allows a possibilistic class membership assignment. The ownership function is influenced by all of the objects in the training set, this in turn means that the number of neighbours does not need to be defined. Other parameters must however be specified for successful operation. In [234], the authors extend the approach but divide the task of classification into four parts. Firstly, using leave-one-out type of strategy the fuzzy-rough ownership value is calculated for each training object for all classes. The ownership value indicates the degree to which other objects support each individual object. Inconsistencies are then filtered from the training data - a high fuzzy-rough ownership value indicates a class other than a known class. Following this, representative points are selected from the processed training data and fuzzy-rough ownership values are refreshed based on mountain clustering. Then, finally test objects are classified using only the representative training data from the previous step using the algorithm proposed in [193].

Other NN classification methods which employ fuzzy-rough hybridisation include [15] which integrates rough uncertainty into the fuzzy kNN classifier using the definitions of fuzzy upper and lower approximations as defined in [52]. The membership of a test object to the upper and lower approximations for every class is determined by k nearest neighbours. Also, a similar approach is used in [141], once again the fuzzy-rough upper and lower approximations are used to determine the membership of test objects to a particular class.

Little research has taken place in the area of fuzzy-rough decision tree induction, although there is much interest in fuzzy decision trees because of their ability to model vagueness. The work on fuzzy-rough decision trees outlined in [14] employs the fuzzy-rough ownership measure from [193] which is used to define a 'fuzzy-roughness' measure and fuzzy-rough entropy measure. The node splitting criterion is determined using the fuzzy-rough entropy measure. In [98] a fuzzy decision tree algorithm based on the well-known fuzzy ID3 approach is described. In this case, fuzzy-rough dependency is employed to decide when node splitting should occur. An approach for rule induction using fuzzy rough sets is proposed in [80] for generating certain and possible rulesets from hierarchical data.

2.6.1.1 Image Data Analysis for Mammographic Risk Assessment

Breast cancer is a major health issue, and the most common amongst women in the EU. It is estimated that 8–13% of all women will develop breast cancer at some point during their lives. Furthermore, in the EU and US, breast cancer is attributed as the leading cause of death of women in their 40s. Although increased incidence of breast cancer has been recorded, so too has the level of early

detection through screening in order to assess the risk of developing cancer using mammographic imaging and expert opinion. However, even expert radiologists can sometimes fail to detect a significant proportion of mammographic abnormalities. In addition, a large number of detected abnormalities are usually discovered to be benign following medical investigation. Existing mammographic Computer Aided Diagnosis (CAD) systems concentrate on the detection and classification of mammographic abnormalities. As breast tissue density increases however, the effectiveness of such systems in detecting mammographic abnormalities is reduced significantly. Also, it is known that there is a strong correlation between mammographic breast tissue density and the risk of development of breast cancer. Automatic classification which has the ability to consider tissue density when searching for mammographic abnormalities is therefore highly desirable.

The approach in [141] describes the application of a number of rough and fuzzy-rough approaches for dealing with mammographic risk assessment data. The objective of this analysis is to determine the risk of developing cancer by classifying each woman or mammogram according to a consensus class which has been agreed upon by three expert radiologists. The actual approach employs a fuzzy-rough framework. There are three steps: feature extraction to extract the features from the raw image data, feature selection which removes noisy irrelevant or redundant features from those extracted features, and classification to classify the mammograms into one of four predefined classes. The work here focuses on a brief review of the fuzzy-rough sets based classification step.

Efficient, and in particular accurate classification of mammographic imaging is of high importance. Any improvement in accuracy for automatic mammographic classification systems can result in a reduction in the amount of required expert analysis thus improving the time taken to perform breast abnormality risk assessment. Also, by reducing inter-expert variation the resulting automatic risk assessments can be more accurate. The data in mammographic imaging is real-valued and can also be noisy. Clearly, any classifier employed must therefore have the ability to deal with such data. Discrete methods require that the real-valued data is discretised and thus may result in significant information loss, however the methods described here require no discretisation, and are based on fuzzy-rough set theory which uses only the information contained within the data.

The fuzzy-rough classifier shown in [141] and [92] is based on the nearest neighbour (NN) classifier technique. It works on the basic principle that the lower and the upper approximations of a decision class, calculated by means of the nearest neighbours of a test object y, provides good clues in order to predict the membership of the test object to that class. The membership of a test object y to each (crisp or fuzzy) decision class is determined via the calculation of the fuzzy

lower and upper approximation. The algorithm outputs the decision class with the resulting best fuzzy lower and upper approximation memberships. The complexity of the algorithm is $O(|C| \cdot (2|\mathbb{U}|))$. Note that, although a value for the parameter k that is employed in the traditional kNN method is not required it can be incorporated into the algorithm.

The FRNN approach is applied to two mammographic imaging datasets, which have been labeled with the consensus opinion of 3 experts. The FRNN algorithm was compared against several other algorithms including fuzzy nearest neighbour [103], a fuzzy-rough nearest neighbour FRNN-O [194] (based on the measure in [193]), and an approach based of VQRS [36] - VQNN vaguely quantified nearest neighbour. The classification accuracies are obtained using 10 x 10-fold cross validation. The FRNN approach performs well compared with the other classifiers achieving accuracies of 91.2% compared with 75.12% for FNN, 82.1% for FRNN-O, and 72% for VQNN for the first dataset. Values for the second dataset also show that FRNN performed better than all of the other approaches [141].

2.6.2 Clustering

The clustering task is the unsupervised classification of data objects (patterns observations, data vectors) into groups or clusters. Clustering has been addressed in many contexts and by researchers of many different disciplines, and this reflects its applicability and popularity as an important step in data analysis. Since both cluster analysis and RST form data groups, it is easy to see the conceptual similarity between the upper and lower approximation constructs of rough sets, and formation of data clusters or groups. This similarity has meant that the rough sets lend themselves easily to the clustering problem. A further advantage that RST offers is that it may also provide scope for the discovery of 'possible' data clusterings through the use of the information contained in the rough set boundary region.

Much of the interest in rough clustering has been relatively recent [76], [77], [180]. The application of rough sets to clustering is not limited to the use of rough indiscernibility [77]. For instance a rough set version of the classical k-means algorithm is proposed in [129]. Similarly in [127], Kohonen SOM (self-organising-maps) were used to generate intervals of clusters based on RST. The authors of [134], propose a rough set clustering algorithm by combining entropy-based thresholding with rough sets

The use of VPRS within the framework of the fuzzy c-means (FCM) algorithm [13], [54] is documented in [7] where VPRS is employed to assign weights to each of the features. The basis for the approach is VPRS but an extension is proposed

for the variable precision fuzzy-rough case. This is demonstrated by applying it to image analysis. VPRS is also used along with fuzzy-rough sets in [261] as part of a fault diagnosis system. As an aid to fuzzy clustering in the general case in [235], VPRS is employed for generating rules from the fuzzy conditional and decision constructs of the fuzzy clustering algorithm. Although not as popular as traditional RST or VPRS, TRSM has been applied to the clustering problem in [101], and [78], where the authors employ an algorithm to cluster documents. Later work [161], also used TRSM in a similar manner for clustering web search results. The traditional rough set approach is extended in [114] by using a tolerance relation to form initial clusters, subsequent clusters are then formed using a constrained similarity relation which is also used as a merging criteria to combine initially identified clusters.

There have been few applications of fuzzy-rough set theory to clustering. Most approaches such as [235] (mentioned previously), and [258] have tended to use both FST and RST but in isolation rather than in terms of fuzzy-rough set theory. Rough-fuzzy sets are employed in [179] for texture separation in imaging, and in [169] the author also describes the application of rough-fuzzy sets for clustering and employs an image segmentation example to demonstrate this. In [31], the authors propose a fuzzy-rough extension of the well-known FCM clustering algorithm and apply it to network security intrusion detection. Another fuzzy-rough approach which is also based on FCM is proposed in [82]. It remains to be seen whether further fuzzy-rough approaches for clustering will be proposed, although it would seem that fuzzy-rough sets are well-suited for such problems.

2.6.2.1 Document Clustering

The clustering of documents is a difficult task for a number of reasons, mainly due to the textual characteristics and unstructured format that every individual document takes. In [79], the authors describe a method to cluster documents using tolerance rough sets. Two algorithms are described: one for hierarchical clustering and another for non-hierarchical clustering.

The approach can be broken down into two stages, the generation of tolerance classes, and the manipulation and generation of the clusters. In the first step shown below in Fig. 2.17, a set of terms (words) is extracted from each document, these are then assigned weights according to occurrence. Each individual term (t_i) is assigned a weight (w_i) which reflects its importance in the document; where $i = 1, 2, 3, ...n$ with n being the number of extracted terms. A document is denoted $d_j = (t_1, w_{1j}; t_2, w_{2j} \ldots; t_n, w_{nj})$ and $wiq \in [0, 1]$. The weights are calculated by means of a frequency function, such that terms that occur often have a lower

weight than those that rarely occur. This ensures that terms that occur in all documents have a zero weight. Each document is represented by a predefined number (R) of its highest weighted terms. All of the terms for all documents denoted T are used in a co-occurrence matrix to determine how terms are related to one another. Using an uncertainty function derived from a tolerance relation, this matrix can then be used to generate tolerance classes of terms in T. It is at this point that the tolerance value (τ) must be specified for the uncertainty function.

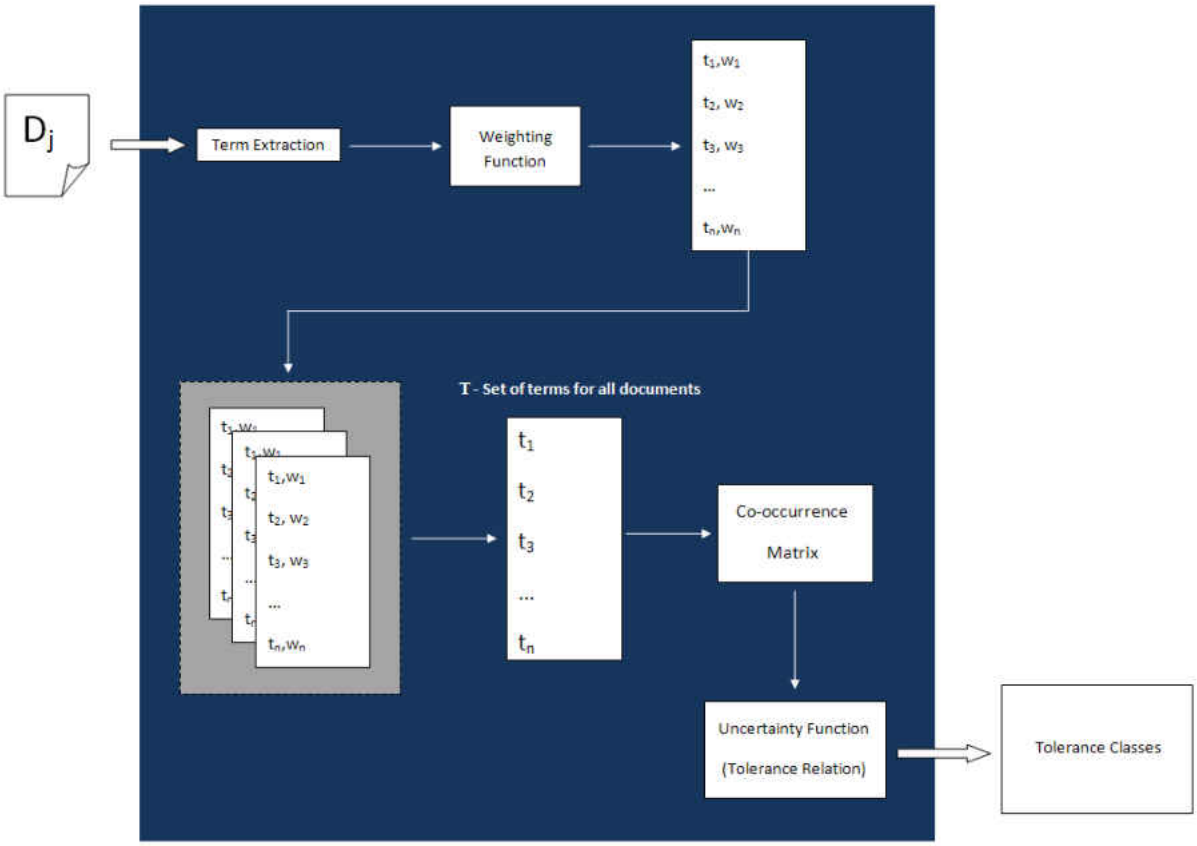

Figure 2.17: Document clustering using tolerance rough sets - stage 1

In the second stage of the approach shown in Fig. 2.18, a concept is defined which is used for the representation of clusters. This representation is what the authors term *polythetic* and must fulfill three properties which relate to the documents under consideration and the terms (words) in each document. Membership of each document to a cluster is defined in terms of a Bayesian minimum error rate and can be used to build each of the clusters. Cluster similarity is carried out in the usual manner, by employing a distance metric. It should be noted that clusters are built using only the upper approximation of the tolerance rough set calculated from a subset of terms $X \subseteq T$.

A number of experiments are conducted using both hierarchical and non-

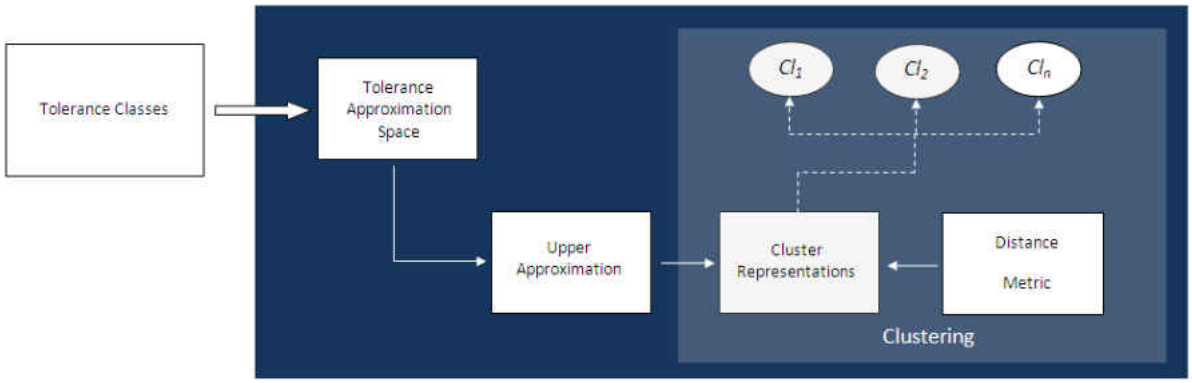

Figure 2.18: Document clustering using tolerance rough sets - stage 2

hierarchical clustering algorithms for both general clustering and information retrieval. In particular, the TRSM-based approach is compared with a vector space model (VSM) approach to clustering for information retrieval. The TRSM-based method demonstrates that it can equal or outperform the VSM method. This however requires that a range of tolerance values are specified for the uncertainty function.

It is interesting to note, there are a number of areas of this approach that could be covered by using fuzzy-rough set theory, thus eliminating the need for the subjective specification of not only the thresholding value of the TRSM but also of a number of other thresholds relating to the number of the terms, R that should be considered for each document.

2.6.3 Feature Selection

Feature selection (FS), which may also be referred to as attribute selection or semantics-preserving attribute reduction, is a term used to describe the problem of selecting input attributes that are most predictive of a given outcome. The FS problem is pervasive and can be encountered in many areas of machine learning, pattern recognition and signal processing. In contrast to other methods for reduction of dimensionality, the feature selection approach preserves the original semantics or meaning of the features following reduction. FS has been applied to tasks that involve datasets which contain very large numbers of features (in the order of tens of thousands) [32]. Without FS, such problems would prove to be computationally intractable.

As RST was originally proposed for supervised learning, it is no surprise therefore that one of the many successful applications of rough set theory has been in the area of FS. The basic tenet of RST which means that only the supplied data is employed for data reduction (with no additional information) has many benefits

in FS. Most other methods require at least some supplementary knowledge. The main disadvantage of rough set-based feature selection in the literature is the restrictive requirement for all data to be crisp, and hence the motivation to extend the rough set model as described in Section 2.4.

There are two main approaches when searching for rough set reducts: the dependency degree approach, and the discernibility matrix approach. Both approaches have been employed for rough set-based FS, although the discernibility matrix approach is computationally expensive for large datasets [98], but some constructs [175] have been proposed to alleviate this problem.

Amongst the earliest rough set-based dependency degree approaches to FS is the Preset algorithm [154], which uses RST to rank features heuristically, within the assumption of a noise free binary domain. In [262], a rough set heuristic filter-based approach is presented. The algorithm starts out by calculating the core of the dataset (attributes that cannot be removed without introducing inconsistency) and then it incrementally adds attributes based on a heuristic measure. A threshold value is required as a stopping criterion to determine when a reduct candidate is sufficiently 'close' to being a reduct. In [32], the authors also present a filter-based method called rough set attribute reduction (RSAR), based on rough set dependency degree. It uses a greedy forward selection technique (starting with an empty subset) that incrementally adds features that result in an increase in the dependency value. Other approaches have also utilised this approach but used other measures such as entropy [94] and a boundary region measure [148] to search for reducts. In terms of the discernibility matrix approach [206], a number of techniques have also been proposed, and algorithms such as that described in [163] adopt this technique to search for reducts. Others also include [166] with specific application to medical problem domains, and [232] which attempts to address the computational complexity associated with discernibility matrices.

Although not as popular as the traditional rough set approach, VPRS has also been applied to the FS problem. In [223] the authors compare VPRS and traditional rough set based FS techniques. A fault-detection process which uses VPRS as a FS step is also described in [122]. The main disadvantage with approaches like VPRS is the specification of additional tunable parameters, in this case β. As mentioned previously, the optimum value can be obtained by repeated experimentation but this may take considerable time depending on the nature of the data being examined.

Applying rough set-based feature selection to domains where the data is real-valued has previously meant that the data must be discretised beforehand. Tolerance rough sets have provided a solution to this problem however, and in [98] the authors demonstrate how this can be achieved. Unfortunately, the tolerance rough

set approach requires a thresholding value which is specified by the user and can only be automatically approximated by repeated experimentation. Human specification of such a threshold however, conflicts with the rough set ideology that only the information in the data should be employed. As mentioned previously, this has resulted in the development of techniques which extend the rough set concepts of the positive region and dependency function through the use of fuzzy sets resulting in a number of fuzzy-rough set approaches [85, 93, 94, 97, 98, 99, 204, 227]. A greedy hill-climbing search mechanism is then employed to search for subsets of features and a new *fuzzy dependency* measure is employed as a stopping criteria. In [83] an approach that employs information measures for fuzzy indiscernibility relations is presented for the computation of feature importance. Reducts are then calculated by employing a greedy selection algorithm. Comprehensive coverage is given to fuzzy-rough FS approaches in [98], which explores all aspects of generation of reducts, and selection and search methods.

2.6.3.1 Feature Selection for Gene Expression Data

The application of techniques such as machine learning, data mining [225], and pattern recognition [167], to areas of medicine and bioinformatics has enjoyed much attention in recent years, and rough sets and their extensions are no exception. One particular area within this field, is the manipulation of gene expression data. Due to the high dimensionality of the sample data, the search space is exponentially large, thus any techniques which are applied to this type of data must be robust. Rough set techniques are therefore an ideal candidate for the examination of such data as demonstrated in [32].

Rough set FS is employed in [156] as a dimensionality reduction step and applied to a number of gene expression datasets. The FS step generates a number of reducts which are then used to reduce the data before it is classified using a nearest-neighbour approach. The approach can be described as a series of individual steps as shown in Fig. 2.19. The first step involves discretising the data such that it can be used with the rough set approach. This discretisation step involves the search for partitions for each attribute domain. These partitions form new intervals to which objects can be assigned. An Bayesian equal-width approach is used in this case, which handles outliers in a sensible fashion, but assumes uniform distribution of the data.

Having discretised the data, the FS step is then implemented, using a heuristic search described below. The approach starts out with an empty set, to which those attributes that have a rough set dependency ($\gamma > 0$) are added incrementally. This generates a set of attributes from which reducts can later be generated. A

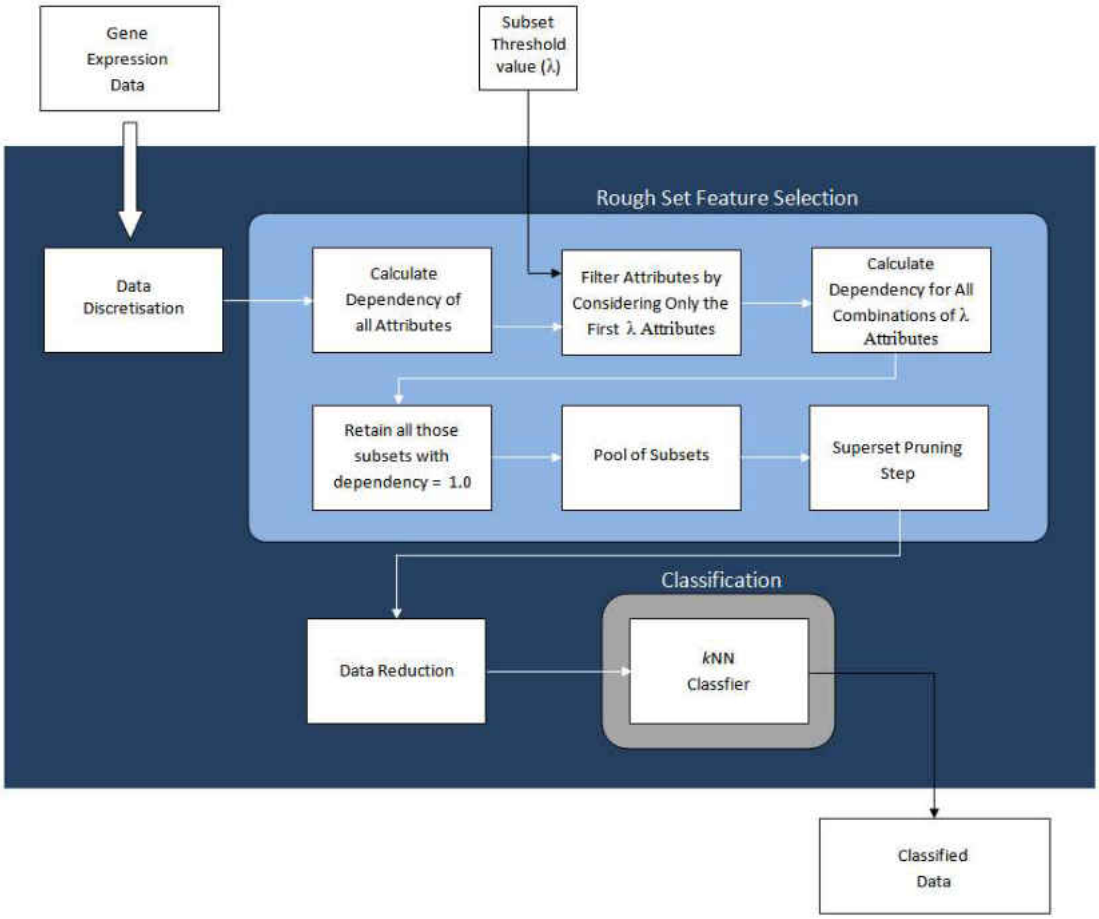

Figure 2.19: Feature selection for gene expression data

thresholding value is also specified at this point termed λ, this value is used to limit the cardinality of all generated reducts. All possible reducts of cardinality λ are then generated, only those of γ =1 are retained. A pruning of all super sets of reducts is then carried out, and the data is reduced prior to the next step.

The next stage is data reduction where all of the reducts are used to minimise the data by selecting the each of the features that appear in a given reduct from the data. Each reduced dataset is then classified. The classifier used here is k nearest neighbours (kNN) [103], which is an object-based classifier learner.

The above process is applied to four publically available datasets relating to various types of cancer. Various values of λ are used to generate the reducts for each of the datasets which are then classified. For the kNN classifier 3, 5 and 7, are selected for values of k, i.e the number of neighbours considered. Discovery of an optimal value for k may take considerable time however. A classification accuracy of 100% for all datasets is achieved, for some but not all of the reducts generated. The process of generating such large numbers of reducts however is computationally expensive. The feature selection approach is compared with two other rough set approaches [202] and [262] which also perform well, however the

authors argue that their method performs better on the basis of classification results.

Again, as with the application example in Section 2.6.2.1, what becomes apparent is the number of tunable parameters, despite very high classification accuracies achieved. The authors mention a parameter for the discretisation of the data, another for the feature selection approach, and of course k for the classification step. Note that, if a hybrid fuzzy-rough approach rather than the current rough set approach were to be employed the discretisation step could be eliminated completely. This would also ensure that any potential loss of information would not occur due to the discretision step.

2.7 Summary

This chapter has in the first case, provided a review of the increasingly important problem of dimensionality reduction (DR). As the amount of available data increases, so too does the need for effective dimensionality reduction. Indeed, in many areas such as machine learning and pattern recognition, it has become unavoidable due to the sheer size of data. Although DR is usually integral to a data preprocessing subsystem, it should not suffer from the same symptoms that learning algorithms do in the presence of large dimensionality. That is to say, it should be able to find minimal or close-to-minimal subsets as discussed in Sections 2.1.2.1 and 2.3.3.

Broadly speaking there are two types of DR, transformation-based, and selection-based. The former is only a valid solution where there is no need to be able to refer to the underlying data, as the original features are transformed and hence the semantics of the data are destroyed. However, quite often DR is employed to *improve* the readability of the data. The selection based approaches such as feature selection avoid this transformation and thus perform DR which preserves the semantics of the underlying data, making the reduced data more transparent to human scrutiny. A particular example of this is rule induction, where rules induced from reduced data may need the same transparency as those from the unreduced data. Feature selection is therefore an important, and valuable technique for DR as it facilitates the reduction of the data to fewer dimensions without the need for transformation.

Rough set theory was originally introduced as an important approach for supervised learning. The significance of RST is reflected in both the level of publications in the area, and the wide number of real-world application problems for which it has been employed. It is no surprise therefore that it has found much success in the area of semantics-preserving DR, or feature selection. This is due largely to

the fact that RST requires no thresholding information and operates only on the data. Furthermore, RST is computationally efficient and only involves simple set operations.

One of the primary disadvantages of RST however, is its inability to deal with real-valued data. This has led to a number of extensions which attempt to address this deficiency. Amongst these are the tolerance rough set approach, fuzzy rough sets, variable precision rough sets, and vaguely quantified rough sets. The utility and applicability of such approaches and others are examined in depth through the use of real-world examples.

A particular point which is apparent from the careful examination of rough set theory is that any extensions that have been proposed tend to be focused in two particular areas; modification of the subset operator, and modification of the equivalence relation. There is however a third area of RST which is often overlooked, and holds much potential without the need to modify the basic underlying rough set model; that of the rough set boundary region. This has provided the motivation for the approaches which follow in Chapter 3 and Chapter 4.

Chapter 3

Exploring the Boundary Region: Rough Sets

Most existing rough set-based FS approaches [70], [75], [120], [121], [154], [165], [221], [262] rely on the information gathered from the lower approximation of a set to minimise data. These approaches have been adopted as the certainty that is embodied in the lower approximation is associated with greater importance in scientific analysis. Although successful, these lower approximation based approaches ignore the information that is contained in the boundary region, or region of uncertainty. Whilst there are also some existing RST approaches which consider the boundary region information [47], [87], they adopt an approach which examines the upper approximation as a whole rather than examining the lower approximation and boundary region as conceptually separate entities. This chapter presents a method which is presented in [148], and [149], and examines both the information in the lower approximation and the information contained in the boundary region for the selection of feature subsets. This can result in the selection of subsets which are smaller than those selected using the information gathered from the lower approximation alone. If the boundary region is empty however the approach will select subsets which are comparable to those of the traditional rough set approach.

There are a number of extensions to the rough set model. However two approaches of note are variable precision rough sets (VPRS) [263] and the tolerance rough sets (TRSM) [209]. These particular extensions have been covered in detail in Chapter 2. They are considered important in the context of this chapter because they extend the rough set approach and utilise the information contained in the boundary region - albeit indirectly. The disadvantages of such approaches however lie in the specification of an additional subjective thresholding value which is necessary for operation.

As discussed previously, almost all techniques for rough set attribute reduction adopt an approach to minimisation of the data that examines only the information contained within the lower approximation of a set. Currently, there are no mechanisms in rough set based methods to deal with the uncertainty of the boundary region. Any useful information that may be contained in the boundary region is therefore lost when only the lower approximation is employed for data minimisation.

The approach described here uses both the information contained in the lower approximation and the information contained in the boundary region to search for reducts. The DMRSAR [148], [149] method uses a distance measure to determine the proximity of objects in the boundary region to those in the lower approximation and assigns a significance value to these distances collectively.

3.1 Distance Metric and Mean Positive Region

The distance metric attempts to qualify the objects in the boundary region of RST with regard to their proximity to the lower approximation. From an intuitive point-of-view, the closer the proximity of an object in the boundary region to objects of the lower approximation, the greater the likelihood that it actually belongs to the set of interest. The central motivation of this approach is illustrated in Fig. 3.1. The granularity of RST enforces a hard or strict perimeter for the lower approximation, this means that objects that may in actuality belong to the set which is being approximated are relegated to the region of uncertainty. Note that this diagram serves only to convey the ideas for the motivation of the approach and is not an attempt to accurately portray all of the concepts of RST. For the method detailed here, all of the distances of objects in the boundary region are calculated. From this, the significance value for a subset can be obtained.

Since calculating the margin of the lower approximation for an n-dimensional space would involve considerable computational effort, a more pragmatic solution is employed - the mean of all object attribute values in the positive region (POS_P) or union of lower approximations is calculated. This can be defined as follows:

$$POS_{P_{MEAN}} = \left\{ \frac{\sum_{o \in \underline{P}X} a(o)}{|POS_P X|} : \forall a \in P \right\} \tag{3.1}$$

Using this definition of the mean of the P positive region, the distance function for the proximity of objects in the boundary region from the P positive region mean can be defined by

$$\delta_P(POS_{P_{MEAN}}, y), \quad y \in BND_P(Q) \tag{3.2}$$

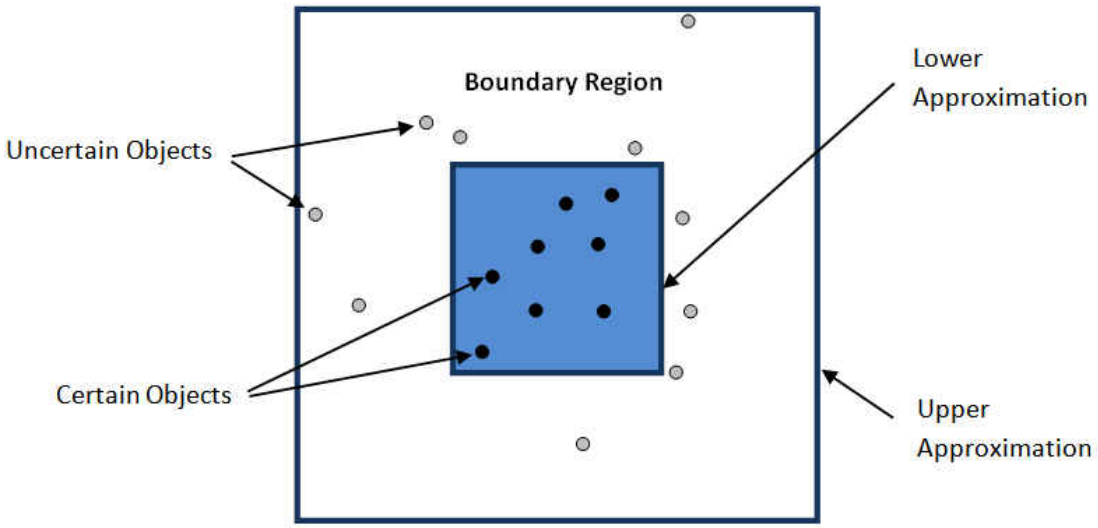

Figure 3.1: Objects of the lower approximation and boundary region

Clearly this definition only holds true if either $POS_{P_{MEAN}}$ or $BND_P(Q)$ is non-empty.

The exact distance function is not defined here as a number of strategies may be employed for the calculation of the distance of objects in the boundary. In the worked example section a Euclidean type distance metric is employed.

In order to measure the quality of the boundary region, a significance value ω for subset P is calculated by obtaining the sum of all object distances and inverting it such that:

$$\omega_P(Q) = \left(\sum_{y \in BND_P(Q)} \delta_P(POS_{P_{MEAN}}, y) \right)^{-1} \tag{3.3}$$

It is important to note that if $POS_P(Q) = \emptyset$ there are no certain objects from which to generate a $POS_{P_{MEAN}}$, in which case no distance function can be defined and hence the significance degree $\omega_P(Q)$ is set to 0. Also, when $BND_P(Q) = \emptyset$ there is no uncertainty about the concept being approximated and so there are no uncertain objects to measure using the distance function, in which case the significance degree $\omega_P(Q)$ is set to its maximum value of 1.

This significance measure is used in conjunction with the rough set dependency value to gauge the utility of attribute subsets in a similar way to that of the rough set dependency measure. As one measure only operates on the objects in the lower approximation and the other only on the objects in the boundary, both entities are considered separately and then combined to create a new evaluation measure M:

$$M_P(Q) = \frac{\omega_P(Q) + \gamma_P(Q)}{2} \tag{3.4}$$

Obviously if $\gamma_P(Q) = 1$, then the concept being approximated has no uncertainty with respect to P and, $\omega_P(Q) = 1$. A mean of both values is obtained as both operate in the range [0,1]. A new feature selection mechanism can be constructed that uses both the significance value and the rough dependency value to guide the search for the best feature subset.

An alternative to the mean positive region concept and distance metric is an approach which uses the *Hausdorff* metric to calculate the distance between non-empty sets. It measures the extent to which each point in a set is located relative to those of another set. The *Hausdorff* metric has been applied to facial recognition [192], image processing [196] and FS [181] with much success. It can be defined as:

$$h(A, B) = \max_{a \in A}\{\min_{b \in B}\{d(a, b)\}\} \tag{3.5}$$

where a and b are points (objects) of sets A and B respectively, and $d(a, b)$ is any distance metric between these points. A basic implementation of this has been incorporated into the above framework using Euclidean distance as a metric. Experimentation using this approach can be seen later. The primary disadvantage to this approach however is the computational overhead involved in calculating the distance for all objects in the boundary region from all of the objects in the lower approximation. For n boundary region objects, this means that $O(n^2)$ distance calculations must be made, unlike the mean positive region which results in $O(n)$ distance calculations.

3.2 Distance Measure-based Selection Algorithm

The illustration in Fig. 3.2 below shows a rough-set based DMQuickReduct algorithm based on the previously described rough set-based approach in Fig. 2.14.

DMQuickReduct is similar to the RSAR algorithm but uses a combined distance and rough-set dependency value of a subset to guide the feature selection process. If the combined value M of the current reduct candidate is greater than that of the previous, then this subset is retained and used in the next iteration of the loop. It is important to point out that the subset is evaluated by examining the value of M, termination only occurs when the addition of any remaining features results in the dependency function value (γ_T) reaching that of the unreduced dataset. The value of M is therefore not used as a termination criterion.

DMQuickReduct($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

(1) $T \leftarrow \{\}, R \leftarrow \{\}$
(2) **do**
(3) $\quad \forall x \in (\mathbb{C} - R)$
(4) $\quad\quad$ **if** $M(R \cup \{x\}) > M(T)$
(5) $\quad\quad\quad T \leftarrow R \cup \{x\}$
(6) $\quad R \leftarrow T$
(7) **until** $\gamma_R(\mathbb{D}) == \gamma_{\mathbb{C}}(\mathbb{D})$
(8) **return** R

Figure 3.2: The rough-set distance metric-based algorithm

The algorithm begins with an empty subset R. The do-until loop works by examining the combined dependency and significance value of a subset and incrementally adding a single conditional feature at a time. For each iteration, a conditional feature that has not already been evaluated will be temporarily added to the subset R. The combined measure of the subset currently being examined (line 6) is then evaluated and compared with that of T (the previous subset). If the combined measure of the current subset is greater, then the attribute added (line 5) is retained as part of the new subset candidate T (line 6).

The loop continues to evaluate in the above manner by adding conditional features, until the dependency value of the current reduct candidate ($\gamma_R(\mathbb{D})$) equals the consistency of the dataset (1 if the dataset is consistent).

3.3 Computational Complexity

As the DMRSAR algorithm is based on a greedy hill-climbing type of search, the computational complexity will be similar to that of other approaches which use this method. However, in addition to the factors which govern the computational complexity of the rough set QuickReduct algorithm [32] demonstrated in Fig. 2.14, other factors must also be taken into account. In the DMRSAR approach objects in the boundary region are also considered and this inevitably adds to the computational overhead. Furthermore, all of those objects in the lower approximation are also considered when calculating a positive region object for each concept - where the objects of the positive region are 'collapsed' to form a single representative object. At this lower level the additional factors that must be considered (also those that are not employed in the rough set case) include: the calculation of the collapsed lower approximation mean, the calculation of the up-

per approximation, and the calculation of the distances of objects in the boundary from the collapsed lower approximation mean.

From a high level point-of-view the DMQuickReduct has an intuitive complexity of $(n^2 + n)/2$ for a dimensionality of n. This is the number of evaluations of the dependency function and distance measure performed in the 'worst case'. For instance if the feature set consists of $\{a_1, a_2\}$, then the DMQuickReduct algorithm will make 3 evaluations, one each for $\{a_1\}$ and $\{a_2\}$, and finally one for $\{a_1, a_2\}$ in the worst case.

In an attempt to compare the complexity of both the RSAR and DMRSAR approaches from an application viewpoint, a number of artificial datasets were generated. These ranged in size from 20 to 350 attributes, and 500 to 8000 objects. The objects in each dataset were created using a simple random number generator program. This program also included a step which generated the class label for each object from a given specified range of labels and ensured that the dataset was consistent - i.e that there were no contradictory object-to-class assignments.

Both FS approaches were applied to these datasets and the time taken to find a reduct was recorded in each case. The results show that there is only a marginal increase in runtime for the DMRSAR approach. There is even a decrease in some cases, but this relates to the fact that DMRSAR found smaller subsets than RSAR in these particular cases. However, Fig. 3.3 and Fig. 3.4 demonstrate that for increased dimensionality and numbers of objects there is little overall difference in runtime between the approaches.

3.4 A Worked Example

To illustrate the operation of the distance measure-based algorithm, a small example dataset is considered, containing discrete-valued conditional and decision attributes. Both crisp and real-valued data is used in the experimentation evaluation in later sections, however crisp data is used in this example to aid explanation of the approach. Note also for brevity, that only the selection of two subsets is shown here.

Table 3.1 contains seven objects. It has four crisp-valued conditional attributes and a single crisp-valued decision attribute.

The fist step is to calculate the lower and upper approximations:

$$\underline{P}X = \{x | [x]_p \subseteq X\}$$

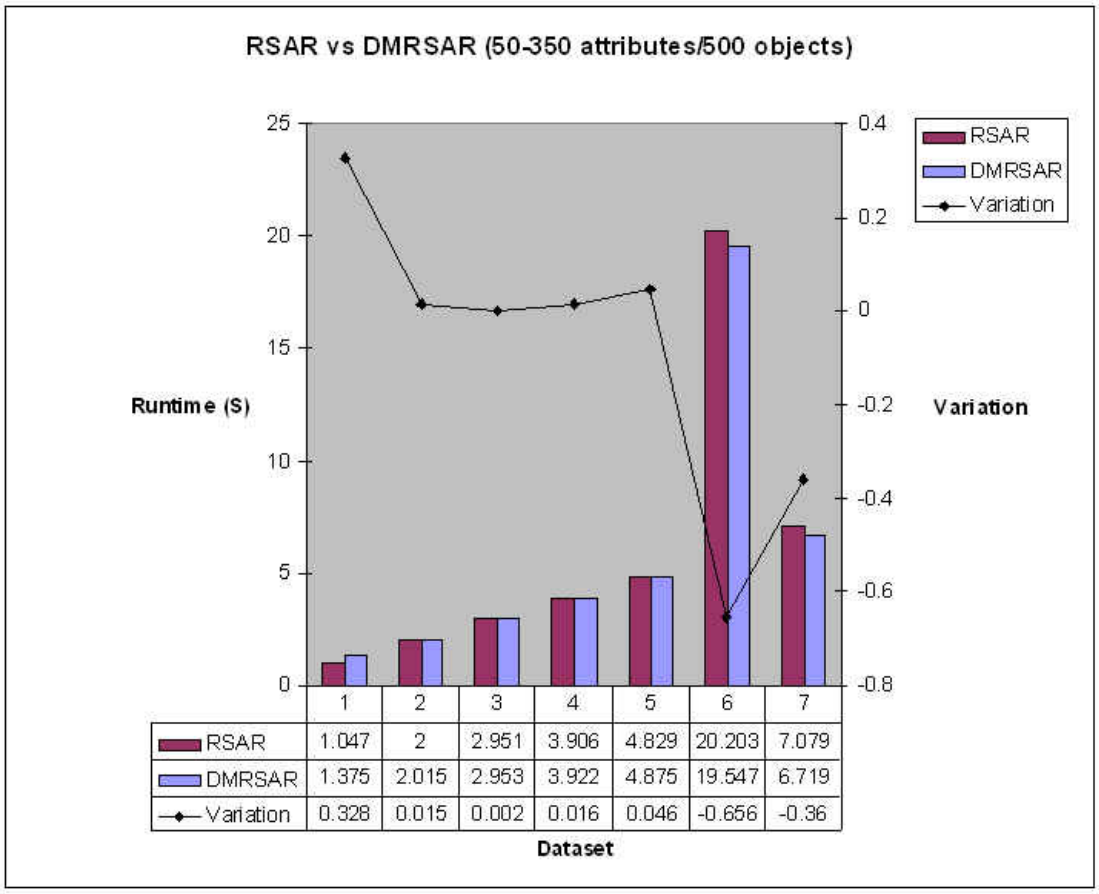

Figure 3.3: RSAR and DMRSAR runtimes for 50-350 attributes

Object	a	b	c	d	e
0	1	0	2	2	0
1	0	1	0	0	2
2	1	0	0	1	1
3	1	0	0	2	2
4	1	2	0	0	1
5	1	2	0	2	0
6	0	1	2	0	1

Table 3.1: Example dataset: crisp attributes

$$\overline{P}X = \{x|\ [x]_P \cap X \neq \emptyset\}$$

Referring to the example in sec. 2.3.1.1, and considering attribute d these can be calculated as:

$\underline{\{d\}} \quad = \quad \{\{\}, \{2\}, \{\}\}$

Similarly for the upper approximation:

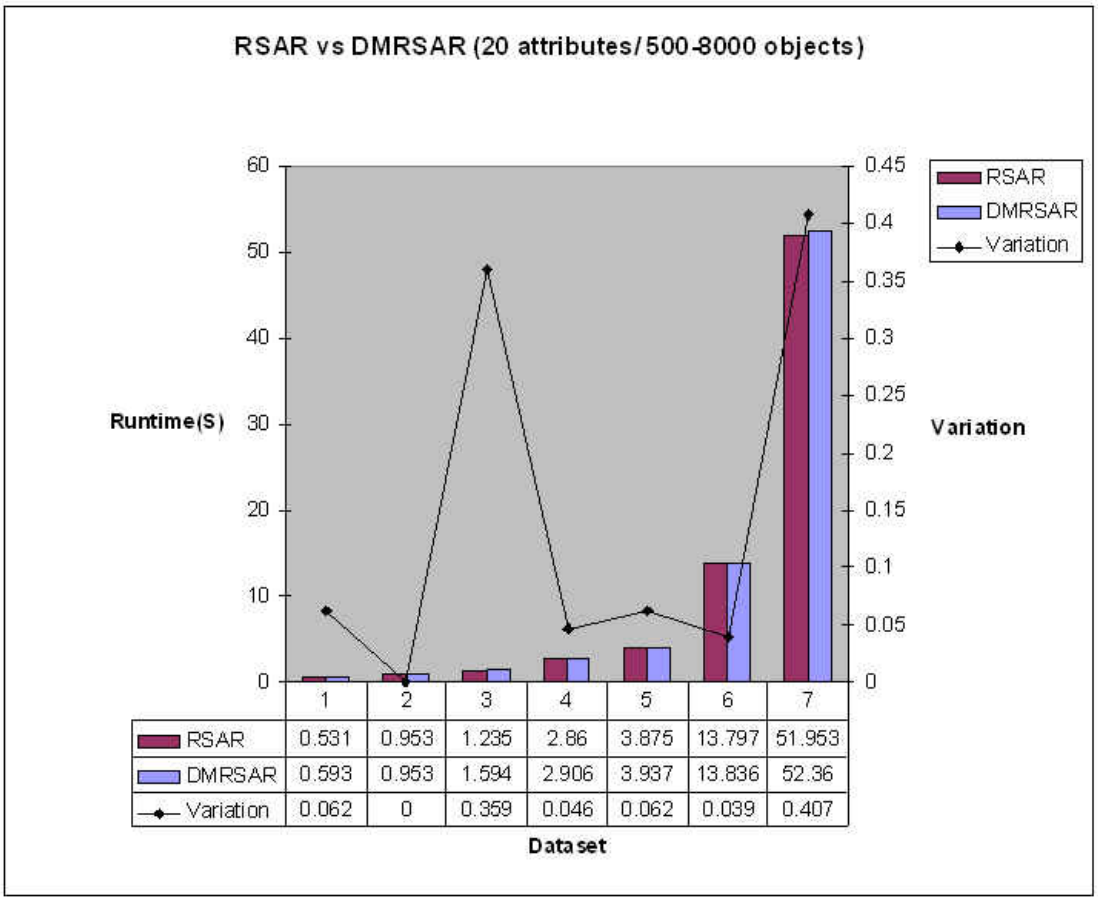

Figure 3.4: RSAR and DMRSAR runtimes for 500-8000 objects

$$\overline{\{d\}} = \{\{0,3,5\},\{1,2,4,6\},\{0,1,3,4,6\}\}$$

Having calculated the upper and lower approximations for $\{d\}$, the positive and boundary regions can be shown to be:

$$\begin{aligned} POS_{\{d\}}(\{e\}) &= \bigcup\{\emptyset,\{2\}\} = \{2\} \\ BND_{\{d\}}(\{e\}) &= \bigcup\{\{0,3,5\},\{2\} \\ &\quad \{1,4,6\},\{1,4,6\}\} \text{ - } \{2\} \\ &= \{0,1,3,4,5,6\} \end{aligned}$$

The rough-set dependency, the positive region mean, and object distances can now all be calculated. As mentioned in the previous section there are many distance metrics which can be applied to measure the distance of the objects in the boundary from the lower approximation mean. For simplicity, a variation of Euclidean distance is used in the approach documented here, and this is defined as:

$$\delta_P(POS_{P_{MEAN}}, y) = \sqrt{\sum_{a \in P} f_a(POS_{P_{MEAN}}, y)^2} \tag{3.6}$$

where:

$$f_a(POS_{P_{MEAN}}, y) = \begin{cases} 1 \iff a(POS_{P_{MEAN}}) \neq a(y) \\ 0 \text{ otherwise} \end{cases}$$

From this, the distances of all of the objects in the boundary region in relation to the lower approximation mean can now be calculated.

As there is only a single object in the lower approximation, the mean of the lower approximation does not need to be calculated in this case. The individual distances of objects in the boundary of $\{d\}$ can be shown to be:

$$\begin{aligned} &obj\ 0\ \sqrt{f_d(POS_{P_{MEAN}}, 0)^2} = 1 \\ &obj\ 1\ \sqrt{f_d(POS_{P_{MEAN}}, 1)^2} = 1 \\ &obj\ 3\ \sqrt{f_d(POS_{P_{MEAN}}, 3)^2} = 1 \\ &obj\ 4\ \sqrt{f_d(POS_{P_{MEAN}}, 4)^2} = 1 \\ &obj\ 5\ \sqrt{f_d(POS_{P_{MEAN}}, 5)^2} = 1 \\ &obj\ 6\ \sqrt{f_d(POS_{P_{MEAN}}, 6)^2} = 1 \end{aligned}$$

Where there is more than one object in the lower approximation of the candidate reduct, calculating the $POS_{P_{MEAN}}$ object can be achieved in the manner described in the previous section i.e. examine all of those attribute values for each of the objects that appear in the lower approximation of the considered subset. For example considering the subset $\{a, d\}$, the lower approximation and boundary regions are:

$$\begin{aligned} POS_{\{a,d\}}(\{e\}) &= \bigcup\{\emptyset, \{2\}, \{4\}\} \\ BND_{\{a,d\}}(\{e\}) &= \bigcup\{\{0,3,5\}, \{0,3,5\}\{1,6\}, \{1,6\}\} \\ &= \{0,1,3,5,6\} \end{aligned}$$

The attribute values for $\{a, d\}$ for objects $\{2, 4\}$ can be obtained by referring to Table 3.1:

$$\begin{aligned} for\ \{a\}: &\quad object\ 2 = '1' \\ &\quad object\ 4 = '1' \end{aligned}$$

$$\begin{aligned} for\ \{d\}: &\quad object\ 2 = '1' \\ &\quad object\ 4 = '0' \end{aligned}$$

This results in:

$$POS_{P_{MEAN}} = \{1, 0.5\} \quad \text{for } \{a, d\}$$

These real-valued numbers however, are not meaningful when dealing with crisp-valued data within the framework of RST (1 is considered as different from 1.1 as it is from 100). The strategy employed to address this problem was to

examine all of the attribute values for the attribute in question and assign it a value which appears in that range of values to which it is closest in terms of magnitude. So as the $POS_{P_{MEAN}}$ value for the attribute a is an existing value, this does not need to be considered; the $POS_{P_{MEAN}}$ value assigned to d however is not in the range of values taken by the attribute d. Values of 0.5 or less are considered to be closer to 0, and thus approximated to $'0'$, and becomes $POS_{P_{MEAN}} = \{1, 0\}$.

Again by utilisation of Euclidean distance and the new $POS_{P_{MEAN}}$, the distances of objects in the boundary region can be calculated:

$$\begin{array}{ll} ob\ 0 & \sqrt{(f_a(POS_{P_{MEAN}}, 0)^2 + f_d(POS_{P_{MEAN}}, 0)^2)} = 1 \\ ob\ 1 & \sqrt{(f_a(POS_{P_{MEAN}}, 1)^2 + f_d(POS_{P_{MEAN}}, 1)^2)} = 1 \\ ob\ 3 & \sqrt{(f_a(POS_{P_{MEAN}}, 3)^2 + f_d(POS_{P_{MEAN}}, 3)^2)} = 1 \\ ob\ 5 & \sqrt{(f_a(POS_{P_{MEAN}}, 5)^2 + f_d(POS_{P_{MEAN}}, 5)^2)} = 1 \\ ob\ 6 & \sqrt{(f_a(POS_{P_{MEAN}}, 6)^2 + f_d(POS_{P_{MEAN}}, 6)^2)} = 1 \end{array}$$

It is perhaps worth noting at this point, that although a form of Euclidean distance is used to calculate the distance of the objects from the $POS_{P_{MEAN}}$, in calculating that distance, the difference between two values is always considered in boolean terms for crisp data. The reason for this is that the values are states rather than real-valued. This means that if the value for a particular attribute in the $POS_{P_{MEAN}}$ happened to be 1 and that of the corresponding attribute value of an object in the boundary region was 1563, the difference between these two states would be $(1 - 1563) = 1$. For real–valued data however, this would not be the case as the values of attributes are real numerical values.

Although the individual distances may be useful in identifying objects that are similar to those in the lower approximation, they are not individually indicative of the subset goodness. A method of achieving this measure is to calculate the sum of all of the distances and invert it, thus giving a significance value to each subset considered for selection. The significance value is real-valued and has membership in the range [0,1] for the purpose of dealing with crisp data.

Thus for $\{a, d\}$:

$$\omega_{\{a,d\}}(\{e\}) \quad = \quad (1 + 1 + 1 + 1 + 1)^{-1} \quad = 0.2$$

Although the significance measure alone can be used to search for subsets, the results from some initial investigation indicated that these were not of equal quality as those returned by RSAR. So the significance value was combined with the rough set dependency value. This results in a combined metric in which both dependency and significance have equal participation. This approach is adopted as it ensures that the subjective specification of a parameter is not required.

By calculating the change in combined significance and dependency value (M) when an attribute is removed from the set of considered conditional attributes, a measure of the goodness of that attribute can be obtained. The greater the change in M the greater the measure of goodness that attribute has attached to it.

Using the previous examples of the DMRSAR method the values for the combined metric can be calculated for all considered subsets of $\mathbb{C}$ using DMRSAR:

$$\begin{array}{ll} M_{\{b\}}(\{e\}) = 0.0 & M_{\{b,d\}}(\{e\}) = 0.3910 \\ M_{\{c\}}(\{e\}) = 0.0 & M_{\{c,d\}}(\{e\}) = 0.3026 \\ M_{\{d\}}(\{e\}) = 0.342 & M_{\{a,b,d\}}(\{e\}) = 0.3026 \\ M_{\{a,d\}}(\{e\}) = 0.2425 & M_{\{b,c,d\}}(\{e\}) = 1.0 \end{array}$$

It is obvious from the above example that the search finds a subset in the manner $\{d\} \rightarrow \{b, d\} \rightarrow \{b, c, d\}$. As $\{a, d\}$ and $\{c, d\}$ and also $\{a, b, d\}$ do not result in the same increase in combined metric these subsets are ignored. Note that larger subsets can result in lower values of M.

3.5 Experimental Evaluation - Comparison with Other Approaches

This section presents the results of experimental studies using both crisp-valued and real-valued datasets. The DMRSAR method is initially compared with a rough set-based feature selection method (RSAR) [32], and Principal Component Analysis (PCA) [48]. Additionally DMRSAR is also compared with fuzzy-rough set-based FS (FRFS)[93] and a tolerance rough set based feature selection method [209] for real-valued data. It is important to note that DMRSAR operates on discretised versions of the real-valued datasets listed. This is related to the fact that the underlying model for DMRSAR is RST, and as discussed previously RST is unable to deal with real-valued data. Chapter 4 presents an approach which is able to overcome this problem.

All of the datasets presented are of the same format as that used in the example of the previous section. All data has been obtained from [5] and [158]. A comparison of the RSAR, FRFS, and distance-based dimensionality reduction techniques is made based on subset size, classification accuracy, and time taken to discover subsets.

3.5.1 Classifier Learners

Three classifier learners were employed for the classification of the data, JRip, J48, and PART [237].

J48 [184] creates decision trees by choosing the most informative features and recursively partitioning the data into subtables based on their values. Each node in the tree represents a feature with branches from a node representing the alternative values this feature can take according to the current subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created, and this classification assigned.

JRip [33] learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, antecedents are added greedily until a termination condition is satisfied. Antecedents are then pruned in the next phase subject to a pruning metric. Once the ruleset is generated, a further optimization is performed where rules are evaluated and deleted based on their performance on randomized data.

PART [236] generates rules by means of repeatedly creating partial decision trees from data. The algorithm adopts a divide-and-conquer strategy such that it removes instances covered by the current ruleset during processing. Essentially, a rule is created by building a pruned tree for the current set of instances; the leaf with the highest coverage is promoted to a rule.

3.5.2 Comparison with RSAR

Results are presented here both in terms of subset size and classification accuracy. The datasets employed range in size from 47 to 2000 objects and between 7 and 57 attributes. Conditional attributes and decision attributes are crisp and discrete-valued.

3.5.2.1 Classification Accuracy

The data presented in Table 3.2 shows the average classification accuracy as a percentage using each of the previously described classifiers. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained, by using the RSAR [32], and DMRSAR dimensionality reduction techniques respectively.

Noting the classification results, the DMRSAR approach performs well and shows increases in classification accuracies for at least one classifier where there has been a corresponding decrease in dimensionality (e.g *credit, exactly*, etc.) Notably for the *exactly* dataset DMRSAR shows an increase of up to 30% whilst si-

	J48			JRip			PART		
Dataset	Unred.	RSAR	DMRSAR	Unred.	RSAR	DMRSAR	Unred.	RSAR	DMRSAR
credit	72.60	70.10	69.59	71.90	72.0	70.30	68.9	68.3	69.4
derm	95.90	80.32	82.51	92.07	77.32	76.77	95.62	78.76	81.96
derm2	95.25	94.41	94.41	93.85	92.73	93.85	92.73	93.85	93.85
ionosphere	85.21	88.26	87.39	86.60	84.34	86.93	86.08	87.39	89.56
exactly	85.5	69.4	98.1	69.30	68.00	91.30	92.10	67.32	99.20
exactly 2	74.9	74.9	73.1	75.0	75.0	74.8	74.2	74.20	78.20
heart	77.89	80.95	81.29	79.59	76.10	77.55	77.21	78.57	81.63
led	100	100	100	100	100	100	100	100	100
lung	84.38	84.38	78.12	68.75	84.38	68.75	71.88	84.38	78.12
m-of-n	100	100	100	97.3	98.60	98.60	100	100	100
monk3	100	100	100	99.76	99.07	99.07	100	100	100
soybean	91.35	89.84	87.59	88.72	88.72	80.89	92.10	87.96	84.21
tic-tac-toe	92.38	88.10	87.89	98.32	91.10	91.44	95.30	87.68	87.68
vote	93.67	93.67	93.67	95.00	93.67	93.67	91.67	93.67	93.67
wq	71.07	64.87	67.37	70.44	68.71	67.51	67.17	65.25	66.02

Table 3.2: Average classification accuracy – crisp data

multaneously demonstrating a reduction in dimensionality. Even where there has been a decrease in the case of some classifiers and datasets which are of similar size to those of RSAR, this decrease is insignificant. Indeed, DMRSAR may sometimes discover subsets of similar size (but contain different features) to RSAR yet demonstrate an increase in classification accuracy (e.g. *derm, ionosphere, heart*).

3.5.2.2 Reduct Size and Run Times

Table 3.3 shows a comparison of reduct size, and runtimes for both the RSAR, and DMRSAR approaches. At the very least DMRSAR can match the performance of RSAR, and shows that there are gains to be made with crisp-valued data, with (*credit, exactly, exactly2, wq*), demonstrating that there is much information contained in the boundary region which if employed for feature selection can improve the approximation ability of RST.

There is little relative increase in runtimes when comparing RSAR with DMRSAR, indeed DMRSAR sometimes demonstrates a reduction in dimensionality along with a reduction in runtime. Considering also that no runtime optimisation has been performed for DMRSAR these results are very encouraging. However, it also suggests that there is some improvement required in terms of the mean positive region calculation which would result in more accurate measurement of distances.

Dataset	Original number of features	Reduct size		Time taken to locate reduct	
		RSAR	DMRSAR	RSAR	DMRSAR
credit	21	9	8	0.937	1.656
derm	35	7	7	0.625	0.625
derm2	35	10	10	0.578	0.640
ionosphere	35	8	8	0.313	0.313
exactly	14	9	8	0.203	0.172
exactly2	14	13	10	0.328	0.235
heart	14	7	7	0.188	0.188
led	25	12	12	2.168	2.375
lung	57	4	4	0.125	0.132
m-of-n	14	8	7	0.171	0.142
monk3	7	3	3	0.063	0.063
soybean	36	12	12	0.797	0.828
tic-tac-toe	10	8	8	0.188	0.203
vote	17	9	9	0.157	0.172
wq	39	15	14	3.250	2.766

Table 3.3: Comparison of reduct size, dependency value, & run times – crisp Data

3.5.3 Comparison with PCA

PCA [48] is a versatile transformation-based DR technique which projects the data onto a new coordinate system of reduced dimensions. This process of linear transformation however also transforms the underlying semantics or meaning of the data. This results in data that is difficult for humans to interpret, but which may still provide useful automatic classification of new data. In order to ensure that the comparison of DMRSAR and PCA is balanced, the same subset sizes discovered for each dataset are also employed in the analysis of PCA. Each of the best number of transformed features are also utilised for PCA.

The results in Table 3.4 show that of the 15 datasets only *credit, derm,* and *tic-tac-toe* demonstrate a small decrease in classification accuracy performance when compared with DMRSAR. These decreases are small in magnitude and DMRSAR outperforms PCA in all other cases, sometimes significantly.

It should be emphasised however, that while PCA might marginally outperform DMRSAR in three instances in terms of classification accuracy, the semantics of the data is irreversibly transformed following dimensionality reduction. This can have consequences where human interpretability of the data is important, which is one of the key reasons for performing feature selection tasks to begin with. As DMRSAR is a *feature selection* approach as opposed to a *feature ranking* method, a predefined threshold is not required; selection is complete as soon as the termination criterion (rough set dependency) is fulfilled. The rough set dependency value is integral to the selection process and as such, in contrast to PCA does not

Dataset	(predefined) subset size	J48	JRIP	PART
credit	8	71.10	71.00	72.00
derm	7	90.40	94.08	93.98
derm2	10	93.29	91.34	93.52
ionosphere	8	81.30	76.95	79.12
exactly	8	67.80	66.70	68.60
exactly2	10	75.90	74.30	75.80
heart	7	77.89	79.58	77.21
led	12	99.38	98.55	99.38
lung	4	71.85	68.75	65.62
m-of-n	7	76.2	73.30	75.30
monk3	3	77.77	76.62	77.31
soybean	12	77.81	72.18	75.18
tic-tac-toe	8	96.18	94.57	95.92
vote	9	89.00	89.00	87.67
wq	14	67.32	67.37	66.41

Table 3.4: Subset size and classification accuracy results for PCA

need to be predefined.

Finally, it is worth noting that PCA is selected for comparison here in recognition of the fact that it is an established approach for dimensionality reduction.

3.5.4 Comparison with FRFS

The real-valued data used in this section comprises of datasets which are small-to-medium in size, with between 120 and 390 objects per dataset and feature sets ranging from 5 to 39. The unreduced data classification is illustrated in Fig. 3.5. The data has been discretised for use with DMRSAR as it is unable to handle real-valued data. The DMRSAR selected subsets are however employed when reducing and classifying the original real-valued data.

	J48			JRip			PART		
Dataset	Unred.	DMRSAR	FRFS	Unred.	DMRSAR	FRFS	Unred.	DMRSAR	FRFS
water 2	85.64	86.67	80.26	83.84	85.89	84.36	83.33	84.36	82.56
water 3	79.48	79.74	79.74	81.28	81.79	82.05	77.43	83.33	78.97
cleveland	50.16	54.20	53.87	52.18	53.53	55.55	51.85	51.51	52.18
glass	67.75	69.15	68.22	67.75	69.62	69.62	67.28	72.89	69.62
heart	73.30	77.78	75.55	77.40	82.22	80.00	76.66	81.82	78.51
ionosphere	86.26	86.10	91.30	86.52	84.78	87.82	87.82	86.10	91.30
olitos	57.50	68.33	62.50	70.83	67.33	70.83	67.50	67.50	67.50
wine	93.82	93.25	93.82	92.69	95.86	88.76	94.33	94.94	92.13

Table 3.5: Classification accuracy of unreduced, DMRSAR reduced, and FRFS reduced, data using JRIP, PART, and J48 classifiers

3.5.4.1 Classification Accuracy

It is interesting to note that where a decrease in classification accuracy is recorded for FRFS, with respect to the unreduced data, the same is also true for DMRSAR. This decrease in classification accuracy is small when comparing both FRFS and DMRSAR approaches to the unreduced data. Also, when comparing classification results, where the DMRSAR approach shows a fall in classification accuracy, the corresponding reduction in dimensionality (shown in Table 3.5) is significantly better than that of FRFS.

3.5.4.2 Subset Size and Runtimes

	Original number of		Subset size		Time taken to locate subset	
Dataset	features	objects	FRFS	DMRSAR	FRFS	DMRSAR
water 2	39	390	11	12	96.58	0.860
water 3	39	390	12	18	158.73	1.266
cleveland	14	297	11	9	24.11	0.219
glass	10	214	9	6	1.61	0.156
heart	14	270	11	10	11.84	0.158
ionosphere	35	230	5	4	0.488	0.512
olitos	26	120	10	8	11.20	0.156
wine	14	178	10	8	1.42	0.125

Table 3.6: Comparison of subset size, dependency value, & run times – FRFS

It is clear also from the runtime figures that DMRSAR runs considerably faster than FRFS. This primarily, can be attributed to the computational complexity of FRFS which is related to the time taken in calculating fuzzy-equivalence classes. Clearly, DMRSAR has a considerable advantage in this respect as the figures in Table 3.5.4.2 demonstrate.

The advantages of the DMRSAR method in terms of subset size are more pronounced when compared with FRFS than those for RSAR. This is a strong indicator that the approach is perhaps more efficient when applied to domains where the data is real-valued, this is borne out by the marked contrast between the subset-size results obtained for both approaches. There are however two datasets where DMRSAR fails to outperform FRFS in terms of subset size – *water 2* and *water 3*- (see chapter summary for further discussion of this). However, it should be noted that FRFS is considerably more mature and refined in terms of both research effort and development.

3.5.5 Comparison with TRSM

In this section an extension of the rough set model - the tolerance rough set model (TRSM) [209] is compared with DMRSAR. TRSM employs a similarity relation to minimise data as opposed to the indiscernibility relation used in classical rough-sets. This allows a relaxation in the way equivalence classes are considered. This flexibility allows a blurring of the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one tolerance class, thus allowing the consideration of real-valued data. Thus, as for FRFS, real-valued data is also employed for the evaluation of this approach.

The ideal tolerance threshold value can be obtained by repeated experimentation for a given dataset. This is where the TRSM diverges from the approaches to which DMRSAR has been compared up until now, which have all been data-driven. Further work which examines a non-data-driven feature selection approach and which utilises the boundary region of the TRSM can be found in [144]. For the comparison of DMRSAR and TRSM, results are presented in the following subsections for two different values of tolerance threshold (τ) - 0.90, and 0.95.

3.5.5.1 Subset Size

The subset sizes for both values of tolerance threshold are outlined in Table 3.7. The results demonstrate that the TRSM method can sometimes outperform both FRFS and DMRSAR in terms of subset size. However, it should be borne in mind that the TRSM is not completely data-driven and much experimentation may be required before optimal results are achieved for each individual dataset. Additionally, the results also demonstrate that the TRSM method does not perform consistently and in some cases returns a larger subset whilst simultaneously displaying a decrease in classification accuracy.

	Original number	TRSM	
Dataset	of features	$\tau = 0.90$	$\tau = 0.95$
water 2	39	8	12
water 3	39	9	12
cleveland	14	11	8
glass	10	3	8
heart	14	12	8
ionosphere	34	6	8
olitos	25	9	6
wine	13	5	5

Table 3.7: Comparison of subset size for each tolerance threshold value

3.5.5.2 Classification Accuracy

The results presented here show that DMRSAR when compared with TRSM performs favourably. The results obtained for both tolerance values, show that for 4 of the 8 datasets for $\tau = 0.9$, the TRSM performs poorly and for the remaining 4 datasets the results are comparable. When $\tau = 0.95$, DMRSAR outperforms TRSM in 6 cases. The TRSM however defeats DMRSAR marginally for the *ionosphere* dataset but the corresponding subset is twice the size. The remaining dataset - *wine*, shows a classification result that is comparable to DMRSAR.

		TRSM	
Dataset	JRIP	PART	J48
water 2	85.38	82.30	87.43
water 3	80.00	81.53	76.67
cleveland	54.20	53.87	52.52
glass	65.88	69.15	68.69
heart	79.25	75.19	78.88
ionosphere	85.65	86.52	85.21
olitos	70.00	65.83	61.66
wine	96.06	94.94	96.62

Table 3.8: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.90$)

		TRSM	
Dataset	JRIP	PART	J48
water 2	82.82	83.07	82.05
water 3	81.02	80.77	81.02
cleveland	50.54	50.84	54.54
glass	69.62	68.22	69.62
heart	80.38	78.57	81.48
ionosphere	86.08	87.39	87.39
olitos	64.16	66.67	64.16
wine	93.25	95.50	96.02

Table 3.9: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.95$)

3.5.6 Hausdorff Metric Implementation

The Hausdorff metric approach to distance measurement has been described previously as an alternative to the mean positive region and Euclidean distance based method which was used to generate the empirical results shown above.

The DMRSAR approach was augmented with the Hausdorff metric to measure the distance between the lower approximation and the boundary region was implemented in order to investigate the performance of this method in terms of

subset size and runtimes. The results of this investigation are included here in Table 3.10.

	DMRSAR	Hausdorff Metric	
Dataset	Subset Size	Subset Size	Runtime
credit	8	10	41.64
derm	7	32	19.343
derm2	10	32	18.437
ionosphere	8	28	7.000
exactly	8	13	17.422
exactly2	10	13	19.250
Heart	7	10	1.734
LED	12	13	566.05
lung	4	5	0.484
m-of-n	7	9	22.03
monk3	3	6	0.422
soybean	12	19	23.518
tic-tac-toe	8	8	5.859
vote	9	9	3.205
wq	14	27	57.031

Table 3.10: DMRSAR – Hausdorff metric implementation subset size and runtimes

It is apparent that this particular implementation of the *Hausdorff* metric fails to capture the useful information of the boundary region in the same way that the mean positive region method does. Examining the results for subset size, it can be seen that the existing DMRSAR approach returns superior results in all cases. Perhaps even more apparent are the results for the runtimes with the LED dataset which takes 566s to run. This was to be expected as there are a large number of distance calculations performed even for small datasets (exponential $O(n^2)$) for n upper approximation objects).

3.6 Summary

In this chapter, a method for feature selection based on the exploitation of the rough set boundary region has been presented. An algorithm for finding feature subsets, based on the new combined dependency and boundary region metric was introduced, and illustrated by means of a simple example.

Several benchmark datasets were also used to evaluate the utility of the DMRSAR algorithm and provide comparisons with other state-of-the-art feature selection algorithms. The results show that the new metric presented here performs better than the use of the rough set dependency measure alone, emphasising the fact that there is much valuable information to be extracted from the rough set

boundary region. Classification accuracy results have been shown to be very similar to those of FRFS, and in some cases the DMRSAR method has even shown an increase whilst simultaneously demonstrating a reduction in dimensionality. Where a decrease has been observed in relation to FRFS, it has been small and, as discussed previously, the actual decrease is statistically insignificant.

Additional comparison with a TRSM based feature selection method has demonstrated that while this method may sometimes marginally outperform DMRSAR, it requires an additional thresholding value. In order to determine the optimal value however, repeated experimentation is required for each dataset. DMRSAR requires no such thresholding value and relies only on the information in the data.

The work described in this chapter proposed an approach which operated on crisp or nominal data. This approach is not suitable for use where the feature values are continuous, as the rigid granular structure cannot handle e.g. a situation where two feature values may only differ as a result of noise. In order to address this deficiency, other rough set methods must be considered. The following chapter explores some extensions that have the ability to handle real-valued data.

Chapter 4

Exploring the Boundary Region: Tolerance Rough Sets and Fuzzy-Rough Sets

In Chapter 3, it was demonstrated how the boundary region of RST can be used to improve the performance of the rough set model for the task of feature selection. However, as discussed previously in Chapter 2, the main disadvantage of RST is its inability to deal with real-valued data. In order to tackle this problem, methods of discretising the data were employed prior to the application of RST. The use of such methods can result in information loss however, and a number of extensions to RST have emerged [52], [209], [263] which have attempted to address this inability to operate on real-valued domains. Perhaps the most significant of these are the tolerance rough set model (TRSM) [209] and fuzzy-rough sets (FRS) [52]. Both approaches have the ability to operate effectively on real-valued (and crisp) data, thus minimising any information loss. This along with the positive results obtained for the crisp approach provided clear motivation for the formulation of an extension to DMRSAR.

Two new methods for feature selection are presented here, which are based on the TRSM and FRS respectively. The first employs a distance metric to examine the uncertain information contained in the boundary region of tolerance rough sets, and uses that information to guide the feature selection process. This uncertain information is normally ignored in the traditional RST and TRSM approaches to FS which can result in information loss. The second approach utilises various metrics such as fuzzy dependency, fuzzy-entropy [110], and fuzzy information gain ratio [184], to guide the FS search process.

The new distance metric-assisted tolerance rough set selection method is demonstrated with a worked example in order to show the approach fully. All exper-

imental evaluation and results for the approach are presented, and comparison is made with the Principal Component Analysis (PCA) dimensionality reduction technique [48], and also four additional FS techniques: CFS [68], consistency-based FS [41], ReliefF [109], and a wrapper FS approach which employs J48 [184] as an evaluation metric.

For the new fuzzy-rough FS method, in-depth experimental comparisons are made with existing FRS approaches as reported in [98] and [99].

4.1 Tolerance-based Feature Selection

The tolerance rough set model (TRSM) [209] can be useful for application to real-valued data. TRSM employs a similarity relation to minimise data as opposed to the indiscernibility relation used in classical rough-sets. This allows a relaxation in the way equivalence classes are considered. Fig. 2.16 attempted to illustrate the effect of employing this relaxation, where the granularity of the rough equivalence classes has been reduced. This flexibility allows a blurring of the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one tolerance class.

In this approach [96], suitable similarity relations must be defined for each feature, although the same definition can be used for all features if applicable. A standard measure for this purpose, given in [209], is:

$$SIM_a(x,y) = 1 - \frac{\mid a(x) - a(y) \mid}{\mid a_{max} - a_{min} \mid} \tag{4.1}$$

where a is a considered feature, and a_{max} and a_{min} denote the maximum and minimum values of a respectively. This similarity relation is employed here as it avoids the situation that may arise with other relations where all similarity values may be potentially unique. When considering the case where there is more than one feature, the defined similarities must be combined to provide an overall measure of similarity of objects. For a subset of features, P, this can be achieved in many ways including the following approaches:

$$(x,y) \in SIM_{P,\tau} \iff \prod_{a \in P} SIM_a(x,y) \geq \tau \tag{4.2}$$

$$(x,y) \in SIM_{P,\tau} \iff \frac{\sum_{a \in P} SIM_a(x,y)}{\mid P \mid} \geq \tau \tag{4.3}$$

where τ is a global similarity threshold and determines the required level of similarity for inclusion within a tolerance class. This framework allows for the specific case of traditional rough sets by defining a suitable similarity measure (e.g. com-

plete equality of features and the use of equation 4.1) and threshold ($\tau = 1$). Further similarity relations are summarised in [163], but are not included here. From this, the so-called tolerance classes that are generated by a given similarity relation for an object x are defined as:

$$SIM_{P,\tau}(x) = \{y \in U \mid (x, y) \in SIM_{P,\tau}\} \tag{4.4}$$

Lower and upper approximations can now be defined in a similar way to that of traditional rough set theory:

$$\underline{P_\tau} X = \{x \mid SIM_{P,\tau}(x) \subseteq X\} \tag{4.5}$$

$$\overline{P_\tau} X = \{x \mid SIM_{P,\tau}(x) \cap X \neq \emptyset\} \tag{4.6}$$

The tuple $\langle \underline{P_\tau} X, \overline{P_\tau} X \rangle$ is known as a tolerance rough set [209]. Using this, the positive region and dependency functions can be defined as follows:

$$POS_{P,\tau}(Q) = \bigcup_{X \in U/Q} \underline{P_\tau} X \tag{4.7}$$

$$\gamma_{P,\tau}(Q) = \frac{\mid POS_{P,\tau}(Q) \mid}{\mid U \mid} \tag{4.8}$$

From these definitions, an attribute reduction method can be formulated that uses the tolerance-based degree of dependency, $\gamma_{P,\tau}(Q)$, to measure the significance of feature subsets (in a similar way to the rough set QUICKREDUCT algorithm described previously). Although this allows the consideration of real-valued data, the inclusion of the tolerance threshold (τ) also now means that TRSM departs from the traditional rough set approach which requires no additional thresholding information.

4.2 Distance Metric-Assisted Tolerance Rough Set Feature Selection

The Distance Metric-Assisted Tolerance Rough Set Feature Selection (DM-TRS) is an extension of the TRSM approach described previously which has the ability to operate on real-valued data. It marries the TRSM with the distance metric assisted rough set approaches. This allows the information of the TRSM boundary region that is otherwise ignored to be examined and used for FS. This ability to deal with real-valued data along with the consideration of the uncertain boundary region information allows a more flexible approach for FS.

4.2.1 Distance Metric-based ToleranceQuickReduct

Following the outline of TRSM in Section 2.4.1, a similarity relation is defined on all features using 4.3. Employing the already defined tolerance lower and upper approximations (see 4.5 & 4.6) definition the boundary region can be computed:

$$BND_{P,\tau}(Q) = \bigcup_{X \in U/Q} \underline{P_\tau} X - \bigcup_{X \in U/Q} \overline{P_\tau} X \tag{4.9}$$

This and the similarity relation form the principal concepts required for the application of the distance metric. However, in an attempt to quantify the value of the boundary region objects, a metric is required. As argued previously in the intuitive sense, by introducing the P-lower approximation mean, the distance function for the calculation of the proximity of objects in the boundary region can be formulated:

$$\delta_P(\underline{P_\tau} X_{MEAN}, y), \quad y \in BND_{P_\tau}(Q) \tag{4.10}$$

Once again, various distance metrics can be employed for this distance function. To measure the quality of the boundary region, a significance value ω is obtained by measuring all of the distances of the objects and combining them such that:

$$\omega_P(Q) = \left(\sum_{y \in BND_P(Q))} \delta_P(\underline{P} X_{MEAN}, y) \right)^{-1} \tag{4.11}$$

An alternative to the mean lower approximation and distance metric is another approach which uses the *Hausdorff* metric to calculate the distance between non-empty sets. It measures the extent to which each point in a set is located relative to those of another set. The *Hausdorff* metric has been applied to facial recognition [192], image processing [196] and FS [181] with much success. It can be defined as:

$$h(A, B) = max_{a \in A}\{min_{b \in B}\{d(a, b)\}\} \tag{4.12}$$

where a and b are points (objects) of sets A and B respectively, and $d(a, b)$ is any metric between these points. A basic implementation of this has been incorporated into the above framework using Euclidean distance as a metric. Experimentation using this approach can be found in Section 4.3.5. The primary disadvantage of this approach however is the computational overhead involved in calculating the distance of all objects in the boundary region from each other. For n boundary region objects, this means that $O(n^2)$ distance calculations must be made, unlike the mean lower approximation which results in $O(n)$ distance calculations.

As with the previously described rough set-based method the significance measure takes values in the interval $[0, 1]$. This measure can now be combined with the tolerance rough set dependency value and used to gauge the utility of attribute subsets, using the same mechanism as defined in 3.4. This ensures that this method is stable and will always return the same subset of features for a given training dataset.

4.1 shows a distance metric tolerance rough set (DM-TRS) algorithm, that implements the ideas presented above, based on the previously described algorithm in Fig. 3.2.

DMTQuickReduct($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

(1) $R \leftarrow \{\}$, $M_{best} \leftarrow 0$, $M_{prev} \leftarrow 0$
(2) $M \leftarrow 0$, $\gamma'_{best} \leftarrow 0$, $\gamma'_{prev} \leftarrow 0$
(3) **do**
(4) $T \leftarrow R$
(5) $M_{prev} \leftarrow M_{best}$
(6) $\gamma'_{prev} \leftarrow \gamma'_{best}$
(7) $\forall x \in (\mathbb{C} - R)$
(8) **if** $M_{R \cup \{x\}}(\mathbb{D}) > M_T(\mathbb{D})$
(9) $T \leftarrow R \cup \{x\}$
(10) $M_{best} \leftarrow M_T(\mathbb{D})$
(11) $\gamma'_{best} \leftarrow \gamma'_T(\mathbb{D})$
(12) $R \leftarrow T$
(13) **until** $\gamma'_{best} == \gamma'_{prev}$
(14) **return** R

Figure 4.1: The DMTQuickReduct algorithm

The algorithm employs the combined significance and dependency value M to choose which features to add to the current reduct candidate. The metric M is only used to select subsets. The termination criterion is the tolerance rough set dependency value; the algorithm terminates when the addition of any single remaining feature does not result in an increase in the dependency.

Whereas the combined evaluation metric determines the utility of each subset, the stopping criteria is automatically defined through the use of the dependency measure and the subset search is complete either; when the addition of further features does not result in an increase in dependency or when it is equal to 1.

4.2.2 Worked Example

To illustrate the operation of the new distance metric-based approach which combines the tolerance rough set and distance metric methods a small example dataset is considered, containing real-valued conditional attributes and crisp decision attributes.

Table 4.2.2 contains six objects. It has three real-valued conditional attributes and a single crisp-valued decision attribute. For this example, the similarity measure is the same as that given in 4.1 for all conditional attributes, with $\tau = 0.8$. The choice of this threshold allows attribute values to differ to a limited degree, with close values considered as though they are identical.

Object	a	b	c	f
0	−0.4	−0.3	−0.5	*no*
1	−0.4	0.2	−0.1	*yes*
2	−0.3	−0.4	−0.3	*no*
3	0.3	−0.3	0	*yes*
4	0.2	−0.3	0	*yes*
5	0.2	0	0	*no*

Table 4.1: Example dataset

Thus by making $A = \{a\}, B = \{b\}, C = \{c\}$ and $F = \{f\}$, the following tolerance classes are generated:

$$\begin{aligned}
\mathbb{U}/SIM_{A,\tau} &= \{\{0,1,2\},\{3,4,5\}\} \\
\mathbb{U}/SIM_{B,\tau} &= \{\{0,2,3,4\},\{1\},\{5\}\} \\
\mathbb{U}/SIM_{C,\tau} &= \{\{0\},\ \{1\},\ \{3,4,5\},\ \{2\}\} \\
\mathbb{U}/SIM_{F,\tau} &= \{\{0,2,5\},\ \{1,3,4\}\} \\
\mathbb{U}/SIM_{\{a,b\},\tau} &= \{\{0,2\},\ \{1\},\ \{3,4\},\ \{3,4,5\},\ \{4,5\}\} \\
\mathbb{U}/SIM_{\{a,c\},\tau} &= \{\{0\},\ \{1\},\ \{2\},\ \{3,4,5\},\ \{3,4,5\}\} \\
\mathbb{U}/SIM_{\{b,c\},\tau} &= \{\{0,2\},\ \{1\},\ \{3,4\},\ \{5\}\} \\
\mathbb{U}/SIM_{\{a,b,c\},\tau} &= \{\{0\},\ \{1\},\ \{2\},\ \{3,4\},\ \{4,5\}\}
\end{aligned}$$

It is apparent that some objects belong to more than one tolerance class. This is a result of employing a similarity measure rather than the strict equivalence of the conventional rough set model. Using these partitions, a degree of dependency can be calculated for attribute subsets, providing an evaluation of their significance in the same way as previously outlined for the crisp rough case.

The algorithm described previously and shown in Fig. 4.1 can now be employed. It considers the addition of attributes to the stored best current subset (initially the empty set) and selects the feature that results in the greatest increase

of the dependency degree. Considering attribute b, the lower approximations of the decision classes are calculated as follows:

$$\begin{aligned} \underline{B_\tau}\ \{0,2,5\} &= \{x \mid SIM_{B,\tau}(x) \subseteq \{0,2,5\}\} = \{5\} \\ \underline{B_\tau}\ \{1,3,4\} &= \{x \mid SIM_{B,\tau}(x) \subseteq \{0,2,5\}\} = \{1\} \end{aligned}$$

Also the upper approximations:

$$\begin{aligned} \overline{B_\tau}\ \{0,2,5\} &= \{x \mid SIM_{B,\tau}(x) \cap \{0,2,5\}\} = \{0,2,5\} \\ \overline{B_\tau}\ \{1,3,4\} &= \{x \mid SIM_{B,\tau}(x) \cap \{0,2,5\}\} = \{1,3,4\} \end{aligned}$$

From the lower approximation, the positive and boundary regions can then be generated:

$$POS_{B,\tau}(F) = \bigcup_{X \in U/F} \underline{B_\tau} X = \{1,5\}$$

$$BND_{B,\tau}(F) = \bigcup_{X \in U/F} \overline{B_\tau} X - POS_{B,\tau}(F) = \{0,2,3,4\}$$

To calculate the distances of the boundary objects from the lower approximation, it is necessary to generate a lower approximation mean object as described previously:

$$\begin{aligned} \underline{P}X_{MEAN} &= \left\{ \frac{\sum_{o \in \underline{P}X} a(o)}{\mid \underline{P}X \mid} : \forall a \in P \right\} \\ &= \left\{ \frac{\sum a(1), a(5)}{\mid 2 \mid} \right\} = 0.1 \end{aligned}$$

There are many distance metrics which can be applied to measure the distance of the objects in the boundary from the lower approximation mean. For simplicity, a variation of Euclidean distance is used in the approach documented here, and this is defined as:

$$\delta_P(\underline{P}X_{MEAN}, y) = \sqrt{\sum_{a \in P} f_a(\underline{P}X_{MEAN}, y)^2} \tag{4.13}$$

where:

$$f_a(x,y) = a(x) - a(y) \tag{4.14}$$

From this, the distances of all of the objects in the boundary region in relation to the lower approximation mean can now be calculated:

$$\begin{array}{llll} obj & 0 & \sqrt{f_b(\underline{P}X_{MEAN}, 0)^2} = 0.4 \\ obj & 2 & \sqrt{f_b(\underline{P}X_{MEAN}, 2)^2} = 0.5 \\ obj & 3 & \sqrt{f_b(\underline{P}X_{MEAN}, 3)^2} = 0.4 \\ obj & 4 & \sqrt{f_b(\underline{P}X_{MEAN}, 4)^2} = 0.4 \end{array}$$

The significance value is therefore:

$$\begin{aligned} \omega_B(F) &= \left(\textstyle\sum_{y \in BND_P(Q))} \delta_P(\underline{P}X_{MEAN}, y)\right)^{-1} \\ &= (\textstyle\sum (0.4, 0.5, 0.4, 0.4))^{-1} = 0.588 \end{aligned}$$

The significance value is combined with the rough set dependency to form a subset measure (M) such that the value for $\{b\}$:

$$M(B) = \frac{\omega_B(F) + \gamma_B(F)}{2} = \frac{0.588 + 0.333}{2} = 0.461$$

By calculating the change in combined significance and dependency value (M) when an attribute is removed from the set of considered conditional attributes, a measure of the goodness of that attribute can be obtained. The greater the change in M the greater the measure of goodness that attribute has attached to it.

The values for the combined metric can be calculated for all considered subsets of conditional attributes using DMRSAR:

$$\begin{array}{ll} M_{\{a\}}(\{f\}) = 0.0 & M_{\{a,c\}}(\{f\}) = 0.498 \\ M_{\{b\}}(\{f\}) = 0.461 & M_{\{b,c\}}(\{f\}) = 1.0 \\ M_{\{c\}}(\{f\}) = 0.805 & M_{\{a,b,c\}}(\{f\}) = 0.492 \end{array}$$

It is obvious from the above example that the search finds a subset in the manner $\{c\} \rightarrow \{b, c\}$. As $\{a\}$ and $\{a, c\}$ and also $\{a, b, c\}$ do not result in the same increase in combined metric these subsets are ignored.

4.3 Experimentation

This section presents the results of experimental studies using 8 real-valued datasets. These datasets are of the same format as that used for the worked example in the previous section.

4.3.1 Experimental Setup

The datasets employed here are well-known machine learning benchmark examples and are small-to-medium in size, with between 120 and 390 objects per dataset and feature sets ranging from 5 to 39. All datasets have been obtained from [5] and [158].

A comparison of both the tolerance rough set algorithm and the distance-metric based tolerance rough set dimensionality reduction techniques is made based on subset size, and classification accuracy. Furthermore, the DM-TRS approach is also compared with five other FS techniques. The comparison is made in terms of both subset size and classification accuracy and also in terms of classification accuracy for each given subset size discovered by the DM-TRS method where applicable.

A range of 4 tolerance values, (0.80–0.95 in intervals of 0.05) were employed when considering the datasets. It should be borne in mind that the ideal tolerance value for any given dataset can only be optimised for that dataset by repeated experimentation. This is true of the TRSM as well as to any extensions applied to it, such as described here. Therefore, the range of values chosen is an attempt to demonstrate the ideal tolerance threshold for each dataset without exhaustive experimentation.

In the generation and discussion of results for classification accuracies, a fuzzy classifier learning method QSBA [188], and three other classifier learners - J48, JRip, and PART [237] - were employed. QSBA is briefly outlined below, the other classifier learners have been covered in detail in Section 3.5.1.

QSBA [188] works by generating fuzzy rules using the fuzzy subsethood measure for each decision class and a threshold to determine what appears in the rule for that decision class. The fuzzy subsethood measure is then used to act as weights, and the algorithm then modifies the weights to act as fuzzy quantifiers.

4.3.2 Comparison of Classification Accuracy

The data presented in Table 4.2 shows the average classification accuracy using the classifiers learned by the four learner methods described previously. The recorded values are expressed as a percentage and obtained using 10-fold cross validation. Classification was initially performed on the unreduced dataset, followed by the reduced datasets, which were obtained by using both the TRS and DM-TRS dimensionality reduction techniques respectively for each of the tolerance values.

Examining the classification values obtained using QSBA, even when the subset size in Table 4.7 is of a similar value to that of the TRS approach, the corresponding classification figures for DM-TRS demonstrate the selection of better quality

	QSBA	$\tau = 0.8$		$\tau = 0.85$		$\tau = 0.90$		$\tau = 0.95$	
Dataset	Unred.	TRS	DM-TRS	TRS	DM-TRS	TRS	DM-TRS	TRS	DM-TRS
water 2	57.94	77.76	77.76	73.16	73.16	74.53	74.53	67.79	76.38
water 3	48.97	63.12	63.12	74.34	74.34	73.56	73.56	68.25	63.83
cleveland	37.46	36.51	39.47	35.78	35.78	43.58	46.61	43.28	43.28
glass	43.65	37.60	37.60	38.51	38.51	25.88	25.88	42.12	39.43
heart	64.07	77.41	77.42	73.33	74.07	70.00	70.00	74.81	74.81
ionosphere	80.67	74.34	74.34	68.26	68.26	68.26	69.14	64.10	65.65
olitos	64.16	61.66	64.16	57.50	86.08	61.66	62.36	54.16	60.01
wine	94.86	85.39	85.39	81.40	81.40	84.11	84.11	83.72	84.10

Table 4.2: Classification accuracy using QSBA

	TRS			DM-TRS		
Dataset	JRIP	PART	J48	JRIP	PART	J48
water 2	83.58	84.61	83.58	83.58	84.61	83.58
water 3	84.61	81.80	83.84	84.61	81.80	83.84
cleveland	52.86	52.18	53.19	55.55	53.53	54.20
glass	50.00	49.53	48.13	50.00	49.53	48.13
heart	73.70	78.89	75.56	73.70	78.89	75.56
ionosphere	89.13	88.26	88.26	89.13	88.26	88.26
olitos	67.50	70.00	64.16	65.83	62.50	59.16
wine	95.50	94.38	81.40	81.40	84.11	84.11

Table 4.3: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.80$)

subsets. In some cases the DM-TRS approach even manages to select a subset of smaller cardinality for a given dataset, whilst also maintaining a similar level of classification as TRS.

Obviously, where DM-TRS discovers identical subsets to those found by TRS, the classification accuracies will also be identical. Where this is not the case however, the results can differ substantially depending on whether fuzzy or crisp classifiers have been employed in obtaining the results e.g. for the *water 3* dataset with ($\tau = 0.95$), the crisp classifiers show an average result for DM-TRS that is better than TRS, whilst the fuzzy classifier shows a result that is poorer than

	TRS			DM-TRS		
Dataset	JRIP	PART	J48	JRIP	PART	J48
water 2	84.61	82.30	84.87	84.61	82.30	84.87
water 3	83.58	82.30	81.02	83.58	82.30	81.02
cleveland	53.87	50.84	54.54	53.87	50.54	54.54
glass	64.95	60.74	68.22	61.93	66.82	68.70
heart	75.55	77.40	82.59	81.85	80.74	82.63
ionosphere	90.42	88.69	86.52	90.42	88.69	86.52
olitos	62.50	60.83	60.00	62.50	65.83	67.50
wine	95.25	95.50	96.06	95.25	95.50	96.06

Table 4.4: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.85$)

	TRS			DM-TRS		
Dataset	JRIP	PART	J48	JRIP	PART	J48
water 2	85.38	82.30	87.43	85.38	82.30	87.43
water 3	80.00	81.53	76.67	80.00	81.53	76.67
cleveland	54.20	53.87	52.52	54.03	55.55	54.88
glass	65.88	69.15	68.69	65.88	69.15	68.69
heart	79.25	75.19	78.88	79.25	75.19	78.88
ionosphere	85.65	86.52	85.21	86.01	89.56	89.13
olitos	70.00	65.83	61.66	59.17	60.84	67.50
wine	96.06	94.94	96.62	96.06	94.94	96.62

Table 4.5: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.90$)

	TRS			DM-TRS		
Dataset	JRIP	PART	J48	JRIP	PART	J48
water 2	82.82	83.07	82.05	84.10	84.10	80.77
water 3	81.02	80.77	81.02	83.59	78.98	81.80
cleveland	50.54	50.84	54.54	50.54	50.84	54.54
glass	69.62	68.22	69.62	65.42	64.95	62.00
heart	80.38	78.57	81.48	80.38	78.57	81.48
ionosphere	86.08	87.39	87.39	85.93	87.82	87.82
olitos	64.16	66.67	64.16	64.16	65.88	64.16
wine	93.25	95.50	96.02	91.57	94.98	97.19

Table 4.6: Classification accuracy using JRIP, PART, and J48 classifiers ($\tau = 0.95$)

TRS. For the same tolerance value (0.95), the *glass* dataset, also demonstrates a small decrease in the order of up to 7% (for all classifiers), however when the corresponding decrease in dimensionality of 37.5% is considered over the TRS method, this decrease is not significant. In all other cases where the crisp classifiers show a decrease in classification accuracy, this is reflected as an increase when QSBA is employed for classification. This is due mainly to the fact that although J48, JRip, and PART are intended to handle real-valued data, they are unable to examine data in the same way that a fuzzy classifier learner such as QSBA can.

4.3.3 Subset Sizes

Table 4.7 presents the results of a comparison of subset size, for both the TRS and DM-TRS approaches, with DM-TRS showing a small but clear advantage in terms of more compact subsets. Note that * indicates a subset whose size was the same as TRS but for which different attributes had been selected.

Examining the results in Table 4.7, the DM-TRS method shows that there is much information contained in the boundary region of a tolerance rough set. This is reflected in the subset sizes obtained. DM-TRS succeeds in finding subsets of

Dataset	Original number of features	Subset size ($\tau = 0.8$) TRS	Subset size ($\tau = 0.8$) DM-TRS	Subset size ($\tau = 0.85$) TRS	Subset size ($\tau = 0.85$) DM-TRS	Subset size ($\tau = 0.90$) TRS	Subset size ($\tau = 0.90$) DM-TRS	Subset size ($\tau = 0.95$) TRS	Subset size ($\tau = 0.95$) DM-TRS
water 2	39	6	6	5	5	8	8	12	12*
water 3	39	5	5	9	9	9	9	12	11
cleveland	14	3	2	2	2	11	10	8	8
glass	10	3	3	5	5*	3	3	8	3
heart	14	4	4	6	8	12	12	8	8
ionosphere	34	3	3	6	6	6	6*	8	8*
olitos	25	8	5	5	5*	9	8	6	6*
wine	13	5	5	4	4	5	5	5	5*

Table 4.7: Comparison of subset size for each tolerance threshold value

cardinality that are at least equal and sometimes smaller than those obtained using the TRS method, with the exception of the *heart* dataset for $\tau = 0.85$. However if the classification results are examined closely, it is clear that although the subset size is of greater cardinality for this particular case, the subset is of greater quality than that obtained using TRS. The results also demonstrate that the nature of the data along with a particular value of τ can mean that there is little or no information in the boundary region and therefore DM-TRS relies purely on the information contained in the lower approximation dependency value. This can in turn, result in subsets that are identical to those discovered by the TRS method.

It is expected that a decrease in τ would reflect a change in performance in terms of subset size for the TRS method alone such that an optimal value is arrived at after a period of experimentation. This occurs as the lower threshold allows a greater flexibility in the membership of data objects to tolerance classes. However the results for subset size demonstrate an interesting trend where the DM-TRS method may actually discover smaller subset sizes than TRS for similar tolerance threshold values. As the DM-TRS method examines the boundary region information, it would be expected that a decrease in τ (thereby increasing the number of objects in the lower approximation and decreasing the number of objects in the boundary region) would result in the DM-TRS performing poorly for the next decrement of threshold value documented above – as there is less information contained in the boundary region for the DM-TRS method to examine. However, if the results in Tables 4.7 and 4.2 are examined for e.g. the dataset *olitos*, it can be seen that DM-TRS selects subsets which are of smaller size and in some cases of better quality. This suggests that, as long as there is some information in the boundary region, regardless of whether the optimal value of τ has been obtained, DM-TRS can select subsets of better quality than TRS.

4.3.4 Comparison with Randomly Selected Subsets

The FS process helps to remove measurement noise as a positive by-product. A valid question therefore is whether other subsets of dimensionality 5 (e.g. for the "water 2" dataset) would perform similarly as those identified by DM-TRS selection. To avoid a biased answer to this, and without resorting to exhaustive computation 30 sets of five features have been randomly chosen in order to see what classification results might be achieved.

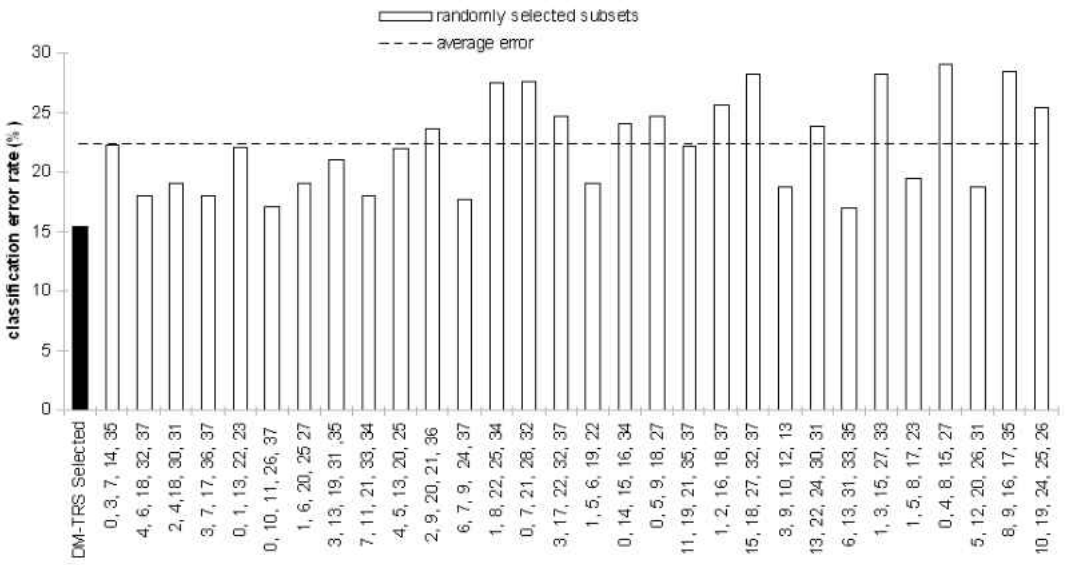

Figure 4.2: DM-TRS vs. randomly selected subsets

Fig. 4.2 shows the error rate of the corresponding 30 classifiers, along with the error rate of the classifier that uses the DM-TRS selected subset. The average error of the classifiers that each employ five randomly selected features is 22.32%, far higher than that attained by the classifier which utilises the DM-TRS selected subset of the same dimensionality. This implies that those randomly selected entail important information loss in the course of feature selection; this is not the case for the DM-TRS selection-based approach.

4.3.5 Hausdorff Metric Implementation

The Hausdorff metric approach to distance measurement has been described previously as an alternative to the mean lower approximation and Euclidean distance based method which was used to generate the empirical results described in the preceding sections.

In this section the DM-TRS approach was augmented with the Hausdorff metric. This metric is used to measure the distance between the lower approximation

and the boundary region, and was implemented in order to investigate the performance of this method in terms of subset size. The results of this investigation are included here in Table 4.3.5. For brevity only the results for a single tolerance value are included.

Dataset	DM-TRS Subset Size	Hausdorff Metric Subset Size
water2	8	10
water3	9	32
cleveland	10	12
glass	3	9
heart	12	13
ionosphere	6	16
olitos	8	14
wine	5	13

Table 4.8: DMRSAR – Hausdorff metric implementation (τ =0.90)

It is clear that this particular implementation of the *Hausdorff* metric fails to capture the useful information of the boundary region in the same way that the mean lower approximation method does. Examining the results for subset size, it can be seen that the existing DM-TRS approach returns superior results in all cases. This approach took a considerable length of time to run, however this was to be expected as there are a large number of distance calculations performed even for small datasets (exponential $O(n^2)$ for n upper approximation objects).

4.3.6 Comparison with Existing FS Methods

In this section further comparison of DM-TRS with some of the more traditional dimensionality reduction and FS techniques demonstrates the approach in a more comprehensive manner. DM-TRS is compared with principal component analysis (PCA) [48], ReliefF [109], CFS [68], consistency-based FS [237], and a wrapper method employing J48 [184] as an evaluation metric.

4.3.6.1 PCA

PCA is a versatile transformation-based DR technique which projects the data onto a new coordinate system of reduced dimensions. This process of linear transformation however also transforms the underlying semantics or meaning of the data. This results in data that is difficult for humans to interpret, but which may still provide useful automatic classification of new data. In order to ensure that the comparison of DM-TRS and PCA is balanced, the same subset sizes discovered

for each dataset and tolerance level are also employed in the analysis of PCA, e.g. *olitos* in Table 4.7 has subsets of size 5, 6, and 8. Each of the best number of transformed features are utilised for PCA, (in this case the best 5, 6, and 8).

The results in Table 4.3.6.1 show that of the eight datasets only *olitos* demonstrates a consistent decrease in classification accuracy performance for DM-TRS (see future work for further discussion). There are other instances where PCA slightly outperforms the DM-TRS method but this is not consistent and in a majority of cases DM-TRS usually shows equal performance or an increase in classification accuracy.

subset size		PCA			DM-TRS		
		J48	JRIP	PART	J48	JRIP	PART
water 2	5	83.33	83.84	83.07	84.61	82.30	84.87
	6	86.41	85.38	87.69	84.87	84.61	82.30
	8	81.02	83.58	83.33	85.38	82.30	87.43
	12	85.89	84.36	81.28	84.10	84.10	80.77
water 3	5	87.94	85.64	83.58	84.61	81.80	83.84
	9	82.30	84.36	81.35	83.58	82.30	81.02
	11	84.35	85.38	83.07	83.59	78.98	81.80
cleveland	2	58.92	53.87	57.23	55.55	53.53	54.20
	8	56.90	57.91	54.20	50.54	50.84	54.54
	10	51.85	52.18	50.16	54.03	55.55	54.88
glass	3	64.48	61.68	65.42	65.88	69.15	68.69
	5	68.61	61.21	66.35	61.93	66.82	68.70
heart	4	82.96	82.59	82.96	73.70	78.89	75.56
	8	79.25	83.33	79.62	81.85	80.74	82.63
	12	82.59	84.07	78.14	79.25	75.19	78.88
ionosphere	3	77.39	77.39	79.56	89.13	88.26	88.26
	6	83.04	86.08	79.56	90.42	88.69	86.52
	8	82.60	85.21	82.17	85.93	87.82	87.82
olitos	5	85.00	80.00	82.50	62.50	65.83	75.56
	6	85.00	81.66	81.66	64.16	65.88	64.16
	8	80.33	75.00	80.33	59.17	60.84	67.50
wine	4	93.25	92.69	93.82	95.25	95.50	96.06
	5	93.25	89.88	94.38	96.06	94.94	96.62

Table 4.9: PCA & DM-TRS – Comparison of classification accuracy

It should be emphasised however that while PCA might outperform DM-TRS in some instances in terms of classification accuracy, the semantics of the data is irreversibly transformed following dimensionality reduction. This can have consequences where human interpretability of the data is important, which is one of the key reasons for performing feature selection tasks to begin with. As DM-TRS is a *feature selection* approach as opposed to a *feature ranking* method, a predefined threshold is not required; selection is complete as soon as the termination criterion (rough set dependency) is fulfilled. The rough set dependency value is

integral to the selection process and as such, in contrast to PCA does not need to be predefined.

Finally, it is worth noting that PCA is selected for comparison here due to recognition of the fact that it is an established approach for dimensionality reduction. However, such comparison uses PCA as a global step prior to classification. This may not maximise the potential of PCA serving as a feature reduction tool. It may be a better approach to include PCA as an intrinsic substep of LDA [183], [73]. However, the FS method employed here is a global preprocessor for semantics-preserving dimensionality reduction and hence PCA is examined in a similar manner.

4.3.6.2 CFS - Correlation-based Feature Selection

CFS [68] is a filter-based approach to FS and uses a search algorithm along with an evaluation metric to decide on the 'goodness' or merit of potential feature subsets. Rather than scoring (and ranking) individual features, the method scores (and ranks) the worth of subsets of features. As the feature subset space is usually large, CFS employs a best-first-search heuristic. This heuristic algorithm takes into account the usefulness of individual features for predicting the class along with the level of intercorrelation amongst features using the premise that good feature subsets contain features that are highly correlated to the class, yet not correlated to each other. CFS calculates a matrix of feature-to-class and feature-to-feature correlations from the training data.

The subset generation technique employed in this case was a greedy-hillclimbing type similar to DM-TRS, where features are added greedily until the termination criteria are fulfilled. The results for subset size and classification values for the three classifier learners are illustrated in Table 4.10.

Dataset	subset size	JRIP	PART	J48
water 2	9	83.33	83.07	84.61
water 3	11	82.30	82.05	81.79
cleveland	7	55.54	57.91	58.92
glass	7	65.42	68.69	69.15
heart	7	77.40	77.03	81.11
ionosphere	11	90.00	90.00	90.00
olitos	16	69.16	71.67	69.16
wine	11	94.38	93.82	94.38

Table 4.10: CFS Subset size and classification accuracy

Unlike DM-TRS, CFS has no tunable parameters which means that it can be quite difficult to compare the results of Tables 4.3–4.7 with those obtained here.

It would be easy just to pick the optimal result for DM-TRS and state that the approach is better based on those performance figures. Two different approaches have therefore been adopted. The first approach is to obtain a mean for all of the subset sizes and classification values for DM-TRS for all values of τ and compare these with CFS. The second is to compare CFS and DM-TRS by finding a subset size in the results for DM-TRS that is comparable to that obtained by CFS and use the associated classification figures. So, if CFS has a subset size of 10 for a particular dataset, find a subset of identical or similar size in the DM-TRS results in Table 4.7 and use this to compare classification accuracy.

Dataset	subset size	JRIP	PART	J48
water 2	7.75	84.29	83.32	84.16
water 3	8.50	82.94	81.15	80.83
cleveland	5.50	53.49	52.61	54.54
glass	3.5	60.80	62.61	61.88
heart	8	78.79	78.34	79.63
ionosphere	5.75	87.87	68.83	87.93
olitos	6	62.91	63.76	64.58
wine	4.75	91.07	92.38	93.49

Table 4.11: Average subset size and classification accuracy for DM-TRS

The results for CFS when compared with the mean values for DM-TRS demonstrate that the DM-TRS method has a clear advantage in terms of subset size. The only exception perhaps is the result for the *heart* dataset, however if Table 4.7 is examined, it can be seen that DM-TRS is capable of reducing this value to *4*. The mean classification values for DM-TRS although not as clear as those for subset size show that the difference in classification accuracy between the two approaches is less than 8% even in the most extreme cases e.g. *olitos* and *glass*. It must be remembered however that the figures are *mean* values, and that DM-TRS outperforms CFS in many of the examples for individual values of τ.

Dataset	subset size	JRIP	PART	J48
water 2	8*	85.38	82.30	87.43
water 3	11	83.59	78.98	81.80
cleveland	8*	50.54	50.84	54.54
glass	5	61.93	66.82	68.70
heart	8*	81.85	80.74	82.63
ionosphere	8*	85.93	87.82	87.82
olitos	8*	59.17	60.84	67.50
wine	5*	96.06	94.94	96.62

Table 4.12: Closest comparable subset size and classification accuracy for DM-TRS

The second approach to comparing CFS with DM-TRS uses information which is derived from Tables 4.3–4.7, perhaps most apparent is the fact that DM-TRS on the whole selects subsets which are more compact than those selected by CFS. The classification values tell a similar story, however some values are lower than those obtained by CFS. The reason for this is related to the fact that suboptimal results must be chosen in order to find a way to compare this approach with CFS, e.g. the *glass* dataset shows comparable classification results to the values recorded in Table 4.12 as it does in Table 4.4 and Table 4.7 but with a subset size of only *5*. Thus it achieves greater reduction in dimensionality yet retains the classification ability, and easily outperforms CFS. It should be noted in the case of Table 4.12 that subsets marked with an asterisk (*) are of size which was not identical to that obtained by CFS but represented the closest available value.

4.3.6.3 Consistency-based Feature Selection

Consistency-based feature selection [41] employs a consistency measure for objects in a dataset. Consistency is measured by comparing the values of a given feature set over a set of objects. There are three steps necessary to calculate the consistency rate for a set of objects: a) Consider two objects where the feature values of both are identical but their respective decision feature classes are not, e.g *object1* ={1 0 1 a}, and *object2* = {1 0 1 b}, (where $a \neq b$) in this case objects 1 and 2 are said to be inconsistent; b) The inconsistency count for an object is the number of times objects with the same feature values appear in the dataset minus the largest number amongst different decision feature classes, e.g. for n objects with identical decision feature values for which $o1$ objects belong to the $d1$ decision feature class, $o2$ to the $d2$ decision feature class, and $o3$ to the $d3$ decision feature class, and $|d1| + |d2| + |d3| = n$ Assume that $|d2|$ is the greatest of all three, the consistency count can be calculated as: $n - |d2|$; c). The consistency rate can then be calculated by summing the consistency counts for the number of groups of objects of given feature values of a subset, divided by the total number of objects.

The FS approach used in this consistency-based method employs a greedy stepwise subset generation technique similar to that of DM-TRS. Again, as with CFS, this method has no tunable parameters, and must be compared with DM-TRS in the same manner as that employed in the previous subsection.

Examining the results in Table 4.13 and comparing them with those of Table 4.11 it is clear that like CFS, the subset sizes obtained for consistency-based FS are greater than the average result obtained using DM-TRS. The classification results show similar performance to CFS with some insignificant increases or decreases with respect to certain datasets, but overall comparable to DM-TRS.

Dataset	subset size	JRIP	PART	J48
water 2	14	84.35	85.60	83.58
water 3	11	83.84	82.56	81.02
cleveland	9	54.54	55.21	56.22
glass	7	65.42	71.96	64.48
heart	10	78.88	74.04	78.88
ionosphere	7	89.56	88.69	89.56
olitos	11	67.50	65.00	68.33
wine	5	90.43	97.19	97.12

Table 4.13: Subset size and classification accuracy results for consistency-based FS

4.3.6.4 ReliefF

ReliefF [109] is an extension of Relief [105] but which has the ability to deal with multiple decision classes. In ReliefF each feature is given a relevance weighting that reflects its ability to discern between the decision class labels. The first threshold, specifies the number of sampled objects used for constructing the weights. For each sampling, an object x is randomly chosen, and its 'near hit' and 'near miss' are calculated. These are x's nearest objects with the same class label and different class label respectively. The user has to supply a threshold which determines the level of relevance that features must surpass in order to be finally chosen.

ReliefF is typically used in conjunction with a feature ranking method employed for the selection of features. In this experimental comparison, the number of nearest neighbours for feature estimation was set to 10, and the other parameter ReliefF requires namely sigma or the influence of nearest neighbours was set to 2. The number of features to select was applied according to the optimal subset sizes obtained for DM-TRS.

Dataset	(predefined) subset size	JRIP	PART	J48
water 2	5	83.33	84.61	84.10
water 3	5	83.84	81.02	81.53
cleveland	2	58.24	58.21	53.87
glass	3	68.22	68.69	65.42
heart	4	78.50	77.77	78.51
ionosphere	7	86.02	87.82	86.52
olitos	5	65.00	70.03	65.00
wine	4	91.00	93.82	89.87

Table 4.14: Subset size and classification accuracy results for ReliefF

The classification results obtained show that despite the improved search method employed by ReliefF, the DM-TRS classification accuracies are comparable with little difference or even a small increase in most cases for DM-TRS.

4.3.6.5 Wrapper FS Employing J48

Although DM-TRS is a filter type FS method, it is interesting to compare it with wrapper-based FS techniques also. Having recognised this, a comparison of the performance of DM-TRS with that of C4.5 [184] which is one of the well known wrapper methods is presented here.

To compare these two approaches meaningfully, the 8 datasets were divided into training and test data respectively. This was accomplished by removing half of the objects from the original data at random and using this data as 'test' data whilst the remainder is used as 'training' data. The results illustrated in Table 4.15 show the classification accuracies recorded having performed FS on the 'test' data.

	C4.5 Wrapper			DM-TRS		
Dataset	JRIP	PART	J48	JRIP	PART	J48
water 2	90.76	91.28	89.74	90.88	91.65	90.10
water 3	83.84	81.02	81.53	88.71	84.61	86.67
cleveland	51.67	47.65	53.60	52.03	54.05	56.67
glass	78.50	74.76	82.24	79.86	74.76	83.85
heart	75.37	76.86	77.61	77.03	77.77	80.27
ionosphere	86.08	85.21	84.34	90.63	92.45	94.44
olitos	61.66	71.66	63.33	65.33	71.78	65.00
wine	88.76	88.76	87.64	96.62	96.62	92.13

Table 4.15: Subset size and classification accuracy results for consistency-based FS

One would expect that the wrapper should outperform any filter method in terms of classification accuracy as the validation step is carried out using a classifier. The results demonstrate however that this is not strictly the case, and DM-TRS shows a clear increase in classification accuracy over the wrapper method. The increase is small and in some cases in the order of a few percent, but the wrapper method has an extremely high computational overhead. This means that execution times are considerably affected as a result.

4.4 The Fuzzy-Rough Set Boundary Region

Fuzzy-rough sets [52], like TRSM also have the ability (which is lacking in traditional RST), to deal with real-valued data. However, TRSM also requires the specification of an additional parameter (τ) in order to operate. This additional parameter is human-specified and repeated experimentation may be required to arrive at an ideal value for any given dataset. The specification of such a parameter is counter to the rough set tenet of using only the information contained in

the data. This has led to the exploration of fuzzy-rough sets as an alternative, which does not require any thresholding information.

In the context of exploration of the boundary region, fuzzy-rough sets offer all of the advantages of traditional RST, but also have the ability to deal with real-valued data. In the very strict sense, it would be incorrect to say that fuzzy-rough does not require any additional human input, as decisions about which fuzzy connectives ans similarity measures need to be made. However, subjective thresholding values are not required. This section presents some new evaluation metrics for fuzzy-rough feature selection, based on the fuzzy entropy measure. These metrics are applied to the fuzzy-rough lower approximation and most importantly to the fuzzy-rough boundary region.

4.4.1 Fuzzy-Rough Feature Selection (FRFS)

In the past, work on fuzzy-rough feature selection used a fuzzy partitioning of the input space [204] in order to determine fuzzy equivalence classes. Alternative definitions for the fuzzy lower and upper approximations can be found in [186], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x,y), \mu_X(y)) \tag{4.15}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) \tag{4.16}$$

Here, I is a fuzzy implicator and T a t-norm. R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu_{R_P}(x,y) = \bigcap_{a \in P} \{\mu_{R_a}(x,y)\} \tag{4.17}$$

$\mu_{R_a}(x,y)$ is the degree to which objects x and y are similar for feature a. Many fuzzy similarity relations can be constructed such as equation 4.1, and others:

$$\mu_{R_a}(x,y) = \exp(-\frac{(a(x)-a(y))^2}{2{\sigma_a}^2}) \tag{4.18}$$

$$\mu_{R_a}(x,y) = \max(\min(\frac{(a(y)-(a(x)-\sigma_a))}{(a(x)-(a(x)-\sigma_a))}, \frac{((a(x)+\sigma_a)-a(y))}{((a(x)+\sigma_a)-a(x))}, 0) \tag{4.19}$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display T-transitivity, the fuzzy transitive closure must be computed for each attribute [45]. The combination of feature relations in equation (4.17) has been shown to

preserve T-transitivity [228].

4.4.1.1 Reduction

In a similar way to the original RSAR approach (Section 2.3.2.1), the fuzzy positive region [99] can be defined as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \tag{4.20}$$

The resulting degree of dependency is:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \tag{4.21}$$

A fuzzy-rough reduct R can be defined as a minimal subset of features that preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough QuickReduct algorithm can be constructed that operates in the same way as Fig. 2.14, but uses equation (4.21) to gauge subset quality. In [99], it has been shown that the dependency function is monotonic and that fuzzy discernibility matrices may also be used to discover reducts.

Core features may be determined by considering the change in dependency of the full set of conditional features when individual attributes are removed:

$$Core(\mathbb{C}) = \{a \in \mathbb{C} | \gamma'_{\mathbb{C}-\{a\}}(Q) < \gamma'_{\mathbb{C}}(Q)\} \tag{4.22}$$

4.4.2 Fuzzy Boundary Region-based FS

The lower approximation contains information regarding the extent of certainty of object membership to a given concept. However, the upper approximation contains information regarding the degree of uncertainty of objects and hence this information can be used to discriminate between subsets. For example, two subsets may result in the same lower approximation but one subset may produce a smaller upper approximation. This subset will be more useful as there is less uncertainty concerning objects within the boundary region.

Following the original rough set approach, the fuzzy-rough boundary region for a concept X can be defined by:

$$\mu_{BND_{R_P}(X)}(x) = \mu_{\overline{R_P}X}(x) - \mu_{\underline{R_P}X}(x) \tag{4.23}$$

When the decision feature is real-valued the same fuzzy similarity measure is employed, resulting in the relation $R_{\mathbb{D}}$ with foresets $D_1, D_2, ..., D_n$. The fuzzy-

rough boundary region then becomes:

$$\mu_{BND_{R_P}(D_j)}(x) = \frac{\mu_{\overline{R_P}D_j}(x) - \mu_{\underline{R_P}D_j}(x)}{|D_j|} \tag{4.24}$$

for decision foreset D_j, where $|D_j|$ stands for the cardinality of D_j.

4.4.2.1 Reduction

As the search for an optimal subset progresses, the object memberships to the boundary region for each concept diminishes until a minimum is achieved. For crisp rough set FS, the boundary region will be zero for each concept when a reduct is discovered. This may not necessarily be the case for fuzzy-rough FS due to the additional imprecise information (ID) involved. The ID for a concept X described using features in P can be calculated as follows:

$$U_P(X) = \frac{\sum_{x \in \mathbb{U}} \mu_{BND_{R_P}(X)}(x)}{|\mathbb{U}|} \tag{4.25}$$

This is the average extent to which objects belong to the fuzzy boundary region for the concept X. The total ID degree for all concepts, given a feature subset P is defined as:

$$\lambda_P(\mathbb{D}) = \frac{\sum_{X \in \mathbb{U}/\mathbb{D}} U_P(X)}{|\mathbb{U}/\mathbb{D}|} \tag{4.26}$$

When the decision feature is fuzzy, this becomes:

$$\lambda_P(\mathbb{D}) = \frac{\sum_{D_j \in R_{\mathbb{D}}} U_P(D_j)}{\sum_{D_n \in R_{\mathbb{D}}} (|D_n|)^{-1}} \tag{4.27}$$

Obviously, this degenerates to the previous definition when dealing with crisp decisions. A QuickReduct-style algorithm can be constructed for locating fuzzy-rough reducts based on this measure. Instead of maximising the dependency degree, the task of the algorithm is to minimize the total uncertainty degree. When this reaches the minimum for the dataset, a fuzzy-rough reduct has been found.

Theorem 1. *B-FRFS monotonicity. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. Then $\lambda_{P \cup \{a\}}(Q) \leq \lambda_P(Q)$.*

Proof. The fuzzy boundary region of a concept X for an object x and set of

features $P \cup \{a\}$ is defined as

$$\mu_{BND_{R_{P\cup\{a\}}}(X)}(x) = \mu_{\overline{R_{P\cup\{a\}}}X}(x) - \mu_{\underline{R_{P\cup\{a\}}}X}(x)$$

For the fuzzy upper approximation component of the fuzzy boundary region:

$$\mu_{\overline{R_{P\cup\{a\}}}X}(x) = \sup_{y\in\mathbb{U}} T(\mu_{R_{P\cup\{a\}}}(x,y), \mu_X(y))$$

It is known from Theorem 1 in [96] that $\mu_{R_{P\cup\{a\}}}(x,y) \leq \mu_{R_P}(x,y)$, so $\mu_{\overline{R_{P\cup\{a\}}}X}(x) \leq \mu_{\overline{R_P}X}(x)$. As $\mu_{\underline{R_{P\cup\{a\}}}X}(x) \geq \mu_{\underline{R_P}X}(x)$, then $\mu_{BND_{R_{P\cup\{a\}}}(X)}(x) \leq \mu_{BND_{R_P}(X)}(x)$. Thus, $U_{P\cup\{a\}}(Q) \leq U_P(Q)$ and therefore $\lambda_{P\cup\{a\}}(Q) \leq \lambda_P(Q)$. □

4.4.2.2 Example

Object	a	b	c	q
1	−0.4	−0.3	−0.5	no
2	−0.4	0.2	−0.1	yes
3	−0.3	−0.4	−0.3	no
4	0.3	−0.3	0	yes
5	0.2	−0.3	0	yes
6	0.2	0	0	no

Table 4.16: Example dataset

To determine the fuzzy boundary region, the lower and upper approximations of each concept for each feature must be calculated. Considering feature a and concept {1,3,6}:

$$\mu_{BND_{R_a}(\{1,3,6\})}(x) = \mu_{\overline{R_a}\{1,3,6\}}(x) - \mu_{\underline{R_a}\{1,3,6\}}(x)$$

For object 4, this is

$$\begin{aligned}
\mu_{BND_{R_a}(\{1,3,6\})}(4) &= \sup_{y\in\mathbb{U}} T(\mu_{R_a}(4,y), \mu_{\{1,3,6\}}(y)) \\
&\quad - \inf_{y\in\mathbb{U}} I(\mu_{R_a}(4,y), \mu_{\{1,3,6\}}(y)) \\
&= 0.699 - 0.0 \\
&= 0.699
\end{aligned}$$

For the remaining objects, this is:

$$\begin{aligned}
\mu_{BND_{R_a}(\{1,3,6\})}(1) &= 1.0 \\
\mu_{BND_{R_a}(\{1,3,6\})}(2) &= 1.0 \\
\mu_{BND_{R_a}(\{1,3,6\})}(3) &= 0.699 \\
\mu_{BND_{R_a}(\{1,3,6\})}(5) &= 1.0 \\
\mu_{BND_{R_a}(\{1,3,6\})}(6) &= 1.0
\end{aligned}$$

Hence, the ID for concept {1,3,6} is:

$$\begin{aligned}
U_a(\{1,3,6\}) &= \frac{\sum\limits_{x\in\mathbb{U}} \mu_{BND_{R_a}(\{1,3,6\})}(x)}{|\mathbb{U}|} \\
&= \frac{1.0+1.0+0.699+0.699+1.0+1.0}{6} \\
&= 0.899
\end{aligned}$$

For concept $\{2,4,5\}$, the ID is:

$$\begin{aligned}
U_a(\{2,4,5\}) &= \frac{\sum\limits_{x\in\mathbb{U}} \mu_{BND_{R_a}(\{2,4,5\})}(x)}{|\mathbb{U}|} \\
&= \frac{1.0+1.0+0.699+0.699+1.0+1.0}{6} \\
&= 0.899
\end{aligned}$$

From this, the total ID for feature a is calculated as follows:

$$\begin{aligned}
\lambda_a(Q) &= \frac{\sum\limits_{X\in\mathbb{U}/Q} U_a(X)}{|\mathbb{U}/Q|} \\
&= \frac{0.899+0.899}{2} \\
&= 0.899
\end{aligned} \tag{4.28}$$

The values of the total ID for the remaining features are:

$$\lambda_{\{b\}}(Q) = 0.640 \quad \lambda_{\{c\}}(Q) = 0.592$$

As feature c results in the smallest total imprecision degree, it is chosen and added to the reduct candidate. The algorithm then considers the addition of the remaining features to the subset:

$$\lambda_{\{a,c\}}(Q) = 0.500 \quad \lambda_{\{b,c\}}(Q) = 0.0$$

The subset $\{b, c\}$ results in the minimal imprecision degree for the dataset, and the algorithm terminates. Interestingly, this is the same subset as that chosen by the fuzzy lower approximation-based method above.

4.4.3 Integration of Fuzzy Entropy

In the above method, the overall uncertainty is evaluated by averaging the uncertainty of all decision concepts. The ID for a concept is itself an average measure of the *belonging* of objects to the fuzzy boundary region. A more appropriate way of measuring the uncertainty within the boundary region of a concept X is to calculate the fuzzy entropy:

$$U'_P(X) = \sum_{x \in \mathbb{U}} -\frac{\mu_{BND_{R_P}(X)}(x)}{|BND_{R_P}(X)|} \log_2 \frac{\mu_{BND_{R_P}(X)}(x)}{|BND_{R_P}(X)|} \tag{4.29}$$

$$\lambda'_P(\mathbb{D}) = \frac{\sum\limits_{D_j \in R_\mathbb{D}} U'_P(D_j)}{\sum_{D_n \in R_\mathbb{D}} (|D_n|)^{-1}} \tag{4.30}$$

This will be minimized when all fuzzy boundary regions are empty, hence $\lambda'_P(\mathbb{D}) = \lambda_P(\mathbb{D}) = 0$ and therefore P must be a fuzzy-rough reduct.

4.4.4 Fuzzy-Rough Reduction with Fuzzy Entropy

Fuzzy entropy itself can be used to find fuzzy-rough reducts [139]. A subset $P \subseteq \mathbb{C}$ induces a fuzzy similarity relation (R_P) with corresponding foresets $F_1, F_2, ..., F_n$. Similarly, the foresets induced by the (fuzzy) decision feature $\mathbb{D}$ are $D_1, D_2, ..., D_n$. The fuzzy entropy for a foreset F_i can be defined as:

$$H(F_i) = \sum_{D_j \in R_\mathbb{D}} \frac{-p(D_j|F_i) \log_2 p(D_j|F_i)}{|D_j|} \tag{4.31}$$

where $p(D_j|F_i)$ is the relative frequency of foreset F_i with respect to the decision D_j, and is defined:

$$p(D_j|F_i) = \frac{|D_j \cap F_i|}{|F_i|} \tag{4.32}$$

Based on these definitions, the fuzzy entropy for an attribute subset P can be defined as follows:

$$E(P) = \sum_{F_i \in R_P} \frac{|F_i|}{\sum_{Y_i \in R_P} |Y_i|} H(F_i) \tag{4.33}$$

This fuzzy entropy is monotonic and can be used to gauge the utility of fea-

ture subsets in a similar way to that of the fuzzy-rough measure. By dividing the entropy by $\log_2(\sum_{D_n \in R_{\mathbb{D}}}(|D_n|)^{-1})$, the measure will be normalized. This can be integrated into a QUICKREDUCT-style algorithm, employing a greedy hill-climbing approach. Again, as the measure monotonically decreases with addition of features, the search algorithm seeks to minimize this value in a manner similar to the boundary region minimization approach.

Theorem 2. *E-FRFS reducts are fuzzy-rough reducts. Suppose that* $P \subseteq \mathbb{C}$. *If* $E(P) = 0$ *then* P *is a fuzzy-rough reduct.*

Proof. Equation (4.20) can be rewritten as [96]:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{D_j} \sup_{F_i} \min(\inf_{y \in \mathbb{U}} I(\mu_{F_i}(y), \mu_{D_j}(y)))$$

If P is a fuzzy-rough reduct, then it must be the case that $F_i \subseteq D_j$ or $F_i \cap D_j = \emptyset \forall F_i, D_j$. If $F_i \subseteq D_j$, then $p(D_j|F_i) = 1$, and if $F_i \cap D_j = \emptyset$, then $p(D_j|F_i) = 0$ $\forall F_i, D_j$. Therefore each $H(F_i) = 0$, and $E(P) = 0$. □

4.4.4.1 Example

Returning to the example dataset in Table 4.16, the fuzzy entropy measure is used to determine fuzzy-rough reducts. The algorithm begins with an empty subset, and considers the addition of individual features. The attribute that results in the greatest decrease in fuzzy entropy will ultimately be added to the reduct candidate. For attribute a, the fuzzy entropy is calculated as follows ($A = \{a\}$):

$$E(A) = \sum_{F_i \in R_A} \frac{|F_i|}{\sum_{Y_i \in R_A} |Y_i|} H(F_i)$$

Each foreset F_i corresponds to one row in the matrix R_A:

F_1	1.0	1.0	0.699	0.0	0.0	0.0
F_2	1.0	1.0	0.699	0.0	0.0	0.0
F_3	0.699	0.699	1.0	0.0	0.0	0.0
F_4	0.0	0.0	0.0	1.0	0.699	0.699
F_5	0.0	0.0	0.0	0.699	1.0	1.0
F_6	0.0	0.0	0.0	0.699	1.0	1.0

Considering F_1, H(F_1) must be calculated:

$$H(F_1) = \sum_{D_j \in R_{\mathbb{D}}} \frac{-p(D_j|F_1) \log_2 p(D_j|F_1)}{|D_j|}$$

Each foreset D_j corresponds to one row in the matrix $R_{\mathbb{D}}$:

D_1	1.0	0.0	1.0	0.0	0.0	1.0
D_2	0.0	1.0	0.0	1.0	1.0	0.0
D_3	1.0	0.0	1.0	0.0	0.0	1.0
D_4	0.0	1.0	0.0	1.0	1.0	0.0
D_5	0.0	1.0	0.0	1.0	1.0	0.0
D_6	1.0	0.0	1.0	0.0	0.0	1.0

For D_1:

$$H(D_1) = \frac{-p(D_1|F_1)\log_2 p(D_1|F_1)}{|D_1|}$$

$$= \frac{-(1.699/2.699)\log_2(1.699/2.699)}{3.0}$$

Calculating this for each D_j produces:

$$H(F_1) = 0.140 + 0.177 + 0.140 + 0.177 + 0.177 + 0.140 = 0.951$$

The procedure is repeated for each remaining foreset:

$$H(F_2) = 0.951, H(F_3) = 0.871, H(F_4) = 0.871,$$
$$H(F_5) = 0.951, H(F_6) = 0.951$$

Hence, the fuzzy entropy is:

$$\begin{aligned} E(A) &= \sum_{F_i \in R_A} \frac{|F_i|}{\sum_{Y_i \in R_A} |Y_i|} H(F_i) \\ &= 0.926 = E(\{a\}) \end{aligned}$$

Repeating this process for the remaining attributes gives:

$E(\{b\}) = 0.921$

$E(\{c\}) = 0.738$

From this it can be seen that attribute c will cause the greatest decrease in fuzzy entropy. This attribute is chosen and added to the potential reduct, $R \leftarrow R \cup \{c\}$. The process iterates and the two fuzzy entropy values calculated are

$E(\{a, c\}) = 0.669$

$E(\{b, c\}) = 0.0$

Adding attribute b to the reduct candidate results in the minimum entropy for the data, and the search terminates, outputting the subset $\{b, c\}$. The dataset can now be reduced to only those attributes appearing in the reduct.

4.4.5 Fuzzy-Rough Reduction with Fuzzy Gain Ratio

The Information Gain (IG) [184] is the expected reduction in entropy resulting from partitioning the dataset objects according to a particular feature. For the fuzzy case this can be expressed as:

$$IG(P \cup \{a\}) = E(P) - E(P \cup \{a\}) \tag{4.34}$$

One limitation of the IG measure is that it favours features with many values. The Gain Ratio (GR) seeks to avoid this bias by incorporating another term, split information, that is sensitive to how broadly and uniformly the attribute splits the considered data. Again, for the fuzzy case this can be expressed as:

$$SP(Q) = \sum_{F_i \in R_Q} \frac{|F_i|}{\sum_{Y_i \in R_Q} |Y_i|} \log_2 \frac{|F_i|}{\sum_{Y_i \in R_Q} |Y_i|} \tag{4.35}$$

The Gain Ratio is then defined as follows:

$$GR(P \cup \{a\}) = \frac{IG(P \cup \{a\})}{SP(P \cup \{a\})} \tag{4.36}$$

When this is minimized, $P \cup \{a\}$ is a fuzzy-rough reduct due to the monotonicity of the fuzzy entropy measure. This metric is applied in the same manner as described previously for the feature selection approach.

4.5 Experimentation

This section presents the initial experimental evaluation of the selection methods for the task of pattern classification, over nine benchmark datasets obtained from [158] with two classifier learners.

4.5.1 Experimental Setup

For the fuzzy-rough methods, the Łukasiewicz fuzzy connectives are used, with fuzzy similarity defined in (4.19). After feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then classified using the relevant classifier learning method.

Two learning mechanisms were employed to create classifiers for the purpose of evaluating the resulting subsets from the feature selection phase: JRip [33] and PART [236, 237].

Dataset	Objects	Features	Reduct size				
			E	B	L	BE	GR
Cleveland	297	14	10	9	9	10	10
Glass	214	10	9	9	10	10	9
Heart	270	14	9	8	8	8	9
Ionosphere	230	35	8	9	9	10	8
Olitos	120	26	6	6	6	6	6
Water 2	390	39	7	7	7	7	7
Water 3	390	39	7	7	7	7	7
Web	149	2557	23	20	21	20	18
Wine	178	14	6	6	6	6	6

Table 4.17: Reduct size and time taken

Dataset	JRip					
	Unred.	E	B	L	BE	GR
Cleveland	52.19	53.53	54.55	54.55	53.20	53.53
Glass	71.50	65.89	65.89	71.50	71.50	65.89
Heart	77.41	80.37	78.52	78.52	78.15	80.37
Ionosphere	86.52	84.37	88.26	88.26	89.15	84.37
Olitos	70.83	67.50	71.67	64.17	65.83	67.50
Water 2	83.85	82.30	85.64	85.64	84.36	83.59
Water 3	82.82	81.29	82.56	81.03	84.10	81.29
Web	58.39	53.02	46.97	55.03	50.37	52.34
Wine	92.70	94.94	95.50	95.50	93.82	91.57

Table 4.18: Resulting classification accuracies JRip (%)

Dataset	PART					
	Unred.	E	B	L	BE	GR
Cleveland	50.17	56.22	53.20	53.20	57.23	56.22
Glass	67.76	70.56	70.56	67.76	67.76	70.56
Heart	73.33	78.51	76.30	76.30	76.30	78.51
Ionosphere	88.26	86.95	86.09	86.09	88.26	86.95
Olitos	57.50	61.67	67.50	58.33	69.16	56.67
Water 2	83.08	83.59	84.62	84.62	84.10	82.31
Water 3	83.33	80.76	81.03	80.77	85.39	80.76
Web	42.95	55.70	55.03	57.72	52.34	53.69
Wine	93.82	94.94	94.38	94.38	94.94	93.82

Table 4.19: Resulting classification accuracies PART (%)

4.5.2 Experimental Results

Table 4.17 compares the reduct size for fuzzy entropy-based FS (E), fuzzy boundary region-based FS (B), fuzzy lower approximation-based FS (L), fuzzy boundary/entropy FS (BE) and fuzzy gain ratio FS (GR). It can be seen that the new entropy-based fuzzy-rough methods find smaller subsets in general (B, BE, GR). The fuzzy boundary region-based method finds smaller or equally-sized subsets than the L. This is to be expected, as B includes fuzzy upper approximation information in addition to that of the fuzzy lower approximation. The entropy-based methods perform similarly, with the fuzzy gain ratio measure finding the smallest subsets in general. This demonstrates the utility of considering the split information when evaluating subset quality.

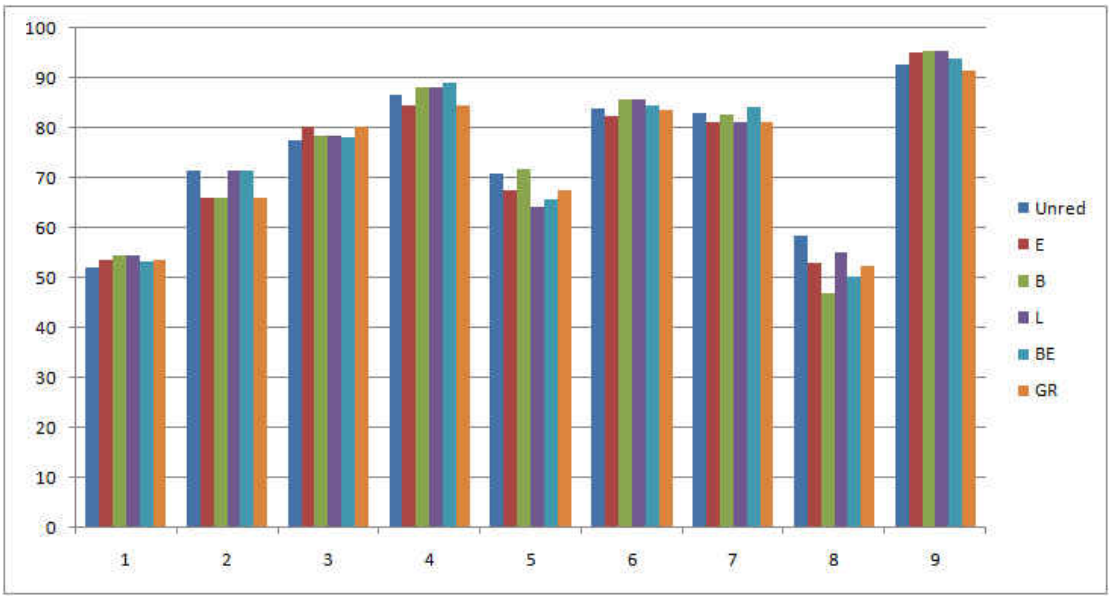

Figure 4.3: Performance: JRip

Table 4.18 shows the average classification accuracy as a percentage obtained using 10-fold cross validation. The classification accuracies are also presented in Figs. 4.3 and 4.4 for each of the nine datasets. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained using the feature selection techniques. All techniques perform similarly, with both the boundary (B) and lower approximation (L) FS approaches showing the most consistent results for both classifier learners. It would appear that the GR approach also generally selects compact subsets at the expense of classification accuracy. The BE approach demonstrates that there is some useful information to be extracted from the fuzzy-rough boundary region for the PART classifier learner. However as this approach only examines the boundary region information, the certain information of the lower approximation is ignored and the results reflect

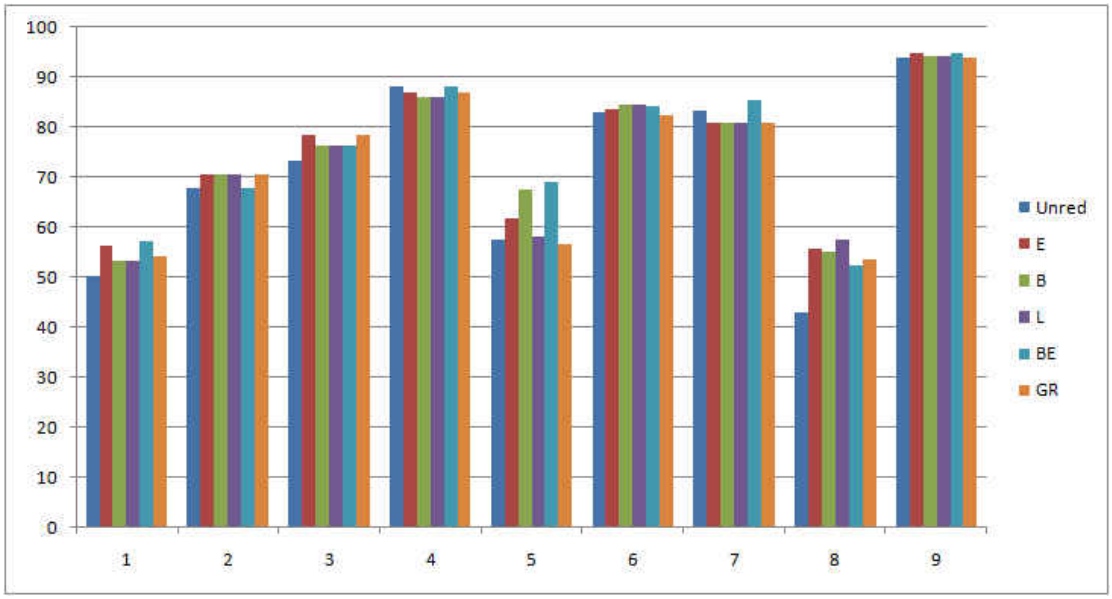

Figure 4.4: Performance: PART

this. The boundary entropy measure attempts to analyse the level of order in the boundary region and use this information to guide the FS task. The level of order in the boundary however may not be always be indicative of the best quality of subsets and hence performs better for some datasets over others - see Fig. 4.3.

4.6 Summary

This chapter has presented a new technique for tolerance rough set-based feature selection and three new techniques for fuzzy-rough feature selection based on the use of fuzzy entropy as an evaluation metric for the fuzzy-rough boundary region (which are also applicable to the fuzzy-rough lower approximation).

Overall, the experimental results presented in this chapter have shown the utility of employing the information contained in the boundary regions of tolerance rough sets and fuzzy-rough sets to guide FS methods.

The work has demonstrated the potential benefits of using the boundary region information in the search for reducts. Moreover, this work has shown that the approach originally proposed in Chapter 3, is capable of being extended to methods which have the ability to handle real-valued data. In all experimental studies, no attempt has been made to discover the optimal value for the parameter τ, as this would involve exhaustive experimentation for each dataset. In the case of fuzzy-rough sets only the fuzzy connectives detailed here were employed for minimisation, as again this would involve considerable experimentation. It is expected that results obtained with such optimisation would show a marked improvement over those that are observed here.

Chapter 5

Association Learning

The nearest neighbour (NN) type of classifier approach is a popular method of learning classifiers. This relates to its relative simplicity, and the fact that concept of 'nearest neighbours' appeals to human intuition. The original kNN approach [53] has seen many developments over the years and a fuzzy extension (FNN) [103], and indeed a fuzzy-rough extension [194]. Despite the fact that the extension in [194] is based on fuzzy-rough sets, no use is made of the fuzzy upper and lower approximation concepts in classifying test objects. Given the great success in using these concepts for feature selection [98], a new NN classifier was developed with this in mind. This chapter describes the development of a fuzzy-rough nearest neighbour (FRNN) classifier learner.

In addition to the work on association learning using the fuzzy-rough lower approximation, another idea regarding the use of fuzzy dependency and its applicability to unsupervised FS is also proposed in this chapter. Since fuzzy dependency has been used to find dependencies between conditional and decisional attributes, it could also be applied to discover dependencies between sets of conditional features. This strategy can be used to learn about the interdependencies of features and eliminate those features which are subsumed by larger subsets, and are therefore redundant.

5.1 Classifier Learning

A classifier learning technique is a systematic approach for building classification models from data. All classifier learning techniques employ a learning algorithm to identify a model or models which describe the relationship between the input data and the class labels as accurately as possible, also termed the level of 'fit'. The model generated by a learning algorithm should fit the input data well and also be able to correctly predict the class labels of objects that it has not previously 'seen'.

The ability of the models which have been learned to generalise well is therefore of primary importance. The performance of a classifier learner can be assessed by classification accuracy, that is the ratio of correctly predicted test objects to the total number of test objects. Classification error may also be used, i.e the measure of the percentage of incorrectly classified objects. Other aspects (when used in conjunction with classifier error/accuracy) such as standard deviation can also be useful indicators for determining the performance of classification algorithms.

5.1.1 Nearest Neighbours Classification

The k-nearest neighbours (kNN) algorithm [53] is a well-known classification technique that assigns a test object to the decision class most common among its k nearest neighbours. In nearest neighbour (NN) classification, the training set is a set of objects in a multidimensional feature space. The feature space is partitioned into regions by locations and labels of the training objects. A point in the space is assigned to the class C if it is the most common class amongst k-nearest training objects. This is demonstrated in the diagram in Fig. 5.1; when $k = 4$ the test object is classified to the class of those red objects, however when $k = 8$ the test object will be classified to the class of blue objects. Euclidean distance is typically used as the distance metric for numeric values. For application to areas such as text classification, metrics like Hamming distance are employed as these can be applied to non-numeric data.

For the algorithm training phase, the feature vectors and class labels of the training objects are used to build a model. Then, for the classification phase, the test object (whose class is unknown) is represented as an object in the feature space. Distances from the new object to all training objects are calculated and the k closest objects are chosen. There are a number of ways to classify the test object to a particular class, one of the most common techniques is to predict the test object with reference to the most common class amongst the k nearest neighbours. A major drawback when using this technique to classify test objects to a particular class is that classes with the most frequent objects tend to dominate the prediction of the test object. This is because they tend to occur in the k nearest neighbours when neighbours are calculated due to their larger number. In order to overcome this, the distance of each of the k nearest neighbours must be considered with respect to the distance from the test object and predict the class of the new vector based on these distances. Note that the special case where a test object class is predicted by the class of the closest single training object ($k = 1$) is known as the nearest neighbour algorithm.

The ideal value for k depends upon the data. In general larger values tend

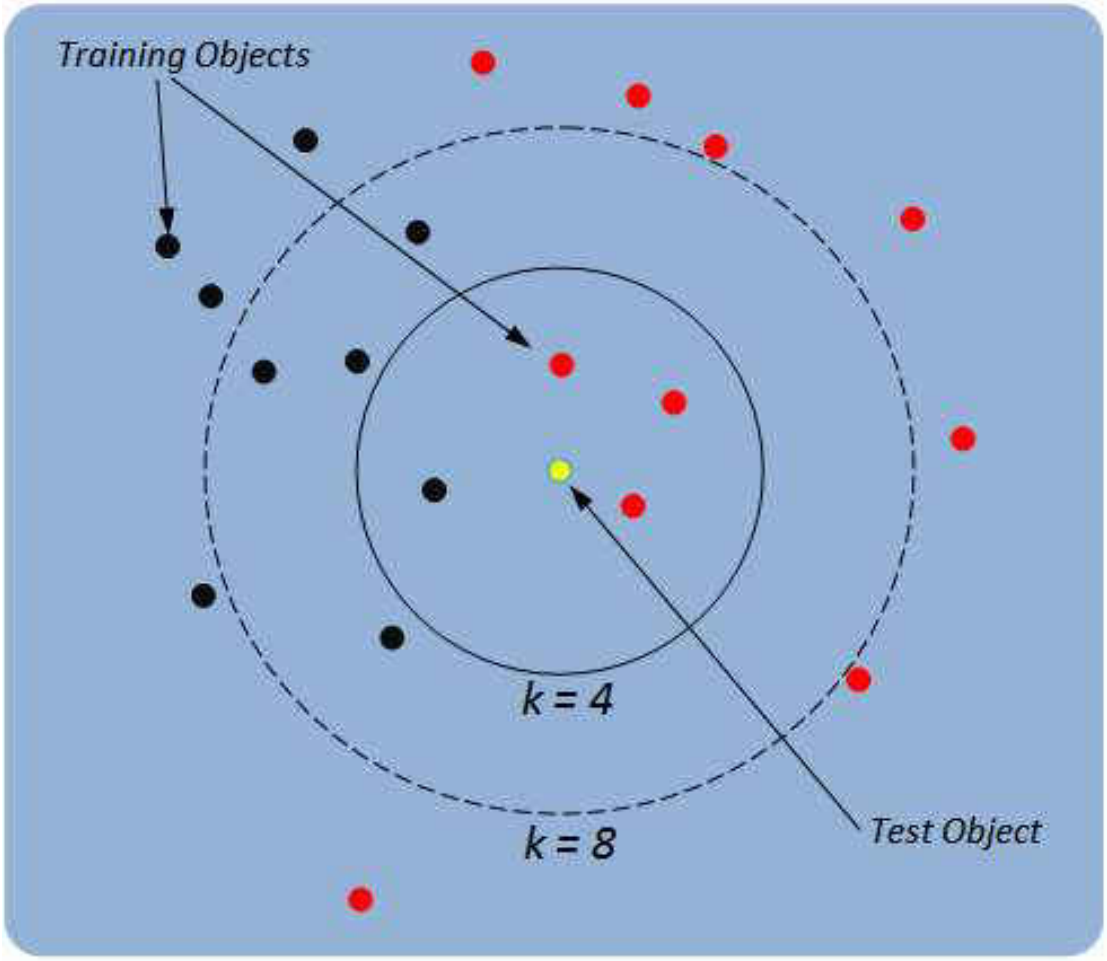

Figure 5.1: Nearest neighbours (NN) classification

to reduce the effects of noise on the classification, but 'blur' the class boundaries for the testing phase. Heuristic techniques may be used as an aid to selecting an optimal value for k, e.g. cross-validation.

The accuracy of the kNN algorithm will degrade severely in the presence of noisy or irrelevant features, or if the feature scaling is not consistent with importance.

5.1.2 Fuzzy Nearest Neighbours Classification

An extension of the kNN algorithm to fuzzy set theory (FNN) was introduced in [103]. It allows partial membership of an object to different classes, and also takes into account the relative importance (closeness) of each neighbour with respect to the test instance. However, as correctly argued in [194], the FNN algorithm has problems dealing adequately with insufficient knowledge. In particular, when every training pattern is far removed from the test object, and hence there are no suitable neighbours, the algorithm is still forced to make clear-cut predictions. This is

because the sum of the predicted membership degrees to the various decision classes is always required to be equal to 1.

The fuzzy k-nearest neighbours algorithm [103] aims to classify test objects based on their similarity to a given number of neighbours and their neighbours' degree of belonging to (crisp or fuzzy) class labels. For the purposes of FNN, the extent to which an unclassified object y belongs to class X is defined as:

$$\mu_X(y) = \sum_{x \in N} \mu_R(x, y)\mu_X(x) \tag{5.1}$$

where N is the set of object y's k-nearest neighbours and $\mu_R(x, y)$ is the fuzzy similarity of y and object x. In the traditional fuzzy kNN approach, this is defined in the following way:

$$\mu_R(x, y) = \frac{||y - x||^{-2/(m-1)}}{\sum_{j \in N} ||y - j||^{-2/(m-1)}} \tag{5.2}$$

where $|| \cdot ||$ denotes the Euclidean norm, and m is a parameter that controls the overall weighting of this fuzzy similarity. The FNN algorithm (Fig. 5.2) employs these definitions to determine the extent to which an object y belongs to each class, typically classifying y to the class with the highest resulting membership. The complexity of this algorithm for the classification of one test pattern is $O(|\mathbb{U}| + k \cdot |C|)$,

FNN($\mathbb{U}$,C,y,k).
$\mathbb{U}$, the training data;
C, the set of decision classes;
y, the object to be classified;
k, the number of nearest neighbours.

(1) $N \leftarrow$ getNearestNeighbours(y,K);
(2) $\forall X \in C$
(3) $\quad \mu_X(y) = \sum_{x \in N} \mu_R(x, y)\mu_X(x)$
(4) **output** $\arg\max_{X \in C} (\mu_X(y))$

Figure 5.2: The fuzzy kNN algorithm

5.2 Fuzzy-Rough Set Theory

In many real-world applications, data is often both crisp and *real-valued*, and this is where traditional rough set theory encounters a problem. It is not possible in the original theory to say whether two attribute values are similar and to what extent

they are the same; for example, two close values may only differ as a result of noise, but RST considers them as different as two values of a dissimilar magnitude. It is, therefore desirable to develop techniques which provide a method for knowledge modelling of crisp and real-value attribute datasets which utilise the extent to which values are similar. This can be achieved through the use of *fuzzy-rough* sets. Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge

Comprehensive coverage of fuzzy-rough sets and definitions for the fuzzy lower and upper approximations concepts can be found in Section 4.4.1, but they are provided here once again for completeness.

A T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X the lower and upper approximations are:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x,y), \mu_X(y)) \tag{5.3}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) \tag{5.4}$$

Here, I is a fuzzy implicator and T a t-norm. R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu_{R_P}(x,y) = T_{a \in P}\{\mu_{R_a}(x,y)\} \tag{5.5}$$

$\mu_{R_a}(x,y)$ is the degree to which objects x and y are similar for feature a, and may be defined in many ways - three common examples are demonstrated in Section 4.4.1.

In a similar way to the original crisp rough set approach, the fuzzy positive region [99] can be defined as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \tag{5.6}$$

An important issue in data analysis is the discovery of dependencies between attributes. The fuzzy-rough dependency degree of $\mathbb{D}$ on the attribute subset P can be defined as:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \tag{5.7}$$

A fuzzy-rough reduct R can be defined as a minimal subset of features which preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough greedy hill-climbing algorithm can be constructed that uses equation (5.7) to gauge subset quality. In [99], it has been shown that the

dependency function is monotonic and that fuzzy discernibility matrices may also be used to discover reducts.

5.2.1 Fuzzy-Rough Ownership kNN

Initial attempts to combine the FNN algorithm with concepts from fuzzy rough set theory were presented in [194, 230] (this approach is denoted FRNN-O here). In these papers, a fuzzy-rough ownership function is constructed that attempts to handle both "fuzzy uncertainty" (caused by overlapping classes) and "rough uncertainty" (caused by insufficient knowledge, i.e. attributes, about the objects). All training objects influence the ownership function, and hence no decision is required as to the number of neighbours to consider, although there are other parameters that must be defined for its successful operation. It should be noted that the algorithm does not use fuzzy lower or upper approximations to determine class membership. The fuzzy-rough ownership function was defined as:

$$\tau_X(y) = \frac{\sum_{x \in \mathbb{U}} \mu_R(x,y)\mu_X(x)}{|\mathbb{U}|} \tag{5.8}$$

This can be modified to consider only the k nearest neighbours as follows:

$$\tau_X(y) = \frac{\sum_{x \in N} \mu_R(x,y)\mu_X(x)}{|N|} \tag{5.9}$$

where N is the set of object y's k-nearest neighbours. When $k = |\mathbb{U}|$ then the original definition is obtained. The fuzzy similarity is determined by:

$$\mu_R(x,y) = \exp(-\sum_{a \in \mathbb{C}} \kappa_a (a(y) - a(x))^{2/(m-1)}) \tag{5.10}$$

where m controls the weighting of the similarity (as in FNN) and κ_a is a parameter that decides the bandwidth of the membership, defined as

$$\kappa_a = \frac{|\mathbb{U}|}{2\sum_{x \in \mathbb{U}} ||a(y) - a(x)||^{2/(m-1)}} \tag{5.11}$$

The algorithm can be seen in Fig. 5.3. Initially, the parameter κ_a is calculated for each attribute and all memberships of decision classes for test object y are set to zero. Next, the weighted distance of y from all objects in the universe is computed and used to update the class memberships of y via equation (5.8). Finally, when all training objects have been considered, the algorithm outputs the class with highest membership. The complexity of the algorithm is $O(|\mathbb{C}||\mathbb{U}|+|\mathbb{U}|\cdot(|\mathbb{C}|+|C|))$. To obtain the k-nearest neighbours version of this algorithm, line (3) should be replaced with $N \leftarrow$ getNearestNeighbours(y,k). The method still requires a choice

of parameter m, which plays a similar role to that in FNN.

FRNN-O($\mathbb{U}$,$\mathbb{C}$,C,y).
$\mathbb{U}$, the training data;
$\mathbb{C}$, the set of conditional features;
C, the set of decision classes;
y, the object to be classified.

(1) $\forall a \in \mathbb{C}$
(2) $\quad \kappa_a = |\mathbb{U}|/2 \sum_{x \in \mathbb{U}} ||a(y) - a(x)||^{2/(m-1)}$
(3) $N \leftarrow |\mathbb{U}|$
(4) $\forall X \in C$, $\tau_X(y) = 0$
(5) $\forall x \in N$
(6) $\quad d = \sum_{a \in \mathbb{C}} \kappa_a (a(y) - a(x))^2$
(7) $\quad \forall X \in C$
(8) $\quad\quad \tau_X(y) += \frac{\mu_X(x) \cdot \exp(-d^{1/(m-1)})}{|N|}$
(9) **output** $\arg\max_{X \in C} \tau_X(y)$

Figure 5.3: The fuzzy-rough ownership NN algorithm

5.2.2 Fuzzy-Rough Nearest Neighbours

At the heart of the FRNN approach described here, is the previously mentioned ability of fuzzy-rough sets to handle real-valued and noisy data. This ability has been exploited with great success for feature selection [98], and given that the classification task is very closely related, the motivation for such an application was clear. Initial work on this has been carried out in [92] and described in further detail here.

To perform classification, the algorithm shown in Fig. 5.4 is used. The rationale behind the algorithm is that the lower and the upper approximation of a decision class, calculated by means of the nearest neighbours of a test object y, provide good clues to predict the membership of the test object to that class.

The membership of a test object y to each (crisp or fuzzy) decision class is determined via the calculation of the fuzzy lower and upper approximation. The algorithm outputs the decision class with the resulting best fuzzy lower and upper approximation memberships. The complexity of the algorithm is $O(|C| \cdot (2|\mathbb{U}|))$. Although k is not required, it can be incorporated into the algorithm by replacing line (1) with "$N \leftarrow$ getNearestNeighbours(y,k)". As $\mu_{R_P}(x, y)$ becomes smaller, x tends to only have a minor influence on $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$.

FRNN($\mathbb{U}$,C,y).
$\mathbb{U}$, the training data;
C, the set of decision classes;
y, the object to be classified.

(1) $N \leftarrow \mathbb{U}$
(2) $\mu_1(y) \leftarrow 0$, $\mu_2(y) \leftarrow 0$, $Class \leftarrow \emptyset$
(3) $\forall X \in C$
(4) $\quad \mu_{\underline{R_P}X}(y) = \inf_{z \in N} I(\mu_{R_P}(y,z), \mu_X(z))$
(5) $\quad \mu_{\overline{R_P}X}(y) = \sup_{z \in N} T(\mu_{R_P}(y,z), \mu_X(z))$
(6) $\quad$ **if** $(\mu_{\underline{R_P}X}(y) \geq \mu_1(y)$ && $\mu_{\overline{R_P}X}(y) \geq \mu_2(y))$
(7) $\quad\quad Class \leftarrow X$
(8) $\quad\quad \mu_1(y) \leftarrow \mu_{\underline{R_P}X}(y)$, $\mu_2(y) \leftarrow \mu_{\overline{R_P}X}(y)$
(9) **output** *Class*

Figure 5.4: The FRNN algorithm

5.2.3 Worked Example

In order to demonstrate the application of the algorithm, a small worked example is presented. This example employs a dataset with 3 real-valued conditional attributes and a single crisp discrete-valued decision attribute as the *training* data, shown in Table 5.1. A further dataset illustrated in Table 5.2 containing 2 objects is used as the *test* data to be classified.

Object	a	b	c	q
1	−0.4	−0.3	−0.5	yes
2	−0.4	0.2	−0.1	no
3	0.2	−0.3	0	no
4	0.2	0	0	yes

Table 5.1: Example training data

Object	a	b	c	q
t1	0.3	−0.3	0	no
t2	−0.3	−0.4	−0.3	yes

Table 5.2: Example test data

Referring to the FRNN algorithm described in the previous section, the first step is to calculate the fuzzy upper and lower approximations for all decision classes. In Table 5.1 there are 4 objects and as noted previously a decision attribute which has 2 classes;*yes*, and *no*.

Using the fuzzy similarity measure as defined in equation 4.1 the similarity of each test object is compared to all of the objects in the training data. For instance, consider the training object $t1$:

$$\mu_{R\{P\}}(t1,1) = T(\mu_{R\{a\}}(t1,1), \mu_{R\{b\}}(t1,1), \mu_{R\{c\}}(t1,1)) = 0$$
$$\mu_{R\{P\}}(t1,2) = T(\mu_{R\{a\}}(t1,2), \mu_{R\{b\}}(t1,2), \mu_{R\{c\}}(t1,2)) = 0.16$$
$$\mu_{R\{P\}}(t1,3) = T(\mu_{R\{a\}}(t1,3), \mu_{R\{b\}}(t1,3), \mu_{R\{c\}}(t1,3)) = 0.83$$
$$\mu_{R\{P\}}(t1,4) = T(\mu_{R\{a\}}(t1,4), \mu_{R\{b\}}(t1,4), \mu_{R\{c\}}(t1,4)) = 0.40$$

These similarity values can then be used to generate the lower and upper approximations. Note that the fuzzy connectives chosen for this example are the Łukasiewicz t-norm ($\max(x+y-1,0)$), and Łukasiewicz fuzzy implicator ($\min(1-x+y),1$).
For the decision concept $X = yes$ these are:

$$\mu_{\underline{R_P}}\, X(t1) = \inf_{y \in \mathbb{U}} \{I(\mu_{R_P}(t1,y), \mu_X(y)\}$$

$$= \inf\{I(0,1), I(0.16,0), I(0.83,0), I(0.4,1)\} = 0.14$$

and,

$$\mu_{\overline{R_P}}\, X(t1) = \sup_{y \in \mathbb{U}} \{I(\mu_{R_P}(t1,y), \mu_X(y)\}$$

$$= \sup\{T(0,1), T(0.16,0), T(0.83,0), T(0.4,1)\} = 0.84$$

Similarly for the decision concept $X = no$:

$$\mu_{\underline{R_P}}\, X(t1) = \inf\{I(0,0), I(0.16,1), I(0.83,1), I(0.4,0)\} = 0.16$$

$$\mu_{\overline{R_P}}\, X(t1) = \sup\{T(0,0), T(0.16,1), T(0.83,1), T(0.4,0)\} = 0.86$$

With reference once again to the FRNN algorithm in Fig. 5.4, it can be seen that the upper and lower approximation membership values for test object $t1$ for the class label $X = no$ are higher than those for when $X = yes$. The algorithm will therefore classify $t1$ as belonging to the class $X = no$. The procedure can be repeated for training object $t2$ which results in upper and lower approximation values for $X = no$:

$$\mu_{\underline{R_P}}\, X(t2) = \inf\{I(0.6,1), I(0.6,0), I(0.17,0), I(0.17,1)\} = 0.4$$

$$\mu_{\overline{R_P}} X(t2) = \sup\{T(0.6,1), T(0.6,0), T(0.17,0), T(0.17,1)\} = 0.6$$

And, $X = yes$:

$$\mu_{\underline{R_P}} X(t2) = \inf\{I(0.6,0), I(0.6,1), I(0.17,1), I(0.17,0)\} = 0.4$$

$$\mu_{\overline{R_P}} X(t2) = \sup\{T(0.6,0), T(0.6,1), T(0.17,1), T(0.17,0)\} = 0.6$$

In this case, both upper and lower approximation membership values for each of the classes $X = no$ and $X = yes$ are identical. However because of line 6 of the FRNN algorithm, $t2$ will be classified as belonging to $X = yes$.

5.3 Unsupervised Feature Selection

Conventional supervised FS methods evaluate various feature subsets using an evaluation function or metric to select only those features which are related to, or lead to, the decision classes of the data under consideration. However, for many data mining applications, decision class labels are often unknown or incomplete, thus indicating the significance of unsupervised feature selection. In a broad sense, two different types of approach to unsupervised FS have been adopted: Those which maximise clustering performance using an index function [42], [168], and those which consider features for selection on the basis of dependency or relevance. The central idea behind the latter, is that any single feature which carries little or no further information than that subsumed by the remaining features is redundant and can therefore be eliminated [37], [68], [153]. The approach described in this work is related to these techniques since it involves the removal of features which are considered to be redundant.

Fuzzy-rough sets are used as a basis for the technique described below. It employs the fuzzy-rough discernibility measure to examine the level of discernibility between a single feature and subsets of other features. Where a single feature can be discerned completely by a subset of features, that single feature is considered to be redundant and can be removed from the feature set. FS is conducted through the removal of features until no further inter-dependency can be found. The resulting subset of original features can then be used to define the original complete feature set.

5.3.1 Unsupervised Fuzzy-Rough Feature Selection

In the previous chapter, it was demonstrated how FRS can be applied to the problem of supervised feature selection. One of the most important aspects relating to feature set reduction is the fuzzy-rough dependency measure (see (5.7)), and it is this measure which is also employed for the new unsupervised fuzzy-rough FS (UFRFS) method described in this section. A worked example is also provided here to illustrate the approach.

5.3.2 Fuzzy Dependency

The discovery of dependencies between attributes, is in general, an important issue in data analysis. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P.

The central idea behind the present work is that, as with supervised fuzzy-rough FS [99], the fuzzy dependency measure can also be used to discover the inter-dependency of features. This can be achieved by simply substituting the decision feature(s) $\mathbb{D}$ in (5.7) of the supervised approach for any given feature or group of features Q such that

$$\gamma'_P(Q) = \frac{\sum\limits_{x \in \mathbb{U}} \mu_{POS_{R_P(Q)}}(x)}{|\mathbb{U}|} \tag{5.12}$$

where $P \cap Q = \emptyset$ and,

$$\mu_{POS_{R_P(Q)}}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{R_P}X}(x) \tag{5.13}$$

Note that the dependency measure is not symmetric, i.e. $(P \rightarrow Q) \neq (Q \rightarrow P)$. This is why a backwards elimination style search has been implemented for the selection of features for the proposed method as oppose to a forward greedy method that is commonly adopted in supervised FS.

Although the above proposal may be applied to evaluate the dependency between any two subsets of features in theory; practically, this may be computationally prohibitive. Fortunately, for unsupervised feature selection it is sufficient to find the dependency between a single feature and other subsets of features. If it has been established that one single feature depends fully on a feature subset then that feature can be removed. Hence, the proposed approach only requires the calculation of the dependency in the specific case of (5.12) where Q contains just a single feature. In light of this observation, equation (5.3) as used in (5.13) becomes:

$$\mu_{\underline{R_P}R_u z}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_{R_u z}(y)) \tag{5.14}$$

where z denotes the single feature under examination and $R_u z$ indicates the tolerance class (or fuzzy equivalence class) for object u of z.

5.3.3 Algorithm

Algorithm 5.5 shows the new unsupervised FS approach. The algorithm starts by considering all of the features contained in the dataset. Each feature is then examined iteratively, and the fuzzy dependency measure is calculated. If the fuzzy dependency (γ') is equal to 1 then that feature can be removed. This process continues until all features have been examined.

UFRQuickReduct($\mathbb{C}$,$\mathbb{D}$)
$\mathbb{F}$, the set of all features;
$\mathbb{R}$, a feature subset.

(1) $R \leftarrow F$
(2) $T \leftarrow F$
(3) **for** $x \in T$
(4) **do** $T \leftarrow T - \{x\}$
(5) **if** $\gamma'_{\{R-\{x\}\}}(\{x\}) = 1$
(6) **then** $R \leftarrow R - \{x\}$
(7) **until** $T = \emptyset$
(8) **return** R

Figure 5.5: The UFRQuickReduct algorithm

If no interdependency exists, the algorithm will return the full set of features. The complexity for the search in the worst case is $O(n)$, where n is the number of original features. This is because the fuzzy dependency calculation is made for every feature with respect to the subset of survival features.

5.3.4 Worked Example

To illustrate the ideas discussed, a small dataset shown in Table 5.3 is employed. As recommended in [45], the Łukasiewicz t-norm (max(x + y - 1, 0)) and the Łukasiewicz fuzzy implicator (min(1 - x + y, 1)) are adopted to implement the fuzzy connectives. Other interpretations may also be used.

Using the fuzzy similarity measure defined in equation (4.19), the resulting relations for each feature in the dataset are shown (for brevity) in Table 5.4.

Object	a	b	c	d
1	0.0	0.1	0.1	0.5
2	0.2	0.1	0.6	0.9
3	0.6	0.7	0.3	0.9
4	0.3	0.4	0.8	0.6
5	0.2	0.7	0.9	0.2

Table 5.3: Example dataset

$R_a(x,y)$					$R_b(x,y)$					$R_c(x,y)$					$R_d(x,y)$				
1.0	**0.83**	0.0	0.50	0.83	1.0	1.0	0.0	0.50	0.0	1.0	0.375	0.75	0.125	0.0	1.0	0.429	0.429	0.857	0.572
0.83	**1.0**	0.33	0.83	1.0	**1.0**	**1.0**	**0.0**	**0.50**	**0.0**	0.375	1.0	0.625	0.75	0.625	0.429	1.0	1.0	0.572	0.0
0.0	**0.33**	1.0	0.50	0.33	0.0	0.0	1.0	0.50	1.0	0.75	0.625	1.0	0.375	0.25	0.429	1.0	1.0	0.572	0.0
0.50	**0.83**	0.50	1.0	0.83	0.50	0.50	0.50	1.0	0.50	0.125	0.75	0.375	1.0	0.375	0.857	0.572	0.572	1.0	0.429
0.83	**1.0**	0.33	0.83	1.0	0.0	0.0	1.0	0.50	1.0	0.0	0.625	0.25	0.375	1.0	0.572	0.0	0.0	0.429	1.0

Table 5.4: Fuzzy similarity relations

Initially, the lower approximations of the concepts of a given feature must be computed for each of the other features in the dataset. This is then used to calculate the dependency degree. For the example dataset, consider the dependency of the feature b on the feature a:

$$\mu_{\underline{R_{\{a\}}}R_u b}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_{\{a\}}}(x,y), \mu_{R_u b}(y)) \tag{5.15}$$

Thus, for a particular instance where object $x = 2$, and $u = 2$, this is (as highlighted in Table 5.4):

$$\mu_{\underline{R_{\{a\}}}R_2 b}(2) = \inf_{y \in \mathbb{U}} I(\mu_{R_a}(2,y), \mu_{R_2 b}(y)) =$$
$$\inf\{I(0.83,1), I(1,1)I(0.33,0)I(0.83,0.5)I(1,0)\} = 0$$

and for the remaining objects regarding a (i.e. $u \in \{1,3,4,5\}$) this is:
$\mu_{\underline{R_{\{a\}}}R_1 b}(2) = \inf\{I(1,1), I(0.83,1)I(0,0)I(0.5,0.5)I(0.83,0)\} = 0.0$
$\mu_{\underline{R_{\{a\}}}R_3 b}(2) = \inf\{I(0,1), I(0.33,1)I(1,0)I(0.5,0.5)I(0.33,0)\} = 0.0$
$\mu_{\underline{R_{\{a\}}}R_4 b}(2) = \inf\{I(0.5,1), I(0.83,1)I(0.5,0)I(1,0.5)I(0.83,0)\} = 0.17$
$\mu_{\underline{R_{\{a\}}}R_5 b}(2) = \inf\{I(0.83,1), I(1,1)I(0.33,0)I(0.83,0.5)I(1,0)\} = 0.0$

This process is repeated for every object regarding b in order to calculate the remaining lower approximations for each object. These can then be used to calculate the positive regions:
$\mu_{POS_{R_{\{a\}}(\{b\})}}(1) = 0.5$
$\mu_{POS_{R_{\{a\}}(\{b\})}}(2) = 0.5$

$\mu_{POS_{R_{\{a\}}(\{b\})}}(3) = 0.67$
$\mu_{POS_{R_{\{a\}}(\{b\})}}(4) = 0.67$
$\mu_{POS_{R_{\{a\}}(\{b\})}}(5) = 0.67$

Therefore the resulting dependency degree is:

$$\gamma'_{\{a\}}(\{b\}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}}(x)}{|\mathbb{U}|} = \frac{3.01}{6} = 0.602$$

In the interests of brevity only the computation of the dependency of feature b upon feature a is illustrated here. However, in the actual implementation of the UFRFS algorithm, the first step is to consider the dependency of $\{a\}$ on the subset $\{b, c, d\}$. For the example dataset this leads to the following result:

$\gamma'_{\{b,c,d\}}(\{a\}) = 1.0$ $\quad (T = \{b, c, d\})$
$\gamma'_{\{c,d\}}(\{b\}) = 0.9569$ $\quad (T = \{c, d\})$
$\gamma'_{\{b,d\}}(\{c\}) = 1.0$ $\quad (T = \{d\})$
$\gamma'_{\{b\}}(\{d\}) = 0.2$ $\quad (T = \emptyset)$

Note that each time $\gamma' = 1$, the feature in question is eliminated resulting in the final subset $\{b, d\}$, after all features have been examined.

5.4 Experimentation

This section presents the experimental evaluation of the unsupervised FS method to support the task of pattern classification, over eight real-valued benchmark datasets (obtained from [158]) with three classifier learners. The approach is also compared with an advanced supervised approach [99] which shares the common mathematical foundations as the present work. Results are presented in terms of selected subset size and classification accuracy.

5.4.1 Experimental Setup

The FS method employs Łukasiewicz fuzzy connectives, with fuzzy similarity defined in (4.19). Following feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then classified using the relevant classifier learning method (as described below) and evaluated with 10-fold cross validation.

Three learning mechanisms were employed to create classifiers for the purpose of evaluating the resulting subsets from the feature selection phase: JRip [33], PART [236] [237], and J48 [184]. For a more detailed description of these classifier learners, see Section 3.5.1.

All of the data which is used in this experimental investigation is labelled. However, before applying UFRFS the decision feature is removed from the data, and the approach operates on the unlabelled data only. When learning classifiers, or applying supervised FRFS the complete dataset including the decision feature is used.

5.4.2 Experimental Results

The results presented in Table 5.5, show the subset sizes discovered by UFRFS and the state of the art supervised FRFS method of [99]. It can be seen that the proposed method manages reduction in all cases and returns substantial levels of dimensionality reduction for some datasets (e.g. *water2*, *water3*, *olitos*). These results compare well with the supervised approach and show that UFRFS may even find smaller subsets in some cases.

	Original number of		FRFS	UFRFS
Dataset	features	objects	Subset size	Subset size
water 2	38	390	7	7
water 3	38	390	7	7
cleveland	13	297	9	11
glass	9	214	9	7
heart	13	270	8	11
ionosphere	34	230	8	9
olitos	25	120	6	6
wine	13	178	10	7

Table 5.5: Subset sizes for UFRFS

The classification results are presented in Table 5.6. These demonstrate that despite the fact that UFRFS does not consider the decision feature for reduction, it retains useful features. This is borne out by comparison to the classification accuracy of the unreduced data, showing that the greatest decrease amongst all of the reduced data is only in the order of 10% overall. There are also cases where the use of UFRFS-reduced data outperforms the unreduced data and that of the FRFS-reduced data.

These promising results demonstrate that the UFRFS method can be used effectively to select features when class labels are not known. The classification accuracies for FRFS and UFRFS are generally comparable.

5.5 Summary

This chapter has presented some novel fuzzy-rough techniques for association learning based on the fuzzy upper and lower approximation concepts. The ap-

	Unreduced Data			FRFS Reduced Data			UFRFS Reduced Data		
Dataset	J48	PART	JRip	J48	PART	JRip	J48	PART	JRip
water 2	83.33	83.08	83.85	80.76	78.97	82.82	80.00	81.53	81.03
water 3	77.43	83.33	82.82	78.55	79.74	80.00	75.64	76.67	78.71
cleveland	51.85	50.17	52.19	55.01	53.19	54.55	53.19	51.51	55.21
glass	67.28	67.76	71.50	65.65	67.76	65.89	64.01	69.62	64.95
heart	76.66	73.33	77.41	78.84	76.30	75.82	71.48	68.69	69.25
ionosphere	87.82	88.26	86.52	83.98	85.23	86.96	83.47	82.17	83.91
olitos	67.50	57.50	70.83	65.00	64.17	63.33	55.00	60.00	55.86
wine	94.38	93.82	92.70	92.20	94.38	88.20	95.50	94.38	94.38

Table 5.6: Unreduced, supervised FRFS, and UFRFS Classification accuracies (%)

proaches are data-driven, and no user-defined thresholds are required. A choice must however be made with regard to similarity relations and fuzzy connectives. Note that these choices must also be made for all existing approaches which share the same underlying mathematical foundations however. A short experimental evaluation for 6 benchmark datasets is presented and the approach is compared with 2 other NN classifiers.

The detail and experimental results presented here offer only a glimpse of the technique. There is much potential for further work with regard to the measures which are currently utilised. Chapter 6 demonstrates the approach more fully where it is applied to the real-world problem of mammographic risk analysis.

Also presented is the fuzzy-rough approach to unsupervised feature selection. The approach employs a backward elimination-type search to remove features from the complete set of original features. As with the FRNN classifier, no thresholding information is required. The approach is compared with an advanced supervised approach and demonstrates that it can effectively remove redundant features. The subsets returned by this unsupervised method are of similar size to that of the supervised approach and classification of the reduced data shows that the method selects useful features which are of comparable quality.

Chapter 6

Application to Mammographic Image Analysis

Breast cancer is a major health issue, and the most common amongst women in the EU. It is estimated that 8–13% of all women will develop breast cancer at some point during their lives [20], [56]. Furthermore, in the EU and US, breast cancer is recognised as the leading cause of death of women in their 40s [20], [23], [56]. Although increased incidence of breast cancer has been recorded, so too has the level of early detection through the screening of potential occurence using mammographic imaging and expert opinion. However, even expert radiologists can sometimes fail to detect a significant proportion of mammographic abnormalities. In addition, a large number of detected abnormalities are usually discovered to be benign following medical investigation.

Existing mammographic Computer Aided Diagnosis (CAD) systems [86, 185] concentrate on the detection and classification of mammographic abnormalities. As breast tissue density increases however, the effectiveness of such systems in detecting mammographic abnormalities is reduced significantly. Also, it is known that there is a strong correlation between mammographic breast tissue density and the risk of development of breast cancer. Automatic classification which has the ability to consider tissue density when searching for mammographic abnormalities is therefore highly desirable. It must be stressed at this point that the problem under consideration here is *mammographic risk analysis* rather than mammographic diagnosis from images, an area where many publications have been written with respect to the application of machine learning techniques [2], [18], [71], [189]. As such, the technique aims to classify image data objects into one of four BIRADS categories [1] shown in Fig. 6.1 which relates to the tissue type found in each mammogram. Therefore, the purpose of the technique is not classify breast tissue abnormalities, but rather give an indication of the tissue density.

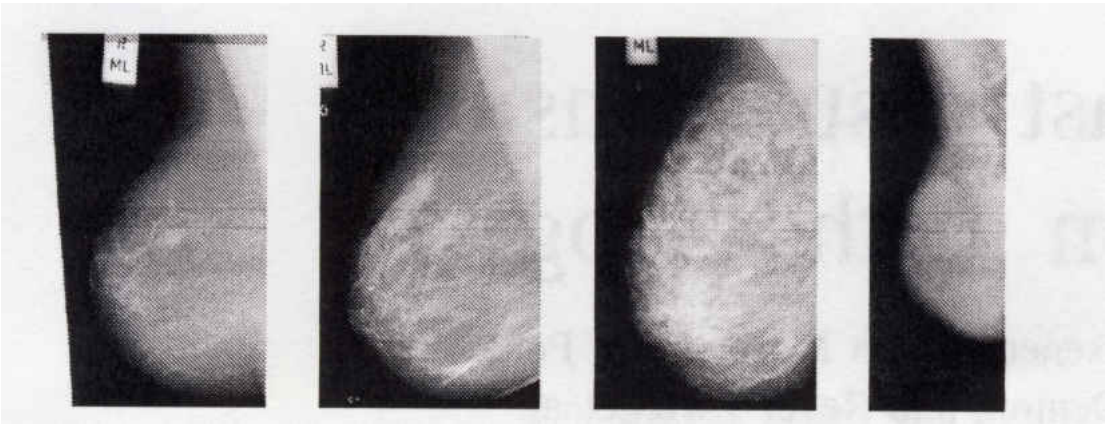

Figure 6.1: Example mammograms where breast tissue density increases from L-R Corresponding to BIRADS class I(far left) to class IV (far right)

In this chapter, a unified approach which employs a number of rough and fuzzy-rough approaches to deal with mammographic data is presented. This has been proposed as there are many and varied methods which can be employed for dealing with complex real-world data, and this can often lead to much confusion when choosing e.g which classifier to use, or which feature selection approach to employ for ease of computation or reduction of feature measurement. In particular, an approach to dealing with mammographic data is presented which considers each step from feature extraction through to data classification, although this chapter focuses primarily on the latter two steps.

Knowledge discovery from images often requires the maximisation of all of the information contained within the image. This means that initially large numbers of features are often extracted from the image. These features typically contain high levels of redundancy, irrelevancy, and noise. However, given that it is not known *a-priori* which features are most valuable and which are not, this is a necessary step. In the unified approach proposed here, a number of rough and fuzzy-rough methods are employed in an attempt to identify the most valuable features such that the process of extracting large amounts of features can be avoided. The selected features can then be fed back into the extraction phase ensuring that only those features need to be identified. The benefits of adopting such an approach include faster identification of relevant features, thus reducing the amount of time and computational effort required in the feature extraction phase. Use of fewer features means that any algorithms employed in both the training and testing phases of the classifier are potentially more accurate as there are fewer noisy features present. Additionally, fewer features means less computational overhead and hence the task is performed in less time. This helps to reduce the demands on experts' time, but most importantly can result in more accurate breast abnormality

risk assessments.

In addition to the proposed unified approach described previously, a fuzzy-rough nearest-neighbours (NN) classifier (described previously in Section 5.1.2) is applied to the image data. This classifier is compared with other approaches and demonstrates a significant increase in performance when compared with existing methods.

An overview of related work is presented and, this forms the basis for the work demonstrated later. The unified fuzzy-rough framework is also demonstrated. Comparative results are presented for a number of dimensionality reduction and classifier learner approaches within the framework discussed earlier.

6.1 System Overview

As mentioned previously, the problem considered in this chapter is that of mammographic risk analysis, where mammographic breast tissue density information extracted from images is used to assess how likely a woman is to develop breast cancer. The basic steps involved are outlined in Fig. 6.2, with detailed background described in [167]. The initial stages involve the segmentation and filtering of the mammographic images: all mammograms are pre-processed to identify the breast region and remove image background, labels, and pectoral muscle areas. This segmentation step results in a very minor loss of skin-line pixels in the breast area, however these pixels are not required for tissue estimation.

Then, a feature extraction step is performed, where the fuzzy c-means (FCM) algorithm [13] is employed which results in the devision of the breast into two clusters. A co-occurrence matrix (which is essentially a 2D histogram) is then used to derive a feature set which results in 10 features to describe morphological characteristics and 216 for the texture information (226 total). This feature set is then labelled using the consensus opinion of 3 experts to assign a label to each object mammogram using the BIRADS [1] classification. This consensus is determined where the classification for a given mammogram, which two or three radiologists agreed upon (majority vote) is selected as the 'consensus class'. If all experts classified a single mammogram differently, the median value is chosen as consensus opinion. The divergence in the opinion of the experts, is a major factor which often frustrates the use of automatic methods. This highlights the need to remove inter-observer (inter-operator) variability through the development of more autonomous approaches.

In this work the classification step is replaced with a dimensionality reduction phase and a classification phase. The existing feature set is used, as is the consensus expert labelling of the data. A unified framework such as that shown in Fig.

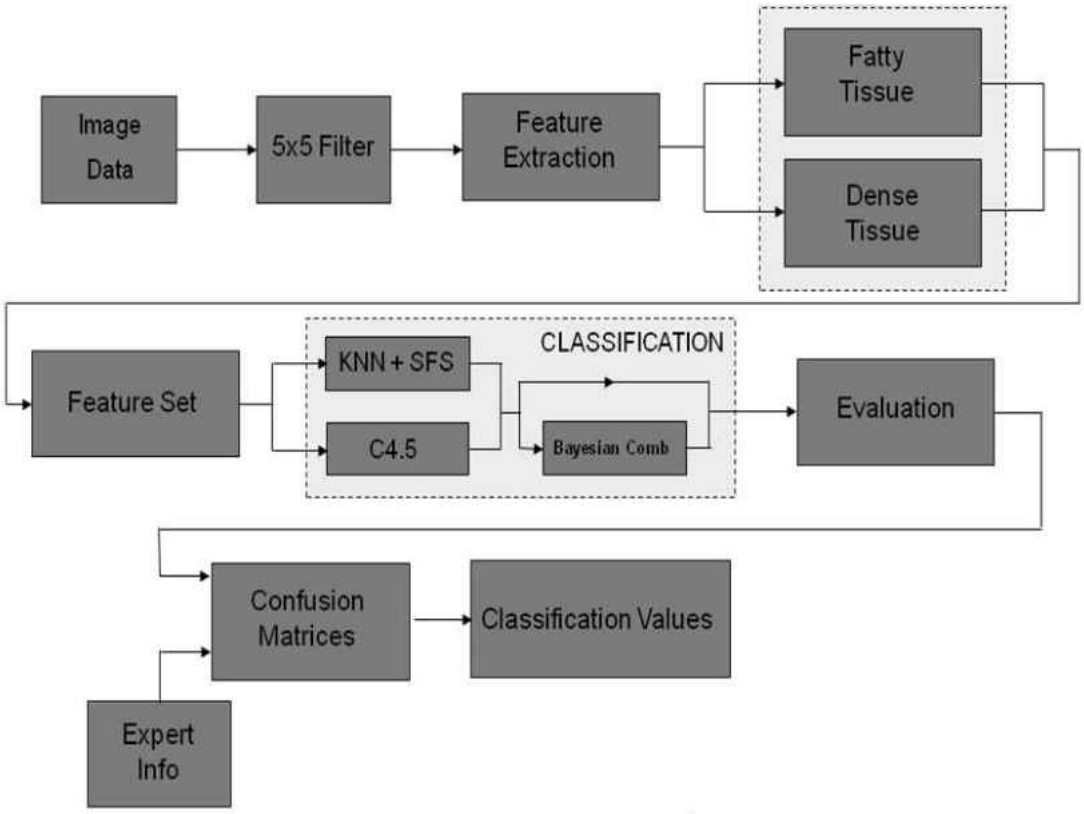

Figure 6.2: Mammographic density classification

6.3 is adopted to simplify the way in which knowledge can be efficiently learned from the (mammographic) training data, and therefore applied to real-world risk assessment problems. In this work, the focus lies in the implementation of rough and fuzzy-rough techniques for the dimensionality reduction and classifier learner steps. The approach for the feature extraction step employed here is documented in [167], however there is no reason why future work could not include a fuzzy-rough method to accomplish this in an effort to unify the underlying mathematical approach (see conclusion chapter for further discussion).

Efficient, and in particular, accurate classification of mammographic imaging is of high importance. Any improvement in accuracy for automatic mammographic classification systems can result in a reduction in the amount of required expert analysis thus improving the time taken to perform breast abnormality risk assessment. Also, by reducing inter-expert variation the resulting automatic risk assessments can be more accurate. The data in mammographic imaging is real-valued and as mentioned previously can be noisy. Clearly, any classifier employed must therefore have the ability to deal with such data. Discrete methods require that the real-valued data is discretised and thus result in information loss, however the methods described in this work require no discretisation and use only the

information contained within the data.

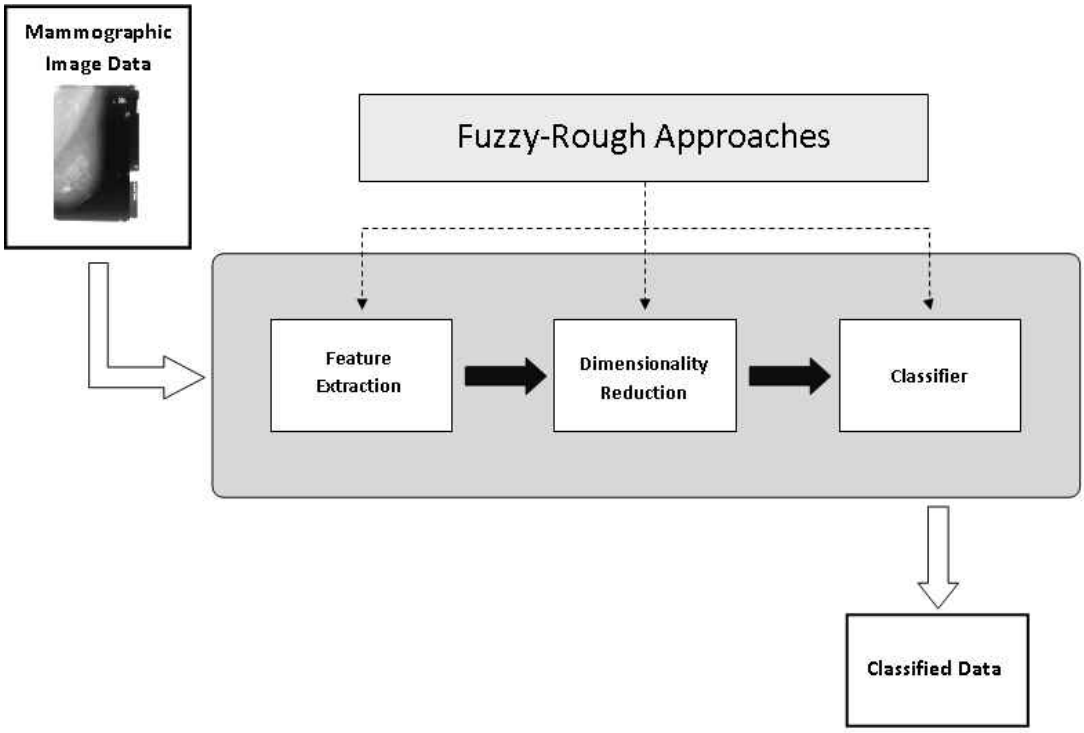

Figure 6.3: Unified fuzzy-rough framework for mammographic data analysis

6.1.1 Dimensionality Reduction

In this work, feature selection (FS) is utilised as the dimensionality reduction technique. This allows the identification of a minimal feature subset from a problem domain while retaining both a suitably high accuracy and the semantics entailed by the original features. In many real world problems, FS is essential due to the level of noisy, irrelevant or misleading features. By removing these factors, techniques for learning from data can benefit greatly. See the earlier literature review in Section 2.1.2 for a more detailed coverage of FS, and some representative applications.

6.1.1.1 Tolerance Rough Set Feature Selection

Unfortunately, one of the main disadvantages of the rough set methodology is its inability to deal with real-valued data unless the data is discretised which can result in information loss. One particular extension which has been proposed to address this shortcoming is the tolerance rough set model (TRSM) [209]. Other extensions such as variable precision rough sets (VPRS) [263] deal with misclassification of objects rather than real-valued data.

The work in this chapter utilises a rough set-based FS approach which has the ability to deal with real-valued data. It implements a version of tolerance rough sets [209] which also takes advantage of the information in the boundary region or region of uncertainty [144]. Section 4.2 contains a detailed study of this approach.

6.1.1.2 Fuzzy-Rough Feature Selection (FRFS)

The requirement of rough set theory to rely on discrete data implies an objectivity in the data that is simply not present. For example, consider an attribute *Blood Pressure* in a medical dataset. In the real world, this is a real-valued measurement but for the purposes of RST must be discretized into a small set of labels such as *Normal*, *High*, etc. Subjective judgments are therefore required to establish boundaries for objective measurements.

A more appropriate way of handling this problem is the use of *fuzzy-rough* sets [52]. Subjective judgments are not entirely removed as fuzzy set membership functions still need to be defined. However, the method offers a high degree of flexibility when dealing with real-valued data, enabling the vagueness and imprecision present to be modeled effectively.

Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge. Vagueness arises due to a lack of distinction or hard boundaries in the data itself. This is typical of human communication and reasoning. Rough sets can be said to model ambiguity resulting from a lack of information through set approximations.

Fuzzy-rough sets and their applications to FS are described in detail in Sections 2.5.2 and 4.4.1.

6.1.2 Nearest Neighbour Classification

In the previous work of [167], conventional crisp classifier learners were employed for the classification of the mammographic data – kNN, C4.5 [184], and a combined Bayesian estimation approach type classifier [53]. In this work a number of hybrid fuzzy set and rough set-based classifiers have been employed to classify the mammographic data.

In this section a number of existing classifiers as well as the application of a hybrid fuzzy-rough classifier described previously in Chapter 5 are examined. These include: FNN [103], a fuzzy version of the well-known kNN algorithm [53]; FRNN-O a fuzzy-rough ownership function based classifier [194, 230]; and VQNN a nearest neighbour (NN) classifier based on the vaguely quantified rough set model [36](discussed in Section 2.4.4).

The kNN algorithm assigns a test object to the decision class most common among its 'k nearest neighbours', i.e., the k training objects that are closest to the test object. An extension of the kNN algorithm to fuzzy set theory (FNN) was introduced in [103]. It allows partial membership of an object to different classes, and also takes into account the relative importance (proximity) of each neighbour with respect to the test instance. However, as correctly argued in [194], the FNN algorithm has problems dealing adequately with insufficient knowledge. In particular, when every training pattern is far removed from the test object, and hence there are no suitable neighbours, the algorithm is still forced to make clear-cut predictions. This is because the sum of the predicted membership degrees to the various decision classes is always required to be equal to 1.

6.1.3 Fuzzy-Rough Ownership kNN

Initial attempts to combine the FNN algorithm with concepts from fuzzy rough set theory were presented in [194, 230] (here denoted FRNN-O). In these papers, a fuzzy-rough ownership function is constructed that attempts to handle both "fuzzy uncertainty" (caused by overlapping classes) and "rough uncertainty" (caused by insufficient knowledge, i.e. attributes, about the objects). All training objects influence the ownership function, and hence no decision is required as to the number of neighbours to consider, although there are other parameters that must be defined for its successful operation. The FRNN-O approach is covered in detail in Section 5.2.1.

Note that the algorithm does not use fuzzy lower or upper approximations to determine class membership unlike the method proposed in Chapter 5. However, the method still requires a choice of parameter m, which plays a similar role to that in FNN.

6.1.4 Vaguely Quantified Rough Sets (VQRS)

Equations (5.3) and (5.4) have been conceived with the purpose of conserving the traditional lower and upper approximations in mind. Indeed, when X and R_P are both crisp, it can be verified that the original crisp rough set definitions are recovered. Note in particular how the inf and sup operations play the same role as the $\forall$ and $\exists$ quantifiers of the classical rough sets approach, and how a change in a single element can thus have a large impact on equations (5.3) and (5.4). This makes fuzzy-rough sets equally susceptible to noisy data (which is difficult to rule out in real-life applications) as their crisp counterparts.

To make up for this shortcoming, the work in [36] proposed to soften the universal and existential quantifier by means of vague quantifiers like *most* and

some. Mathematically, the vague quantifiers were modeled in terms of Zadeh's notion of a regularly increasing fuzzy quantifier Q: an increasing $[0,1] \rightarrow [0,1]$ mapping that satisfies the boundary conditions $Q(0) = 0$ and $Q(1) = 1$.

Examples of fuzzy quantifiers can be generated by means of the following parametrised formula, for $0 \leq \alpha < \beta \leq 1$, and x in $[0,1]$,

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \tag{6.1}$$

For instance, $Q_{(0.1,0.6)}$ and $Q_{(0.2,1)}$ might be used respectively to reflect the vague quantifiers *some* and *most* from natural language.

Once a couple (Q_l, Q_u) of fuzzy quantifiers is fixed, the Q_l-upper and Q_u-lower approximation of a fuzzy set A under a fuzzy relation R are defined by

$$\mu_{\underline{R_P}X}^{Q_u}(y) = Q_u\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{6.2}$$

$$\mu_{\overline{R_P}X}^{Q_l}(y) = Q_l\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{6.3}$$

for all y in $\mathbb{U}$. In other words, an element y belongs to the lower approximation of X if most of the elements related to y are included in X. Likewise, an element belongs to the upper approximation of X if some of the elements related to y are included in X. Notice that when X and R_P are a crisp set and a crisp equivalence relation respectively, the approximations may still be non-crisp.

The algorithm given in Fig. 5.4 can be adapted to perform VQRS-based nearest neighbours (VQNN) classification by replacing $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$ with $\mu_{\underline{R_P}X}^{Q_u}(y)$ and $\mu_{\overline{R_P}X}^{Q_l}(y)$. The computational complexity of this approach is similar to that of classical rough set approach.

6.2 Fuzzy-Rough Nearest Neighbours

This section provides a description of the new fuzzy-rough nearest neighbour algorithm. The need for such a classification technique arose from the fact that although the FRNN-O algorithm proposed in [194] uses a fuzzy-rough framework; no use is made of the fuzzy upper and lower approximations to determine class membership. This has prompted the development of an approach which was built upon the existing fuzzy-rough techniques which had been applied successfully to the feature selection problem [99]. As both the FS problem and the classification

problem are similar is many ways, the motivation was therefore quite clear.

The intuitive basis for the approach is that the lower and the upper approximation of a decision class, calculated by means of the nearest neighbours of a test object y, provide good clues to predict the membership of the test object to that class. Thus, by calculating the upper and lower approximation of a given decision class these can be employed as a metric for the test object in determining class membership.

6.2.1 FRNN Algorithm

The membership of a test object y to each (crisp or fuzzy) decision class is determined via the calculation of the fuzzy lower and upper approximation. The algorithm outputs the decision class with the resulting best fuzzy lower and upper approximation memberships. More specifically, if the membership of y to the fuzzy lower approximation of class C is high, it means that all of $y's$ neighbours belong to class C, while a high membership value of the fuzzy upper approximation of C indicates that at least one neighbour or neighbours belong to that class. The algorithm iterates through all of class concepts (X) in the training data. The decision class which results in the highest upper and lower approximation membership values is assigned to the test object.

The algorithm works by examining each of the classes in the training data in-turn. It computes the membership of a test object to the fuzzy upper and lower approximations. These values are then compared with the highest existing values. If the approximation membership values for the currently considered class are higher, then both are assigned these values and that class label is assigned to the test object. If not, the algorithm continues to iterate through all remaining decision classes. Classification accuracy is calculated by comparing the output with the actual class labels of the test objects. Further detailed description as well as a worked example are presented in Section 5.1.2.

6.3 Experimentation

In this section the results of applying the previously described classifiers and FS preprocessors are presented. Initially the classifiers are applied to the unreduced extracted feature data - i.e. data on which FS has not been performed. Classification is then performed on data which has been reduced by two previously described FS preprocessors DMTRS [144], and FRFS [99] as shown in Fig. 6.4. The results are then assessed for both FS methods used in conjunction with each of the individual classifiers. Additionally the results obtained in this work are

briefly compared with those reported in [167].

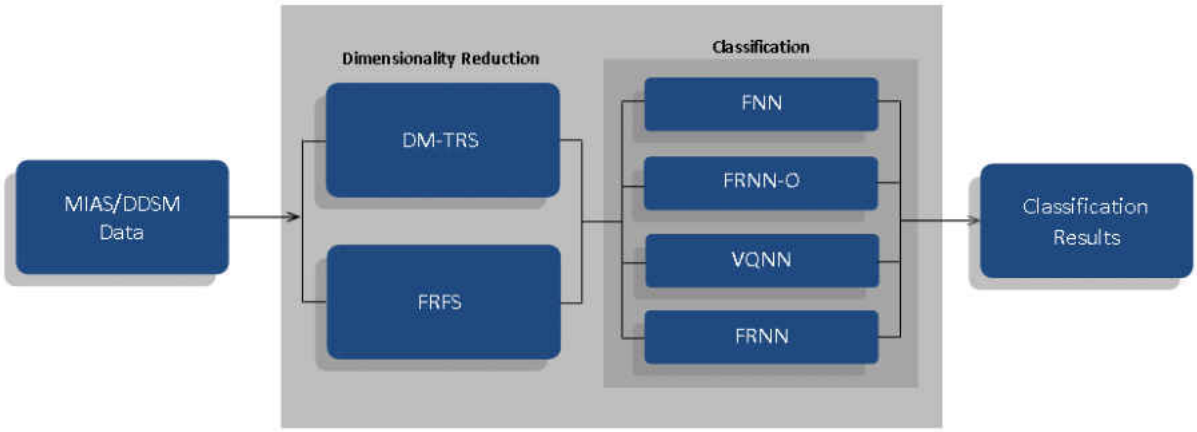

Figure 6.4: Experimental setup

Initially, the fuzzy-rough set classification techniques are applied to the *unreduced* datasets, and the 10-fold-cross validation (10-fold CV) approach is used to generate the classification models. The dimensionality of the data is then reduced and a summary of the average classification values achieved for each FS method is used to compare the methods.

6.3.1 Mammographic Risk Analysis Data

There are two datasets considered here, and both are available in the public domain: the Mammographic Image Analysis Society (MIAS) database [216], and the Digital Database of Screening Mammography (DDSM) [74]. The MIAS dataset is composed of Medio-Lateral-Oblique (MLO) left and right mammograms from 161 women (322 objects). Each mammogram object is represented by 281 features extracted using the process described earlier in Section 6.1 and in further detail in [167]. The spatial resolution of the images is $50\mu m \times 50\mu m$ and quantized to 8 bits with a linear optical density in the range $0 - 3.2$.

The DDSM database provides four mammograms, comprising left and right Medio-Lateral-Oblique (MLO) and left and right Cranio-Caudal (CC) views, for most women. To avoid bias only the right MLO mammogram for each woman is selected. The dataset contains 832 mammograms (objects) and again 281 features obtained in the same manner as those for the MIAS dataset above.

The class labels for each mammogram are assigned by three experts consensus opinion as described previously in Section 6.1. There are four discrete labels ranging from 1 to 4 relating to the BIRADS classification [1], where 1 represents a breast that is entirely fatty and 4 represents a breast that is extremely dense.

6.3.2 Experimental Setup

The FRFS preprocessor employs the fuzzy similarity defined in equation (4.19) along with the Łukasiewicz t-norm ($\max(x+y-1,0)$) and the Łukasiewicz fuzzy implicator ($\min(1-x+y,1)$). It has been shown that these work particularly well when used for fuzzy-rough feature selection [99].

The DM-TRS preprocessor used 4 different tolerance values (τ) – 0.97 and 0.98 for the MIAS dataset, while for the DDSM dataset the values 0.98 and 0.99 were chosen. These were the values that empirically demonstrated the best level of dimensionality reduction for each of the datasets respectively.

For each of the classifier learners the value of k is initialised as 30 and then decremented by 1 each time, resulting in 30 experiments for each dataset. Such a wide range of values for k ensures a comprehensive exploration and comparison of each of the classifiers. Cross validation of 10 times 10-fold cross-validation (10-fold CV) is performed for each experiment. Note that the k parameter is essential only for FNN and is not required for the other classifier learners. However, for ease of comparison, the other approaches have been adapted such that a k value can be specified. This is achieved by calculating the test objects k nearest neighbours rather than using all of the objects in the training set. For FNN and FRNN-O, m is set to 2. The VQNN approach was implemented using the commonly adopted $Q_l = Q_{(0.1,0.6)}$ and $Q_u = Q_{(0.2,1.0)}$, according to the general formula in equation (6.1).

For the new classifier approach, although there are no parameters to tune, decisions about which fuzzy relations and implicators must still be made. For the purpose of the experimentation documented here, the fuzzy relation given in equation (4.19) was chosen for simplicity. In the FRNN approach, the min t-norm and the Kleene-Dienes implicator I (defined by $I(x,y) = \max(1-x,y)$) were used.

6.3.3 Unreduced Data

The classification accuracy results for the unreduced data are presented in this section. This was achieved by applying each of the four classifiers to both of the datasets which gives a background against which to make subsequent comparative studies.

Considering the classification accuracy results illustrated in Fig. 6.5, it can be seen that there is little variation in the performance for the MIAS dataset. The FRNN-O approach seems to have a slight advantage, however this is only in the order of 2-6% for all values of k. The results for the DDSM dataset tell a slightly different story with VQNN achieving a small but clear advantage. FNN also appears to marginally outperform FRNN, and FRNN-O following a similar trend

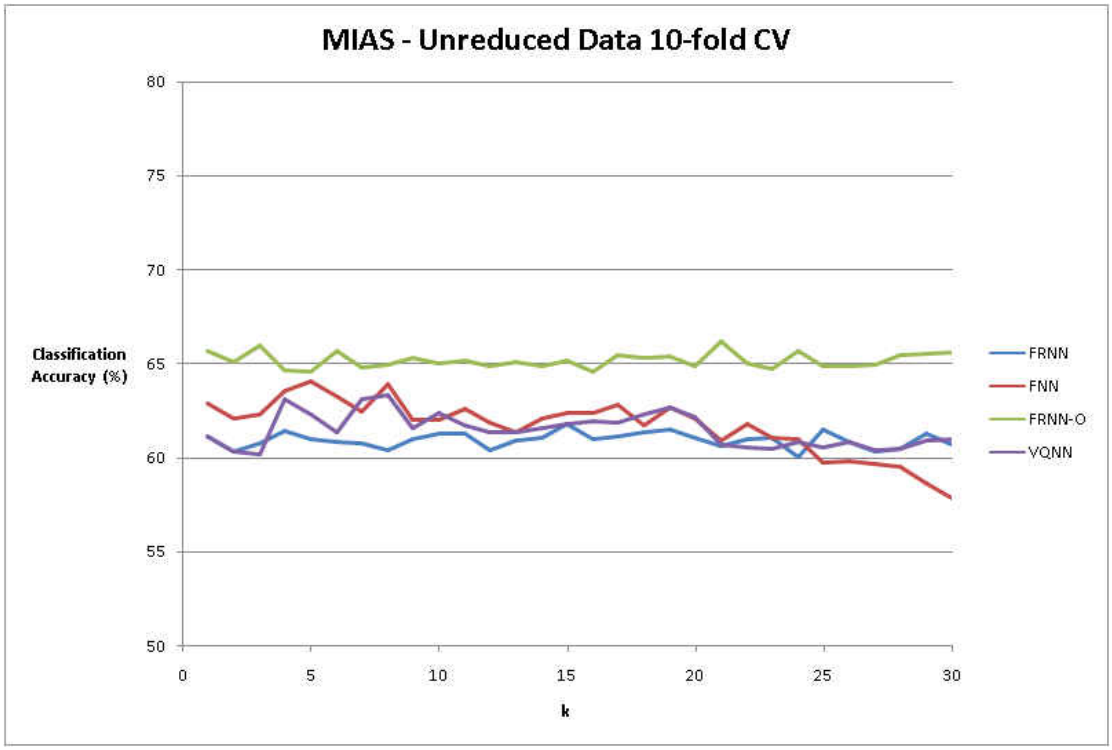

Figure 6.5: Classification accuracy: Unreduced MIAS data for the four methods and different values of k – 10CV

to that of VQNN. Generally, as the number of objects in the dataset increases, so too does the potential for measurement noise. The noise-tolerant characteristics of VQNN and the fact that the DDSM dataset has many more objects than the MIAS dataset may explain why VQNN performs particularly well in this case.

It is important to note at this point that the levels of performance shown for the FRNN approach are of little importance in this section as the data prior to reduction with FS contains much redundancy, irrelevance, and noise.

6.3.4 Reduced Data

In this section the results of classifying the MIAS and DDSM datasets following feature selection are presented. Classification accuracy results are provided for both DMTRS and FRFS, using both 10-fold CV and LOOCV. In Table 6.1, the subset sizes obtained following FS are presented. It is interesting to note that a substantial level of dimensionality reduction is achieved for both approaches. A reduction of 97.15% and 97.5% were achieved for the MIAS dataset, while the DDSM dataset (Table 6.2) achieved 97.15%, and 98.22%.

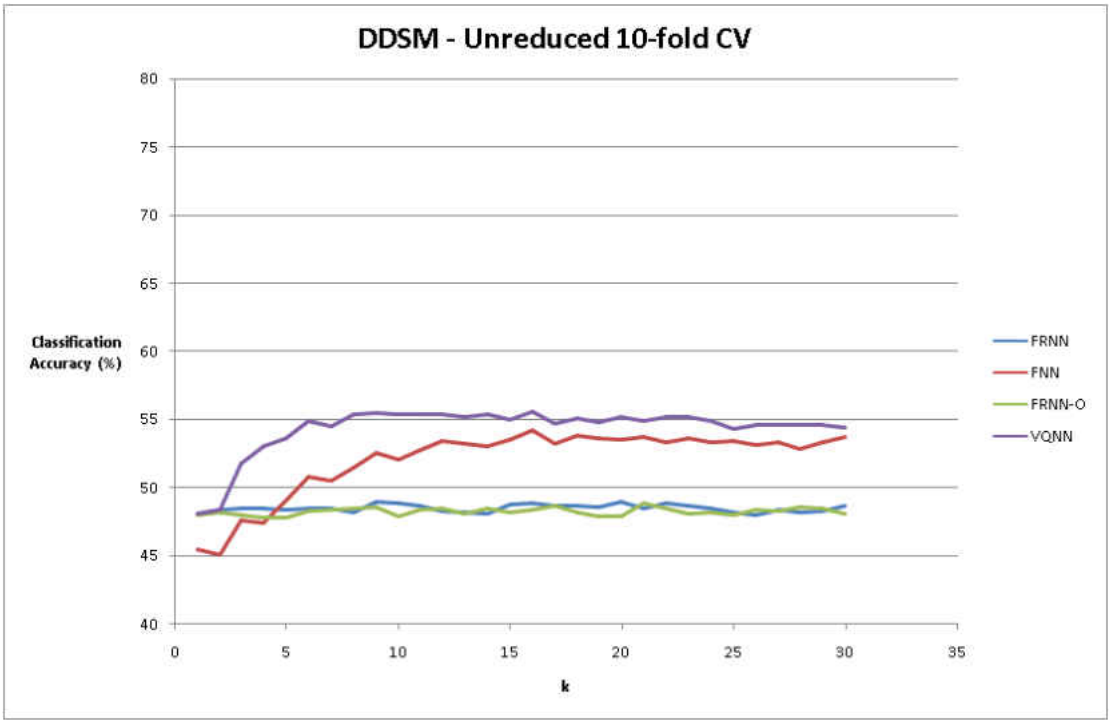

Figure 6.6: Classification accuracy: Unreduced DDSM data for the four methods and different values of k – 10CV

Orig No. of feats	DMTRS (τ=0.97)	DMTRS (τ=0.98)	FRFS
281	8	7	7

Table 6.1: Reduct sizes for the MIAS dataset following the application of DMTRS and FRFS

6.3.4.1 DMTRS Reduced Data

The results presented in this section illustrate the classification accuracies obtained when using DMTRS as a FS preprocessing step. There are a total of four diagrams (Fig. 6.7 – 6.10), two of which represent the tolerance values for the MIAS dataset (0.97 and 0.98), and the remaining two represent the values for the DDSM dataset (0.98 and 0.99).

Orig No. of feats.	DMTRS (τ=0.98)	DMTRS (τ=0.99)	FRFS
281	8	5	8

Table 6.2: Reduct sizes for the DDSM dataset following the application of DMTRS and FRFS

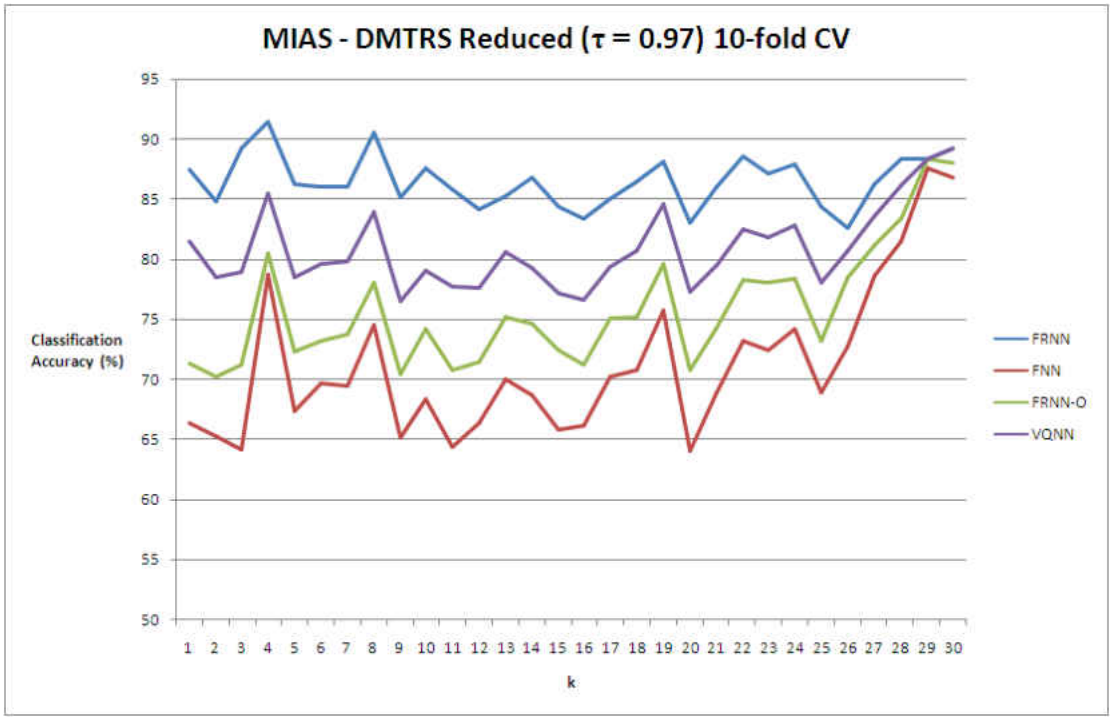

Figure 6.7: Classification accuracy: DMTRS reduced MIAS data for the four methods and different values of k – 10-fold CV

6.3.4.2 FRFS Reduced Data

The results shown in Fig. 6.11 and Fig. 6.12 are those obtained when applying the classifiers to the data following the application of FRFS to reduce the data.

Perhaps the most obvious aspect of the results demonstrated here is the increase in classification accuracy for all classifiers following the use of FS. The advantages of applying FS are manifold, however in this case the level of dimensionality reduction and the aforementioned increase in classification accuracy are borne out in Figs. 6.7–6.12.

6.3.5 Investigation I: Unreduced Data for All Classifiers

As clearly demonstrated in Figs. 6.7–6.12 employing either method for FS results in a significant increase in classification accuracy. Importantly, the FRNN technique proposed in Chapter 4 performs best for both the MIAS and DDSM datasets, with the VQNN approach closely mirroring the performance of FNN. FRNN-O also seems to show similar accuracy for some values of k to FRNN but fails to do so consistently.

Figs. 6.7 – 6.12 present the classification accuracy results following the application of both the FRFS and DMTRS feature selection pre-processors. What is most noticeable about these results is the overall increase in classification accu-

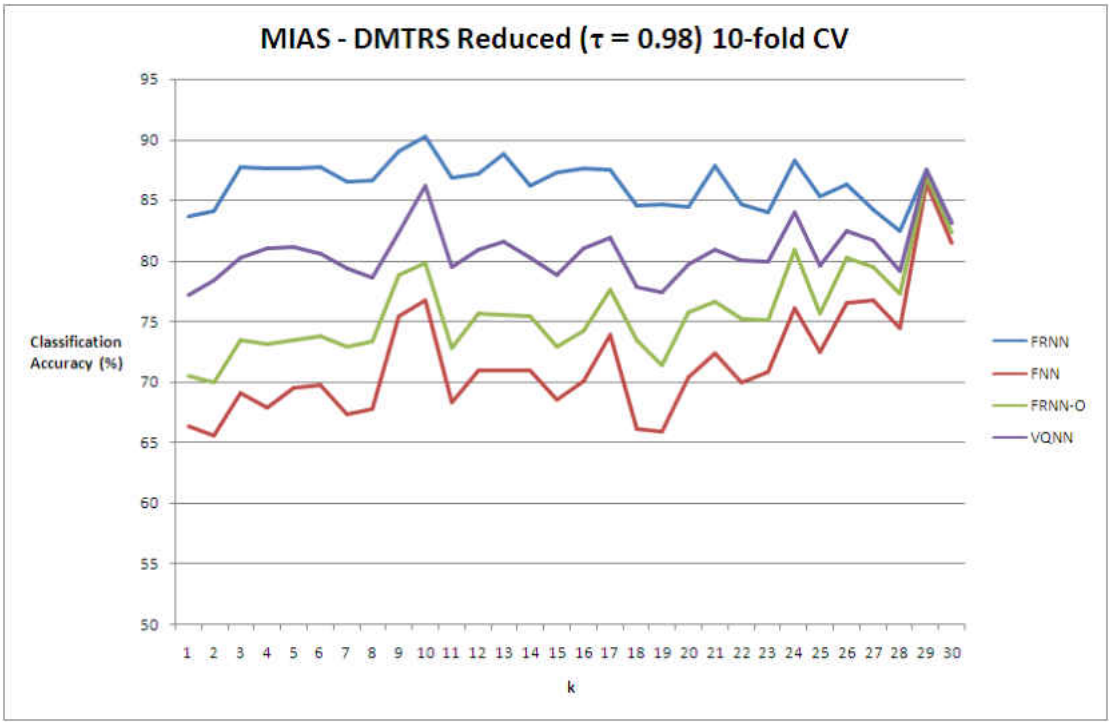

Figure 6.8: Classification accuracy: DMTRS reduced MIAS data for the four methods and different values of k – 10-fold CV

racy when FS has been employed. This not only highlights the level of redundant features in the original (unreduced) dataset, but also the ability of fuzzy-rough FS methods to reduce the data dimensionality considerably.

In order to aid comparison, the classification results of Figs. 6.7–6.12 have been summarised in Tables 6.3 and 6.4. Note that this summary is of *average* classification accuracy values. It is interesting to note the subset sizes obtained for each FS approach. For example in Table 6.1, the DMTRS approach achieves a subset sizes of 7 and 8 for the MIAS dataset. In Table 6.3, it can be seen that there is little difference in average classification accuracy between each of the tolerance values for DMTRS. Similarly, FRFS produces average classification results which are comparable with those of DMTRS for all classifiers. For the DDSM dataset however the DMTRS method manages better classification accuracies than FRFS for τ=0.98 for all classifiers except FRNN-O. Indeed the standard deviation value for this DMTRS subset is also lower than that achieved by FRFS. For the subset selected when τ=0.99, which is of size 5 compared to that of FRFS which is 8, there is little to separate FRFS and DMTRS in terms of average classification accuracy despite the greater level of dimensionality reduction.

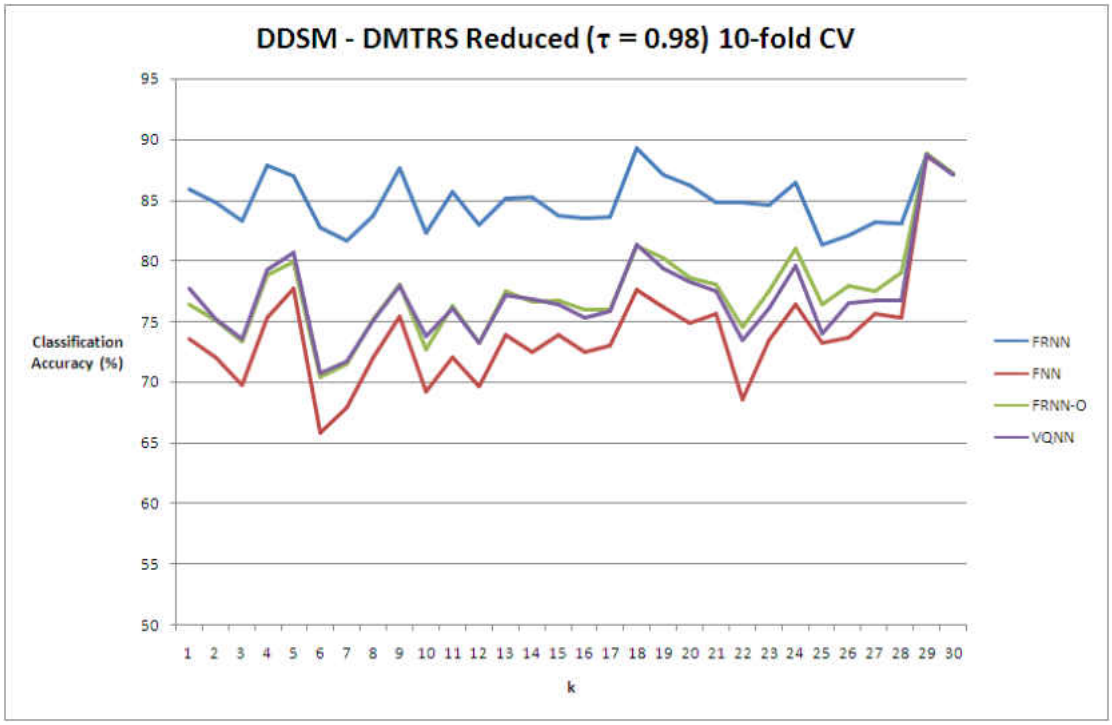

Figure 6.9: Classification accuracy: DMTRS reduced DDSM data for the four methods and different values of k – 10-fold CV

Classifier	FRFS	st.dev	DMTRS ($\tau = 0.97$)	st.dev	DMTRS ($\tau = 0.98$)	st.dev
FRNN	86.99	6.85	86.69	7.16	86.30	7.07
FNN	75.78	8.65	71.18	10.13	71.61	10.03
FRNN-O	82.21	7.42	75.77	8.77	75.78	8.53
VQNN	76.85	8.34	80.85	8.10	80.75	7.98

Table 6.3: MIAS - Average classification accuracy, and standard deviation results using 10-fold CV

6.3.6 Investigation II: Comparison with Current State-of-the-art

When comparing the results obtained here with those of [167], which represents the current state-of-the-art in automated mammographic breast density classification, it is clear that there is a significant improvement in classification accuracy. In [167], for the MIAS dataset classification rates of 77%, 72%, and 86% are achieved respectively for each of the classifier learners employed - namely SFS+kNN, C4.5, and a Bayesian classifier (although this is an approach which combines the previous two methods). Leave-one-out cross validation (LOOCV) is employed for cross validation in the paper in question, and k=7, for the kNN classifier. Additionally, the SFS+kNN approach employs a 'wrapper' type FS approach to select a subset

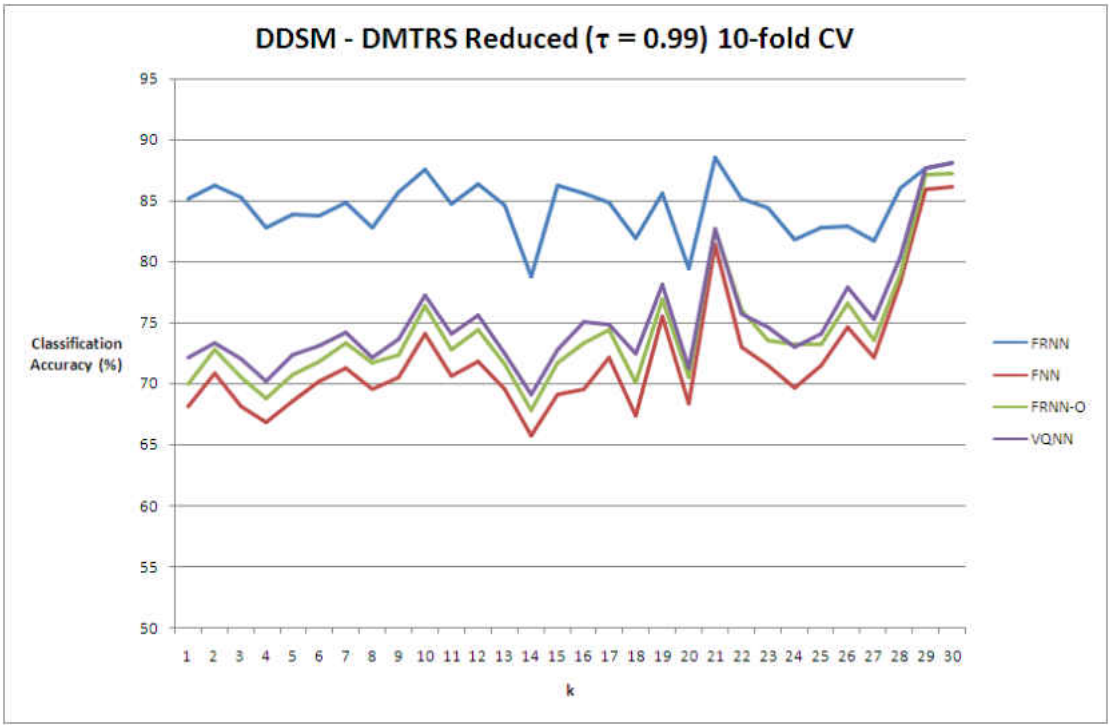

Figure 6.10: Classification accuracy: DMTRS reduced DDSM data for the four methods and different values of k – 10-fold CV

Classifier	FRFS	st.dev	DMTRS ($\tau = 0.98$)	st.dev	DMTRS ($\tau = 0.99$)	st.dev
FRNN	82.60	7.98	84.85	6.75	84.52	7.36
FNN	72.98	9.43	74.07	8.83	72.08	10.29
FRNN-O	81.14	8.69	77.39	7.67	74.14	9.44
VQNN	72.81	9.13	77.05	8.08	75.20	9.13

Table 6.4: DDSM - Average classification accuracy, and standard deviation results using 10-fold CV

of size 9 for MIAS and 9 also for the DDSM data.

Both DMTRS and FRFS feature selection approaches achieve results of 8 and 7 for MIAS and 5 and 8 for DDSM. Both of these approaches find smaller subset sizes when compared to the approach noted above whilst simultaneously leading to a significant increase in both average classification accuracy values albeit using 10-fold CV - see Fig. 6.6 - 6.8. As demonstrated previously, more optimum values can be achieved for individual values of k (Figs. 6.7 - 6.12) rather than considering only those average classification accuracy results.

Both FS techniques employed here are data-driven and do not require any normalisation or transformation of the data. In the work of [167] however, the data has to be normalised prior to the application of wrapper FS using kNN.

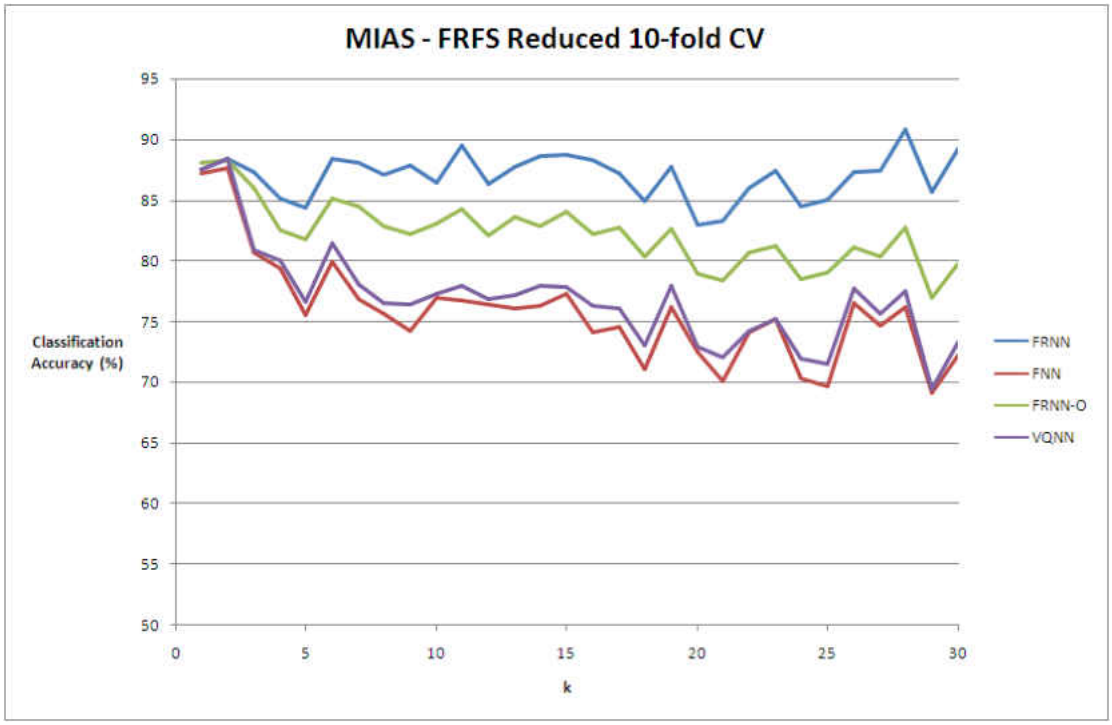

Figure 6.11: Classification accuracy: FRFS reduced MIAS data for the four methods and different values of k – 10-fold CV

Classifier	FRFS	st.dev	DMRSAR ($\tau = 0.98$)	st.dev	DMRSAR ($\tau = 0.99$)	st.dev
FRNN	84.12	35.62	86.41	33.10	86.50	32.92
FNN	72.54	43.96	75.17	42.04	73.02	42.98
FRNN-O	82.27	37.07	78.13	40.44	75.12	41.87
VQNN	74.46	42.94	77.42	40.75	76.20	41.36

Table 6.5: DDSM - Average classification accuracy, and standard deviation results using LOOCV

This may have the effect of information loss since it involves subjective human intervention when dealing with the data.

Considering the FRNN results obtained here for the MIAS dataset, classification accuracies of 91.4%, 90.28%, and 90.81%, were achieved for DMTRS($\tau = 0.97$), DMTRS($\tau = 0.98$), and FRFS reduced data respectively. Indeed, if the results from Table 6.3 are examined, it can be seen that even the average classification accuracies are considerably better in most cases than those obtained in [167].

For the DDSM dataset where classification accuracies of 70%, 72%, and 77% have been achieved in the previous work [167], considerably improved results have also been obtained using the new fuzzy and fuzzy-rough methods - 89.24%, 88.51%, and 85.84%. Again the average classification results of Table 6.4 reflect what has

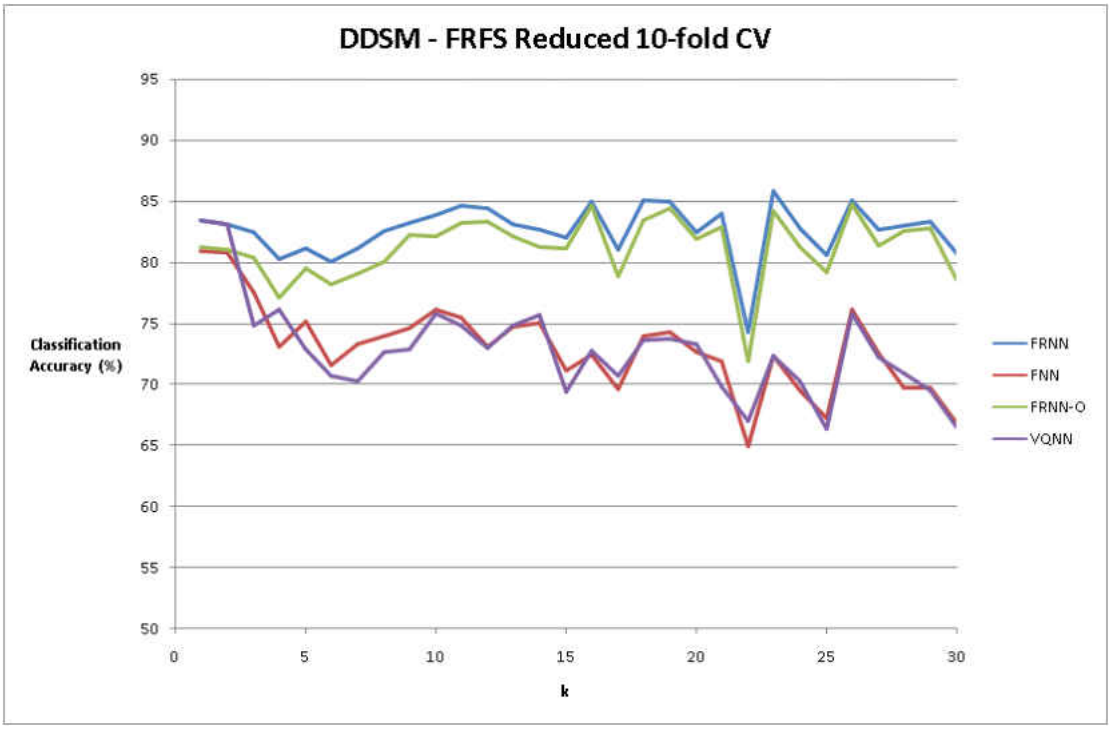

Figure 6.12: Classification accuracy: FRFS reduced DDSM data for the four methods and different values of k – 10-fold CV

Classifier	FRFS	st.dev	DMRSAR ($\tau = 0.97$)	st.dev	DMRSAR ($\tau = 0.98$)	st.dev
FRNN	87.34	32.76	86.69	33.01	87.21	32.95
FNN	72.82	43.74	71.79	43.83	69.94	44.84
FRNN-O	81.09	38.80	76.49	41.50	75.42	42.34
VQNN	78.15	40.86	80.76	38.54	80.00	39.50

Table 6.6: MIAS - Average classification accuracy, and standard deviation results using LOOCV

also been demonstrated in the case of the MIAS dataset.

The results obtained above show that the work described here can easily outperform that in [167] despite the use of 10-fold CV. LOOCV is also employed for cross validation in this section, such that both approaches can be compared directly. The subset sizes are identical for those discovered previously, only the cross-validation technique for the classifier learners has been altered to LOOCV.

Considering the results obtained in this work, values of 96.58%, 88.92%, 90.37%, and 92.23% using the four hybrid fuzzy-rough set-based classifiers described earlier. Similar results are obtained regardless of the FS approach employed. Indeed, if the results from Tables 6.4 and 6.5 are examined, it can be seen that even the average classification values are considerably better in most cases than those obtained in [167].

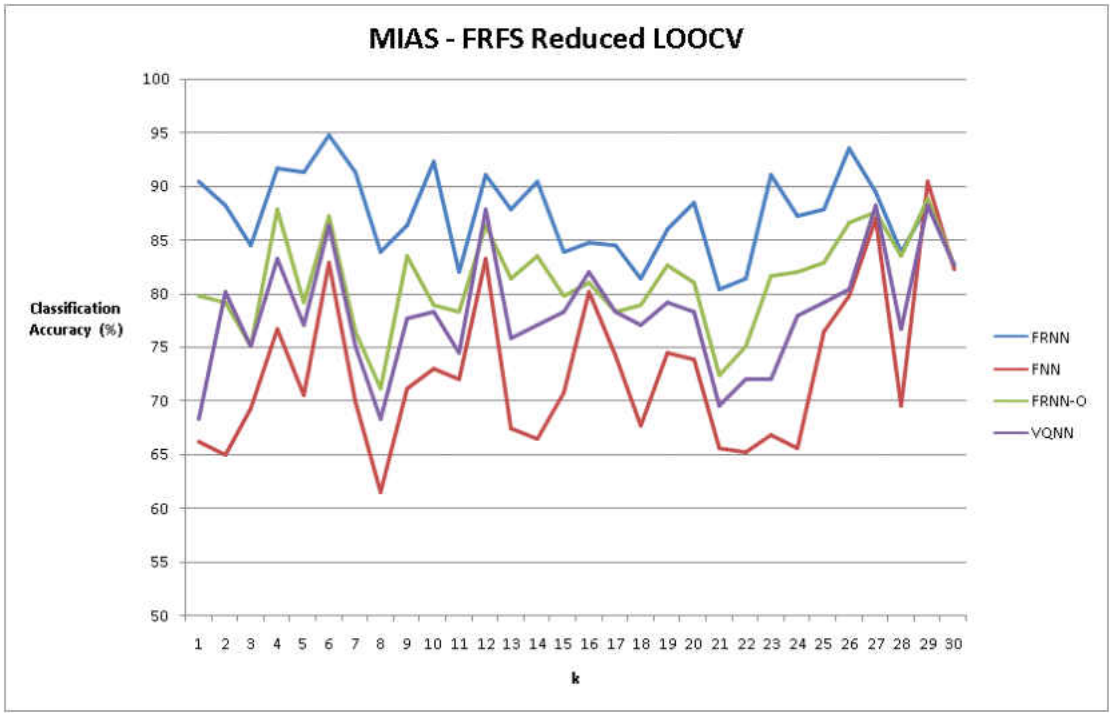

Figure 6.13: Classification accuracy: FRFS reduced MIAS data for each of the four methods and different values of k – LOOCV

For the DDSM dataset where classification accuracies of 70%, 72%, and 77% have been achieved in previous work, considerably improved results have also been obtained using the new fuzzy and fuzzy-rough methods - 96.63%, 94.07%, 95.60%, and 94.71%. Again the average classification results of Tables 6.4 and 6.5 reflect what has also been demonstrated in the case of the MIAS dataset.

6.4 Summary

This chapter has demonstrated the application of fuzzy-rough methods to data for mammographic risk analysis. It has also introduced a new NN classification approach and demonstrated how this can be applied for the analysis of mammographic data. In particular, it has demonstrated how the classification accuracy for mammographic risk-analysis can be increased significantly by employing fuzzy classifiers which have the ability to handle real-valued data.

Most importantly however, the value of adopting a unified approach has been highlighted. This is clearly shown in the large improvement of classification accuracy over the unreduced data for all classifier methods and also the significant reduction in dimensionality, which has a direct impact on the time taken to classify mammographic density. The use of FS to identify information-rich features whilst minimising feature measurement noise from the many initially extracted features

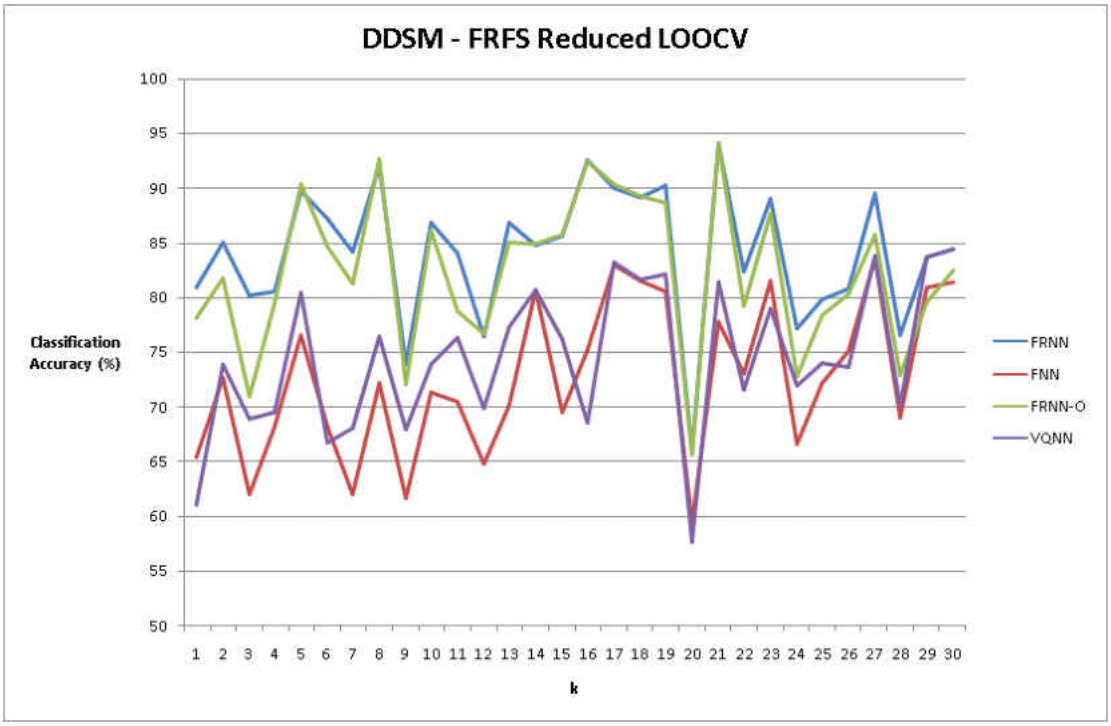

Figure 6.14: Classification accuracy: FRFS reduced DDSM data for each of the four methods and different values of k – LOOCV

is important. It can be used as an indicator to identify the same information in previously unseen mammograms thus, reducing the time needed in extracting many irrelevant, redundant and noisy features. Increases in classification accuracy for diagnosis means a benefit not only for the patient but also a reduction in expert analysis thus the minimising inter-observer variation. Additionally, correct initial identification of breast density can potentially mean that further additional screening of the same woman is not required, reducing the physical demands and stresses of further examination.

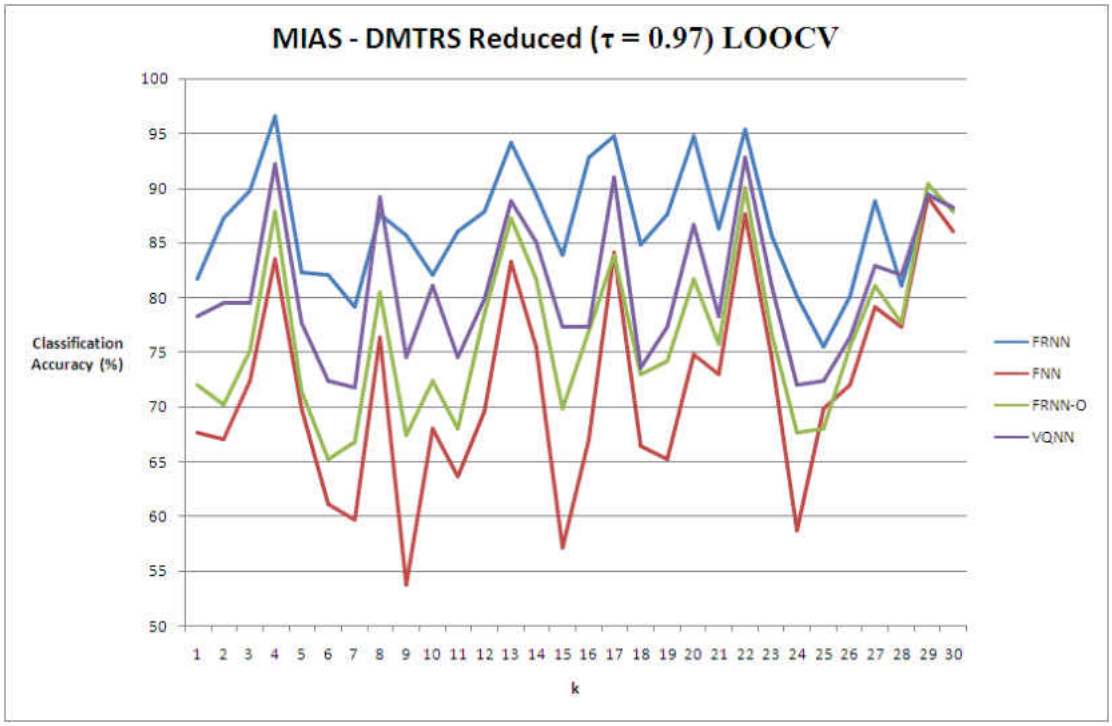

Figure 6.15: Classification accuracy: DMTRS (τ=0.97) reduced MIAS data for the four methods and different values of k – LOOCV

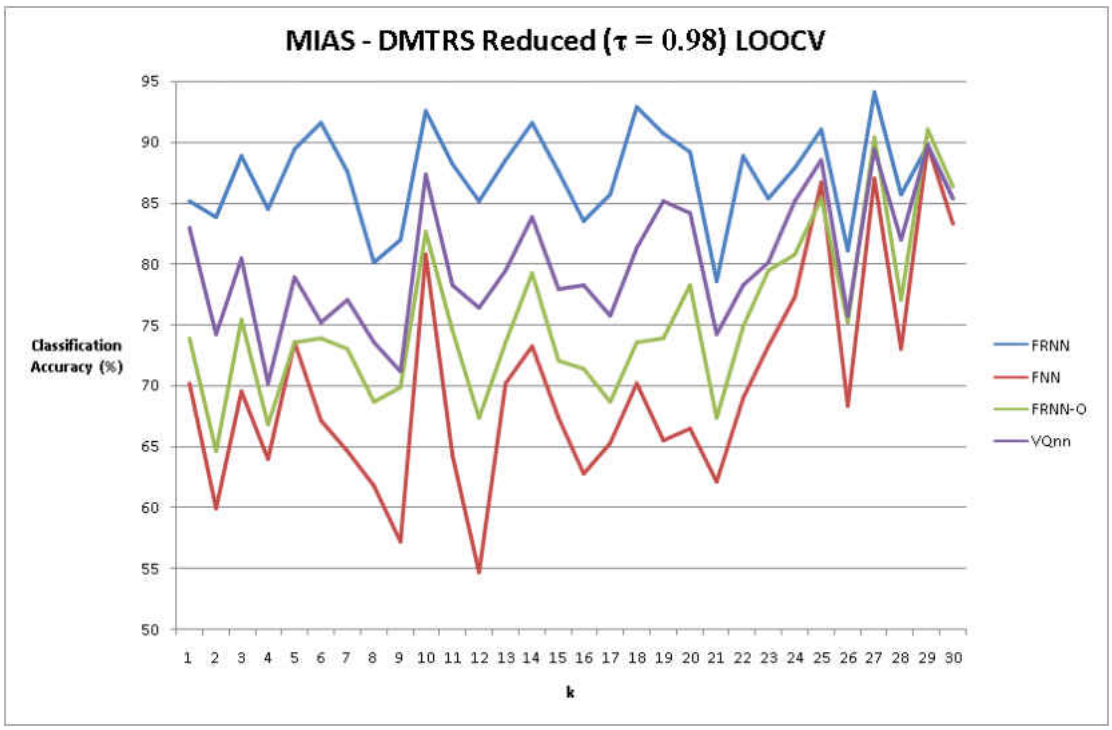

Figure 6.16: Classification accuracy: DMTRS (τ=0.98) reduced MIAS data for the four methods and different values of k – LOOCV

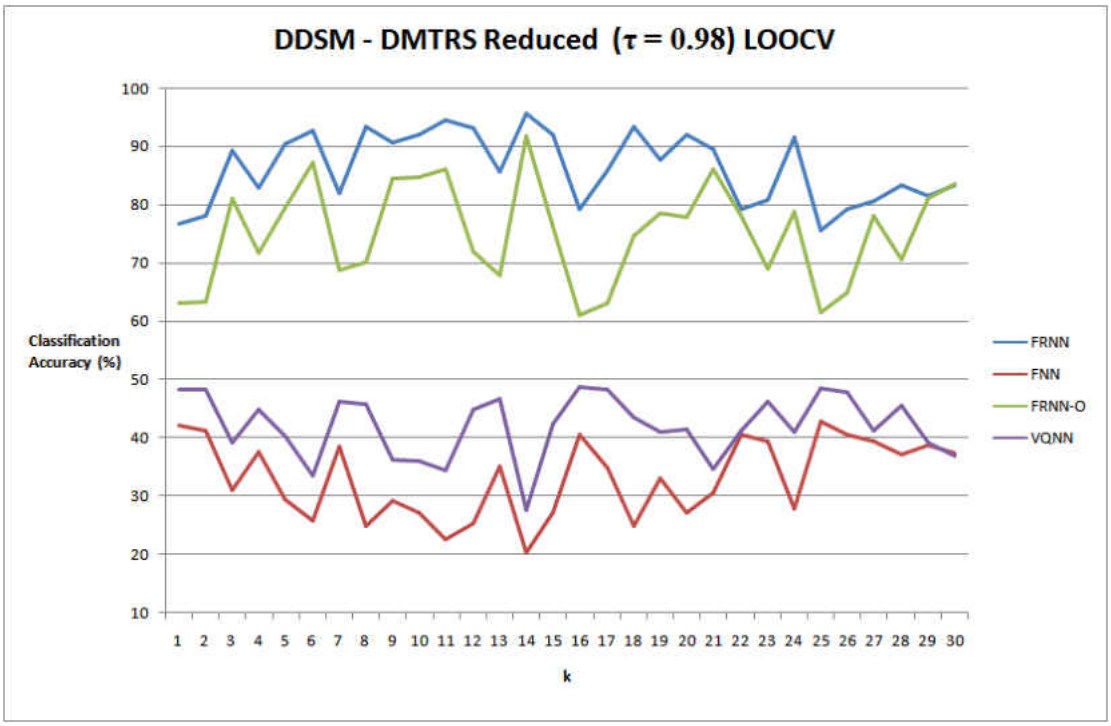

Figure 6.17: Classification accuracy: DMTRS (τ=0.98) reduced DDSM data for the four methods and different values of k – LOOCV

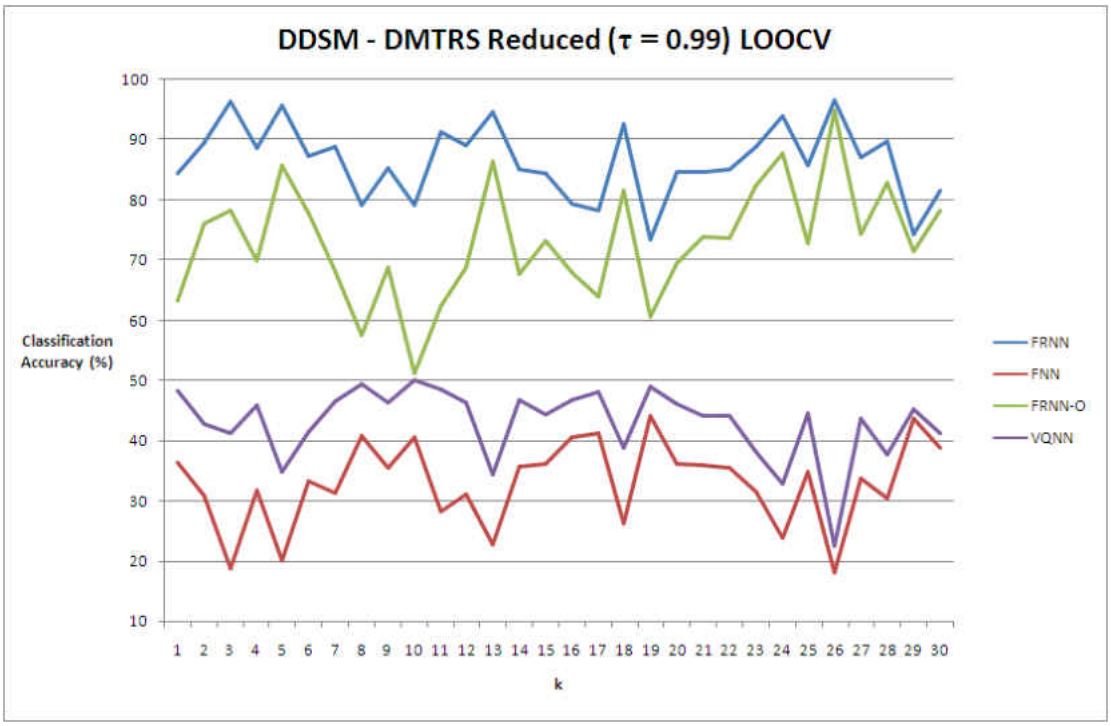

Figure 6.18: Classification accuracy: DMTRS (τ=0.99) reduced DDSM data for the four methods and different values of k – LOOCV

Chapter 7

Application to Plant Monitoring

The ever-increasing demand for dependable, trustworthy intelligent diagnostic and monitoring systems, as well as knowledge-based systems in general, has focused much of the attention of researchers on the knowledge-acquisition bottleneck. The task of gathering information and extracting knowledge from it is known to be the most difficult part of creating a knowledge-based system. Complex application problems, such as reliable monitoring and diagnosis of industrial plant, are likely to present large numbers of features, many of which will be redundant for the task at hand [178, 195]. Additionally, inaccurate and/or uncertain values cannot be ruled out. Such applications typically require convincing explanations about the inference performed, therefore a method to allow automated generation of knowledge models of clear semantics is highly desirable.

In this chapter the UFRFS method described in Section 5.3.1 is applied to the problem of selecting information-rich features for a water treatment plant monitoring system. The work here aims to reduce the dimensionality of the existing data in order to simplify rulesets induced from historical descriptions of domain features which are often of high dimensionality. The UFRFS method demonstrates that it can eliminate redundant features using the fuzzy dependency measure and thus reduce dimensionality. The UFRFS reduced data is compared with the unreduced data and the complexity of the rules induced from both are examined. Also the classification accuracies are examined once again for both the reduced and unreduced data.

7.1 Rule Induction

The most common approach to developing expressive and human readable representations of knowledge is the use of if-then production rules [116]. In order to speed up rule induction learning algorithms (RIA) and reduce rule complexity,

a preprocessing step is required. This is particularly important for tasks where learned rulesets need to be regularly updated to reflect the changes in the description of domain features. This step reduces the dimensionality of potentially very large feature sets while minimising the loss of information needed for rule induction. It has an advantageous side-effect in that it removes redundancy from the historical data. This also helps to simplify the design and implementation of the actual pattern classifier itself, by determining what features should be made available to the system. In addition, the reduced input dimensionality increases the processing speed of the classifier, leading to better response times. Most significant, however, is the fact that the technique employed here (UFRFS - see Section 5.3.1) preserves the semantics of the surviving features following the removal of any redundancy. This is essential in satisfying the requirement of user interpretability of the generated knowledge model, as well as ensuring the transparency of the pattern classification process.

There exist a number of approaches relevant to the rule induction task at hand, both from the point of view of applications and that of computational methods. For example, the FAPACS (*Fuzzy Automatic Pattern Analysis and Classification System*) algorithm documented in [6, 26] is able to discover fuzzy association rules in relational databases. It works by locating pairs of features that satisfy an 'interestingness' measure that is defined in terms of an adjusted difference between the observed and expected values of relations. A similar method, [90] has proposed modifications to decision trees to combine traditional symbolic decision trees with approximate reasoning, offered by fuzzy representation. This approach redefines the methodology for knowledge inference, resulting in a method best suited to relatively stationary problems.

A common disadvantage of these techniques is their sensitivity to high dimensionality. This may be remedied using conventional work such as Principal Components Analysis (PCA) [48]. Unfortunately, whilst efficient, PCA irreversibly destroys the underlying semantics the data, as discussed in Section 2.1.1. Most semantics-preserving dimensionality reduction (or feature selection) approaches tend to be domain specific, however, relying on the use of well-known features of the relevant application domain problems.

Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets [254]) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge [52] and can be applied to the problem of FS. This chapter is based on the most recent work as reported in [146] (and also described in Section 5.3.1 of this thesis), presents such a method which employs fuzzy-rough sets to improve the handling of this uncertainty. The theoretical domain independence of the approach allows it to be employed for different rule

induction algorithms and classifier learners. Furthermore, this method uses only the information of the conditional features to reduce the data - no information about the class labels is considered. In light of this, the present work is developed in a modular manner. Note that both of the approaches given in [203] and [204] are similar to the work described here, however supervised approaches are used for selection of features.

7.2 Unsupervised Fuzzy-Rough Feature Selection

In the previous chapter, it was demonstrated how FRS can be applied to the problem of supervised feature selection. One of the most important aspects relating to feature set reduction is the fuzzy-rough dependency measure (see Section 4.4.1), and it is this measure which is also employed for the new unsupervised fuzzy-rough FS (UFRFS) method described in this section. A short worked example is also provided here to illustrate the approach.

7.2.1 Fuzzy Dependency

The central idea behind the work described in Section 5.3.1 is that, as with supervised fuzzy-rough FS [99], the fuzzy dependency measure can also be used to discover the inter-dependency of features. This can be achieved by simply substituting the decision feature(s) of the supervised approach for any given feature or group of features.

Although the above proposal may be applied to evaluate the dependency between any two subsets of features in theory, practically, this may be computationally prohibitive. Fortunately, for unsupervised feature selection it is sufficient to find the dependency between a single feature and other subsets of features. If it has been established that one single feature depends fully on a feature subset then that feature can be removed. Hence, the proposed approach only requires the calculation of the dependency in the specific case of (5.12) where Q contains just a single feature. In light of this observation, equation (5.3) as used in (5.13) becomes:

$$\mu_{\underline{R_P}R_u z}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_{R_u z}(y)) \tag{7.1}$$

where z denotes the single feature under examination and $R_u z$ indicates the tolerance class (or fuzzy equivalence class) for object u of z.

7.2.2 Classification and Rule Induction

To show the potential utility of unsupervised fuzzy-rough feature selection, the UFRFS method is applied as a pre-processor to some existing fuzzy classifiers and rule induction algorithms (RIA). The classification algorithms used are: FRNN [141], [143] (described in Chapter 5 of this thesis also), and QSBA [188]. QSBA works by generating fuzzy rules using the fuzzy subsethood measure for each decision class and a threshold to determine what appears in the rule for that decision class. The fuzzy subsethood measure is then used to act as weights, and the algorithm then modifies the weights to act as fuzzy quantifiers.

Rule induction algorithms include JRIP, and PART, which are described in detail in Section 3.5.1. These techniques have been shown to produce highly competitive results [237] in terms of both classification accuracy, and number of rules generated. However, as is the case for most rule induction algorithms, the resultant rules may be unnecessarily complex due to the presence of redundant or misleading features. Unsupervised Fuzzy-Rough Feature Selection may be used to significantly reduce dataset dimensionality, removing redundant features that would otherwise increase rule complexity and reducing the time for the induction process itself.

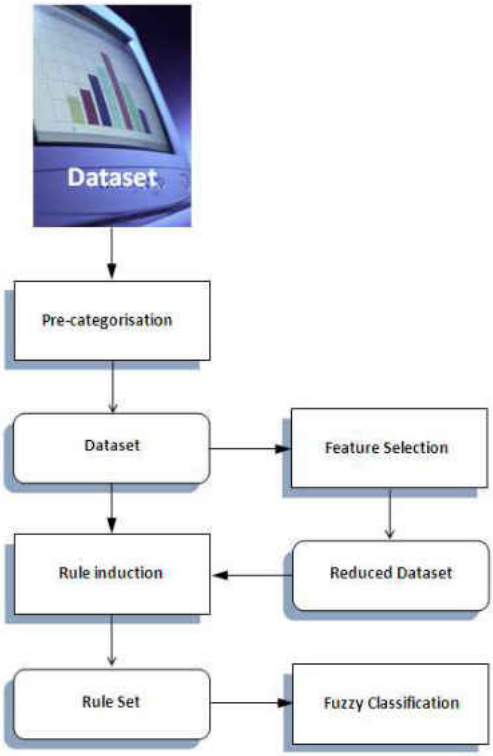

Figure 7.1: Modular decomposition of the implemented system

Note that the original monitoring system as shown in Fig. 7.1, and developed in [202] consisted of several modules. This modular structure is retained for the

work described here, but the feature selection step is replaced with UFRFS. A pre-categorisation step was used originally such that the data could be discretised and used with traditional rough set theory. This step is now redundant as UFRFS has the ability to deal with real-valued data, and thus reduce the potential for information loss.

7.3 A Realistic Application

In order to evaluate the utility of the UFRFS approach and to illustrate its domain-independence, a challenging test dataset was chosen, namely the Water Treatment Plant Database [158]. The dataset itself is a set of historical data recorded over 521 days (or data objects), with 38 different input features measured each day. Each daily object is classified into one of thirteen categories depending on the operational status of the plant. However, these can be collapsed into just two or three categories (i.e. *Normal* and *Faulty*, or *OK*, *Good* and *Faulty*) for plant monitoring purposes as many classifications reflect similar performance. Because of the efficiency of the actual plant the measurements were taken from, all faults appear for short periods (usually single days) and are dealt with immediately. This does not allow for a lot of training examples of faults, which is a clear drawback if a monitoring system is to be produced. However, it should be emphasised that for the work presented here, that the conditional features are not considered for the reduction of the data.

The 38 conditional features account for the following five aspects of the water treatment plant's operation (see Fig 7.2):

1. Input to plant (9 features)
2. Input to primary settler (6 features)
3. Input to secondary settler (7 features)
4. Output from plant (7 features)
5. Overall plant performance (9 features)

It is likely that not all of the 38 input features are required to determine the status of the plant, hence the dimensionality reduction step. However, choosing the most informative features is a difficult task as there will be many dependencies between subsets of features. There is also a number factors associated with the monitoring of large numbers of inputs. Firstly there is the equipment cost since more gauges and measuring equipment means increased initial expenditure on

this equipment, as well as on-going maintenance costs. Furthermore, the larger the amount of measuring equipment involved, the more inherently unreliable the system becomes. It is soon clear that a reduction in the number of measurements is highly desirable.

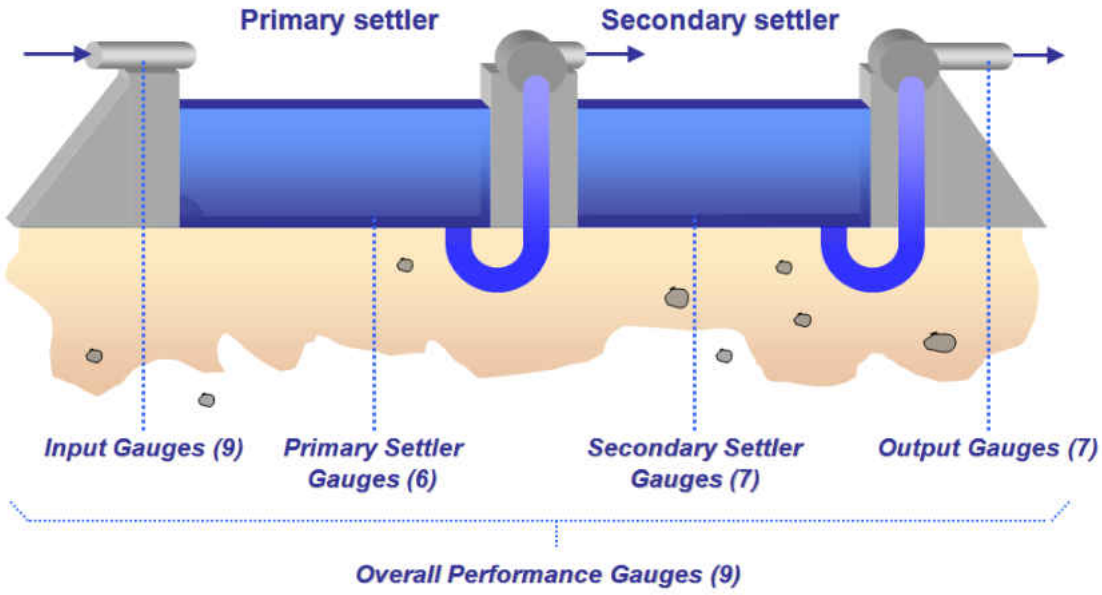

Figure 7.2: Water treatment plant overview

7.4 Experimental Results

This section firstly provides the results for the UFRFS-based approach compared with the unreduced data in using the previously mentioned classifier learners. Note that all classifiers use 10-fold cross validation in generating the classification results. The FRNN algorithm does not employ a k value but instead uses all of the objects in the training data. Next, a comparative experimental study is carried out employing the RIA; and the ruleset complexity and sizes are compared for the unreduced and reduced data.

7.4.1 Classification of Unreduced and Reduced Features

First of all, it is important to show that, at least, the use of features selected does not significantly reduce the classification accuracy as compared to the use of the full set of original features. The results which demonstrate the dimensionality reduction are presented in Table 7.1 for both datasets. They show a significant reduction in dimensionality from the original feature sizes.

Table 7.2 compares the classification accuracies of the reduced and unreduced datasets. As can be seen, there is only a small difference between the UFRFS

Dataset	Orig No. of feats	UFRFS reduced
water 2-class	38	8
water 3-class	38	7

Table 7.1: Subset sizes obtained using UFRFS

results and the unreduced data (in the order of less than 10%) for both classifiers. This is not significant given the corresponding level of dimensionality reduction in each case. Indeed, for the FRNN for the classifier there is a slight increase for the 3-class data.

	Unreduced		UFRFS Reduced	
Dataset	QSBA	FRNN	QSBA	FRNN
water 2-class	85.38	84.61	75.87	80.00
water 3-class	82.30	78.46	72.21	79.25

Table 7.2: Classification accuracy results: reduced and unreduced data

7.4.2 Rule Induction

Table 7.3 compares the resulting rule arity of the two approaches. It is evident that rules induced using UFRFS as a preprocessor are simpler, albeit with a little loss in classification accuracy. In fact, the simple rules produced regularly outperform the more complex ones generated by the unreduced approach. The rule arity of the UFRFS reduced 3-class data is 3.0 which is less than that of the unreduced optimum, 7.0 for the PART RIA. Although the rule arity values are 3.0 for both the unreduced and reduced 3-class data using JRIP, if the number of rules is examined in Table 7.4, it can be seen that the unreduced data generates 7 rules, whilst for the resulting reduced data only 3 rules are generated.

	Unreduced		UFRFS Reduced	
Dataset	JRIP	PART	JRIP	PART
water 2-class	4	6	1	2
water 3-class	3	7	3	3

Table 7.3: Maximum rule arity for unreduced and UFRFS reduced data

It is important to compare not only the rule arity but also the number of rules generated to assess the impact of employing a dimensionality reduction step such as UFRFS. The numbers of rules generated for each RIA are shown in Table 7.4. The results show that UFRFS reduced data generates fewer rules than the

unreduced data, with the exception of PART for the 3-class data which results in 16 rules as opposed to 13 rules for the unreduced data. However, if the corresponding arity value is examined, it can be see that although the UFRFS reduced data results in the generation of more rules, they are less complex than those of the unreduced data.

	Unreduced		UFRFS Reduced	
Dataset	JRIP	PART	JRIP	PART
water 2-class	3	12	2	2
water 3-class	7	13	3	16

Table 7.4: Number of rules generated for unreduced and UFRFS reduced data

These results show that UFRFS is useful not only in removing redundant feature measures but also in dealing with the noise associated with such measurements. To demonstrate that the resulting rules are comprehensible, two sets of rules produced by the induction mechanism JRIP for the 3-class data are given in Fig. 7.3. The rules produced are reasonably short and understandable. However, when semantics-destroying dimensionality reduction techniques are applied, such interpretability is lost.

7.4.3 Comparison with Other FS Techniques

The above comparison ensured that little information loss is incurred due to UFRFS. The question now is whether any other feature selection approaches would perform similarly. To answer to this, a number of other methods are examined here in order to see what classification results and subset sizes might be achieved.

The FS approaches employed, in contrast to UFRFS are supervised and can take advantage of the class information. It may seem to disadvantage UFRFS to be compared with such approaches but the results demonstrate the utility of UFRFS and its ability to select useful features.

Three different FS algorithms have been chosen for comparison: CFS [68], Consistency-based feature selection [41], and a C4.5 wrapper [184]. The reduced data is classified using the same classifier learners that were used in the last section - QSBA [188] and FRNN [141].

The results for the subset size returned by the 3 approaches are shown in Table 7.5. Both Consistency-based and CFS return subsets which are significantly larger than those obtained by UFRFS. But, the C4.5 wrapper obtains very small subsets. However, the corresponding classification results (Table 7.6)show that this method fails to find informative features and the results for UFRFS are considerably better.

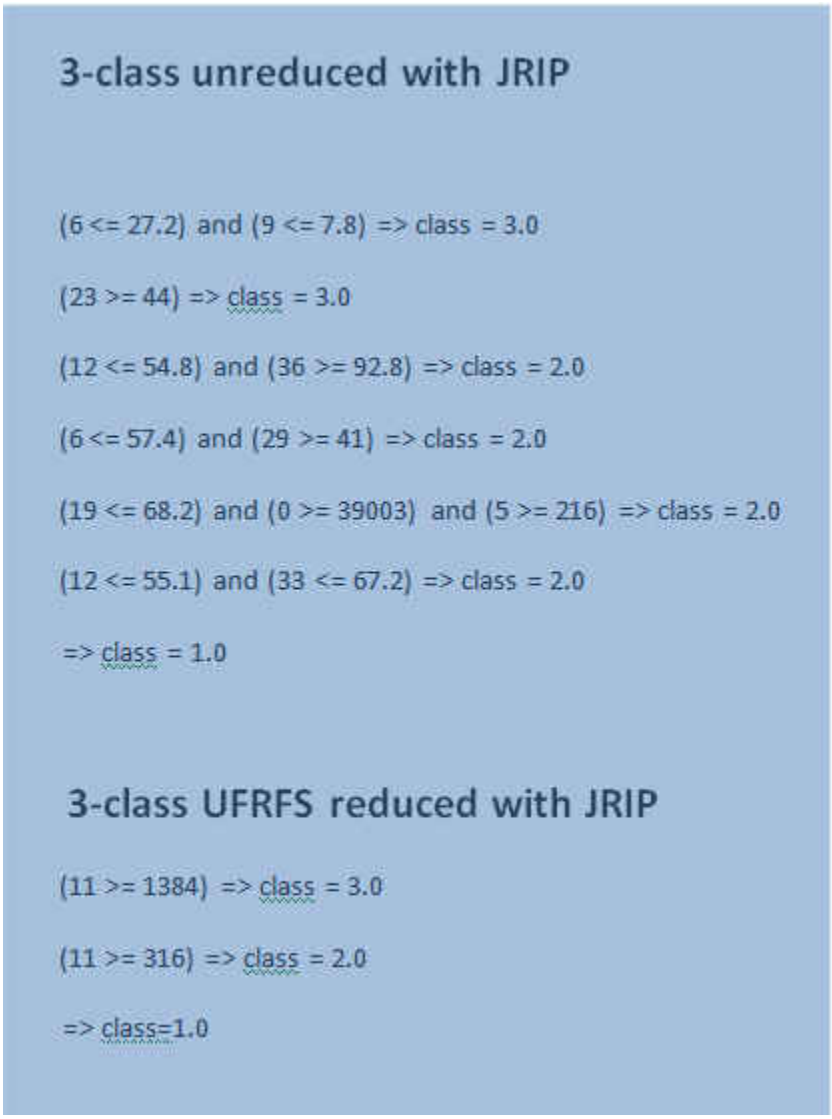

Figure 7.3: Example rules generated from the 3-class data: unreduced and reduced

Dataset	Consistency	CFS	C4.5 wrapper
water 2-class	14	10	3
water 3-class	12	12	4

Table 7.5: Subset sizes returned by consistency-based, CFS, and C4.5 wrapper, feature selection methods

When comparing UFRFS with the results obtained in Table 7.6, it can be seen that, Consistency-based and CFS methods outperform UFRFS although the differences are once again small (less than 10%). However, the corresponding subset sizes obtained for both of these approaches are much larger than UFRFS, as demonstrated in Table 7.5. Additionally, UFRFS outperforms the C4.5 wrapper approach easily, demonstrating that useful features are retained by this method. It is worth emphasising once again that all 3 methods are supervised and can take

	Consistency		CFS		C4.5 Wrapper	
Dataset	QSBA	FRNN	QSBA	FRNN	QSBA	FRNN
water 2-class	85.12	85.64	85.38	84.35	71.02	44.36
water 3-class	80.00	82.01	80.32	85.12	74.23	78.33

Table 7.6: Subset size and classification accuracy results for consistency-based FS

advantage of the class information, UFRFS cannot.

The most important discovery from the experimentation in this work, is that it is possible to identify informative features for plant monitoring using UFRFS. Indeed, this unsupervised approach even outperforms the C4.5 wrapper method in selecting features.

7.5 Summary

Automated generation of feature pattern-based if-then rules is essential to the success of many intelligent pattern classifiers, especially when their inference results are expected to be directly human-comprehensible. This work has presented such an approach which integrates rule induction algorithms with a fuzzy-rough method for unsupervised feature selection. Unlike semantics-destroying approaches such as PCA, this approach maintains the underlying semantics of the feature set, thereby ensuring that the resulting models are interpretable and the inference explainable. The rules are simplified by the use of UFRFS, and the resulting classification accuracies are comparable to the unreduced data. This method alleviates important problems encountered by traditional RSFS such as dealing with noise and real-valued features.

In all experimental studies there has been no attempt to optimize the fuzzifications or the classifiers employed. It can be expected that the results obtained with optimization would be even better than those already observed. The generality of this approach should enable it to be applied to other domains. The ruleset generated by the RIAs were not processed by any post-processing tools so as to allow its behaviour and capabilities to be revealed fully. By enhancing the induced ruleset through post-processing, performance should also improve.

Chapter 8

Conclusion

This chapter concludes the thesis. A summary of the research is presented, with a focus on the main contribution: exploiting the rough set boundary region information for feature selection, and its applications. Examining the existing literature, it was demonstrated that many rough set techniques and their extensions rely on the information in the lower approximation to perform feature selection.

Although the classical rough set-based approaches may only operate on crisp datasets, the introduction of the new boundary measure employs a distance metric allows the importance of uncertain data objects to be assessed for the task of feature selection. This measure is based on a unary fuzzy set and has shown that there is much useful information to be extracted from the boundary region. Additionally, it has been demonstrated that the boundary region of fuzzy-rough sets can offer useful information for guiding the feature selection process. Three different techniques in this area have been presented in this thesis; in Section 3.2, Section 4.2 and Section 4.4.1.

Other additional work that has been carried out includes a comprehensive review of rough sets and some of their applications [145]. This review allows an in-depth view of existing techniques as well as the most recent developments and hybridisations of rough sets with other approaches. It also offers some suggestions and identifies areas for further exploration. The application of fuzzy and rough techniques for the classification of mammographic risk analysis data represents a new area of application for the techniques proposed earlier in this thesis. The fuzzy-rough nearest neighbour classification algorithm (FRNN) employed is described in Section 5.2.2 and takes advantage of the important fuzzy upper and lower approximation concepts. The improved performance as a result of utilising these rough and fuzzy-rough techniques is then demonstrated in Section 6.2.1.

Another new method for unsupervised feature selection based on fuzzy-rough sets (UFRFS) is proposed and described in detail in Section 5.3.1. This method examines the interdependency of features using fuzzy-rough sets. The selection

takes place by comparing single features to subsets of features and using the fuzzy dependency measure to eliminate redundant features. This is a useful technique to employ for feature selection when the decision feature(s) are missing or unknown. This is demonstrated in Section 7.3, by application to a real-world plant monitoring problem where it is shown that UFRFS retains information-rich features.

8.1 Distance Measure Assisted Rough Set Feature Selection

This thesis has been concerned primarily with the utility of distance measure assisted rough set feature selection as a way of utilising the uncertain information in the boundary region of rough sets. The approach achieves this by examining the uncertain objects, and quantifying them in terms of their proximity from the lower approximation concept. This method results in better performance when choosing which features to select. This is done without the need to modify the underlying rough set mathematical model.

The approach was also extended to handle real-valued data through the use of tolerance rough sets and fuzzy-rough sets and proved in these cases also, that there is much information to be extracted from the boundary region. This information has in the past largely been ignored as the certainty that is embodied in the lower approximation is associated with greater importance in scientific analysis.

The DMRSAR and DM-TRS methods have proven that the uncertain information in the boundary region can also be used to guide the feature selection process more effectively than through the use of the dependency measure alone. This is reflected in the results obtained through experimental evaluation, and the real-world application of mammographic data analysis.

The new approaches have been evaluated experimentally by comparing them with other state-of-the-art dimensionality reduction/FS techniques such as PCA [48], Consistency-based FS [41], CFS [68], ReliefF [109], and a J48-based wrapper [184].

In the case of the fuzzy-rough set boundary region, not only has the work in this thesis attempted to use this uncertain information but some new interpretations of existing measures have also been proposed such as the boundary entropy measure, and boundary region reduction measure. These measures have demonstrated that they are useful for finding fuzzy-rough reducts from the information of the fuzzy boundary region as well as the traditional use of the lower approximation. This is supported by the experimentation documented in Section 4.5, where existing methods which only use the lower approximation for the task of feature selection

are compared with those which utilise the information of the boundary region over a number of benchmark datasets.

8.2 Unsupervised Fuzzy-Rough Feature Selection

In unsupervised learning decision class labels are not provided. This poses some relevant questions, such as: which features should be retained? and, why not use all of the information? The problem is that not all features are important. Some of the features may be redundant, and others may be irrelevant and noisy. In the work described in this thesis, a new fuzzy-rough set-based unsupervised feature selection approach (UFRFS) was proposed. The approach does not require thresholding information, and uses the fuzzy-rough indiscernibility measure to select features, which results in a significant reduction in dimensionality whilst retaining the data semantics. This new approach was compared with an advanced supervised feature selection technique. UFRFS returned similar results to the supervised method both in terms of subset sizes and classification accuracy, despite the absence of class labels.

8.3 Mammographic Image Analysis

The knowledge discovery obstacle is a significant problem that impedes the development of intelligent systems for mammographic data analysis. The generation of accurate classifier learning techniques for this task is extremely difficult. This is particularly true where expert opinion differs. Machine learning techniques are of great benefit to this area by providing strategies to automatically extract useful knowledge, given sufficient historical data.

For many techniques, the high dimensionality of the domain attributes makes many of the problems computationally intractable. In addition, when applying classifier learners that can cope with this size of data, the resulting knowledge may be of poor quality. A semantics-preserving dimensionality reduction step is required to alleviate this problem, such that the resulting subset is interpretable by humans. This is essential for medical applications.

DM-TRS was applied to this domain for mammographic risk analysis to show that classification can be improved significantly with feature selection, whilst reducing the dimensionality by over 95%. The identification of only those important features from such vast data means that the process of extracting large amounts of features can be avoided. The selected features can then be fed back into the extraction phase ensuring that only those features need to be identified in future.

The benefits of adopting such an approach include faster identification of relevant features, thus reducing the amount of time and computational effort required in the feature extraction phase. For the end user (the patient) this means more swift and accurate diagnosis, and less screening. The DM-TRS method was shown to perform very well against other feature selection methods for this task. Additionally, the FRNN classification technique was applied and demonstrated improved performance over existing classifiers.

8.4 Industrial Plant Monitoring

Complex domain application problems, such as the reliable monitoring and diagnosis of industrial processes, are likely to result in large numbers of features, many of which are redundant for the task at hand. Also, inaccurate and/or noisy values cannot be ruled out. Such applications usually require convincing explanations about the inference performed, therefore a method which enables the automated generation of knowledge models of clear semantics is highly desirable.

In previous work [202], [204], supervised feature selection was used for the removal of noisy and redundant features. Here, UFRFS has been employed. The use of an unsupervised method demonstrates how redundancy can be removed from large feature sets which have incomplete or missing class labels without affecting those valuable or information-rich features. The rules induced from the UFRFS reduced data are shown to be less complex, and more easily interpreted, than those induced from the unreduced data. Additionally, the classification accuracies, (despite the lack of class label information) are shown to be comparable to the unreduced data.

8.5 Future Work

There is much scope for future work for all of the developments presented in this thesis. A summary of the main points of this can be divided into two categories: work that could be completed if additional time was available, and longer-term work that would involve considerable research effort.

8.5.1 Short-term Developments

Amongst the topics in the first category are ideas which relate to the DMRSAR/DM-TRS feature selection approaches. Further development and re-evaluation of how the mean lower approximation is calculated, may prove beneficial. Implementation of a more accurate calculation of the lower approximation boundary would

mean that distances of objects in the boundary region could be more accurately measured.

Additionally the significance measure which is employed for DMRSAR is rudimentary, and considers the boundary region as a single value which is expressed as membership value of a unary fuzzy set. By redefining this as a number of fuzzy sets, the boundary region could be quantified more accurately by expressing object membership in terms of weights in the boundary in relation to distance from the lower approximation. Apart from the use of extra fuzzy sets, the way in which objects in the boundary are related is another area which is worthy of investigation. By examining the correlation of objects and their individual distances, it may be possible to provide more accurate information on individual objects and the extent to which they belong to the concept under consideration. The use of this information for other areas e.g. classification may result in improved performance.

Other aspects which warrant investigation include the distance metric and also the application area of the approach. For the worked examples described in this thesis a Euclidean distance metric is employed. Other metrics such as Mahalanobis distance [132], fuzzy Hausdorff distance [27], and others could also be considered. Other measures such as entropy or information gain ratio could be employed in place of rough dependency, these have proven useful for the classical rough set case [93], [98] and should also yield similar results when combined with the distance metric. Additionally, the distance-based rough set approach could be extended to other application areas such as clustering or unsupervised FS with some further development. Most rough classification techniques like FS, currently rely on the certain information of the lower approximation, where the boundary is considered the distances of individual objects to the lower approximation are not taken into account. This is where the distance measure assisted approach could prove useful and offer additional information for guiding the classification of test objects. Similarly, for the task of clustering the distance measure could be used to assess the distance of individual unclustered objects from the lower approximations of existing cluster prototypes.

The unsupervised fuzzy-rough feature selection method is another approach that could be extended significantly. There are a considerable number of unsupervised FS approaches which investigate the validity of the generated clusters [42], [89], [168]. However, this aspect of UFRFS has not been investigated as yet. Given the positive results obtained during experimentation and the additional tentative work in this area, it is expected that this method would demonstrate an improvement in performance over existing methods. The complimentary nature of fuzzy-rough sets enables the UFRFS approach to consider both the vagueness and indiscernibility of real-valued data. This is an aspect that is lacking in other

approaches.

The FRNN algorithm [92], although proposed for classification could be equally applicable to FS by implementing a nearest-neighbour search when selecting features. Further developments of this idea could extend to the use of the fuzzy-boundary region as well as the lower approximation to select features as demonstrated in [142].

In addition There are also a wealth of application areas where the approaches described in this thesis could be applied such as crime investigation [245] where feature selection could be used to identify hidden patterns in data, or for image scene analysis [49] to identify important features which may be related to the specific patterns of human behaviour.

8.5.2 Long-term Developments

This section describes future developments that could form the basis either individually or collectively for a large project (e.g. PhD project or group project).

As reported previously the areas of FS, classification and clustering are all closely related. Given that fuzzy-rough sets have been employed with much success for supervised learning, the development of fuzzy-rough clustering approaches would represent a further step to addressing the need for additional work on unsupervised learning. Clustering is essentially unsupervised classification, and the initial work on supervised learning, and unsupervised classification described in this thesis form a useful starting point for such further investigation.

In Chapter 6 the application of fuzzy-rough methods to the field of mammographic data analysis highlights the need for a fuzzy-rough approach to feature extraction/clustering. At present, a fuzzy c-means (FCM) [13] algorithm is employed for image segmentation and feature extraction in the approach described in Section 6.1. The development of a new fuzzy-rough feature extraction/clustering technique would unify the underlying mathematical model for the approach demonstrated initially in this thesis.

FCM requires the subjective specification of a number of parameters including a 'neighbourhood' value, a 'fuzziness' value, and a 'termination threshold' value. By utilising a fuzzy-rough method, the number of these parameters could be reduced to the specification of a single value for the number of clusters. Additionally, only the information contained in the data would be required for clustering rendering the specification of fuzziness and a termination threshold obsolete. This would perhaps extend the previously suggested work in the area of clustering.

By enhancing the algorithm with a rough extension in the same way as the fuzzy nearest neighbour algorithm for instance, any new method could take advan-

tage of the benefits that a fuzzy-rough hybrid has offered in other areas. FS and classification are good examples of this. This is due to the fact that fuzzy-rough sets are better equipped to deal with the uncertainty and vagueness present in real-world data. In particular, looking closely at the fuzzy-rough approach outlined in [99] and also in Sections 5.2.2 and 5.3.1, it can be seen that the definition for the fuzzy lower and upper approximation concepts lend themselves well to unsupervised learning. Examining the definitions below, in informal terms the fragment $\mu_{R_P}(x, y)$ represents the fuzzy similarity of two objects x and y, or the extent to which two distinct objects are similar. Also, $\mu_X(y)$ is the extent that y belongs to the concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_X(y)) \tag{8.1}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x, y), \mu_X(y)) \tag{8.2}$$

These definitions could be extended to the non-hierarchical c-means clustering framework by replacing the set X with the cluster n, where n in the number of predefined cluster concepts. The choice of fuzzy similarity measures and connectives would also provide much scope for further experimentation. In addition to this, given that FRNN has been successful for supervised classification a similar approach could be adopted for agglomerative clustering where a nearest-neighbour type search strategy could be employed. This strategy would involve the use of the fuzzy lower and upper approximation concepts as a measure to agglomerate or merge clusters as the algorithm progresses. Given that the number of clusters is not required in the first instance, this method would not require any subjective parameter specification and could be data-driven exclusively.

Another particular area worthy of investigation for FS in general, and in particular for the methods proposed in this thesis is that of search methods. Approaches such as particle swarm optimisation (PSO) [233], and artificial immune systems (AIS) [58] although not optimal or complete do provide useful areas for exploration as an alternative to more conventional search methods for the task of feature selection. Moreover, the utility of the boolean propositional satisfiability (SAT) technique [43] to search for feature subsets offers much scope for development. This search method is complete. Although the SAT problem is $NP - complete$ [34], in practice the technique is computationally efficient, does not involve an exhaustive examination of the search space, and can guarantee the minimality of discovered solutions. Indeed, SAT has been employed for FS for the discovery of rough set reducts with much success [98], but its use for fuzzy-rough set FS has yet to be investigated. It is felt that this would offer a significant performance increase

in the area of (supervised and unsupervised) fuzzy-rough set FS - eliminating the need to search for reducts down sub-optimal paths of the search space. Moreover, the minimality of the reducts discovered can be guaranteed, which ensures that only the most compact reduct or subset will result.

To conclude, it is appropriate to emphasise the value of the uncertain information contained within the boundary region of rough sets and rough set extensions and the role it can play in improving the performance of the existing methods. Given the ever-growing availability of information (and the size of data in general), the performance of tools for knowledge discovery such as FS, classification, and clustering are becoming more and more important. The series of investigations and experimentation documented in this thesis, demonstrate the potential utility of employing the boundary region information of rough and fuzzy-rough sets for the task of feature selection. Those issues which affect the use of rough sets and their extensions for feature selection and areas which require further research have been highlighted and discussed.

Appendix A

Publications Arising from the Work in this Thesis

A.1 Published or Accepted for Publication

All publications presented in chronological order.

1. N. Mac Parthaláin, R. Jensen and Q. Shen. Fuzzy Entropy-assisted Fuzzy-Rough Feature Selection. Proceedings of the 15th International Conference on Fuzzy Systems (FUZZ-IEEE'06), 2006.

2. N. Mac Parthaláin, R. Jensen, and Q. Shen. Comparing Fuzzy-Rough and Fuzzy Entropy-assisted Fuzzy-Rough Feature Selection. Proceedings of the 6th Annual Workshop on Computational Intelligence (UKCI'06), 2006.

3. N. Mac Parthaláin, Q. Shen, and R. Jensen. Distance Measure Assisted Rough Set Feature Selection. Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE'07), pp. 1084–1089. 2007.

4. N. Mac Parthaláin, R. Jensen, and Q. Shen. Finding Fuzzy-Rough Reducts with Fuzzy Entropy. Proceedings of the 2008 IEEE Conference on Fuzzy Systems, (FUZZ-IEEE'08), 2008.

5. N. Mac Parthaláin, R. Jensen and Q. Shen. Rough and Fuzzy-Rough Methods for Mammographic Data Analysis. Proceedings of the 8th Annual UK Workshop on Computational Intelligence (UKCI'08), 2008.

6. N. Mac Parthaláin, and Q. Shen, Exploring the Boundary Region of Tolerance Rough Sets for Feature Selection, *Pattern Recognition*, vol. 42, no. 5, pp. 655–667, 2009.

7. N. Mac Parthaláin, Q. Shen and R. Jensen. A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction. To appear in *IEEE Transactions on Knowledge and Data Engineering*.

8. N. Mac Parthaláin, and Q. Shen, On Rough Sets, their Recent Extensions and Applications, To appear in *Knowledge Engineering Review*.

9. N. Mac Parthaláin, R. Jensen, Q. Shen, and R. Zwiggelaar. Rough and Fuzzy-Rough Methods for Mammographic Data Analysis. To appear in *Intelligent Data Analysis*.

10. N. Mac Parthaláin, and R. Jensen. Measures for Unsupervised Fuzzy-Rough Feature Selection. To appear in Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA'09).

A.2 Currently under Review

1. N. Mac Parthaláin, and Q. Shen, A Fuzzy-Rough Approach to Unsupervised Feature Selection, Under review for potential journal publication.

2. N. Mac Parthaláin, and Q. Shen, Unsupervised Fuzzy-Rough Feature Selection and its Application for Complex Systems Monitoring, Under review for potential journal publication.

Appendix B

Glossary of Terms

A list of commonly used terms and acronyms in the text of this thesis are described below.

DMRSAR (Distance Measure Rough Set Attribute Reduction)
DMRSAR [148], [149] is a feature selection technique based on rough sets. It attempts to qualify the objects in the boundary region of rough set theory with regard to their proximity to the lower approximation. From an intuitive point-of-view, the closer the proximity of an object in the boundary region to objects of the lower approximation, the greater the likelihood that it actually belongs to the set of interest.

DM-TRS (Distance Metric-Assisted Tolerance Rough Set Feature Selection)
DM-TRS [144] is an extension of the TRSM approach which has the ability to operate on real-valued data. It marries the TRSM with the distance metric assisted rough set approaches. This allows the information of the TRSM boundary region that is otherwise ignored to be examined and used for FS.

DR (Dimensionality Reduction)
DR is the reduction of the data to a size which is computationally tractable, without information loss. It is usually included as part of a data preprocessing system.

DRSA (Dominance-based Rough Set Approach) The DRSA [62] is an extension of RST for multi-criteria decision analysis. In contrast to traditional RST, DRSA employs a dominance relation instead of an equivalence relation. This allows DRSA to deal with the inconsistencies which are typical of criteria and preference-ordered decision classes.

Feature
A feature, (may also be known as a variable or attribute) is a single dimension of a data object. A data object may have many features which are used to describe it. For example the object *car* may have the features: 4, 2, and red to describe the number of wheels, doors, and colour.

FS (Feature Selection)
Feature selection [130] is a commonly used approach in machine learning (may also be known as feature subset selection, variable selection, or attribute reduction) and can be considered as the process of selecting the input attributes of a dataset that most closely define a particular outcome.

Fuzzy Set Theory
Fuzzy sets [254] are sets whose elements have degrees of membership. Fuzzy set theory was introduced by Lotfi A. Zadeh in 1965 as an extension of the classical set. In traditional set theory, the membership of elements in a set is defined in binary terms according to a hard condition an element either belongs to the set or an element does not belong to the set. In contrast, fuzzy set theory allows gradual membership of elements in a set; this is described by employing a membership function in the real unit interval [0, 1].

PCA (Principal Component Analysis)
PCA [48] is a versatile transformation-based DR technique which projects the existing data onto a new coordinate system of reduced dimensions. This process of linear transformation however also transforms the underlying semantics or meaning of the data. This results in data that is difficult for humans to interpret, but which may still provide useful automatic classification of new data.

Reduct
A reduct C is a subset of original attributes or features of a dataset which provides an equivalent predictive characteristic (γ) as the complete set of conditional features (R) with regard to the decision feature (D). A reduct can be defined as a subset of the conditional attribute set such that $\gamma R(D) = \gamma C(D)$.

RST (Rough Set Theory)
Rough set theory proposed by Pawlak [172] is a tool used to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information. RST is the formal approximation

of a crisp or classical set in terms of a pair of sets which give the lower and the upper approximation of the original set. The lower approximation contains those objects which definately belong to the concept (set) of interest, while the upper approximation contains those objects which definately belong to the subset of interest *and* those objects which possibly belong to the subset of interest.

Supervised Learning
Supervised learning is a technique in machine learning for deducing a function from training data. The training data usually consist of pairs of input objects (typically vectors), along with desired outputs. The output of the function can predict a class label of the input object - this is known as classification. The task of the supervised learning mechanism is to predict the value of the function for any input object having 'seen' a number of training examples - i.e. pairs for input and target output.

TRSM (Tolerance Rough Set Model)
TRSM [208] employs a similarity relation to minimise data as opposed to the indiscernibility relation used in classical rough-sets. This allows a relaxation in the way equivalence classes are considered. The effect of employing this relaxation, means that the granularity of the rough equivalence classes has been blurred slightly. This flexibility enables a change to occur in the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one so-called tolerance class which is TRSM equivalent of a rough set equivalence class.

UFRFS (Unsupervised Fuzzy-Rough Feature Selection)
A feature selection technique based upon fuzzy-rough set theory that minimises the feature set based on the interdependencies between sets of features [138], [146].

Unsupervised Learning
Unsupervised learning is the task of attempting to determine how data is organised. It can be differentiated from supervised learning because the learning mechanism uses only unlabeled data objects. Unsupervised learning is closely related to the problem of density estimation in statistics. However, unsupervised learning also encompasses many other techniques that also attempt to summarise and explain key aspects of the data. One particular example of unsupervised learning is the task of clustering.

VPRS (Variable Precision Rough Sets)
The variable precision rough sets (VPRS) approach [263] extends rough set theory

by relaxing the subset operator. It was originally proposed in order to analyse and identify data patterns which represent statistical trends rather than those which are functional. At the heart of VPRS, is the idea of allowing objects to be classified with an error smaller than a given predefined level or threshold.

Bibliography

[1] American College of Radiology. Illustrated Breast Imaging Reporting and Data System BIRADS. 3rd Edition, American College of Radiology, 1998.

[2] F. Aghdasi, R.K. Ward, J. Morgan-Parkes, and B. Palcic. Feature selection for classification of mammographic microcalcification clusters. Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 58–59, 1993.

[3] H. Almuallim, and T.G. Dietterich. Learning Boolean Concepts in the Prescence of Many Irrelevant Features. Artificial Intelligence, vol. 69, no. 1 & 2, pp. 279–305, 1994.

[4] S. Asharaf and M.N. Murty. An adaptive rough fuzzy single pass algorithm for clustering large data sets, Pattern Recognition, vol. 36, no. 12, pp. 3015–3018, 2004.

[5] C. Armanino, R. Leardi, S. Lanteri, and G. Modi. Chemom. Intell. Lab.Syst., vol. 5, pp. 343–354, 1989.

[6] W.H. Au and K.C.C. Chan. An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases. Proceedings of the 7th IEEE International Conference on Fuzzy Systems, pp. 1314–1319, 1998.

[7] Z. Bao, B. Han, and S. Wu. A novel clustering algorithm based on variable precision rough-fuzzy sets, Proceedings of the International conference on intelligent computing, ICIC 2006, Kunming, China, August 16-19, pp. 284-289, 2006.

[8] J. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, and J. Wroblewski. Rough Set Algorithms in Classification Problem. In L. Polkowski, S. Tsumoto and T.Y. Lin (eds.): Rough Set Methods and Applications. Physica-Verlag, Heidelberg, New York, pp. 49–88, 2000.

[9] D.A. Bell and J.W. Guan. Computational Methods for Rough Classification and discovery, Journal of the American Society for Information Science, vol. 5, pp. 403–414, 1998.

[10] R.E. Bellman. Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.

[11] M.J. Beynon. An investigation of β-reduct selection within the variable precision rough sets model, In Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing (RSCTC 2000), pp. 114–122, 2000.

[12] M.J. Beynon. Reducts within the Variable Precision Rough Sets Model: A Further Investigation. European Journal of Operational Research, vol. 134, no. 3, pp. 592–605, 2001.

[13] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[14] R.B. Bhatt and M. Gopal. FRID: Fuzzy-Rough Interactive Dichotomizers, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE‘ 04), pp. 1337–1342, 2004.

[15] H. Bian and L. Mazlack. Fuzzy-Rough Nearest-Neighbor Classification Approach, Proceeding of the 22nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS), pp. 500–505, 2003.

[16] R.L. Birdwell, D.M Ikeda, K.F. O'Shaughnessy, and E.A Sickles. Mammographic Characteristics of 115 Missed Cancers Later Detected with Screening Mammography and the Potential Utility of Computer-Aided Detection, Radiology, vol. 219, pp. 192–202, 2001.

[17] D. Boixader, J. Jacas, and J. Recasens, Upper and lower approximations of fuzzy sets, International Journal of General Systems, vol. 29, no. 4, pp. 555–568, 2000.

[18] M. Brady, and R. Highnam. Mammographic Image Analysis, Kluwer Series on Medical Image Understanding, 1999.

[19] G. Brassard, and P. Bratley. Fundamentals of Algorithms. New Jersey, Prentice Hall, 1996.

[20] F. Bray, P. McCarron, and D.M. Parkin. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Research, vol. 6, pp. 229–239, 2004.

[21] C. Bregler and S.M. Omoundro. Nonlinear image interpolation using manifold learning. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), Advances in Neural Information Processing Systems 7, pp. 973–980, 1995.

[22] C. Browne, I. Düntsch, G. Gediga. IRIS revisited, A comparison of discriminant enhanced rough set data analysis. Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Software Systems. Physica-Verlag, Polkowski and Skowron (eds.), pp. 347–370, 1998.

[23] S. Buseman, J. Mouchawar, N. Calonge, and T. Byers. Mammography screening matters for young women with breast carcinoma. Cancer, vol. 97, no. 2, pp. 352–358, 2003.

[24] M.A. Carreira-Perpinñán. Continuous latent variable models for dimensionality reduction and sequential data reconstruction. PhD thesis, University of Sheffield, UK, 2001.

[25] R. Caruna, and D. Freitag. Greedy Attribute Selection. Proceedings of the 11th International Conference on Machine Learning, pp. 28–36, 1994.

[26] K. Chan and A. Wong. APACS: A System for Automatic Analysis and Classification of Conceptual Patterns. Computational Intelligence, vol. 6, pp. 119–131, 1990.

[27] B.B Chaudhuri and A. Rosenfeld, A modified Hausdorff distance between fuzzy sets, Information Sciences, vol. 118, no. 1–4, pp. 159–171, 1999.

[28] D. Chen, W.X. Zhang, D. Yeung, and E.C.C. Tsang, Rough approximations on a complete completely distributive lattice with applications to generalized rough sets, Information Sciences, vol. 176, no. 13, pp. 1829–1848, 2006.

[29] T-H. Cheng, C-P. Wei, and V.S. Tseng. Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. 19th IEEE International Symposium on Computer-Based Medical Systems, pp.165–170, 2006.

[30] S. Chimphlee, N. Salim, M.S.B. Ngadiman, W. Chimphlee, and S. Srinoy. Independent Component Analysis and Rough Fuzzy based Approach to Web Usage Mining, Proceedings Artificial Intelligence and Applications, 2006.

[31] W Chimphlee, A. H. Abdullah, M. N. M. Sap, S. Srinoy, S. Chimphlee. Anomaly-Based Intrusion Detection using Fuzzy Rough Clustering, International Conference on Hybrid Information Technology (ICHIT'06), vol. 1, pp.329–334, 2006.

[32] A. Chouchoulas and Q. Shen, Rough set-aided keyword reduction for text categorisation, Applied Artificial Intelligence, vol. 15, no. 9, pp. 843–873, 2001.

[33] W.W. Cohen. Fast effective rule induction. In Machine Learning: Proceedings of the 12th International Conference, pp. 115–123, 1995.

[34] S. Cook. The complexity of theorem proving procedures, Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151158, 1971.

[35] O. Cordon, F. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena. Ten years of genetic fuzzy systems: current framework and new trends, Fuzzy Sets and Systems, vol. 141, pp. 5–31, 2001.

[36] C. Cornelis, M. De Cock and A. Radzikowska. Vaguely Quantified Rough Sets, Proc. of the 11th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC2007), Lecture Notes in Artificial Intelligence vol. 4482, pp. 87–94, 2007.

[37] S.K. Das. Feature Selection with a Linear Dependence Measure, IEEE transactions on Computers, pp. 1106–1109, 1971.

[38] S. Das. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. Proceedings of the 18th International Conference on Machine Learning, pp. 74–81, 2001.

[39] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature Selection for Clustering - a Filter Solution. Proceedings of the second international conference on Data Mining, pp. 115–122, 2002.

[40] M. Dash and H. Liu. Feature Selection for Classification. Intelligent Data Analysis, vol. 1, no. 3, pp. 131–156, 1997.

[41] M. Dash, H. Liu, and H. Motoda, Consistency Based Feature Selection, Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 98-109, 2000.

[42] M. Dash, and H. Liu, Unsupervised Feature Selection, Proceedings of the Pacific and Asia Conference on Knowledge Discovery and Data Mining, pp. 110–121, 2000.

[43] M. Davis, G. Logemann, and D. Loveland. A machine program for theorem proving. Communications of the ACM. vol. 5, pp. 394–397, 1962.

[44] M. De Cock, C. Cornelis, and E.E. Kerre, Fuzzy rough sets: beyond the obvious, IEEE International Conference on Fuzzy Systems, vol. 1, pp. 103–108, 2004.

[45] De Cock, M., Cornelis, C. and Kerre, E.E., 2007. Fuzzy Rough Sets: The Forgotten Step, IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp. 121–130, 2007.

[46] J. S. Deogun, V. V. Raghavan, and H. Sever. Rough set based classification methods and extended decision tables, Proceedings of the International Workshop on Rough Sets and Soft Computing, (San Jose, California), pp. 302–309, 1994.

[47] J.S. Deogun, V.V. Raghavan, and H. Sever, Exploiting upper approximations in the rough set methodology, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Quebec, Canada, pp. 1–10, 1995.

[48] P. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach, Prentice Hall, 1982.

[49] V. Devendran, A. K. Hemalatha Thiagarajan, Santra, and Amitabh Wahi, Feature Selection for Scene Categorization Using Support Vector Machines, Congress on Image and Signal Processing (CISP), vol. 1, pp. 588–592, 2008.

[50] J. Doak. An Evaluation of Feature Selection Methods and Their Application to Computer Security. technical report, University of California at Davis, Dept. of Computer Science, 1992.

[51] D. Dubois and H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems, vol. 17, pp. 191–209, 1990.

[52] D. Dubois and H. Prade, Putting Rough Sets and Fuzzy Sets Together, Intelligent Decision Support, pp. 203–232, 1992.

[53] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification, 2nd ed. John Wiley and Sons, New York, 2001.

[54] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics, vol. 3, pp. 32–57, 1973.

[55] J.G. Dy, and C.E. Brodley. Feature Subset Selection and Order Identification for Unsupervised Learning. Proceedings of the 17th international Conference on Machine Learning, pp. 247–254, 2000.

[56] Eurostat. Health statistics atlas on mortality in the European Union. Official Journal of the European Union, 2002.

[57] B.S. Everitt. An Introduction to Latent Variable Models, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1984.

[58] J.D. Farmer, N. Packard and A. Perelson, The immune system, adaptation and machine learning, Physica D, vol. 2, pp. 187–204, 1986.

[59] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases, AI Magazine, vol. 17, pp 37–54, 1996.

[60] D. Gering. Linear and nonlinear data dimensionality reduction. Technical report, Massachusetts Institute of Technology, 2002.

[61] M. Glymin and W. Ziarko. Rough set approach to spam filter learning, Proceedings of Rough Sets and Emerging Intelligent Systems Paradigms, RSEISP'07, Lecture Notes in Artificial Intelligence vol. 4585, pp. 350–359, 2007.

[62] S. Greco, B. Matarazzo, and R. Slowiński. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, vol. 129, no.1 pp. 1–47, 2001.

[63] D. M. Grzymala-Busse, and J. W. Grzymala-Busse. The usefulness of machine learning approach to knowledge acquisition, Computational Intelligence, vol. 11, pp. 268–279, 1995.

[64] J.W. Grzymala-Busse, and C.P.B. Wang. Classification methods in rule Induction, Proceedings of the 5th Intelligent Information Systems Workshop, pp. 120–126, 1996.

[65] J.W. Grzymala-Busse. A Comparison of Three Strategies to Rule Induction from Data with Numerical Attributes, Proceedings of the International Workshop on Rough Sets in Knowledge Discovery (RSKD 2003), pp. 132–140, 2003.

[66] J. W. Grzymala-Busse. Rough Set Theory with Applications to Data Mining, Chapter in M. Negoita, B. Reusch (Eds.), Real World Applications of Computational Intelligence, Studies in Fuzziness and Soft Computing series, vol. 179, pp. 223–244, 2006.

[67] I. Guyon, and A. Elisseeff. An Introduction to Variable and Feature Selection. Journal of Machine Learning, vol. 3, pp. 1157–1182, 2003.

[68] M.A. Hall, Correlation-based feature selection machine learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[69] J. Han, X. Hu, T.Y.Lin. Feature Subset Selection Based on Relative Dependency between Attributes. Rough Sets and Current Trends in Computing: 4th International Conference, RSCTC 2004, Uppsala, Sweden, June 1-5, pp. 176–185, 2005.

[70] A. Hassanien. Rough set approach for attribute reduction and rule generation: a case of patients with suspected breast cancer. Jour. Am. Soc. Inf. Sci. Technol. vol. 55, no. 11, pp. 954–962, 2004.

[71] A. Hassanien. Fuzzy rough sets hybrid scheme for breast cancer detection. Image and Vision Computing. vol. 25, no.2, pp. 172-183, 2007.

[72] I. Hayashi, T. Maeda, A. Bastian and L.C. Jain. Generation of Fuzzy Decision Trees by Fuzzy ID3 with Adjusting Mechanism of AND/OR Operators, Proceedings of the 7th IEEE International Conference on Fuzzy Systems, pp. 681–685, 1998.

[73] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistcal Learning: Data Mining, Inference, and Prediction, Springer, 2001.

[74] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P.J. Kegelmeyer. The Digital Database for Screening Mammography. Proceedings of the International Workshop on Digital Mammography, pp. 212–218, 2000.

[75] A. Hedar, J. Wang and M. Fukushima, Tabu search for attribute reduction in rough set theory, Technical Report 2006-008, Department of Applied Mathematics and Physics, Kyoto University, 2006

[76] S. Hirano and S. Tsumoto. Rough Clustering and its application to medicine, Journal of Information Sciences, vol. 124, pp. 125–137, 2000.

[77] S. Hirano, and S. Tsumoto, Indiscernibility-based clustering : Rough clustering. In International Fuzzy Systems Association World Congress, LNCS Springer-Verlag, Heidelberg, pp. 378–386, 2003.

[78] B. Ho and N.B. Nguyen. Nonhierarchical document clustering based on a tolerance rough set model, International Journal of Intelligent Systems, vol. 17, no. 2, pp. 199-212, 2002.

[79] T.B. Ho, S. Kawasaki, and N.B. Nguyen. Documents Clustering using tolerance Rough Set Model and Its Application to Information Retrieval, Studies In Fuzziness And Soft Computing, Intelligent Exploration of the Web, pp. 181–196, 2006.

[80] T.P. Hong, Y.L. Liou, and S.L. Wang. Learning with Hierarchical Quantitative Attributes by Fuzzy Rough Sets, Proceedings of the Joint Conference on Information Sciences, Advances in Intelligent Systems Research, 2006.

[81] Q-H. Hu, and D-R. Yu. Variable precision dominance based rough set model and reduction algorithm for preference-ordered data, Proceedings of the 2004 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2279–2284, 2004.

[82] Q-H. Hu and D-R. Yu. Fuzzy rough C-means clustering, World congress on fuzzy logic, soft computing, computational intelligence: theories and applications (IFSA2005), Tsinghua. Beijing, Springer Lecture notes, 2005.

[83] Q. Hu, D. Yu, and Z. Xie. Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters, vol. 27, no. 5, pp. 414–423, 2006.

[84] Q. Hu, H. Zhao, Z. Xie, and, D. Yu. Consistency based attribute reduction. PAKDD 2007, LNAI 4426, Yang (Ed.), vol. 4426, pp. 96–107, 2007.

[85] Q. Hu, Z. Xie, D. Yu. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition vol. 40, pp. 3509–3521, 2007.

[86] iCAD Second Look "http://www.icadmed.com" *accessed: 10/06/2008*

[87] M. Inuiguchi, and M. Tsurumi. Measures Based on Upper Approximations of Rough Sets for Analysis of Attribute Importance and Interaction, International Journal of Innovative Computing Information and Control, vol. 2, no. 1, pp. 1–12, 2006.

[88] C. Jacobsen, U. Zscherpel, P. Perner. A comparison between neural networks and decision trees, Machine Learning and Data Mining in Pattern Recognition, Springer, Berlin, pp. 144–158, 1999.

[89] A. K. Jain, M. N. Murty and P.J. Flynn. Data clustering: a review, ACM Computing Surveys, vol.31, no.3, pp. 264–323, 1999.

[90] C.Z. Janikow. Fuzzy Decision Trees: Issues and Methods. IEEE Transactions on Systems, Man and Cybernetics — Part B: Cybernetics, vol. 28, pp. 1–14, 1998.

[91] J. Jelonek, K. Krawiec, R. Slowinski, J. Stefanowski, and, J. Szymas. Rough sets as an intelligent front-end for the neural network. Proc. of the First National Conference on Neural Networks their Applications 2, Poland, pp. 116–122, 1994.

[92] R. Jensen and C. Cornelis. A New Approach to Fuzzy-Rough Nearest Neighbour Classification. Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing, pp. 310–319, 2008.

[93] R. Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches. IEEE Transactions on Knowledge and Data Engineering, 16(12): 1457–1471, 2004.

[94] R. Jensen, and Q. Shen, Fuzzy-Rough Attribute Reduction with Application to Web Categorization. Fuzzy Sets and Systems, vol. 141, no. 3, pp. 469-485, 2004.

[95] R. Jensen and Q. Shen. Fuzzy-Rough Data Reduction with Ant Colony Optimization. Fuzzy Sets and Systems, vol. 149, no. 1, pp. 5–20, 2005.

[96] R. Jensen and Q. Shen, Tolerance-based and Fuzzy-Rough Feature Selection, Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE'07), pp. 877–882, 2007.

[97] R. Jensen and Q. Shen. Fuzzy-Rough Sets Assisted Attribute Selection. IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp. 73–89, 2007.

[98] R. Jensen and Q. Shen. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, IEEE Press and Wiley & Sons, 2008.

[99] R. Jensen and Q. Shen. New Approaches to Fuzzy-Rough Feature Selection. To appear in IEEE Transactions on Fuzzy Systems.

[100] L-R. Jian and M-Y. Li. An Extension of VPRS Model Based on Dominance Relation, Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), vol.3, pp. 113–118, 2007.

[101] S. Kawasaki, N.B. Nguyen, and T. B. Ho. Hierarchical document clustering based on tolerance rough set model, In Principles of Data Mining and Knowledge Dis- covery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings (2000), D. A. Zighed, H. J. Komorowski, and J. M. Zytkow, Eds., vol. 1910 of Lecture Notes in Computer Science, Springer, 2000.

[102] L. Ke, Z. Feng and Z. Ren. An efficient ant colony optimization approach to attribute reduction in rough set theory. Pattern Recognition Letters, vol. 29, pp. 1351–1357, 2008.

[103] J.M. Keller, M.R. Gray and J.A. Givens. A fuzzy K-nearest neighbor algorithm. IEEE Trans. Systems Man Cybernet., vol. 15, no. 4, pp. 580–585, 1985.

[104] D. Kim, S-Y. Bang. A Handwritten Numeral Character Classification Using Tolerant Rough Set, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 9, pp. 923–937, 2000.

[105] K. Kira and L.A. Rendell, The feature selection problem: Traditional methods and a new algorithm, Proceedings of Ninth National Conference on Artificial Intelligence, pp. 129–134, 1992.

[106] R. Kohavi, and G.H. John. Wrappers for Feature Subset Selection. Artifical Intelligence, vol. 97, nos. 1-2, pp. 273–324, 1997.

[107] D. Koller, and M. Sahami. Toward Optimal Feature Selection. In: Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, 1996.

[108] J. Komorowski, Z. Pawlak, L. Polkowski and A. Skowron. Rough Sets: A Tutorial. In Rough-Fuzzy Hybridization: A New Trend in Decision Making S.K. Pal and A. Skowron (Eds.). Springer Verlag, Singapore. 1999, pp. 3–98, 1999.

[109] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, Proceedings of the European Conference on Machine Learning, L. De Raedt and F. Bergadano (eds.), Springer-Verlag: Catania, pp. 171–182, 1994.

[110] B. Kosko. Fuzzy entropy and conditioning. Information Sciences, vol. 40, no. 2, pp. 165–174, 1986.

[111] W. Kotłowski, K. Dembczyński, S. Greco, and R Słowiński. Stochastic dominance-based rough set model for ordinal classification, International Journal of Information Sciences, vol. 178, no. 21, pp. 4019–4037, 2008.

[112] M.A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal, vol. 37, no. 2, pp. 233–243, 1991.

[113] M. Kryszkiewicz. Maintenance of reducts in the variable precision rough sets model. ICS Research Report 31/94,Warsaw University of Technology, 1994.

[114] P. Kumar, P.R. Krishna, R.S. Bapi, and S.K. De. Rough clustering of sequential data, Data & Knowledge Engineering, vol. 63, no. 2, pp. 183–199, 2007.

[115] B.C. Kuo and D.A. Landgrebe. Nonparametric Weighted Feature Extraction for Classification, GeoRS, vol.42, no. 5, pp. 1096–1105, 2004.

[116] N. Kwak and C.H. Choi. Input feature selection for classification problems. IEEE Trans. Neural Networks, vol. 13, no. 1, pp. 143–159, 2002.

[117] P. Langley. Selection of relevant features in Machine Learning. Proceedings of the AAAI Fall Symposium on Relevance, pp. 140–144, 1994.

[118] J.A. Lee, M. Verleysen. Nonlinear Dimensionality Reduction Springer Information Science and Statistics Series, 2007.

[119] W. Lee, S.J. Stolfo, and K.W. Mok. Adaptive Intrusion Detection: A Data Mining Approach. AI Review, vol.14, no.6, pp. 533–567, 2000.

[120] H.R. Li, W.X. Zhang.: Applying Indiscernibility Attribute Sets to Knowledge Reduction, Lecture Notes in Artifical Intelligence, pp. 816–821, 2005.

[121] K. Li, Y. Liu. Rough set based attribute reduction approach in data mining. Proceedings of the 2002 International Conference on Machine Learning and Cybernetics. vol. 1, pp. 60–63, 2002.

[122] M. Li, C. Wu, Y. Zhang, and Y. Yue. An Improved BP Network Classifier Based on VPRS Feature Reduction, The Sixth World Congress on Intelligent Control and Automation (WCICA 2006), vol. 2, pp. 9677–9680, 2006.

[123] R. Li, Z.-O. Wang, Mining classification rules using rough sets and neural networks. European Journal of Operational Research, vol. 157, pp. 439–448, 2004.

[124] P. Lingras, Rough neural networks. Proceedings of the Sixth International Conference. Information Processing Management of Uncertainty in Knowledge-Based Systems (IPMU'96), vol.2, pp. 1445–1450, 1996.

[125] P. Lingras (1997), Comparison of neofuzzy rough neural networks. In: Wang (ed.) Proceedings of the Fifth International Workshop on Rough Sets Soft Computing (RSSC'97), pp. 259–262, 1997.

[126] P. Lingras and C. Davies. Applications of Rough Genetic Algorithms, Computational Intelligence, vol. 17, no. 3, pp. 435–445, 2001.

[127] P. Lingras, M. Hogo, and M. Snorek. Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets, Web Intelligence and Agent Systems, vol. 2, no. 3, pp. 217–230, 2004.

[128] P. Lingras and R. Jensen. Survey of Rough and Fuzzy Hybridization. Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE'07), pp. 125-130, 2007.

[129] P. Lingras and C. West. Interval Set Clustering of Web Users with Rough K-means, Journal of Intelligent Information System, vol. 23, no. 1, pp. 5–16, 2004.

[130] H. Liu, and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining Springer International Series in Engineering and Computer Science, vol. 454, 1998.

[131] H. Liu and H. Motoda (eds), Computational Methods of Feature Selection, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2008.

[132] P.C. Mahalanobis, On the generalised distance in statistics, Proceedings National Institute of Science, India, 1936.

[133] B. Mak, T. Munakata, Rule extraction from expert heuristics: a comparative study of rough sets with neural networks and ID3, European Journal of Operational Research, vol. 136, pp. 212–229, 2002.

[134] D. Malyszko and J. Stepaniuk. Standard and Fuzzy Rough Entropy Clustering Algorithms in Image Segmentation, Rough Sets and Current Trends in Computing, vol. 5306/2008, pp. 409–418, 2008.

[135] T. McKee, and T. Lensberg. Genetic programming and rough sets: a hybrid approach to bankruptcy classification, European Journal of Operational Research, vol. 140, no. 2, pp. 436–51, 2002.

[136] S.J. Messick and R.P. Abelson. The additive constant problem in multidimensional scaling. Psychometrika, vol. 21, pp. 1–17, 1956.

[137] A.J. Miller. Selection of subsets of regression variables. Journal of the Royal Statistical Society, vol A, no. 147, 1984.

[138] N. Mac Parthaláin, and R. Jensen. Measures for Unsupervised Fuzzy-Rough Feature Selection. To appear in Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA'09).

[139] N. Mac Parthaláin, R. Jensen and Q. Shen. Fuzzy entropy-assisted fuzzy-rough feature selection. Proceedings of the 15th International Conference on Fuzzy Systems (FUZZ-IEEE'06), 2006.

[140] N. Mac Parthaláin, R. Jensen, and Q. Shen. Comparing Fuzzy-Rough and Fuzzy Entropy-assisted Fuzzy-Rough Feature Selection. Proceedings of the 6th Annual Workshop on Computational Intelligence (UKCI'06), 2006.

[141] N. Mac Parthaláin, R. Jensen, Q. Shen, and R. Zwiggelaar. Rough and fuzzy-rough methods for mammographic data analysis. To appear in Intelligent Data Analysis.

[142] N. Mac Parthaláin, R. Jensen, and Q. Shen. Finding Fuzzy-Rough Reducts with Fuzzy Entropy. Proceedings of the 2008 IEEE Conference on Fuzzy Systems (FUZZ-IEEE'08), Hong Kong, 2008.

[143] N. Mac Parthaláin, R. Jensen and Q. Shen. Rough and fuzzy-rough methods for mammographic data analysis. Proceedings of the 8th Annual UK Workshop on Computational Intelligence (UKCI'08), 2008.

[144] N. Mac Parthaláin, and Q. Shen, Exploring the boundary region of tolerance rough sets for feature selection, Pattern Recognition, vol. 42, no. 5, pp. 655–667, 2009.

[145] N. Mac Parthaláin, and Q. Shen, On Rough Sets, their Recent Extensions and Applications, To appear in Knowledge Engineering Review.

[146] N. Mac Parthaláin, and Q. Shen, A Fuzzy-Rough Approach to Unsupervised Feature Selection, Under review for potential journal publication.

[147] N. Mac Parthaláin, and Q. Shen, Unsupervised Fuzzy-Rough Feature Selection and its Application for Complex Systems Monitoring, Under review for potential journal publication.

[148] N. Mac Parthaláin, Q. Shen, and R. Jensen. Distance Measure Assisted Rough Set Feature Selection. Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE'07), pp. 1084-1089. 2007.

[149] N. Mac Parthaláin, Q. Shen and R. Jensen. A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction. To appear in IEEE Transactions on Knowledge and Data Engineering, 2009.

[150] J.S. Mi and W.X. Zhang, An axiomatic characterization of a fuzzy generalization of rough sets, Information Sciences, vol. 160, no. 1–4, pp. 235–249, 2004.

[151] S. Mitra, and M. Banerjee, Knowledge based neural net with rough sets. In T. Yamakawa et al. (Eds.), Methodologies for the Conception, Design, Application of Intelligent Systems, Proceedings of the Fourth International Conference on Soft Computing (IIZUKA'96), Iizuka Japan, World Scientific, pp. 213–216, 1996.

[152] P. Mitra and S. Mitra. Staging of Cervical Cancer with Soft Computing, IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 934–940, 2000.

[153] pal02 P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised Feature Selection Using Feature Similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 1–13, 2002.

[154] M. Modrzejewski. Feature selection using rough sets theory, In Proceedings of the 11th International Conference on Machine Learning, pp. 213–226, 1993.

[155] L.C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: a survey and experimental evaluation, Proceedings of ICDM02, pp. 306–313, 2002.

[156] Momin, B. F., Mitra, S., and Gupta, R. D. Reduct Generation and Classification of Gene Expression Data. Proceedings of the 2006 international Conference on Hybrid information Technology (ICHIT06), vol. 1, pp. 699–708, 2006.

[157] N.N. Morsi and M. M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets and Systems, vol. 100, no. 1–3, pp. 327–342, 1998.

[158] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[159] P. Narendra, and K. Fukunaga. A Branch and Bound Algorithm for Feature Subset Selection. IEEE Transactions on Computers, vol. C-26, no.9, pp. 917–922, 1977.

[160] A.Y. Ng. On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples. Proceedings of the 15th International Conference on Machine Learning, pp. 404–412, 1998.

[161] C.L. Ngo, and H.S. Nguyen. A tolerance rough set approach to clustering web search results. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (Pisa, Italy, September 20 - 24, 2004). J. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, (eds.)

Lecture Notes In Computer Science, vol. 3202, Springer-Verlag New York, New York, NY, pp. 515–517, 2004.

[162] S.H. Nguyen, and D. Slezak. Approximate Reducts and Association Rules Correspondence and Complexity Results, Lecture Notes in Computer Science (LNCS), vol. 1711/2004, pp. 137–145, 2004.

[163] S.H. Nguyen, and A. Skowron, Searching for Relational Patterns in Data, Proceedings of the first European Symposium on Principles of Data Mining and Knowledge Discovery, pp. 265–276, 1997.

[164] S.H. Nguyen, and A. Skowron. Boolean Reasoning for Feature Extraction Problems, Proceedings of the 10th International Symposium on Methodologies for Intelligent Systems, pp. 117–126, 1997.

[165] R. Nie, J. Yue, An Attribute Reduction Method Based on Rough Set and SVM and with Application in Oil-Gas Prediction, Proceedings of the 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007), pp. 502-506, 2007.

[166] A. Øhrn. Discernibility and Rough Sets in Medicine: Tools and Applications, Department of Computer and Information Science, Trondheim, Norway, Norwegian University of Science and Technology, vol. 239, 1999.

[167] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E.R.E. Denton, R. Zwiggelaar. A Novel Breast Tissue Density Classification Methodology. IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp. 55–65, 2008.

[168] S.K. Pal, R.K. De, and J. Basak. Unsupervised Feature Evaluation: A Neuro-Fuzzy approach, IEEE Transactions in Neural Networks, vol. 11, pp. 366–376, 2000.

[169] S.K. Pal. Pattern Recognition Algorithms for Data Mining, Chapman and Hall, 2004.

[170] P. Pattaraintakorn, and N. Cercone. Integrating rough set theory and medical applications, Applied Mathematics Letters, vol. 21, no. 4, pp. 400–403, 2007.

[171] Z. Pawlak. Some Issues on Rough Sets. LNCS Transactions on Rough Sets, vol. 1, pp. 1–53, 2003.

[172] Z. Pawlak. Rough sets, International Journal of Computing and Information Sciences, vol. 11, pp. 341–356. 1982.

[173] Z. Pawlak. Rough Classification, International Journal of Man-Machine studies, vol. 20, pp. 469–483, 1984.

[174] Z. Pawlak, K. Slowinski, and R. Slowinski. Rough Classification of Patients After Highly Selective Vagotomy for Duodenal Ulcer, International Journal of Man Machine Studies, vol. 24, pp. 413–433, 1986.

[175] Z. Pawlak. Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishing, 1991.

[176] Z. Pawlak and A. Skowron. A rough set approach for decision rules generation, ICS Research Report 23/93, Warsaw University of Technology, Proceedings of the IJCAI'93 Workshop W12: The Management of Uncertainty in AI, France, 1993.

[177] W. Pedrycz. Shadowed sets: bridging fuzzy and rough sets. In S.K. Pal and A. Skowron (eds.) Rough-Fuzzy Hyridisation, Springer Verlag, Sngapore, pp. 179–199, 1999.

[178] W. Pedrycz and G. Vukovich. Feature analysis through information granulation. Pattern Recognition, vol. 35, no.4, pp. 825–834, 2002.

[179] A. Petrosino and M. Ceccarelli. Unsupervised Texture Discrimination Based on Rough Fuzzy Sets and Parallel Hierarchical Clustering, Proceedings of the international Conference on Pattern Recognition, vol. 3 (September 03 - 08, 2000), ICPR, IEEE Computer Society, Washington, DC, 7100, 2000.

[180] J.F. Peters, A. Skowron, Z. Suraj, W. Rzasa, and M. Bokowski. Clustering: A rough set approach to constructing information granules, Proceedings of the 6th International Conference on Sofft Computing and Distributed Processing pp. 57–61, 2002.

[181] S. Piramuthu, The Hausdorff Distance Measure for Feature Selection in Learning Applications. in Procs of the 32nd Annual International Conference on System Sciences Hawaii, vol. 6, 1999.

[182] L. Polkowski, and A. Skowron. Rough sets: A perspective. In: L. Polkowski and A. Skowron, (Eds.), Rough Sets in Knowledge Discovery 1: Methodology and Applications, Physica-Verlag, pp. 31–56, 1998.

[183] K. Qina and Z. Pei, On the topological properties of fuzzy rough sets, Fuzzy Sets and Systems, vol. 151, no. 3, pp. 601–613, 2005.

[184] J.R. Quinlan, C4.5: Programs for Machine Learning, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA., 1993.

[185] R2 ImageChecker, "http://www.r2tech.com" *accessed: 10/06/2008*

[186] A.M. Radzikowska and E.E. Kerre. A comparative study of fuzzy rough sets. Fuzzy Sets and Systems, vol. 126, no. 2, pp. 137–155, 2002.

[187] A.M. Radzikowska and E.E. Kerre, Fuzzy Rough Sets Based on Residuated Lattices, Transactions on Rough Sets II, Lecture Notes in Computer Science (LNCS), vol. 3135/2004, pp. 278–296, 2004.

[188] K.A. Rasmani and Q. Shen, Data-driven fuzzy rule generation and its application for student academic performance evaluation, Applied Intelligence, vol. 25, no. 3, pp. 305–319, 2006.

[189] M. Roffilli. Advanced machine learning techniques for digital mammography, Technical Report UBLCS-2006-12, University of Bologna (Italy), 2006.

[190] A. Mieszkowicz-Rolka, and L. Rolka. Fuzzy Implication Operators in Variable Precision Fuzzy Rough Sets Model. Lecture Notes in Computer Science (LNCS), vol. 3070/2004, pp. 498–503, 2004.

[191] S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, vol. 290, no. 5500, pp. 2323–2326, 2000.

[192] W. Rucklidge. Efficient Visual Recognition Using the Hausdorff Distance. vol. 1173, Lecture notes in computer science. Springer, 1996.

[193] M. Sarkar. Fuzzy-Rough nearest neighbors algorithm. Proceedings of the IEEE conference on Systems Man and Cybernetics, pp. 3556–3561, 2000.

[194] M. Sarkar. Fuzzy-Rough nearest neighbors algorithm. Fuzzy Sets and Systems, vol. 158, pp. 2123–2152, 2007.

[195] M. Sebban and R. Nock. A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recognition, vol. 35, no. 4, pp. 835–846, 2002.

[196] B Sendov, Hausdorff distance and image processing, Russian Math Surveys, 59 (2), pp. 319–328, 2004.

[197] G. Shafer. A Mathematical Theory of Evidence, Princeton University Press, 1976.

[198] D. Shan, N. Ishii, Y. Hujun, N. Allinson, R. Freeman, J. Keane, and S. Hubbard. Feature weights determining of pattern classification by using a rough genetic algorithm with fuzzy similarity measure, Proceedings of Intelligent data engineering and automated learning, pp. 544–550, 2002.

[199] C. Shang and Q. Shen. Rough feature selection for neural network based image classification, International Journal of Image and Graphics, vol. 2, no. 4, pp. 541–555, 2002.

[200] M-W. Shao, and W-X. Zhang. Dominance relation and rules in an incomplete ordered information system, International Journal of Intelligent Systems, vol. 20, no. 1, pp. 13–20, 2004.

[201] L. Shen, F. E. H. Tay, L. Qu and Y. Shen. Fault diagnosis using Rough Sets Theory, Computers in Industry, vol. 43, pp. 61–72, 2000.

[202] Q. Shen and A. Chouchoulas. A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. Engineering Applications of Artificial Intelligence, vol. 13, no. 3, pp. 263–278, 2000.

[203] Q. Shen and A. Chouchoulas. A rough-fuzzy approach for generating classification rules. Pattern Recognition, vol. 35, no. 2, pp. 2425–2438, 2002.

[204] Q. Shen and R. Jensen. Selecting Informative Features with Fuzzy-Rough Sets and its Application for Complex Systems Monitoring, Pattern Recognition, vol. 37, no. 7, pp. 1351-1363, 2004.

[205] A. Skowron, Z. Pawlak, J. Komorowski and L. Polkowski. A rough set perspective on data and knowledge. Handbook of data mining and knowledge discovery, pp. 134–149, Oxford University Press, 2002.

[206] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, R. Slowinski (Ed.), Intelligent Decision Support: Handbook of Applications and Advances to Rough Sets Theory, Kluwer Academic, Dordrecht, pp. 331–362, 1992.

[207] A. Skowron. Boolean reasoning for decision rules generation, Proceedings of the 7th International Symposium ISMIS'93, Trondheim, Norway 1993, In Komorowski J. and Ras Z. (eds.), Lecture Notes in Artificial Intelligence, vol. 689. Springer- Verlag, pp. 295–305, 1993.

[208] A. Skowron, J. Stepaniuk. Generalized approximation spaces. Proceedings of the 3rd International Workshop on Rough Sets and Soft Computing, pp. 156–163, 1994.

[209] A. Skowron, and J. Stepaniuk, Tolerance Approximation Spaces, Fundamenta Informaticae, vol. 27, pp. 245–253, 1996.

[210] R. Slowinski and D. Vanderpooten. Similarity Relations As a Basis for Rough Approximations, Advances in Machine Intelligence and Soft- Computing, P.P. Wang, (ed.), Raleigh, NC. Bookwrights, pp. 17–33, 1997.

[211] R. Slowinski and D. Vanderpooten. A Generalized Definition of Rough Approximations Based on Similarity, IEEE Trans. on Knowl. and Data Eng., vol. 12, no. 2, pp. 331–336, 2000.

[212] Slowinski, K., Stefanowski, J., and Siwinski, D. Application of rule induction and rough sets to verification of magnetic resonance diagnosis. Fundamenta Informaticae, vol. 53, no. 3/4, pp. 345–363, 2002.

[213] E.P.M. de Sousa, C. Traina, A.J.M. Traina, L. Wu, C. Faloutsos. A fast and effective method to find correlations among attributes in databases. Data Mining and Knowledge Discovery, vol. 14, pp. 367–407, 2007.

[214] P. Srinivasan, M.E. Ruiz, D.H. Kraft, and J. Chen. Vocabulary mining for information retrieval: rough sets and fuzzy sets, Information Processing & Management, vol. 37, no. 1, pp. 15–38, 1998.

[215] J. Stefanowski. On rough set based approaches to induction of decision rules. In A. Skowron, L. Polkowski (eds.), Rough Sets in Knowledge Discovery vol. 1, Physica Verlag, Heidelberg, pp. 500–529, 1998.

[216] J Suckling, J. Partner, D.R. Dance, S.M. Astley, I. Hutt, C.R.M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S.L. Kok, P. Taylor, D. Betal, and J. Savage. The Mammographic Image Analysis Society digital mammogram database. International Workshop on Digital Mammography, pp 211–221, 1994.

[217] D.L. Swets and J.J. Weng. Efficient Content-Based Image Retrieval Using Automatic Feature Selection. IEEE International Symposium on Computer Vision, pp. 85–90, 1995.

[218] R. Swiniarski, F. Hunt, D. Chalvet, and D. Pearson. Intelligent data processing and dynamic process discovery using rough sets, statistical reasoning and neural networks in a highly automated production system. Proc. of the First

European Conference on Application of Neural Networks in Industry, Helsinki, Finland, 1995.

[219] R. Swiniarski. Rough sets Bayesian methods applied to cancer detection. In Polkowski and Skowron (eds.) Proceeding of the First International Conference on Rough Sets and Soft Computing (RSCTC'98), Springer-Verlag, LNAI, vol.1424, pp. 609–616, 1998.

[220] R. Swiniarski. Rough Sets and Principal Component Analysis and Their Applications in Data Model Building and Classification. In S. K. Pal and A. Skowron (Eds.), Rough Fuzzy Hybridization: New Trends in Decision Making, Springer Verlag, Singapore, 1999.

[221] R. Swiniarski and A. Skowron. Rough Set Methods in Feature Selection and Recognition, Pattern Recognition Letters, vol. 24, no. 6, pp. 83–849, 2003.

[222] J.B. Tenenbaum, V. de Silva, J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, vol. 290, no. 5500, pp. 2319–2323, 2000.

[223] K. Thangavel, A. Pethalakshmi, and P. Jaganathan. A Comparative Analysis of Feature Selection Algorithms Based on Rough Set Theory, International Journal of Soft Computing, vol.1, no. 4, pp. 288–294, 2006.

[224] H. Thiele, Fuzzy rough sets versus rough fuzzy sets - an interpretation and a comparative study using concepts of modal logics, Technical report no. CI-30/98, University of Dortmund, 1998.

[225] C. Tjortjis, M. Saraee, B. Theodoulidis, J. A. Keane. Using T3, an Improved Decision Tree Classifier, for Mining Stroke-Related Medical Data, Methods of Information in Medicine, vol. 46, no.5, pp. 523–529, 2007.

[226] W.S. Torgerson. Psychometrika, vol. 17, pp. 401–419, 1952.

[227] E.C.C. Tsang, D. Chen, D.S. Yeung, X-Z Wang, J. Lee. Attributes Reduction Using Fuzzy Rough Sets, IEEE Transactions on Fuzzy systems, vol. 16, no. 5, pp. 1130–1141, 2008.

[228] Wallace, M., Avrithis, Y., and Kollias, S., 2006. Computationally efficient sup-t transitive closure for sparse fuzzy binary relations, Fuzzy Sets and Systems, vol. 157, no. 3, pp. 341–372, 2006.

[229] P. Wojdyllo, Wavelets, rough sets arificial neural networks in EEG analysis. In: Polkowski and Skowron (eds.) Proceedings of the First International

Conference on Rough Sets and Soft Computing (RSCTC'98), Springer-Verlag, LNAI vol. 1424, pp. 444–449, 1998.

[230] Y. Wang, M. Ding, C. Zhou, T. Zhang. A hybrid method for relevance feedback in image retrieval using rough sets and neural networks, International Journal of Computational Cognition, vol.3, no.1, 2005.

[231] Z. Wang, X. Shao, G. Zhang, and H. Zhu. Integration of Variable Precision Rough Set and Fuzzy Clustering: An Application to Knowledge Acquisition for Manufacturing Process Planning, Proc. 10th International Conference, RSFDGrC 2005, pp. 585–593, 2005.

[232] J. Wang, and J. Wang. Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method. Journal of Computer Science & Technology, vol. 16, no. 6, pp. 489–504, 2001.

[233] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen. Feature Selection based on Rough Sets and Particle Swarm Optimization. Pattern Recognition Letters, vol. 28, no. 4, pp. 459–471, 2007.

[234] X. Wang, J. Yang, X. Teng and N. Peng. Fuzzy-Rough Set Based Nearest Neighbor Clustering Classification Algorithm, Lecture Notes in Computer Science, vol. 3613/2005, pp. 370–373, 2005.

[235] Z. Wang, X. Shao, G. Zhang, and H. Zhu. Integration of Variable Precision Rough Set and Fuzzy Clustering: An Application to Knowledge Acquisition for Manufacturing Process Planning, Proceedings of the 10th conference on Rough sets, fuzzy sets, data mining, and granular computing (RSFDGrC 2005), 2005.

[236] I.H. Witten and E. Frank. Generating Accurate Rule Sets Without Global Optimization. In Machine Learning: Proceedings of the 15th International Conference, Morgan Kaufmann Publishers, San Francisco, 1998.

[237] I.H. Witten and E. Frank. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000.

[238] J. Wróblewski (1995). Finding minimal reducts using genetic algorithms. Proceedings of the International Workshop on Rough Sets Soft Computing at Second Annual Joint Conference on Information Sciences (JCIS'95), pp. 186–189, 1995.

[239] W.Z. Wu, J.S. Mi, and W.X. Zhang, Generalized fuzzy rough sets, Information Sciences, vol. 151, pp. 263–282, 2003.

[240] W.Z. Wu and W.X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Information Sciences, vol. 159, no.3–4, pp. 233–254, 2004.

[241] W.Z. Wu, Y. Leung, and J.S. Mi, On characterizations of (I,T)-fuzzy rough approximation operators, Fuzzy Sets and Systems, vol. 154, no. 1, pp. 76–102, 2005.

[242] W.Z. Wu, A Study on Relationship Between Fuzzy Rough Approximation Operators and Fuzzy Topological Spaces, L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3613, pp. 167–174, 2005.

[243] M. Wygralak, Rough sets and fuzzy sets - some remarks on interrelations, Fuzzy Sets and Systems, vol. 29, no. 2, pp. 241–243, 1989.

[244] E. Xing, M. Jordan, and R. Carp. Feature Selection for High Dimensional Genomic Microarray Data, Proceedings of the 15th International Conference on Machine Learning, pp. 601–608, 2001.

[245] Y. Xue and D.E. Brown. Decision Based Spatial Analysis of Crime, Lecture Notes in Computer Science, Springer, vol. 2665/2003, pp. 153–167, 2003.

[246] Yahia, M., Mahmod, R., Sulaiman, N. and Ahmad, F. Rough neural expert systems, Expert Systems with Applications, vol. 27, no. 2, pp. 87–99, 2000.

[247] Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML 97), pp. 412–420, 1997.

[248] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, vol. 13, no. 1, pp. 44–49, 1998.

[249] Y.Y. Yao, Combination of rough and fuzzy sets based on α-level sets, in: T.Y. Lin, N. Cereone (Eds.), Rough Sets and Data Mining: Analysis of Imprecise Data, Kluwer Academic Publishers, pp. 301–321, 1997.

[250] G. Yi, H. Hu, and Z. Lu. Web Document Classification Based on Extended Rough set. Proceedings of the Sixth international Conference on Parallel and Distributed Computing Applications and Technology (PDCAT), IEEE Computer Society, Washington, DC, pp. 916–919, 2005.

[251] D.S. Yeung, D. Chen, E.C.C. Tsang, J.W.T. Lee, and W. Xizhao. On the Generalization of Fuzzy Rough Sets, IEEE Transactions on Fuzzy Systems, vol. 13, no. 3, pp. 343–361, 2005.

[252] F.W. Young and R.M. Hamer. Theory and Applications of Multidimensional Scaling. Eribaum Associates. Hillsdale, NJ, 1994.

[253] O. Yun, and J. Ma. Land cover classification based on tolerant rough set. International Journal of Remote Sensing, vol. 27, no. 14, pp. 3041–3047, 2006.

[254] L.A. Zadeh. Fuzzy sets. Information and Control, vol. 8, no.3, pp. 338–353, 1965.

[255] L.A. Zadeh. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning-1, Information Sciences vol.8 pp. 199–249, 1975.

[256] Y. Zhao, H. Zhang, and Q. Pan. Classification Using the Variable Precision Rough Set, Proceedings of RSFDGrC 2003, Chongqing, China, vol. 2639, pp. 350–353, 2003.

[257] S. Zhao and Z. Zhang. A generalized definition of rough approximation based on similarity in variable precision rough sets. Proc. of 2005 International Conference on Machine Learning and Cybernetics, pp. 3153–3156, 2005.

[258] Y. Zhao, X. Zhou, and G. Tang. A rough set-based fuzzy clustering, Proc. Second Asia information retrieval symposium, pp. 401–409, 2005.

[259] W.Q. Zhao and Y.L. Zhu. Classifying email using variable precision rough set approach, Lecture Notes In Artificial Intelligence, vol. 4062, pp. 766–771, 2006.

[260] S. Zhao, and E.C.C. Tsang. On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences, vol. 178, pp.3163–3176, 2008.

[261] X. Zheng and J.Wang. Power transformer fault diagnosis based on variable precision rough set, Proceedings of the 3rd International Conference on Electric Utility Deregulation and Restructuring and Power Technologies 2008, pp. 1353–1358, 2008.

[262] N. Zhong, J. Dong, and S. Ohsuga. Using Rough Sets with Heuristics for Feature Selection, Journal of Intelligent Information Systems, vol. 16, no. 3, pp. 199–214, 2001.

[263] W. Ziarko. Variable Precision Rough Set Model, JCSS, vol. 46, no. 1, pp. 39–59, 1993.

[264] W. Ziarko. Acquisition of hierarchy-structured probabilistic decision tables and rules from data, Expert Systems, vol. 20, no. 5, pp. 305–310, 2003.