2018-10-19

Learning and Reasoning with Imperfect Data

Janith Heendeni P. Don
University of Miami, janithnadun@gmail.com

UNIVERSITY OF MIAMI

LEARNING AND REASONING WITH IMPERFECT DATA

By

Janith Nadun Anuja Heendeni Pathiranage Don

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2018

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

LEARNING AND REASONING WITH IMPERFECT DATA

Janith Nadun Anuja Heendeni Pathiranage Don

Approved:

Kamal Premaratne, Ph.D.
Professor of Electrical and
Computer Engineering

Manohar N. Murthi, Ph.D.
Associate Professor of Electrical and
Computer Engineering

Xiaodong Cai, Ph.D.
Professor of Electrical and
Computer Engineering

Jie Xu, Ph.D.
Assistant Professor of Electrical and
Computer Engineering

Dilip Sarkar, Ph.D.
Associate Professor of Computer
Science

Guillermo Prado, Ph.D.
Dean of the Graduate School

HEENDENI PATHIRANAGE DON, JANITH N.A. (Ph.D. Electrical and Computer Engineering)

Learning and Reasoning with Imperfect Data (December 2018)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Kamal Premaratne.
No. of pages in text. (143)

The need for increased automation of learning, knowledge discovery, reasoning, and inference from the rapid growth of the availability of a multitude of various types of sensor/data feeds and databases has generated renewed interest in machine learning (ML). The practical utility of ML algorithms and their effectiveness greatly depend on how well one may learn the relevant parameters from data, and the parameter learning phase of modern ML environments has emerged as a significant challenge because of the increasing complexity of the data being gathered.

Adequate representative statistical training data are often too costly to obtain or are simply unavailable; available real-world data are usually rife with incomplete, unknown, or missing entries due to a host of reasons, including simple data entry errors, security and privacy concerns, difficulty in obtaining data corresponding to infrequent events, and others. Data imputation strategies being employed to deal with data "missingness" run the gamut from interpolating the missing value from values of other variables, to using a data "missingness" probability distribution to estimate the missing value, to simply ignoring data records possessing missing values and using only the "clean" records to learn the parameters. Interpolating or employing a "missingness" distribution for data imputation constitutes a recipe for making impaired decisions lacking trustworthiness when there is little or no evidence to sup-

port the assumptions made; disregarding data records possessing imperfections has the potential to destroy critical evidence.

The main objective of this research work is to develop a comprehensive strategy that can model and account for a wider variety of data imperfections, including those that are generated from human-generated “soft” data; incorporate and propagate the information contained in these data imperfections throughout the decision-making process; conduct the learning, knowledge discovery, reasoning, and inference processes in a computationally efficient manner and generate conclusions that are appropriately calibrated to reflect the underlying uncertainties.

The approach we take is based on a framework that employ interval-valued (i.v.) probability functions. They are better suited and offer more flexibility for handling a wider variety of uncertainties and they are what naturally arise in partial elicitation (when insufficient knowledge is available), when it is too time consuming to gather the necessary knowledge to estimate exact probabilities. We do not insist on any monotonicity condition on the i.v. probability functions we utilize, and take the viewpoint that these i.v. probabilities, which we refer to as *PrBounds,* emerge from a single underlying probability distribution about which agents have only partial information. With a fresh perspective of the i.v. counterpart notions of conditioning and independence, we then propose a framework which allows parameter learning, knowledge discovery, reasoning, and inference in a computationally efficient manner in much the same way as one would with probabilistic graphical models.

We show how PrBounds could be extracted from imperfect datasets where the values of different attributes may be dependent and embrace more general evidential uncertainty. When the attribute values are unknown/missing or are known to lie within a set of values, PrBounds can be learned via a computationally tractable and efficient frequency counting method. The probabilities associated with an arbitrary imputation strategy, including the underlying "true" probabilities, are guaranteed to lie within the PrBounds learned in this manner. We also develop new Demspter-Shafer (DS) belief theoretic and PrBounds-based models of an imperfect implication rule which are consistent with Bayesian and classical logic models. We demonstrate how it can be fused with an imperfect antecedent to generate the PrBounds associated with the rule consequent. Finally, inspired by deep learning neural network architectures but operating within the proposed PrBounds-based framework, we develop what we refer to as a *deep fusion network (DFN)* which allows one to automate fusion of evidence from input data, fusion parameter selection, and classification of potentially uncertain data generated from multi-modal sensors.

To my parents.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Professor Kamal Premaratne for the continuous support of my Ph.D. study, for his patience, motivation, encouragement and knowledge. His guidance helped me in all the time of research, academic courses, writing of academic papers and this dissertation. I am also thankful to Professor Manohar Murthi for his thoughtful inputs and guidance to my research as the co-advisor. I would like to thank the rest of my dissertation committee members Professor Dilip Sarkar, Professor Xiaodong Cai, and Professor Jie Xu, for their insightful inputs, comments, and valuable time.

I would like to acknowledge the U.S. Office of Naval Research (ONR) which supported my research work via grant #N00014-10-1-0140. I would like to thank Dr. Jake Vanderplas, Dr. Peter Lucas, and all down at Sloan Digital Sky Survey and UCI Machine Learning Repository for helping and providing us with datasets.

I'm grateful to my family; my father, my brother, my late mother, and my uncles and aunts, for their enlightenment, caring and support throughout my life. I'm also grateful to all of my teachers and instructors for their knowledge, wisdom and guidance.

My sincere thanks also goes to my friends Dr. Buddhika Samarakoon, Dr. Randil Gajasinghe, Dr. Ranga Dabarera, Lalintha Polpitiya, Kusumitha Perera and their family members for their support and for all the fun we have had in the last six years. I'm also thankful to my past and present lab mates and fellow students Dr. Thanuka Wickramaratne, Dr. Sayan Maity, Dr. Yilin Yan, Rafael Nunez, May Zar Lin, Saad Sadiq, and Olga Sanchez.

I thank all of my Sri Lankan friends in Miami. A very special gratitude goes out to Mrs. Milina Herath and her family, Mrs. Manel Vass and her family for their support with food and lodging.

Last but not least, I would like to thank all the university staff members who have supported me along the way.

Janith Nadun Anuja Heendeni Pathiranage Don

University of Miami

December 2018

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Machine Learning

"Machine Learning is the field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to 'learn' [from data]" [Kohavi and Provost, 1998]. What has been learned from examples or *training data* is then used for the knowledge discovery, reasoning, and inference processes. The increased availability of a multitude of streaming/stored sensor and data feeds and databases does not necessarily imply that the learning, knowledge discovery, reasoning, and inference processes are easier [Premaratne et al., 2009], because one must now grapple with *heterogeneous* data (meaning that attributes may be numerical or categorical, and they may be scalar- or vector-valued), *multi-scope* data (meaning that sources may not possess identical "scopes of expertise" because not all sources can be expected to have access to global knowledge), and significantly higher volumes of streaming and stored data. Increased automation of the learning, knowledge discovery, reasoning, and inference processes is an overriding requirement for the purpose of easing the immense burden associated with dealing with such data. This explains the renewed

interest in effective and analytically tractable models for representation of data and computationally viable machine learning (ML) techniques and algorithms.

1.2 Challenges

1.2.1 Real Data Are Imperfect

The practical utility of these ML algorithms and their effectiveness depend on how well one may learn the relevant parameters from training data. The parameter learning phase has become a significant challenge in modern ML environments mainly because of the increased complexity of the data that are available and the data that are being gathered for knowledge discovery, reasoning, and inference.

Adequate representative statistical training data are often too costly to obtain or are simply unavailable, and when available, real-world training data are often incomplete, unknown or missing, due to absence or errors in data entry (e.g., in subject surveys), security and privacy concerns, difficulty in obtaining data corresponding to infrequent events, etc. [Liao and Ji, 2009]. There are other causes of data imperfections as well: reliability (or lack thereof) of sources providing the data, ambiguities created by attempts to "equalize" the scopes of multi-scope data, differing opinions of domain experts whose expertise is being sought (for the purpose of classifying the training set, for example), etc. [Janez and Appriou, 1998, Premaratne et al., 2007].

1.2.2 Dealing With Data Imperfections

Judging by the datasets that are publicly available, one may be inclined to dispute this claim that real-world data are rife with imperfections. But data are typically decontaminated of unknown or missing values before they are utilized in ML algorithms or before they are made publicly available.

A comprehensive survey of methods for handling data "missingness" appears in [Liao and Ji, 2009]. A "piece" of data could be

(a) *not missing at random (NMAR),* i.e., the missingness could be related to both its own value and the values of other variables,

(b) *missing at random (MAR),* i.e., the missingness is unrelated to its own value but could be related to the values of other variables, or

(c) *missing completely at random (MCAR),* i.e., the missingness is unrelated to either its own value or the values of other variables.

Data imputation strategies for dealing with data "missingness" run the gamut from employing known attribute values to interpolate the missing attribute value, to assuming that a probability distribution which dictates the value of the missing attribute is known so that it can be used to estimate the missing attribute value, to using only the "clean" data records to learn relevant parameters and ignoring data records possessing missing attribute values.

1.2.3 Difficulties

The strategy of ignoring data records possessing missing attribute values and learning relevant parameters from only the "clean" data records destroys what potentially could have been critical evidence. Moreover, as the following example illustrates, such a strategy turns out to be inherently flawed.

Example 1 *Consider the 7 data records* $(a_1, b_1, ?)$*,* $(a_2, ?, c_2)$*,* (a_1, b_1, c_1)*,* (a_2, b_2, c_2)*,* $(?, ?, c_1)$*,* $(?, b_1, c_2)$*,* $(a_2, ?, c_2)$*, each attribute having 2 states,* (a_1, a_2)*,* (b_1, b_2)*, and* (c_1, c_2)*; "?" denotes missing attribute values.*

In the absence of priors, the probability $P(a_1, b_1, c_1)$ *of the occurrence of* (a_1, b_1, c_1) *must satisfy*

$$P(a_1, b_1, c_1) \in [1/7, 3/7],$$

regardless of the imputation strategy. If we count only the clean data records, we get

$$P(a_1, b_1, c_1) = 1/2,$$

a value no *imputation strategy can yield and hence unacceptable. Unless prior evidence justifies it, committing to* one *imputation strategy, which is equivalent to picking* one *value from [1/7, 3/7], can also be misleading (e.g., in classification).* ■

Expectation Maximization (EM) algorithm [Dempster, 1968, Lauritzen, 1995], Gibbs sampling methods [Geman and Geman, 1984], AI&M scheme [Jaeger, 2006], and other related techniques [Cowell, 1999, Ramoni and Sebastiani, 2001], are often the methods of choice for handling data "missingness". However, these cater mainly to the NMAR and MAR cases where imputation is guided by the values of other variables. The MCAR case, or the case when the relationships between variables are unknown or indeterminate (for example, when adequate evidence is lacking to justify the use of an underlying distribution related to data "missingness"), is more challenging. One is then compelled to harness human-generated *soft* data. In fact, to deal with MCAR data, the method in [Liao and Ji, 2009] allows the more qualitative subjective information of domain experts to provide interval-valued probabilities which, as it turns out, are better suited to capturing soft evidence.

Interpolating and employing an underlying distribution of “missingness” for the purpose of data imputation can severely impair the decision-making process when there is no evidence to justify the assumptions made [Hewawasam et al., 2005]. This is specially problematic in medical/healthcare, defense, cybersecurity, and other critical application scenarios where the trustworthiness of the decisions made is of paramount importance.

Indeed, development of better techniques of handling data imperfections for learning, extracting knowledge, reasoning, and inference can be considered a chronic problem hampering data-driven studies that attempt to discern among competing hypotheses [Motro and Smets, 1997]. It remains one of the most challenging problems confronting the application of ML techniques in real-world application domains.

1.3 Main Objective and Our Approach

1.3.1 Main Objective

Our main objective is to develop a comprehensive strategy that can

(a) model and account for a wider variety of data imperfections, including those that one would encounter in soft evidence;

(b) incorporate the information contained in these imperfections (as opposed to ignoring them or employing a data imputation strategy based on unjustifiable assumptions regarding data “missingness”) into ML algorithms and propagate this information throughout the decision-making process;

(c) conduct the learning, knowledge discovery, reasoning, and inference processes so that the conclusions provided to the decision-maker are appropriately calibrated to reflect the underlying uncertainties; and of course

(d) carry out these tasks in a computationally efficient manner.

1.3.2 Our Approach

Conventional approaches may not be well equipped to handle the variety of uncertainties that inhabit data, especially those in nuanced soft evidence which may play a critical role in dealing with data imperfections. Being beholden to a-priori assumptions regarding the underlying distributions and priors, it is questionable whether the Bayesian framework is well suited to the task of modeling such uncertainties [Benavoli et al., 2008, Sambhoos et al., 2008, Premaratne et al., 2009, Khaleghi et al., 2013, Núñez et al., 2013, Heendeni et al., 2014].

1.3.2.1 Interval-Valued (I.V.) Probabilities

On the other hand, *interval-valued (i.v.) probabilities* because they are better suited and offer more flexibility for handling these data uncertainties. In addition to being better suited for the purpose at hand, i.v. probabilities

(a) overcome the difficulty of a single distribution (as used in the Bayesian approach) to distinguish between *uncertainty* and *ignorance* (or between *certainty* and *confidence*) [Levi, 1990, Tassem, 1992];

(b) are a more reasonable way to describe confidence (instead of using a single point) [Chrisman, 1996b];

(c) may arise from incomplete or partial elicitation (e.g., when insufficient knowledge is available) or when it is too time consuming to obtain the necessary knowledge to estimate exact probabilities [Good, 1962, Fertig and Breese, 1993];

(d) are useful for studying sensitivity and robustness in probabilistic inference [Berger, 1982, Walley, 1991, Wasserman and Kadane, 1992];

(e) can be used to weigh computational precision against modeling precision [Cozman and Krotkov, 1996];

(f) arise in group decision problems [Seidenfeld et al., 1989] and in axiomatic approaches to uncertainty when the axioms of probability are weakened [Giron and Rios, 1980, Walley, 1991];

(g) arise when determining constraints on probabilities given only the probabilities of only a finite set of other events [Nilsson, 1986];

(h) may result from the abstraction of more detailed probabilistic models [Chrisman, 1992, Chrisman, 1996a, Haddawy and Suwandi, 1994].

1.3.2.1.1 Dempster-Shafer (DS) Theoretic Functions *Dempster-Shafer (DS) belief theoretic functions* constitute the special class of ∞-*monotone* i.v. probability functions. So, not too surprisingly, and in contrast to fuzzy sets, rough sets, and similar uncertainty handling frameworks, the DS theoretic (DST) framework bears a closer relationship to the probabilistic framework [Fagin and Halpern, 1990]. In fact,

(a) the inner and outer measures of a non-measurable event turn out to be closely related to the DST notions of belief and plausibility, respectively [Fagin and Halpern, 1990, Halpern and Fagin, 1992]; and

(b) DST models converge to probability mass function (p.m.f.) models in the limiting case [Shafer, 1976], thus enabling one to swiftly generalize legacy probabilistic techniques.

Therefore, DST functions can be viewed as generalizations of p.m.f.s [Dempster, 1968, Shafer, 1976, Halpern and Fagin, 1992, Smets, 1992, Smets, 1994]. DST models are also more intuitive in the way they capture probabilistic and possibilistic data and the types of uncertainties and nuances of "soft" data [Yager et al., 1994, Blackman and Popoli, 1999, Vannoorenberghe, 2004, Hewawasam et al., 2007, Wickramarathne et al., 2014, Dabarera et al., 2016]. For example, completely missing/unknown data can be captured via the DST "vacuous" model and data ambiguities precipitated by the inability to discern between hypotheses due to lack of evidence can be captured by allocating evidential support for non-singleton propositions. These factors explain the popularity and the widespread use of the DST framework for dealing with different types of data imperfections.

However, the utility of DS theoretic (DST) models for learning, knowledge discovery, reasoning, and inference in a computationally efficient manner is still a cause for concern. The expressive power of DST models comes at the expense of a higher computational burden. While significant advances have been made for mitigating the computational complexity, e.g., [Bauer, 1997, Wilson, 2001, Wickramarathne et al., 2013], and the very recent work in [Polpitiya et al., 2016, Polpitiya et al., 2017], this computational burden constitutes the main criticism that the DST approach has drawn over the years.

1.3.2.1.2 Probability Bounds Instead, in this work, we develop a framework which is grounded on i.v. probabilities which are not required to be monotonic, or, for convenience of reference, *probability bounds.* This is in stark contrast previous work on i.v. probability functions which are required to be monotonic (see the works quoted in Section 1.3.2.1).

Relinquishing the requirement of monotonicity yields a framework that enables parameter learning, knowledge discovery, reasoning, and inference in much the same manner as one would carry out these tasks with probabilities (as in a Bayesian network, for example).

We pay special emphasis to datasets where attribute values could be unknown, missing, known to lie within a set of values or are most general evidential data. For such datasets, we show that an intuitive frequency counting method can be employed to learn interval-valued parameters which are guaranteed to capture the underlying probabilities. As it turns out, the parameters so learned are DS belief theoretic functions.

1.4 Organization of the Document

Here we provide a brief overview of the contents of each chapter of this document.

1.4.1 Chapter 2 Preliminaries

Chapter 2 introduces the basic notions related to i.v. probabilities and DS theory that are essential for the work being proposed.

1.4.2 Chapter 3 Imperfect Implication Rules

Implication rules, which take the form “if A, then B” or, as is often expressed, $R : A \Longrightarrow B$, constitute the backbone of reasoning and inference engines. A large volume of existing work addresses the extraction of such rules from databases and their use in various application scenarios [Agrawal et al., 1993, Agrawal and Srikant, 1994, Liu et al., 1998, Li et al., 2001, Nanavati et al., 2001]. However, most of these works do not allow evidence/information to be imperfect. In reality, the rule consequent B and the rule $R : A \Longrightarrow B$ itself are imperfect. And, of course, one cannot expect to get “perfect” rules when only *finite* databases are available for parameter and rule extraction.

Probabilistic and fuzzy models are perhaps the two most commonly used approaches to capture imperfect rules [Dubois et al., 2001, Nguyen et al., 2002]. Several previous works provide DST models of imperfect rules: in [Ginsberg, 1984, Hau and Kashyap, 1990], DST fusion/combination strategies are employed to get results that are similar to ours, but the general bounds and inequalities that we derive are absent and the approach taken is different; in [Benavoli et al., 2008], emphasis is placed on satisfying the material implications of propositional logic statements; in [Nunez et al., 2013], a complete uncertain logic framework (imperfect rules being a special case) which is compatible with classical (perfect) logic [Nguyen et al., 2002] is provided. We take a different view: we do not impose compatibility with classical logic in imperfect domains; rather, we expect compatibility only when the domain is perfect, so that our model is very general and all probability and classical logic models follow as special cases.

Our model is based on the DST Fagin-Halpern (FH) conditional [Fagin and Halpern, 1990]. While the use of the Bayesian conditional has been criticized as a model of probabilistic imperfect rules [Lewis, 1976, Benavoli et al., 2008], we demonstrate that the DST FH conditionals can be used as an effective i.v. model of an imperfect rule which can then be fused with an imperfect antecedent. Given the uncertainty intervals associated with the rule antecedent and the rule itself, we derive explicit lower and upper bounds for the uncertainty interval of the rule consequent. Then we explicitly show its consistency with Bayesian inference and classical logic.

1.4.3 Chapter 4 PrBounds: A Framework Based on Probability Bounds

From Bayesian networks (BNs) to factor graphs, graphical structures offer efficient algorithms to deal with functions of many variables by exploiting how they can be factorized into a product of functions of a smaller number of variables [Kschischang et al., 2001]. For example, Bayesian networks (BNs) explicitly "code" and capitalize on conditional independence properties to factor a joint probability distribution. This information is then exploited for reasoning and inference with uncertain knowledge [Pearl, 1988, Lauritzen and Wermuth, 1989, Daly et al., 2011, Whittaker, 1990, Castillo et al., 1997]. The recent work in [Coden et al., 2016, D'Addabbo et al., 2016, Leicester et al., 2016, Roudposhti et al., 2016, Villalba et al., 2016, Das et al., 2017, Kourou et al., 2017] is representative of the widespread utility of graphical models in a broad spectrum of application scenarios.

Existing works of graphical models that allow reasoning with i.v. probabilities consist of two views. *Bayesian sensitivity analysis* views i.v. probabilities as the lower/upper bounds corresponding to a *set* of underlying probabilities which immediately raises the difficulty of how to capture the notion of independence (including conditional independence). Bayesian analysis interpretation-based work, including work related to *credal networks,* the imprecise probability version of BNs [van der Gaag, 1990, Tassem, 1992, Jaffray, 1992, Cozman, 2000], often utilizes *strong independence,* a strong assumption requiring all the extreme distributions to display independence [Augustin et al., 2014], as a surrogate of the notion of independence.

The alternate *behavioral interpretation,* which can be developed with no recourse to probabilities, is the theory of coherent lower/upper previsions (or "expectations") [Walley, 1991, Walley, 1996, Miranda, 2008]. This work, including work on credal networks viewed through the lens of behavioral interpretation, employ another surrogate notion of independence called *epistemic independence* which enables factorizibility properties to be utilized as in BNs [de Cooman et al., 2010, de Cooman et al., 2011, Augustin et al., 2014]. However, epistemic independence may not be a good indicator of probabilistic independence when dealing with imperfect datasets. In addition, *natural extension,* the mathematical procedure required to preserve "coherence", requires additional computations [Walley, 1991, Walley, 1996, de Cooman et al., 2010, de Cooman et al., 2011, Augustin et al., 2014].

These different surrogate notions of independence are *defined* in terms of lower/upper bounds, and no two notions are necessarily consistent with each other [de Campos and Moral, 1995, Chrisman, 1996a, de Cooman et al., 2011, Augustin et al., 2014].

We prefer to view i.v. probabilities as emerging from a *single* underlying true probability distribution instead of from a *set* of underlying distributions. It is important to draw a distinction between this view and the presumption of a single underlying distribution (which would essentially dictate how data imputation is to be carried out), a strategy which has drawn criticism [Cozman, 1997, Zaffalon, 2002b, Halpern and Leung, 2016]. We interpret i.v. probabilities — which, for convenience, we call *PrBounds* — as how an agent captures and quantifies the underlying distribution when it has access to only partial information about it. So the set of PrBounds an agent generates is tacitly taken to be "consistent" in that it contains the underlying distribution. In a multiple agent scenario, one then encounters multiple sets of PrBounds, each agent generating its own set of consistent PrBounds depending on the evidence it has access to. With this vantage point of a single underlying distribution, we take a fresh look at the i.v. conditional and independence notions and demonstrate how PrBounds could be maneuvered for parameter learning and reasoning with computational complexity comparable to what is required in BNs. Of course, when it comes to parameter learning from a dataset, we assume that any subjective knowledge about variables (e.g., independence and conditional independence) is reflected within the true dataset (which of course may not be available).

This viewpoint of a single underlying distribution, which is in fact a special case of the Bayesian analysis interpretation, is not new and appears in, for example, [Quinlan, 1983, Grosof, 1985, van der Gaag, 1990]. The i.v. probability notion employed in [Quinlan, 1983] impose additional constraints (called *inferno propagation constraints*). The work in [Grosof, 1985] utilizes an inequality paradigm to arrive at the i.v. probabilities associated with a single underlying distribution (although no explicit mention of a single underlying distribution is made). The work in [van der Gaag, 1990] addresses a different issue in that it provides a linear algebraic strategy to find unknown probabilities when the joint distribution is known only partially. Our work differs from this work in that we avoid imposing any additional constraints on PrBounds, and in how we use and exploit i.v. conditionals and independence notions and in how we attribute these explicitly to the underlying distribution.

1.4.4 Chapter 5 Learning Parameters From Imperfect Data

We pay special attention to a type of uncertainty that is most commonly encountered in realistic datasets: attribute values that are unknown/missing or that are known to lie within a set of values but otherwise cannot be discerned further (Definition 7). For datasets populated with only this type of uncertainty, we give a computationally more tractable alternative to compute all the probability bounds for each attribute (Corollary 3), an efficient way to obtain lower/upper probability bounds for each data record (Lemma 6), and an intuitive frequency counting method to learn the lower/upper bounds of probability and conditional probability parameters that are needed for our i.v. graphical models. Importantly, the underlying probabilities are guaranteed to be constrained within the bounds learned in this manner (Corollary 5).

Example 2 (Example 1 Revisited) *The results we develop (see Example 9) indeed show that, in Example 1,*

$$P(a_1, b_1, c_1) \in [1/7, 3/7].$$

Our results apply to conditional probabilities as well as intermediate results generated in, for example, graphical models. For example, what is the probability $P(a1|c1)$ *that attribute-1 takes the value* a_1 *given that attribute-3 takes the value* c_1*? If attribute-3 of* $(a_1, b_1, ?)$ *is* c_1*,* $P(a_1|c_1)$ *is bounded by [2/3, 1]; if not, it is bounded by [1/2, 1]. So,* $P(a_1|c_1)$ *is bounded by [1/2, 1], exactly what our results yield.* ■

1.4.5 Chapter 6 Deep Fusion Networks (DFNs)

Neural networks and/or deep learning architectures that can handle imperfect or uncertain data with DS theory have already appeared in the literature [Denoeux and Bjanger, 2000, Soua et al., 2016, Wang et al., 2016, Itkina and Kochenderfer, 2017]. The work in [Denoeux and Bjanger, 2000] employs the Dempster's combination rule (DCR) for evidence combination and allows only uncertainties in the label variables and not in the attributes. The work in [Soua et al., 2016] does not use uncertain attributes or uncertain labels, but it uses Dempster's rule of conditioning to combine two network outputs before making its final decision. The work in [Wang et al., 2016] uses DS masses as inputs and certain combination strategies that are related to DCR at the neurons. The work in [Itkina and Kochenderfer, 2017] uses the DCR to combine different types of occupancy grid information before feeding the fused outputs to the neural network. A

In this chapter, we develop what we refer to as a *deep fusion network (DFN),* a PrBounds-based deep learning architecture which can be used to automate fusion of input data streams, fusion parameter selection, and classification of potentially uncertain data emanating from multi-modal sensors. This architecture consists of a *fusion layer,* where data fusion of the input data streams occurs. The initial stages of the network operates on the lower and upper PrBounds in parallel, and it utilizes new activation functions that are more appropriate for PrBound pairs. It also incorporates a layer to increase the system resilience to sensor failures. With these innovations, the proposed DFN is able deal with uncertain data and deliver higher performance compared to conventional methods used in deep learning.

CHAPTER 2

Preliminaries

2.1 Basic Notation

We use $\mathbb{N}$ and $\mathbb{R}$ to denote the integers and reals, respectively. We will attach a subscript to these to restrict their domain of definition. For example, $\mathbb{R}_{\geq 0}$ denotes the non-negative reals; $\mathbb{R}_{[0,1]}$ denotes the reals taking values in $[0, 1]$.

We use Θ to denote the sample or state space of outcomes of an experiment. Given $A \subseteq \Theta$, $\overline{A}$ is its set theoretic complement (in Θ), i.e., $\overline{A} = \Theta \setminus A$, and $|A|$ is its cardinality. The power set of all possible subsets of Θ is denoted by $2^{\Theta} = \{A \mid A \subseteq \Theta\}$.

2.2 Interval-Valued (I.V.) Probabilities

Consider the probability space $(\Theta, \mathcal{X}, p)$, where $\mathcal{X}$ is an event space or σ-algebra (over Θ), and p is a probability measure. Then the *inner and outer measures* can be thought of as the "best" probability interval one may allocate to a non-measurable event [Fagin and Halpern, 1990].

When multiple probability measures are defined for the same sample and event space pair $(\Theta, \mathcal{X})$, they generate a *lower/upper probability envelope pair.* A notion that is easier to characterize is the *lower/upper probability function pair* which is defined as a primitive concept with no recourse to an underlying set of probability measures [Chateauneuf and Jaffray, 1989, Cozman, 1997, Huber and Ronchetti, 2009]:

Definition 1 (Lower/Upper Probability Function Pair) *A K-*monotone lower probability function *for $(\Theta, \mathcal{X})$ is any function $\mathbb{L}^{(K)}(\cdot) : \mathcal{X} \mapsto [0,1]$ s.t. $\mathbb{L}^{(K)}(\emptyset) = 0$, $\mathbb{L}^{(K)}(\Theta) = 1$, and*

(i) for $K = 1$: $\forall A_1, A_2 \subseteq \Theta$, $L^{(1)}(A_1) \leq L^{(1)}(A_2)$ whenever $A_1 \subseteq A_2$; and

(ii) for $K \geq 2$: $\forall A_i \subseteq \Theta$,

$$\mathbb{L}^{(K)}\left(\bigcup_{i=1}^{K} A_i\right) \geq \sum_{\substack{\mathcal{I} \subseteq \{1,\ldots,K\} \\ \mathcal{I} \neq \emptyset}} (-1)^{|\mathcal{I}|+1} \mathbb{L}^{(K)}\left(\bigcap_{i \in \mathcal{I}} A_i\right).$$

*The corresponding K-*monotone upper probability function $\mathbb{U}^{(K)}(\cdot) : \mathcal{X} \mapsto [0,1]$ *is $\mathbb{U}^{(K)}(A) = 1 - \mathbb{L}^{(K)}(\overline{A})$, $\forall A \subseteq \Theta$. A function which is K-monotone for all $K \geq 1$ is said to be* ∞-monotone. ■

One can think of 1- and ∞-monotonicity as the “weakest” and “strongest” monotonicity conditions, respectively, because a K-monotone function is K'-monotone for all $1 \leq K' \leq K$ [Chateauneuf and Jaffray, 1989]. It turns out that lower/upper probability envelopes are 1-monotone [Chrisman, 1996a] and inner measures are ∞-monotone [Choquet, 1954, Chateauneuf and Jaffray, 1989, Fagin and Halpern, 1990]. In practice, higher monotone probability functions are preferred because they can lead to mathematically more tractable and “cleaner” results [Chrisman, 1996a, Chrisman, 1996b].

Given a probability function pair $\{\mathbb{L}^{(K)}(\cdot), \mathbb{U}^{(K)}(\cdot)\}$, the p.m.f. $P(\cdot)$ is said to be *consistent* with it if [Chateauneuf and Jaffray, 1989]

$$0 \leq \mathbb{L}^{(K)}(A) \leq P(A) \leq \mathbb{U}^{(K)}(A) \leq 1, \ \forall A \subseteq \Theta. \tag{2.1}$$

An ∞-monotone pair is guaranteed to possess at least one consistent p.m.f. [Chateauneuf and Jaffray, 1989].

2.3 Demspter-Shafer (DS) Belief Theory

It turns out that ∞-monotone probability functions are essentially DS theoretic (DST) belief functions [Chateauneuf and Jaffray, 1989,Fagin and Halpern, 1990,Chrisman, 1996a,Choquet, 1954].

Suppose $\Theta = \{\theta_1, \ldots, \theta_N\}$ is a finite set of N mutually exclusive and exhaustive outcomes. In our work, we restrict our attention to the case of a finite number of outcomes, i.e., $|\Theta| = N$, $N \in \mathbb{N}_{\geq 0}$. The lowest level of discernible information is captured by the elementary outcomes $\theta_i \in \Theta$, which we refer to as *singletons.* In DS theory, Θ is usually called the *frame of discernment (FoD)* [Shafer, 1976].

2.3.1 Basic DST Notions

2.3.1.1 Basic Belief or Mass Assignment, Belief, Plausibility

Definition 2 (Basic Belief or Mass Assignment, Belief, Plausibility) *Consider the FoD Θ.*

(i) The mapping $m : 2^\Theta \mapsto [0,1] : A \mapsto m(A)$ is referred to as a basic belief assignment (BBA) *or* mass assignment *if*

$$\sum_{A \subseteq \Theta} m(A) = 1 \textit{ and } m(\emptyset) = 0.$$

(ii) The mapping $Bl : 2^\Theta \mapsto [0,1] : A \mapsto Bl(A)$, where

$$Bl(A) = \sum_{B \subseteq A} m(B),$$

is referred to as the corresponding belief function.

(iii) The mapping $Pl : 2^{\Theta} \mapsto [0,1] : A \mapsto Pl(A)$*, where*

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B),$$

is referred to as the corresponding plausibility function. ■

Remarks:

1. The mass $m(A)$ can be thought of as a measure of the "support" being assigned to the proposition A, and A only. On the other hand, the belief $Bl(A)$ can be thought of as the support for all propositions that are certain to be implied by A, and the plausibility $Pl(A)$ can be thought of as the support for all propositions that may be implied by A. Alternately, $Bl(A)$ represents the total support committed to A without also being committed to its complement $\overline{A}$, and $Pl(A)$ represents the total belief that does not contradict A.

2. Note that,

$$0 \leq Bl(A) \leq Pl(A) \leq 1, \; \forall A \subseteq \Theta, \tag{2.2}$$

and

$$Bl(A) + Pl(\overline{A}) = 1, \; \forall A \subseteq \Theta. \tag{2.3}$$

3. The interval $Un(A) = [Bl(A), Pl(A)]$ is referred to as the *uncertainty interval* associated with $A \subseteq \Theta$; $|Un(A)| = Pl(A) - Bl(A)$ is the *uncertainty interval width.*

2.3.1.2 Focal Elements, Core, Body of Evidence

A *focal element* is a proposition that receives non-zero mass and the *core* $\mathfrak{F}$ is the set of focal elements, i.e,

$$\mathfrak{F} = \{A \subseteq \Theta \mid m(A) > 0\}. \tag{2.4}$$

A focal element can be a singleton or a non-singleton. For instance, the mass $m(\theta_i, \theta_j)$ allocated to the doubleton (θ_i, θ_j), $\theta_i, \theta_j \in \Theta$, represents ignorance or lack of evidence to differentiate between the occurrence of singleton θ_i or θ_j. The *body of evidence (BoE)* is the triple $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$.

The *vacuous BoE* 1_Θ, which has Θ as its only focal element, captures the state of complete ignorance. A BoE is called *Bayesian* (or *probabilistic*) if its core consists of only singletons. For a Bayesian BoE, the BBA, belief, and plausibility, all reduce to the same probability (i.e., p.m.f.) assignment.

One can find a consistent probability mass function (p.m.f.) for the DST pair $\{Bl(\cdot), Pl(\cdot)\}$ because it is ∞-monotone [Chateauneuf and Jaffray, 1989]. One such consistent p.m.f. is the *pignistic probability* [Smets, 1999]

$$BetP(\theta_i) = \sum_{\theta_i \in A \subseteq \Theta} \frac{m(A)}{|A|},\ \theta_i \in \Theta. \tag{2.5}$$

2.3.1.3 Computational Considerations

The expressive power wielded by the DST framework demands a high computational cost. For a given FoD Θ, where $|\Theta| = N$, a DST model allocates $2^N - 2$ mass assignments; in contrast, only $N - 1$ probability assignments are required for a p.m.f. Much advances have been made for the purpose of mitigating the associated computational burden, e.g., [Bauer, 1997, Wilson, 2001, Wickramarathne et al., 2013], and the more recent work in [Polpitiya et al., 2016, Polpitiya et al., 2017].

A special DST model which retains the ability to capture complete ignorance with only a slight increase in computational complexity is the *Dirichlet BoE* (so named because of its close relationship with Dirichlet probability distributions [Josang, 2010]). The core of a Dirichlet BoE can only consist of the singletons $\{\theta_i\}$ and Θ only, thus requiring only N mass assignments.

2.3.2 Conditioning

The conditional operation plays perhaps the most pivotal role in evidence updating and fusion, and in general, in reasoning under uncertainty. When it comes to DST functions, notable among the various conditional notions that have been proposed over the years are the *Dempster's conditional* [Shafer, 1976, Klawonn and Smets, 1992, Nguyen and Smets, 1993, Xu and Smets, 1996, Smets, 2002] and the *Fagin-Halpern (FH) conditional* [Fagin and Halpern, 1990].

Definition 3 *Consider the BoE $\mathcal{E} = \{\Theta, \mathfrak{F}, m\}$ and $B \subseteq \Theta$.*

(i) For $Pl(B) > 0$, the Dempster's conditional belief and plausibility of A given B *respectively are [Shafer, 1976]*

$$Bl(A\!:\!B) = \frac{Bl(A \cup \overline{B}) - Bl(\overline{B})}{Pl(B)}; \quad Pl(A\!:\!B) = \frac{Pl(A \cap B)}{Pl(B)}.$$

(ii) For $Bl(B) > 0$, the Fagin-Halpern (FH) conditional belief and plausibility of A given B *respectively are [Fagin and Halpern, 1990]*

$$Bl(A|B) = \frac{Bl(A \cap B)}{Bl(A \cap B) + Pl(\overline{A} \cap B)}; \quad Pl(A|B) = \frac{Pl(A \cap B)}{Pl(A \cap B) + Bl(\overline{A} \cap B)}. \quad \blacksquare$$

The Dempster's conditional and the FH conditional are related as [Fagin and Halpern, 1990].

$$Bl(A|B) \leq Bl(A\!:\!B) \leq Pl(A\!:\!B) \leq Pl(A|B). \tag{2.6}$$

It is well-known that the Dempster's conditional may not be reconcilable with probability theory [Smets, 1992, Smets, 1994, Smets, 1999, Heendeni et al., 2016]. On the other hand, of the various notions of DST conditionals that abound in the literature, the FH conditional offers a unique probabilistic interpretation and constitutes a natural transition to the Bayesian conditional notion because of its close connection with the inner/outer conditional probability measures [Fagin and Halpern, 1990, Wickramarathne et al., 2013]. The *conditional approach,* a newer strategy for updating and fusion of DST evidence, is in fact based on this FH conditional [Premaratne et al., 2009, Wickramarathne et al., 2011b, Wickramarathne et al., 2011a, Ferdous et al., 2012, Sarathy et al., 2017, Shi et al., 2017, Zhang et al., 2017].

2.3.2.1 Computational Considerations

When it comes to the Dempster's conditional, a thorough discussion on how one may carry out the conditional computation is provided in [Klawonn and Smets, 1992, Smets, 2002]. As for the FH conditional, The conditional core theorem [Wickramarathne et al., 2013] can be utilized to directly identify the conditional focal elements. Perhaps the first work that directly deals with the FH conditional computation appears in [Polpitiya et al., 2017].

CHAPTER 3

Imperfect Implication Rules

In this chapter we describe how imperfect implication rules can be modeled so that they could be utilized within rule-based systems. This work also forms the basis on which one may develop systems that work with imperfect logic.

3.1 Model

3.1.1 Rule Uncertainty

Consider the implication rule $R:\ A \Longrightarrow B$, where A denotes the antecedent, B denotes the consequent, and $\Longrightarrow$ denotes the implication. In situations where A and B belong to two different BoEs, we assume that a common BoE has been established (e.g., the cross-product BoE) so that both the antecedent and consequent belong to the same BoE.

As for the uncertainty interval associated with the rule $R:\ A \Longrightarrow B$ itself, let us explore using the quantities

$$Bl(R) = Bl(B|A); \quad Pl(R) = Pl(B|A). \tag{3.1}$$

With the results we develop, we argue that these conditionals $Bl(R) = Bl(B|A)$ and $Pl(R) = Pl(B|A)$ capture the uncertainty associated with the rule $R:\ A \Longrightarrow B$

reasonably well. We also show that, if we have additional evidence regarding the rule $\overline{R}: \overline{A} \Longrightarrow B$, we may use the uncertainty associated with $\overline{R}: \overline{A} \Longrightarrow B$ (captured via $Bl(B|\overline{A})$ and $Pl(B|\overline{A})$) to obtain more refined results regarding the uncertainty interval associated with the consequent B of $R: A \Longrightarrow B$.

3.1.2 Consequent Uncertainty

What is the uncertainty interval associated with the consequent B of the rule $R: A \Longrightarrow B$ given the uncertainty intervals associated with its antecedent A and the rule R itself? To proceed, let us use the following notation:

$$\begin{aligned} Bl(A) &= \alpha_1; \; Pl(A) &= \beta_1; \; Bl(B) &= \alpha_2; \; Pl(B) &= \beta_2; \\ Bl(B|A) &= \alpha_R; \; Pl(B|A) = \beta_R; \; Bl(B|\overline{A}) &= \alpha_{\overline{R}}; \; Pl(B|\overline{A}) &= \beta_{\overline{R}}. \end{aligned} \tag{3.2}$$

We then have the following result:

Theorem 1 (Consequent Uncertainty: General Bounds) *The uncertainty interval $[\alpha_2, \beta_2]$ associated with the consequent B of the rule $R: A \Longrightarrow B$ satisfies*

$$0 \leq \alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}} \leq \alpha_2 \leq \beta_2 \leq \alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + Un(A) \leq 1.$$

Here, $[\alpha_R, \beta_R]$ and $[\alpha_{\overline{R}}, \beta_{\overline{R}}]$ refer to the uncertainty intervals associated with the rules $R: A \Longrightarrow B$ and $\overline{R}: \overline{A} \Longrightarrow B$, respectively, and $Un(A) = (\beta_1 - \alpha_1)$ is the uncertainty interval width associated with the antecedent A. □

Proof. Note that

$$Bl(B) = Bl(B \cap A) + Bl(B \cap \overline{A}) + \sum_{\substack{\emptyset \neq P \subseteq (B \cap A) \\ \emptyset \neq Q \subseteq (B \cap \overline{A})}} m(P \cup Q), \tag{3.3}$$

From (3.3), we get

$$Bl(B \cap A) + Bl(B \cap \overline{A}) \leq Bl(B). \tag{3.4}$$

We also know that $Bl(A) \leq Bl(B \cap A) + Pl(\overline{B} \cap A)$ [Kulasekere et al., 2004] . This, together with the FH conditionals (where $Bl(A) \neq 0$), then lead us to

$$Bl(A)\, Bl(B|A) \leq Bl(B \cap A). \tag{3.5}$$

This inequality holds true for $Bl(A) = 0$ as well. Substitute $\overline{A}$ for A in (3.5):

$$Bl(\overline{A})\, Bl(B|\overline{A}) \leq Bl(B \cap \overline{A}). \tag{3.6}$$

Upper bound on α_2: Use (3.4), (3.5), and (3.6), and use the fact that $Bl(\overline{A}) = 1 - Pl(A)$, and the notation in (3.2) to get

$$\alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}} \leq \alpha_2.$$

It is easy to verify that $0 \leq \alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}}$.

Lower bound on β_2: Substitute $\overline{B}$ for B in (3.4),(3.5), and (3.6):

$$Bl(\overline{B}) \geq Bl(A)\, Bl(\overline{B}|A) + Bl(\overline{A})\, Bl(\overline{B}|\overline{A}).$$

Use $Bl(\overline{B}) = 1 - Pl(B)$, $Bl(\overline{B}|A) = 1 - Pl(B|A)$, and $Bl(\overline{B}|\overline{A}) = 1 - Pl(B|\overline{A})$, and the notation in (3.2) to get

$$\beta_2 \leq \alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + (\beta_1 - \alpha_1).$$

It is easy to verify that $\alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + (\beta_1 - \alpha_1) \leq 1$. ■

Remarks:

1. Table 3.1 illustrates how the uncertainty interval of the consequent as given by the general bounds in Theorem 1 are affected by the incorporation/removal of information on each of the rules $R:\ A \Longrightarrow B$ and/or $\overline{R}:\ \overline{A} \Longrightarrow B$.

Table 3.1: Uncertainty Interval Associated With the Rule Consequent: Effect of the Rules $R: \; A \Longrightarrow B$ and $\overline{R}: \; \overline{A} \Longrightarrow B$.
Note: Columns 3 and 4 give bounds (not exact values) for α_2 and β_2, respectively.

Available Information	**Parameters**	**Lower Bound on α_2**	**Upper Bound on β_2**
None	$\alpha_R = \alpha_{\overline{R}} = 0;$ $\beta_R = \beta_{\overline{R}} = 1$	0	1
Ralaxed Bounds: R only	$\alpha_{\overline{R}} = 0;$ $\beta_{\overline{R}} = 1$	$\alpha_1 \alpha_R$	$1 - \alpha_1(1 - \beta_R)$
$\overline{R}$ only	$\alpha_R = 0;$ $\beta_R = 1$	$(1 - \beta_1)\, \alpha_{\overline{R}}$	$1 - \beta_1(1 - \beta_{\overline{R}})$
General Bounds: Both R and $\overline{R}$		$\alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}}$	$\alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + Un(A)$

(a) Note that the bounds corresponding to *both* rules cannot be wider than the bounds corresponding to *one* rule, i.e., the incorporation of more information allows us to narrow the uncertainty interval $[\alpha_2, \beta_2]$ of the consequent.

(b) We will refer to the bounds obtained when information regarding only the rule R is available as *relaxed bounds,* viz.,

$$0 \leq \alpha_1 \alpha_R \leq \alpha_2 \leq \beta_2 \leq 1 - \alpha_1(1 - \beta_R) \leq 1. \tag{3.7}$$

2. We define the least commitment (LC) choice; that we are rely on the available intervals which are wider than the actual intervals. Accordingly, with the LC choice, we may select the following values for α_2 and β_2 (therefore when LC choice is used, α_2 and β_2 provide not exact values, but lower and upper boundaries for $Bl(B)$ and $Pl(B)$ respectively):

   (a) *General bounds:*

$$\alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}} = \alpha_2 \leq \beta_2 = \alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + (\beta_1 - \alpha_1). \tag{3.8}$$

   (b) *Relaxed bounds:*

$$\alpha_1 \alpha_R = \alpha_2 \leq \beta_2 = 1 - \alpha_1(1 - \beta_R). \tag{3.9}$$

   As a comparison, the work in [Benavoli et al., 2008] gives $\alpha_1 \alpha_R = \alpha_2 \leq \beta_2 = 1 - (1 - \beta_1)(1 - \beta_R)$.

Since the relaxed bounds in (3.7) considers the implication $R: A \Longrightarrow B$ only, we propose to employ these to capture the uncertainty associated with the rule R:

Definition 4 (Consequent Uncertainty of an Implication Rule) *Consider the implication rule $R: A \Longrightarrow B$ where the uncertainty associated with the rule R and its* *antecedent A are $[\alpha_R, \beta_R]$ and $[\alpha_1, \beta_1]$, respectively. Then the uncertainty associated with the consequent B is $[\alpha_2, \beta_2]$, where*

$$\alpha_1\alpha_R \leq \alpha_2 \leq \beta_2 \leq 1 - \alpha_1(1 - \beta_R).$$

With the LC choice, we use

$$\alpha_1\alpha_R = \alpha_2 \leq \beta_2 = 1 - \alpha_1(1 - \beta_R). \qquad \blacksquare$$

3.1.2.1 Interpretation of the Consequent Uncertainty Interval

The general bounds in Theorem 1 yield the following upper bound on the consequent uncertainty interval $Un(B) = \beta_2 - \alpha_2$:

$$Un(B) \leq Un(A) + \alpha_1 Un(R) + (1 - \beta_1)\, Un(\overline{R}), \tag{3.10}$$

where $Un(R) = \beta_R - \alpha_R$ and $Un(\overline{R}) = \beta_{\overline{R}} - \alpha_{\overline{R}}$ are the uncertainty intervals associated with the rules R and $\overline{R}$, respectively, and $Un(A) = \beta_1 - \alpha_1$ is the antecedent uncertainty interval. Note that $0 \leq Un(B) \leq 1$.

This upper bound of the consequent uncertainty interval has an interesting intuitive interpretation: the uncertainty interval $Un(B)$ of the consequent is bounded above by the uncertainty interval $Un(A)$ of the antecedent plus the uncertainty intervals of the rules $R: \; A \Longrightarrow B$ and $\overline{R}: \; \overline{A} \Longrightarrow B$ weighted by their corresponding belief terms $Bl(A) = \alpha_1$ and $Bl(\overline{A}) = 1 - \beta_1$, respectively.

3.2 Consistency With Probability

For p.m.f.s, notice the following:

(i) The last summation term in (3.3) is absent. Therefore, the inequality (3.4) reduces to an equality, thus yielding

$$Pr(B) = Pr(B \cap A) + Pr(B \cap \overline{A}). \tag{3.11}$$

(ii) The inequality in (3.5) reduces to an inequality, thus yielding

$$Pr(A)\, Pr(B|A) = Pr(B \cap A). \tag{3.12}$$

So, continuing through the proof of Theorem 1, we get the following "equalities" instead of the "inequalities" in Theorem 1 for α_2 and β_2:

$$0 \leq \alpha_1 \alpha_R + (1 - \beta_1)\, \alpha_{\overline{R}} = \alpha_2 \leq \beta_2 = \alpha_1 \beta_R + (1 - \beta_1)\, \beta_{\overline{R}} + (\beta_1 - \alpha_1) \leq 1. \tag{3.13}$$

Note that these correspond to the LC choice associated with the general bounds (see (3.8)).

We now obtain Table 3.2 which illustrates the situation when the antecedent and/or the rules are probabilistic. Note the following:

(i) When the rules are probabilistic, the uncertainty interval width of the antecedent propagates through to the consequent.

(i) When both the antecedent and the rules are probabilistic, so is the consequent, and

$$\alpha_2 = \beta_2 = \alpha_1 \alpha_R + (1 - \alpha_1)\, \alpha_{\overline{R}} \tag{3.14}$$

corresponds to the probabilistic relationship

$$Pr(B) = Pr(A)\, Pr(B|A) + Pr(\overline{A})\, Pr(B|\overline{A}). \tag{3.15}$$

Table 3.2: Uncertainty Interval Associated With the Rule Consequent: Probabilistic Case.
Note: Columns 3 and 4 give exact values (not bounds) for α_2 and β_2, respectively.

Probabilistic Information	**Parameters**	α_2 **Value**	β_2 **Value**
Antecedent	$\alpha_1 = \beta_1$	$\alpha_1\alpha_R + (1-\alpha_1)\,\alpha_{\overline{R}}$	$\alpha_1\beta_R + (1-\alpha_1)\,\beta_{\overline{R}}$
Rules	$\alpha_R = \beta_R$; $\alpha_{\overline{R}} = \beta_{\overline{R}}$	$\alpha_1\alpha_R + (1-\beta_1)\,\alpha_{\overline{R}}$	$\alpha_1\alpha_R + (1-\beta_1)\,\alpha_{\overline{R}} + Un(A)$
Both antecedent and rules	$\alpha_1 = \beta_1$; $\alpha_R = \beta_R$ $\alpha_{\overline{R}} = \beta_{\overline{R}}$	$\alpha_1\alpha_R + (1-\alpha_1)\,\alpha_{\overline{R}}$	$\alpha_1\alpha_R + (1-\alpha_1)\,\alpha_{\overline{R}}$

3.3 Consistency with Classical Logic

To explore the relationship between our implication rule model and what classic logic yields, we associate the two cases $\alpha_1 = \beta_1 = 1$ and $\alpha_1 = \beta_1 = 0$ with the logical **True** and logical **False** in classical logic. For example, we may interpret $\alpha_1 = \beta_1 = 1$ and $\alpha_1 = \beta_1 = 0$ as the occurrence or non-occurrence of proposition A with 100% confidence.

3.3.1 Classical Logic

With $\alpha_1 = \beta_1 = \{0, 1\}$ and $\alpha_2 = \beta_2 = \{0, 1\}$, Table 3.3 shows the truth table for $R:\ A \Longrightarrow B$ in classical logic.

Table 3.3: Truth Table for $A \Longrightarrow B$ in Classical Logic

A	B	$A \Longrightarrow B$ $(\alpha_R = \beta_R)$
0	0	1
0	1	1
1	0	0
1	1	1

3.3.2 Proposed Rule Model

With $\alpha_1 = \beta_1 = \{0, 1\}$, $\alpha_R = \beta_R = \{0, 1\}$, and $\alpha_{\overline{R}} = \alpha_{\overline{R}} = \{0, 1\}$ in Theorem 1, Table 3.4 shows the truth table for $R: A \Longrightarrow B$ associated with the proposed rule model.

Table 3.4: Truth Table for $R: A \Longrightarrow B$ Associated With the Proposed Rule Model. *Note:* Information regarding both $R: A \Longrightarrow B$ and $\overline{R}: \overline{A} \Longrightarrow B$ are assumed to be available.

$\alpha_1 = \beta_1$	$\alpha_R = \beta_R$	$\alpha_{\overline{R}} = \beta_{\overline{R}}$	$\alpha_1\alpha_R + (1 - \beta_1)\,\alpha_{\overline{R}}$	$\alpha_1\beta_R + (1 - \beta_1)\,\beta_{\overline{R}}$ $+Un(A)$	$\alpha_2 = \beta_2$
0	0	0	0	0	0
0	0	1	1	1	1
0	1	0	0	0	0
0	1	1	1	1	1
1	0	0	0	0	0
1	0	1	0	0	0
1	1	0	1	1	1
1	1	1	1	1	1

Remarks:

1. Similar to the probabilistic case, the bounds in Theorem 1 reduce to equalities in the classical logic case.

2. The FH conditionals are not defined when $\alpha_1 = \beta_1 = 0$ [Fagin and Halpern, 1990]. However, by taking a limiting argument where $\alpha_1 = \beta_1 \to 0$, it is easy to show that the bounds in Theorem 1 remain valid with the formal subsstitution of $\alpha_1 = \beta_1 = 0$.

The truth table in Table 3.4 can be expressed as

$$\begin{aligned}(A \wedge (A \Longrightarrow B)) \vee (\neg A \wedge (\neg A \Longrightarrow B)) &= (A \wedge (\neg A \vee B)) \vee (\neg A \wedge (A \vee B)) \\ &= B. \end{aligned} \tag{3.16}$$

With $\alpha_1 = \beta_1 = \{0, 1\}$ and $\alpha_R = \beta_R = \{0, 1\}$ (and assuming that information regarding the rule $\overline{R} :\ \overline{A} \Longrightarrow B$ is unavailable) in Definition 4, Table 3.5 shows the truth table for $R :\ A \Longrightarrow B$ associated with the proposed rule model.

Table 3.5: Truth Table for $R :\ A \Longrightarrow B$ Associated With the Proposed Rule Model. *Note:* Information regarding only $R :\ A \Longrightarrow B$ is assumed to be available.

$\alpha_1 = \beta_1$	$\alpha_R = \beta_R$	$\alpha_1\alpha_R$	$1 - \alpha_1(1 - \beta_R)$	α_2	β_2
0	0	0	1	0	1
0	1	0	1	0	1
1	0	0	0	0	0
1	1	1	1	1	1

Let us compare the entries of Table 3.5 (obtained from the relaxed bounds in (3.7) or Definition 4) and Table 3.3 (obtained from Table 3.3, the truth table for $A \Longrightarrow B$ in classical logic).

(a) *Antecedent is true:* See lines 3-4, where both tables show identical behavior.

(b) *Antecedent is false:* See lines 1-2, where the tables behave differently.

(b.1) *When rule is true:* The consequent can take 0 or 1 in both tables. Note that the information in lines 1-2 of Table 3.3 are captured in line 2 of Table 3.5 which explains that, when the antecedent is false and the implication rule is true, the consequent can be either true or false.

(b.2) *When rule is false:* This case does not appear in Table 3.3 whereas line 1 of Table 3.5 not only allows this, but it also allows the consequent to be either true of false.

Tables 3.4 and 3.5 show that the model we have proposed is consistent with classical logic. In addition, it provides a better explanation of the behavior of implication rules than what the classical logic model in Table 3.3 is able to provide.

In concluding this section, we wish to point out the following result which we can use to refine the uncertainty interval associated with the consequent:

Lemma 1 *Regarding the FH conditional,*

$$\min\{Bl(B|A), Bl(B|\overline{A})\} \leq Bl(B) \leq Pl(B) \leq \max\{Pl(B|A), Pl(B|\overline{A})\}. \qquad \square$$

Proof: For convenience, let

$$M = \sum_{\substack{\emptyset \neq P \subseteq (B \cap A) \\ \emptyset \neq Q \subseteq (B \cap \overline{A})}} m(P \cup Q).$$

Then,

$$\begin{aligned}
Bl(B) &= Bl(B \cap A) + Bl(B \cap \overline{A}) + M \\
&= Bl(B|A)\,[Bl(B \cap A) + Pl(\overline{B} \cap A)] + Bl(B|\overline{A})\,[Bl(B \cap \overline{A}) + Pl(\overline{B} \cap \overline{A})] \\
&\qquad + M \\
&\geq \min\{Bl(B|A), Bl(B|\overline{A})\}\,[Bl(B \cap A) + Pl(\overline{B} \cap A) + Bl(B \cap \overline{A}) \\
&\qquad\qquad + Pl(\overline{B} \cap \overline{A})] + M \\
&\geq \min\{Bl(B|A), Bl(B|\overline{A})\}\,[Bl(B \cap A) + Pl(\overline{B} \cap A) + Bl(B \cap \overline{A}) \\
&\qquad\qquad + Pl(\overline{B} \cap \overline{A}) + M] \\
&= \min\{Bl(B|A), Bl(B|\overline{A})\}\,[Bl(B) + Pl(\overline{B} \cap A) + Pl(\overline{B} \cap \overline{A})].
\end{aligned}$$

But, we know that

$$Pl(\overline{B} \cap A) + Pl(\overline{B} \cap \overline{A}) \geq Pl(\overline{B}).$$

So,

$$\begin{aligned}
Bl(B) &\geq \min\{Bl(B|A), Bl(B|\overline{A})\}\,[Bl(B) + Pl(\overline{B})] \\
&= \min\{Bl(B|A), Bl(B|\overline{A})\}.
\end{aligned}$$

Substitute $\overline{B}$ instead of B in this lower bound $Bl(B)$:

$$Bl(\overline{B}) \geq \min\{Bl(\overline{B}|A), Bl(\overline{B}|\overline{A})\}.$$

So,

$$1 - Bl(\overline{B}) \leq 1 - \min\{Bl(\overline{B}|A), Bl(\overline{B}|\overline{A})\},$$

i.e.,

$$\begin{aligned}
Pl(B) &\leq \max\{1 - Bl(\overline{B}|A), 1 - Bl(\overline{B}|\overline{A})\} \\
&= \max\{Pl(B|A), Pl(B|\overline{A})\}. \qquad \blacksquare
\end{aligned}$$

We may now use Lemma 1 to get

$$\min\{\alpha_R, \alpha_{\overline{R}}\} \leq \alpha_2 \leq \beta_2 \leq \max\{\beta_R, \beta_{\overline{R}}\}, \tag{3.17}$$

which are also consistent with both Tables 3.4 and 3.5. So, together with the general bounds in Theorem 1, one may employ the narrower bounds

$$\begin{aligned}
0 &\leq \max\{\alpha_1 \alpha_R + (1-\beta_1)\, \alpha_{\overline{R}}, \min\{\alpha_R, \alpha_{\overline{R}}\}\} \\
&\leq \alpha_2 \leq \beta_2 \\
&\qquad \leq \min\{\alpha_1 \beta_R + (1-\beta_1)\, \beta_{\overline{R}} + Un(A), \max\{\beta_R, \beta_{\overline{R}}\}\} \leq 1.
\end{aligned} \tag{3.18}$$

3.4 Illustrative Example

As an illustrative simple example, consider 3 urns A, B, and C containing red and black balls. Table 3.6 describes the contents of each urn.

Table 3.6: 3-Urn Example: Contents of the Urns A, B, and C

	Red Balls	**Black Balls**	**Red or Black Balls**
Urn	**(RB)**	**(BB)**	
A	3	5	2
B	5	2	3
C	3	5	2

Consider the following experiment consisting of two trials:

(a) *Trial 1:* Select urn A and randomly take out a ball.

(b) *Trial 2:* If in Trial 1 we get a red ball (RB), take out a ball from urn B; otherwise, if we get a black ball (BB), take out a ball from urn C.

What are the belief and plausibility values of getting a RB in Trial 2?

In the DST framework, consider the following belief/plausibility pairs:

$$\begin{aligned}
[\alpha_1, \beta_1] \; &= \text{drawing a RB in Trial 1;} \\
[\alpha_R, \beta_R] &= \text{drawing a RB in Trial 2 given that Trial 1 yields a RB;} \\
[\alpha_2, \beta_2] \; &= \text{drawing a RB in Trial 2.}
\end{aligned} \tag{3.19}$$

Therefore,

$$\begin{aligned}
[\alpha_1, \beta_1] \; &= [0.3, 0.5]; \;\; [\overline{\alpha}_1, \overline{\beta}_1] \; = [0.5, 0.7]; \\
[\alpha_R, \beta_R] &= [0.5, 0.8]; \;\; [\alpha_{\overline{R}}, \beta_{\overline{R}}] = [0.3, 0.5].
\end{aligned} \tag{3.20}$$

Then, accounting for all the possibilities, we get

$$[\alpha_2, \beta_2] = [0.36, 0.65]. \tag{3.21}$$

Table 3.7 show what our results yield. Note that the general bounds are much narrower than the relaxed bounds (which ignore the information in $[\alpha_{\overline{R}}, \beta_{\overline{R}}]$).

Table 3.7: 3-Urn Example: Results

General bounds in Theorem 1:	$0.30 \leq \alpha_2 \leq \beta_2 \leq 0.69$
Relaxed bounds in Definition 4:	$0.15 \leq \alpha_2 \leq \beta_2 \leq 0.94$
Bounds in (3.17):	$0.30 \leq \alpha_2 \leq \beta_2 \leq 0.80$
Combined bounds in (3.18):	$0.30 \leq \alpha_2 \leq \beta_2 \leq 0.69$
True bounds in (3.21):	$0.36 \leq \alpha_2 \leq \beta_2 \leq 0.65$

3.5 Summary

Given the belief and plausibility pairs corresponding to the antecedent A and the implication rule $R: \; A \Longrightarrow B$, the work in this chapter provides bounds for the uncertainty interval of the rule' consequent B. We also show that the uncertainty interval of the consequent B can be further tightened by incorporating information, if available, from the rule pair $R:\ A \Longrightarrow B$ and $\overline{R}:\ \overline{A} \Longrightarrow B$.

The difficulty of capturing rule uncertainty via probability (e.g., using the conditional $Pr(B|A)$) is well documented [Lewis, 1976, Benavoli et al., 2008]. Instead, for our work, we model the implication rule in terms of the DST FH conditional pair $[Bl(B|A), Pl(B|A)]$. The justification of this model comes from the fact that probability and classical logic models emerge as special cases of the proposed model. Therefore, our model can be considered more general and more flexible than what appears in previous works.

CHAPTER 4

PrBounds: A Framework Based on Probability Bounds

4.1 Introductory Remarks

As mentioned in Section 1.4.3, our PrBound notion differ from the existing work of interval valued probabilities. The main contributions of our work, which essentially yields a framework for carrying out learning and reasoning with i.v. probabilities from the vantage point of a single underlying distribution, are the following:

(a) No monotonicity or other constraint is imposed on the PrBound pairs (Definition 5). This allows one to carry out critical operations (e.g., evidence updating in Bayesian inference) without having to ensure that intermediate results satisfy the same constraints, i.e., it is not necessary to ensure that operations are "closed under monotonicity" (Section 4.2.2.2).

(b) We develop PrBounds for conditionals in terms of the PrBounds of the underlying unconditioned probability distribution (Theorem 2). At first glance, these expressions we derive may appear trifling in that they take the same "structural" form as the many different i.v. conditional notions that have appeared elsewhere. But these existing i.v. conditional notions require the lower bound of the interval associated with the conditioning proposition to be strictly positive, viz., the expressions for $\{P_L(A|B), P_U(A|B)\}$, the lower/upper bounds for the conditional probability $P(A|B)$, apply only when $P_L(B) > 0$ [Fagin and Halpern, 1990, Zaffalon, 2002a]. However, when one presupposes the existence of a single underlying distribution, i.v. conditionals should exist as long as $P(B) > 0$ (note that, while $P(B) = 0$ implies $P_L(B) = 0$, the converse is not necessarily true). Our i.v. conditional expressions are indeed valid whenever $P(B) > 0$ (Theorem 2 and Corollary 1). To our knowledge, such expressions have not appeared elsewhere (see [Walley, 1991] for some comments regarding this exact issue.).

(c) Our vantage point allows us to take the stance that independence among variables is governed by *the* underlying distribution, and not by any i.v. probability notion; rather, it emerges from the underlying distribution (Theorem 3). In contrast, when one presupposes a *set* of underlying probabilities, one is obliged to resort to *defining* independence in terms of surrogate notions involving lower/upper bounds associated with i.v. probabilities, thus generating notions that are neither consistent with each other nor rooted in probability. Indeed, epistemic independence associated with the behavioral interpretation [de Cooman et al., 2010, de Cooman et al., 2011, Augustin et al., 2014] may not reflect probabilistic independence when dealing with imperfect data (Example 5). So the argument that Bayesian sensitivity analysis interpretation may be "unnecessary" [Walley, 1996] has to be stated with the caveat to exclude the case of a single underlying distribution.

(d) The absence of additional constraints on PrBound pairs and, more importantly, how the notion of independence is viewed, allow us to generalize graphical networks in that they operate much like their conventional counterparts [Pearl, 1988] but with PrBound pairs instead of probabilities (Section 4.3).

(e) The convenience, intuitiveness, and versatility that Dempster-Shafer (DS) belief theoretic models offer can be harnessed to capture a wide variety of data imperfections in a principled manner [Blackman and Popoli, 1999]. The proposed framework has the capability to learn PrBounds from more general evidential datasets which are based on these DS theoretic (DST) models [Anand et al., 1996, Vannoorenberghe, 2004, Hewawasam et al., 2007, Wickramarathne et al., 2011a] (Section 5.2.1). However, because Dempster's Combination Rule (DCR), the popular DST fusion strategy [Shafer, 1976], may not be reconcilable with probability [Smets, 1992, Smets, 1994, Smets, 1999, Heendeni et al., 2016, Núñez et al., 2018], we avoid the DCR and its derivatives (e.g., the Dempster's conditional [Shafer, 1976]). This sets our work apart from work on DS theory, the transferrable belief model, and Bayesian inference generalizations [Dempster, 1967, Dempster, 1968, Shafer, 1976, Smets, 1994].

(f) Previous work has explored how i.v. bounds could be learned in the presence of a type of uncertainty that is most commonly encountered in real datasets, viz., attribute values that are unknown or missing or that are known to lie within a set of values but otherwise cannot be discerned further [Tassem, 1992, Cozman, 2000, Zaffalon, 2002b, Zaffalon, 2002a, Augustin et al., 2014]. We also explore this scenario (Definition 7), but from the vantage point of a single underlying distribution. For such datasets, we give a computationally efficient and more tractable way to compute the PrBounds for each attribute and each data record (Corollary 3 and Lemma 6) and an intuitive frequency counting method to learn the PrBound and conditional PrBound parameters (Section 5.3.2.1). The probabilities associated with an *arbitrary* imputation strategy, including the underlying "true" probabilities, are guaranteed to be contained within the PrBounds so learned (Section 5.3.2.2). This bestows a clear meaning to the PrBounds thus introducing essentially a caveat to the argument that Bayesian analysis interpretation of i.v. probabilities may be unable to offer a "useful meaning" to the underlying *set* of distributions [Walley, 1996].

Our approach does share several similarities with existing imprecise probability approaches. For instance, upper/lower envelope computational methods, enumeration-optimization algorithms, and decision making strategies employed within credal set approaches to get bounds related to the credal set of distributions [Tassem, 1992,Cozman, 2000,Zaffalon, 2002b,Zaffalon, 2002a,Augustin et al., 2014] can also be employed within the proposed framework, but now to get bounds associated with the single underlying distribution.

4.2 Bounding the Probabilities

4.2.1 PrBounds

Definition 5 (PrBound Pairs) *Suppose* $P(\cdot)$ *is a p.m.f. defined over* Θ*. A* lower PrBound for $P(\cdot)$ *is any function* $\mathbb{L}(\cdot) : 2^{\Theta} \mapsto [0, 1]$ *s.t.* $\mathbb{L}(\emptyset) = 0$*,* $\mathbb{L}(\Theta) = 1$*, and* $\mathbb{L}(A) \leq P(A)$*,* $\forall A \subseteq \Theta$*; an* upper PrBound for $P(\cdot)$ *is any function* $\mathbb{U}(\cdot) : 2^{\Theta} \mapsto [0, 1]$

s.t. $\mathbb{U}(\emptyset) = 0$, $\mathbb{U}(\Theta) = 1$, *and* $P(A) \leq \mathbb{U}(A)$, $\forall A \subseteq \Theta$. *Then,* $\{\mathbb{L}(A), \mathbb{U}(A)\}$ *is said to be a* PrBound pair *for* $P(A)$, *and we denote as* $P(A) \triangleleft \{\mathbb{L}(\cdot), \mathbb{U}(\cdot)\}$.

We say that the PrBound pair $\{\mathbb{L}''(\cdot), \mathbb{U}''(\cdot)\}$ *is* narrower *than the PrBound pair* $\{\mathbb{L}'(\cdot), \mathbb{U}'(\cdot)\}$ *at* $A \subseteq \Theta$ *if* $\mathbb{L}'(A) \leq \mathbb{L}''(A)$ *and* $\mathbb{U}'(A) \geq \mathbb{U}''(A)$. ■

Remarks.

1. Suppose $P(A) \triangleleft \{\widetilde{\mathbb{L}}(A), \widetilde{\mathbb{U}}(A)\}$ and $P(\overline{A}) \triangleleft \{\widetilde{\mathbb{L}}(\overline{A}), \widetilde{\mathbb{U}}(\overline{A})\}$. But the fact $P(A) + P(\overline{A}) = 1$ yields $1 - \widetilde{\mathbb{U}}(\overline{A}) \leq P(A) \leq 1 - \widetilde{\mathbb{L}}(\overline{A})$ and $1 - \widetilde{\mathbb{U}}(A) \leq P(\overline{A}) \leq 1 - \widetilde{\mathbb{L}}(A)$, meaning that we can use the narrower bound pairs

$$\begin{aligned} \{\mathbb{L}(A), \mathbb{U}(A)\} &= \Big\{\max\{\widetilde{\mathbb{L}}(A), 1 - \widetilde{\mathbb{U}}(\overline{A})\}, \min\{\widetilde{\mathbb{U}}(A), 1 - \widetilde{\mathbb{L}}(\overline{A})\}\Big\}; \\ \{\mathbb{L}(\overline{A}), \mathbb{U}(\overline{A})\} &= \Big\{\max\{\widetilde{\mathbb{L}}(\overline{A}), 1 - \widetilde{\mathbb{U}}(A)\}, \min\{\widetilde{\mathbb{U}}(\overline{A}), 1 - \widetilde{\mathbb{L}}(A)\}\Big\}. \end{aligned} \tag{4.1}$$

   Note that $\mathbb{L}(A) + \mathbb{U}(\overline{A}) = 1$ and $\mathbb{U}(A) + \mathbb{L}(\overline{A}) = 1$. So, with no loss of generality, we take a PrBound pair to satisfy

$$\mathbb{L}(A) + \mathbb{U}(\overline{A}) = 1, \ \forall A \subseteq \Theta. \tag{4.2}$$

2. Naturally, one should use the narrowest PrBounds. In fact, if the probability $P(A)$ of $A \subseteq \Theta$ is known, one should use $\{\mathbb{L}(A), \mathbb{U}(A)\} = \{P(A), P(A)\}$. In particular, since it is always true that $P(\emptyset) = 0$ and $P(\Theta) = 1$, it is unnecessary to explicitly state the PrBounds for $\emptyset$ and Θ in Definition 5.

3. Any monotone probability function pair, including a DST belief/plausibility function pair (which is ∞-monotone [Pearl, 1990, Halpern and Fagin, 1992]) constitutes a valid PrBound pair. This enables us to learn PrBound parameters from evidential datasets which are founded on DST models' ability to capture

a wide variety data imperfections. These learned PrBound parameters can then be utilized within various operations (e.g., evidence updating, fusion, etc.) without having to ensure that the resultants remain ∞-monotone.

4.2.2 Some Features of PrBound Pairs

We now highlight some features of the PrBounds as defined in Definition 5.

4.2.2.1 Synthesizing New PrBound Pairs

PrBound pairs are not unique in that multiple sets of PrBound pairs may bound a given underlying p.m.f. When presented with multiple sets of PrBound pairs $P(\cdot) \lhd \{\mathbb{L}_i(\cdot), \mathbb{U}_i(\cdot)\},\ i \in \overline{1,n}$, one can easily synthesize one "fused" set of narrower PrBound pairs as $P(\cdot) \lhd \{\{\mathbb{L}(\cdot), \mathbb{U}(\cdot)\}$, where

$$\mathbb{L}(A) = \max_{i \in \overline{1,n}} \{\mathbb{L}_i(A)\};\ \ \mathbb{U}(A) = \min_{i \in \overline{1,n}} \{\mathbb{U}_i(A)\}. \tag{4.3}$$

Note that $\mathbb{L}(A) + \mathbb{U}(\overline{A}) = 1$.

4.2.2.2 Jettisoning the Monotonicity Requirement

We do not require PrBounds to satisfy any monotonicity constraint because it can be too restrictive and too unwieldy a property to maintain.

Example 3 *Consider the underlying "true" probability*

$$P(A) = 0.6;\ \ P(B) = 0.4,\ \textit{with } P(A \cup B) = 0.7;\ \ P(A \cap B) = 0.3.$$

Suppose we desire to "fit" a 2-monotone lower probability function $P_L(\cdot)$ for $P(\cdot)$. In the absence of any evidence, we must use $P_L(X) = 0,\ \forall X \subseteq \Theta$.

Suppose we are now told that $0.5 \leq P(A)$ *and* $0.3 \leq P(B)$. *To maintain 2-monotonicity, the updated lower bounds must satisfy*

$$P_L(A \cup B) \geq P_L(A) + P_L(B) - P_L(A \cap B) = 0.8,$$

which violates the true probability $P(A \cup B) = 0.7$. *In other words, we must ignore the 2-monotonicity condition if we are to harness the new evidence.*

How do PrBounds handle this scenario? Note that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \geq \mathbb{L}(A) + \mathbb{L}(B) - \mathbb{U}(A \cap B).$$

With $\mathbb{L}(A) = 0.5$, $\mathbb{L}(B) = 0.3$, *and* $\mathbb{U}(A \cap B) = 1$ *(because we do not have any evidence regarding* $P(A \cap B)$*), we get* $P(A \cap B) \geq -0.2$, *and hence* $\mathbb{L}(A \cap B) = 0$, *a trivial yet non-contradictory statement.*

However, with PrBounds, we have the opportunity to get a narrower set of PrBound pairs. Indeed, note that

$$P(A \cup B) \geq \max\{P(A), P(B)\} \geq \max\{\mathbb{L}(A), \mathbb{L}(B)\} = 0.5,$$

which yields $P(A \cup B) \lhd \{\mathbb{L}(A \cup B), \mathbb{U}(A \cup B)\} = \{0.5, 1\}$. ■

Monotonicity renders the bounds for each proposition interdependent, thus making it difficult or impossible to update one (or a few) bounds without having to update the bounds of other propositions.

Example 4 *Take two events* $A, B \subseteq \Theta$ *for which we have no priors. So, as before, we use the lower bounds* $P_L(X) = 0$, $\forall X \subseteq \Theta$.

Suppose an expert now informs us that $0.2 \leq P(A \cap B)$. *If we are to maintain monotonicity, then we cannot simply update* $P_L(A \cap B)$ *without updating* $P_L(A)$

as well (although there is no direct evidence provided for $P(A)$). For instance, for 1-monotonicity, we must have $P_L(A \cap B) \leq P_L(A)$ and therefore we cannot keep $P_L(A) = 0$ if we are to use $P_L(A \cap B) = 0.2$. This interdependence that monotonicity compels leads to higher computational cost.

If we need, how could PrBounds be used to find a lower bound for $P(A)$? With the given evidence, we would have $\mathbb{L}(A \cap B) = 0.2$ and $\mathbb{L}(A) = 0$. Then, if $P(\cdot)$ is the underlying probability,

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) \geq \mathbb{L}(A \cap B) + \mathbb{L}(A \cap \overline{B}) \geq 0.2.$$

So if we need, we can update $\mathbb{L}(A) = 0$ as $\mathbb{L}(A) = 0.2$. ■

In this manner, provided that the required bounds are available, we can endow PrBounds with up to ∞-monotonicity (because the underlying probability $P(\cdot)$ is ∞-monotone). This process of using the underlying probability to endow PrBounds with monotonicity can be thought of as analogous to how natural extension is used in imprecise probability formalisms to maintain coherence [Walley, 1991].

Jettisoning monotonicity also unshackles us from having to ensure that only operations that are closed under monotonicity are employed within the learning and reasoning processes. For instance, consider propositions B_i, $i \in \overline{1, n}$, that are conditionally independent given A. The naïve Bayes classifier exploits this information to extract the posterior $P(A|B)$, where $B = (B_1, \ldots, B_n)$, from the likelihoods $P(B_i|A)$, $i \in \overline{1, n}$, and the prior $P(A)$. Various i.v. versions of this naïve Bayes classifier have appeared in the literature (e.g., see [Zaffalon, 2002b, Heendeni et al., 2016]); in Section 4.3.2.1 we develop its PrBound version. For now, take the lower bound corresponding to the posterior in the naïve Bayes classifier in [Heendeni et al., 2016]:

$$P_L(A|B) = \frac{P_L(A)\prod_{i=1}^{n} P_L(B_i|A)}{P_L(A)\prod_{i=1}^{n} P_L(B_i|A) + P_U(\overline{A})\prod_{i=1}^{n} P_U(B_i|\overline{A})}. \tag{4.4}$$

If the $P_L(\cdot)$ and $P_U(\cdot)$ terms in the right-hand side of this equation are ∞-monotone lower and upper functions respectively, then there is no guarantee that the computed posterior $P_L(\cdot|B)$ is ∞-monotone. Similarly, if $P_L(\cdot)$ and $P_U(\cdot)$ are lower and upper envelopes respectively, then there is no guarantee that the computed posterior $P_L(\cdot|B)$ is the lower envelope of a conditional p.m.f. One may of course employ enumeration techniques to ensure that the computed posterior retains the appropriate property, but this would invariably add to the required computational complexity.

4.2.3 Conditioning

Hereafter, for the underlying p.m.f. $P(\cdot)$, we assume that $P(\cdot)\lhd\{\mathbb{L}(\cdot), \mathbb{U}(\cdot)\}$. When referring to $P(A|B)$, the assumption $P(B) > 0$ is implicit.

I.V. conditional notions have been developed in many different forms, including the DST conditional belief in [Fagin and Halpern, 1990] (see Definition 3), the conditional inner measure in [Fagin and Halpern, 1990], and the lower conditional credal set bound in [Zaffalon, 2002b]. They all take a very similar "structural" form to that of DST conditional belief $Bl(A|B)$ in Definition 3.

Another common thread running through all these i.v. conditional notions is that they require the lower bound of the interval associated with the conditioning proposition to be strictly positive (e.g., $Bl(B) > 0$ in Definition 3). In this respect, the conditional PrBounds that we develop below differ. The notion of PrBounds rests on the premise that there is a single underlying true p.m.f. $P(\cdot)$, and in turn, for the conditional $P(\cdot|B)$ to exist, all one needs is $P(B) > 0$, a less conservative condition than $\mathbb{L}(B) > 0$.

First, here are the expressions for the conditional PrBounds:

Theorem 2 *Let*

$$\Delta_{\mathbb{L}(A|B)} = \mathbb{L}(A \cap B) + \mathbb{U}(\overline{A} \cap B) \text{ and } \Delta_{\mathbb{U}(A|B)} = \mathbb{U}(A \cap B) + \mathbb{L}(\overline{A} \cap B).$$

For $\Delta_{\mathbb{L}(A|B)} + \Delta_{\mathbb{U}(A|B)} > 0$, *define* $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$ *as*

$$\mathbb{L}(A|B) = \begin{cases} \dfrac{\mathbb{L}(A \cap B)}{\Delta_{\mathbb{L}(A|B)}}, & \text{for } \Delta_{\mathbb{L}(A|B)} > 0,\ \Delta_{\mathbb{U}(A|B)} \geq 0; \\ 0, & \text{for } \Delta_{\mathbb{L}(A|B)} = 0,\ \Delta_{\mathbb{U}(A|B)} > 0; \end{cases}$$

$$\mathbb{U}(A|B) = \begin{cases} \dfrac{\mathbb{U}(A \cap B)}{\Delta_{\mathbb{U}(A|B)}}, & \text{for } \Delta_{\mathbb{L}(A|B)} \geq 0,\ \Delta_{\mathbb{U}(A|B)} > 0; \\ 1, & \text{for } \Delta_{\mathbb{L}(A|B)} > 0,\ \Delta_{\mathbb{U}(A|B)} = 0. \end{cases}$$

Then, for $P(B) > 0$, $P(A|B) \lhd \{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$. □

Proof: For $P(B) > 0$, the pair $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$ is well defined.

(a) Suppose $\Delta_{\mathbb{L}(A|B)} > 0$ and $\Delta_{\mathbb{U}(A|B)} > 0$. This is case 1 for $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$.

(a.1) Suppose $\mathbb{L}(A \cap B) > 0$. Then $P(A \cap B) > 0$ and we can write

$$P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P(\overline{A} \cap B)} = \frac{1}{1 + P(\overline{A} \cap B)/P(A \cap B)}.$$

Now use the bounds $\mathbb{L}(A \cap B) \leq P(A \cap B) \leq \mathbb{U}(A \cap B)$ and $\mathbb{L}(\overline{A} \cap B) \leq P(\overline{A} \cap B) \leq \mathbb{U}(\overline{A} \cap B)$ to get $1/[1 + \mathbb{U}(\overline{A} \cap B)/\mathbb{L}(A \cap B)] \leq P(A|B) \leq 1/[1 + \mathbb{L}(\overline{A} \cap B)/\mathbb{U}(A \cap B)]$, which yields $\mathbb{L}(A|B) \leq P(A|B) \leq \mathbb{U}(A|B)$.

(a.2) Suppose $\mathbb{L}(A \cap B) = 0$. Then $\mathbb{U}(\overline{A} \cap B) > 0$ and $\mathbb{L}(A|B) = 0$. If $P(A \cap B) > 0$, then we can write $P(A|B) = 1/[1 + P(\overline{A} \cap B)/P(A \cap B)] \leq 1/[1 +$

$\mathbb{L}(\overline{A} \cap B)/\mathbb{U}(A \cap B)] = \mathbb{U}(A|B)$. So, clearly, $\mathbb{L}(A|B) = 0 \leq P(A|B) \leq \mathbb{U}(A|B)$. If on the other hand $P(A \cap B) = 0$, then $P(A|B) = P(A \cap B)/P(B) = 0$, and so $\mathbb{L}(A|B) = 0 \leq P(A|B) = 0 \leq \mathbb{U}(A|B)$.

(b) Suppose $\Delta_{\mathbb{L}(A|B)} = 0$ and $\Delta_{\mathbb{U}(A|B)} > 0$. This is case 2, i.e., $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\} = \{0, 1\}$.

(c) Suppose $\Delta_{\mathbb{L}(A|B)} > 0$ and $\Delta_{\mathbb{U}(A|B)} = 0$. This is case 3, i.e., $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\} = \{0, 1\}$. In both these cases, it is trivially true that $\mathbb{L}(A|B) = 0 \leq P(A|B) \leq \mathbb{U}(A|B) = 1$.

Verify that $\{\mathbb{L}(\emptyset|B), \mathbb{L}(\Theta|B)\} = \{0, 1\}$ directly. ■

Remarks.

1. Note the following: When $\Delta_{\mathbb{L}(A|B)} > 0$ and $\Delta_{\mathbb{U}(A|B)} = 0$, we must have $\mathbb{U}(A \cap B) = \mathbb{L}(A \cap B) = 0$ and $\mathbb{U}(\overline{A} \cap B) > 0$ so that $\mathbb{L}(A \cap B)/\Delta_{\mathbb{L}(A|B)} = 0$. On the other hand, when $\Delta_{\mathbb{L}(A|B)} = 0$ and $\Delta_{\mathbb{U}(A|B)} > 0$, we must have $\mathbb{U}(\overline{A} \cap B) = \mathbb{L}(\overline{A} \cap B) = 0$ and $\mathbb{U}(A \cap B) > 0$ so that $\mathbb{U}(A \cap B)/\Delta_{\mathbb{U}(A|B)} = 1$.

2. Observe that $\mathbb{L}(A|B) + \mathbb{U}(\overline{A}|B) = 1$, $\forall A \subseteq \Theta$.

Note that the conditional PrBound pair $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$ in Theorem 2 is well defined whenever $P(B) > 0$ (and not $\mathbb{L}(B) > 0$). In fact, as we now show, the conditional PrBounds expressions in Theorem 2 explicitly capture the condition $P(B) > 0$:

Corollary 1 *$P(B) > 0$ iff $\Delta_{\mathbb{L}(A|B)} + \Delta_{\mathbb{U}(A|B)} > 0$, or equivalently, $P(B) = 0$ iff $\Delta_{\mathbb{L}(A|B)} = \Delta_{\mathbb{U}(A|B)} = 0$.* □

Proof: We will show that $P(B) = 0$ iff $\Delta_{\mathbb{L}(A|B)} = \Delta_{\mathbb{U}(A|B)} = 0$.

Necessity. Suppose $\Delta_{\mathbb{L}(A|B)} = \Delta_{\mathbb{U}(A|B)} = 0$. This implies that $\Delta_{\mathbb{L}(A|B)} = \Delta_{\mathbb{U}(A|B)} = 0 \implies \mathbb{L}(A \cap B) = \mathbb{U}(\overline{A} \cap B) = \mathbb{U}(A \cap B) = \mathbb{L}(\overline{A} \cap B) = 0$, meaning that $P(B) = P(A \cap B) + P(\overline{A} \cap B) \leq \mathbb{U}(A \cap B) + \mathbb{U}(\overline{A} \cap B) = 0$, i.e., $P(B) = 0$.

Sufficiency. Suppose $P(B) = 0$. This implies that $P(A \cap B) = P(\overline{A} \cap B) = 0 \implies \mathbb{L}(A \cap B) = \mathbb{L}(\overline{A} \cap B) = 0$ and $\mathbb{U}(A \cap B) = \mathbb{U}(\overline{A} \cap B) = 0$ (Remark 2 section 4.2.1), meaning that $\Delta_{\mathbb{L}(A|B)} = \Delta_{\mathbb{U}(A|B)} = 0$. ■

4.2.4 Independence

In probability, the two events A and B are said to be independent if either of the following two equivalent conditions is true:

$$\text{Factorizability: } P(A \cap B) = P(A)\,P(B); \quad \text{Irrelevance: } P(A|B) = P(A). \tag{4.5}$$

With i.v. probability functions, the notion of independence is hardly a settled issue. As we elaborated upon earlier, when one presumes that the i.v. probabilities are generated by a *set* of underlying p.m.f.s, no *one* consistent notion of independence can be employed. In fact, neither factorizability nor irrelevance of the i.v. probability bounds, viz.,

$$\begin{aligned} &\text{Factorizability: } \mathbb{L}(A \cap B) = \mathbb{L}(A)\,\mathbb{L}(B) \text{ and } \mathbb{U}(A \cap B) = \mathbb{U}(A)\,\mathbb{U}(B); \\ &\text{Irrelevance: } \quad \mathbb{L}(A|B) \; = \mathbb{L}(A) \quad \text{and } \mathbb{U}(A|B) \; = \mathbb{U}(A), \end{aligned} \tag{4.6}$$

is a good indicator of independence with respect to the underlying p.m.f.

Example 5 *Consider the two datasets (5a) and (5b) in Table 4.1. Each attribute can assume two states, viz., $A = \{a, \overline{a}\}$ and $B = \{b, \bar{b}\}$.*

Dataset (5a): $P(a) = 0.5$, $P(b) = 0.5$ and $P(ab) = 0.25$. So,

$$P(ab) = P(a)\,P(b); \quad P(a|b) = \frac{P(ab)}{P(b)} = P(a),$$

Table 4.1: Datasets for Examples 5 and 6

Example 5				**Example 6**			
(5a)		**(5b)**		**(6a)**		**(6b)**	
A	**B**	**A**	**B**	**A**	**B**	**A**	**B**
$\overline{a}$	$\overline{b}$	$\overline{a}$	$\overline{b}$	a	□	a	b
a	b	a	□	$\overline{a}$	b	$\overline{a}$	b
$\overline{a}$	b	$\overline{a}$	b	a	b	a	b
a	$\overline{b}$	□	□	$\overline{a}$	□	$\overline{a}$	$\overline{b}$

revealing an independence relation between A and B.

Dataset (5b): This is the same dataset in (5a), except that some attribute values are missing. We get $\{\mathbb{L}(a), \mathbb{U}(a)\} = \{0.25, 0.50\}$, $\{\mathbb{L}(b), \mathbb{U}(b)\} = \{0.25, 0.75\}$, $\{\mathbb{L}(ab), \mathbb{U}(ab)\} = \{0, 0.50\}$, and $\{\mathbb{L}(\overline{a}b), \mathbb{U}(\overline{a}b)\} = \{0.25, 0.5\}$. So,

$$\mathbb{L}(ab) \neq \mathbb{L}(a)\,\mathbb{L}(b); \quad \mathbb{U}(ab) \neq \mathbb{U}(a)\,\mathbb{U}(b),$$

and

$$\mathbb{L}(a|b) = \frac{\mathbb{L}(ab)}{\mathbb{L}(ab) + \mathbb{U}(\overline{a}b)} \neq \mathbb{L}(a); \quad \mathbb{U}(a|b) = \frac{\mathbb{U}(ab)}{\mathbb{U}(ab) + \mathbb{L}(\overline{a}b)} \neq \mathbb{U}(a).$$

So, independence is not being reflected in either condition in (4.6). ■

Example 6 *Consider the two datasets (6a) and (6b) in Table 4.1. As before, each attribute can assume two states, viz., $A = \{a, \overline{a}\}$ and $B = \{b, \overline{b}\}$.*

Dataset (6a): Data reveal that $\{\mathbb{L}(a), \mathbb{U}(a)\} = \{0.5, 0.5\}$, $\{\mathbb{L}(b), \mathbb{U}(b)\} = \{0.5, 1\}$, and $\{\mathbb{L}(ab), \mathbb{U}(ab)\} = \{0.25, 0.5\}$. So, $\mathbb{L}(ab) = \mathbb{L}(a)\,\mathbb{L}(b)$ and $\mathbb{U}(ab) = \mathbb{U}(a)\,\mathbb{U}(b)$.

Dataset (6b): Here we have "filled" in the missing entries in dataset (6a). It yields $P(a) = 0.5$, $P(b) = 0.75$, and $P(ab) = 0.5$, i.e., no independence relationship exists between A and B. ■

We maintain that independence is determined by the underlying p.m.f. $P(\cdot)$, and not an i.v. probability pair. Regarding the PrBound pairs of independent events, we have the following important result:

Theorem 3 *Suppose $P(\cdot) \lhd \{\mathbb{L}(\cdot), \mathbb{U}(\cdot)\}$ for the p.m.f. $P(\cdot)$ defined over Θ. If $A, B \subseteq \Theta$ are independent, then the following are true:*

$$\begin{aligned} &\textit{Factorizability:} && P(A \cap B) = P(A)\,P(B) \lhd \{\max\{\mathbb{L}(A \cap B), \mathbb{L}(A)\,\mathbb{L}(B)\}, \\ &&& \qquad\qquad\qquad\qquad \min\{\mathbb{U}(A \cap B), \mathbb{U}(A)\,\mathbb{U}(B)\}\}\,; \\ &\textit{Irrelevance:} && P(A|B) \;\; = P(A) \qquad \lhd \{\max\{\mathbb{L}(A|B), \mathbb{L}(A)\}, \\ &&& \qquad\qquad\qquad\qquad \min\{\mathbb{U}(A|B), \mathbb{U}(A)\}\}\,. \end{aligned}$$ □

Proof:

Factorizability. For A and B independent, we have $P(A \cap B) = P(A)\,P(B)$. Using the PrBounds for $P(A \cap B)$, $P(A)$, and $P(B)$ we immediately get

$$\begin{aligned} \max\{\mathbb{L}(A \cap B), \mathbb{L}(A)\,\mathbb{L}(B)\} &\leq P(A \cap B) = P(A)\,P(B) \\ &\leq \min\{\mathbb{U}(A \cap B), \mathbb{U}(A)\,\mathbb{U}(B)\}. \end{aligned}$$

First, note that we must have $\mathbb{L}(A \cap B) \leq \mathbb{U}(A)\,\mathbb{U}(B)$; otherwise, we would have $P(A)\,P(B) \leq \mathbb{U}(A)\,\mathbb{U}(B) < \mathbb{L}(A \cap B) \leq P(A \cap B)$, which is impossible since $P(A \cap B) = P(A)\,P(B)$. Similarly, we must have $\mathbb{L}(A)\,\mathbb{L}(B) \leq \mathbb{U}(A \cap B)$.

What remains to be established are the "boundary" cases, viz., $P(A \cap B = \emptyset) \lhd \{0, 0\}$ and $P(A \cap B = \Theta) \lhd \{1, 1\}$.

Suppose $A \cap B = \emptyset$, $P(A \cap B) = P(A)\,P(B) = 0$: This yields $\mathbb{L}(A \cap B) = \mathbb{U}(A \cap B) = 0$ and either $P(A) = 0$, which yields $\mathbb{L}(A) = \mathbb{U}(A) = 0$, or $P(B) = 0$, which yields $\mathbb{L}(B) = \mathbb{U}(B) = 0$. So, for $A \cap B = \emptyset$, we have $\max\{\mathbb{L}(A \cap B), \mathbb{L}(A)\,\mathbb{L}(B)\} = \min\{\mathbb{U}(A \cap B), \mathbb{U}(A)\,\mathbb{U}(B)\} = 0$.

Suppose $A \cap B = \Theta$: We must have $A = B = \Theta$ and $P(A \cap B) = P(A)\, P(B) = 1$. This yields $\mathbb{L}(A \cap B) = \mathbb{U}(A \cap B) = 1$, $\mathbb{L}(A) = \mathbb{U}(A) = 1$ and $\mathbb{L}(B) = \mathbb{U}(B) = 1$. So, for $A \cap B = \Theta$, we have $\max\{\mathbb{L}(A \cap B), \mathbb{L}(A)\, \mathbb{L}(B)\} = \min\{\mathbb{U}(A \cap B), \mathbb{U}(A)\, \mathbb{U}(B)\} = 1$.

Irrelevance. For A and B independent, we have $P(A|B) = P(A)$ whenever $P(B) > 0$. Using the PrBounds for $P(A|B)$ and $P(A)$ we immediately get

$$\max\{\mathbb{L}(A|B), \mathbb{L}(A)\} \leq P(A|B) = P(A) \leq \min\{\mathbb{U}(A|B), \mathbb{U}(A)\}.$$

As before, we can show that $\mathbb{L}(A|B) \leq \mathbb{U}(A)$ and $\mathbb{L}(A) \leq \mathbb{U}(A|B)$.

We can also directly verify the “boundary” cases as before. ■

Remark. As Examples 5 and 6 demonstrate, independence with respect to the underlying p.m.f. cannot be ascertained from the factorizability and/or irrelevance relationships in Theorem 3.

4.2.5 Bayesian Inference

In Bayesian inference, one updates the probabilities by computing the posteriors from the likelihoods and priors via the Bayes’ rule:

$$P(B|A) = \frac{P(A|B)\, P(B)}{P(A|B)\, P(B) + P(A|\overline{B})\, P(\overline{B})}, \text{ for } 0 < P(B) < 1,\ P(A) > 0. \qquad (4.7)$$

A PrBound-based i.v. counterpart to Bayes’ rule which allows one to *update the PrBounds* by computing the posterior PrBounds from likelihood and prior PrBounds is the following result:

Lemma 2 *For* $0 < P(B) < 1,\ P(A) > 0,\ P(B|A) \lhd \{\mathbb{L}(B|A), \mathbb{U}(B|A)\}$, *where*

$$\mathbb{L}(B|A) = \frac{\mathbb{L}(A|B)\, \mathbb{L}(B)}{\mathbb{L}(A|B)\, \mathbb{L}(B) + \mathbb{U}(A|\overline{B})\, \mathbb{U}(\overline{B})};$$
$$\mathbb{U}(B|A) = \frac{\mathbb{U}(A|B)\, \mathbb{U}(B)}{\mathbb{U}(A|B)\, \mathbb{U}(B) + \mathbb{L}(A|\overline{B})\, \mathbb{L}(\overline{B})}. \qquad \square$$

Proof: The pairs $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$ and $\{\mathbb{L}(B|A), \mathbb{U}(B|A)\}$ are well defined for $0 < P(B) < 1$ and $P(A) > 0$. We prove the claim for the $\mathbb{L}(A \cap B) > 0$ case only. In this case, (4.7) yields $P(B|A) = 1/[1 + P(A|\overline{B})\, P(\overline{B})/P(A|B)\, P(B)]$. Now make use of Theorem 2 to get $1/[1 + \mathbb{U}(A|\overline{B})\, \mathbb{U}(\overline{B})/\mathbb{L}(A|B)\, \mathbb{L}(B)] \leq P(B|A) \leq 1/[1 + \mathbb{L}(A|\overline{B})\, \mathbb{L}(\overline{B})/\mathbb{U}(A|B)\, \mathbb{U}(B)]$, which yields $\mathbb{L}(B|A) \leq P(B|A) \leq \mathbb{U}(B|A)$.

Verify that $\{\mathbb{L}(\emptyset|A), \mathbb{L}(\Theta|A)\} = \{0, 1\}$ directly. ■

4.3 Propagation of Bounds

4.3.1 Bayesian Network (BN)

A BN is a directed acyclic graph (DAG) $\mathcal{N} = (\mathcal{V}, \mathcal{E}, P)$, where the set of nodes $\mathcal{V}$ denotes the N random variables $\{v_1, \ldots, v_N\}$. The p.m.f. $P(\cdot)$ denotes the joint distribution over the random variables in $\mathcal{V}$. The set of directed edges $\mathcal{E}$ captures the conditional dependence between pairs of nodes; $e_{ij} \in \mathcal{E}$, $i, j = 1, \ldots, N$, is the directed edge from node v_j to node v_i. Then,

$$P(\mathcal{V}) = \prod_{i=1}^{N} P(v_i|\widehat{\mathcal{V}}(v_i)), \tag{4.8}$$

yields the joint probability distribution of the random variables $\{v_1, \ldots, v_N\}$. Here, $\widehat{\mathcal{V}}(v_i)$ are the *parents* of node $v_i \in \mathcal{V}$, i.e., $\widehat{\mathcal{V}}(v_i) = \{v_j \in \mathcal{V} \mid e_{ij} \in \mathcal{E}\}$. This representation enables one to use a BN for reasoning and inference with uncertain knowledge in a computationally efficient manner.

One may think of the graphical structure which “codes” the conditional dependence between variables as the qualitative aspect of the BN and the conditional probabilities at each node as its quantitative aspect.

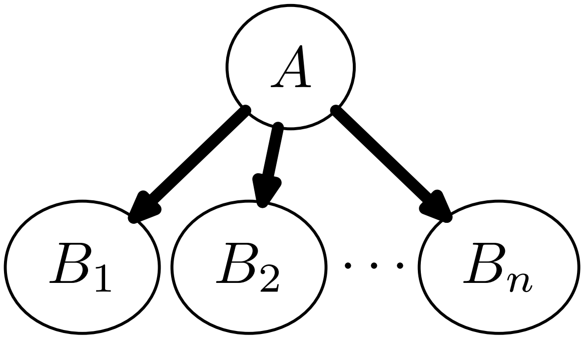


Figure 4.1: Naïve Bayes' Classifier. The variables B_i, $i = 1, \dots, n$, are conditionally independent given A.

4.3.2 PrBound-Based I.V. Bayesian Network (I.V. BN)

From (4.8) it follows that we can use PrBound pairs for $P(v_i|\widehat{\mathcal{V}}(v_i))$, $i = 1, \dots, N$, to generate $P(\mathcal{V}) \lhd \{\mathbb{L}(\mathcal{V}), \mathbb{U}(\mathcal{V})\}$, where

$$\mathbb{L}(\mathcal{V}) = \prod_{i=1}^{N} \mathbb{L}(v_i|\widehat{\mathcal{V}}(v_i)); \;\; \mathbb{U}(\mathcal{V}) = \prod_{i=1}^{N} \mathbb{U}(v_i|\widehat{\mathcal{V}}(v_i)). \tag{4.9}$$

Now, each term in a probabilistic expression arising within $\mathcal{N} = (\mathcal{V}, \mathcal{E}, P)$ can be simply substituted with its corresponding PrBound pair that yields the narrowest interval with the evidence at hand. We will refer to the graphical structure resulting from $\mathcal{N}$ when it is endowed with the relevant PrBound pairs as an *interval-valued Bayesian network (i.v. BN)* $\overline{\mathcal{N}} = ((\mathcal{V}, \mathcal{E}), \{\mathbb{L}, \mathbb{U}\}) = (\mathcal{N}, \{\mathbb{L}, \mathbb{U}\})$.

With the results we have developed in Section 4.2, we are now able to propagate these PrBound pairs within an i.v. BN in much the same way as probabilities are propagated within a BN. In other words, our framework allows the utility of i.v. BNs with all the advantages that BNs offer.

4.3.2.1 Naïve Bayes' Classifier

To explain, consider the naïve Bayes' classifier in Fig. 4.1 which codes the fact that B_i, $i = 1, \ldots, n$, are conditionally independent given A, i.e.,

$$P(B|A) = \prod_{i=1}^{n} P(B_i|A), \text{ where } B = (B_1, \ldots, B_n). \tag{4.10}$$

4.3.2.1.1 BN When treated as the BN $\mathcal{N}$, the parameters hosted at the nodes are the following: $P(A)$ at node A and $P(B_i|A)$ at node B_i. For $P(A) > 0$, the joint p.m.f. is given by

$$P(A \cap B) = P(A) \prod_{i=1}^{n} P(B_i|A). \tag{4.11}$$

For $0 < P(A) < 1$ and $P(B_i) > 0$, $\forall i = 1, \ldots, n$, we have

$$P(A|B) = \frac{P(A) \prod_{i=1}^{n} P(B_i|A)}{P(A) \prod_{i=1}^{n} P(B_i|A) + P(\overline{A}) \prod_{i=1}^{n} P(B_i|\overline{A})}. \tag{4.12}$$

4.3.2.1.2 I.V. BN When endowed with the PrBound pairs, the parameters hosted at nodes of the corresponding i.v. BN $\overline{\mathcal{N}}$ are the following: $\{\mathbb{L}(A), \mathbb{U}(A)\}$ at node A and $\{\mathbb{L}(B_i|A),\ \mathbb{U}(B_i|A)\}$ at node B_i. As an i.v. counterpart to (4.11), we have $P(A \cap B) \triangleleft \{\mathbb{L}(A \cap B), \mathbb{U}(A \cap B)\}$, where

$$\{\mathbb{L}(A \cap B), \mathbb{U}(A \cap B)\} = \left\{ \mathbb{L}(A) \prod_{i=1}^{n} \mathbb{L}(B_i|A), \mathbb{U}(A) \prod_{i=1}^{n} \mathbb{U}(B_i|A) \right\}. \tag{4.13}$$

Next we get the following result:

Lemma 3 *As an i.v. counterpart to* (4.12)*, we have* $P(A|B) \triangleleft \{\mathbb{L}(A{\downarrow}B), \mathbb{U}(A{\downarrow}B)\}$*, where*

$$\mathbb{L}(A{\downarrow}B) = \frac{\mathbb{L}(A)\prod_{i=1}^{n}\mathbb{L}(B_i|A)}{\mathbb{L}(A)\prod_{i=1}^{n}\mathbb{L}(B_i|A) + \mathbb{U}(\overline{A})\prod_{i=1}^{n}\mathbb{U}(B_i|\overline{A})};$$

$$\mathbb{U}(A{\downarrow}B) = \frac{\mathbb{U}(A)\prod_{i=1}^{n}\mathbb{U}(B_i|A)}{\mathbb{U}(A)\prod_{i=1}^{n}\mathbb{U}(B_i|A) + \mathbb{L}(\overline{A})\prod_{i=1}^{n}\mathbb{L}(B_i|\overline{A})},$$

when $0 < P(A) < 1$ *and* $P(B_i) > 0$, $\forall i = 1, \ldots, n$. □

Proof: Write (4.12) as

$$P(A|B) = \frac{1}{1 + P(\overline{A})\prod_{i=1}^{4}P(B_i|\overline{A})/P(A)\prod_{i=1}^{4}P(B_i|A)}.$$

Now substitute each term by its corresponding PrBound pair to show that $P(A|B) \triangleleft \{\mathbb{L}(A{\downarrow}B), \mathbb{U}(A{\downarrow}B)\}$.

Verify that $\{\mathbb{L}(\emptyset{\downarrow}B), \mathbb{L}(\Theta{\downarrow}B) = \{0, 1\}$ directly. ■

4.4 Revisiting Imperfect Implication Rules

We now describe how imperfect implication rules, developed in Chapter 3 based on DS theory, can be expressed, and hence generalized, in terms of PrBounds. Note that we only need to establish Theorem 1 and Lemma 1 of Chapter 3 in terms of PrBounds.

We will henceforth utilize the following PrBound-based notation:

$$\begin{aligned}
\mathbb{L}(A) &= \alpha_1; \;\; \mathbb{U}(A) &= \beta_1; \;\; \mathbb{L}(B) &= \alpha_2; \;\; \mathbb{U}(B) &= \beta_2; \\
\mathbb{L}(B|A) &= \alpha_R; \;\; \mathbb{U}(B|A) = \beta_R; \;\; \mathbb{L}(B|\overline{A}) = \alpha_{\overline{R}}; \;\; \mathbb{U}(B|\overline{A}) = \beta_{\overline{R}}. \qquad (4.14)
\end{aligned}$$

Recall that, as mentioned in Chapter 3, A is the antecedent and B is the consequent.

4.4.1 PrBounds-Based Statement of Theorem 1

The PrBounds-based statement of Theorem 1 of Chapter 3 is as follows:

Theorem 4 (Consequent Uncertainty: General Bounds) *A PrBound pair corresponding to the consequent B of rule $R:\; A \Longrightarrow B$ is*

$$P(B) \lhd \{\alpha_1\alpha_R + (1-\beta_1)\,\alpha_{\overline{R}},\, \alpha_1\beta_R + (1-\beta_1)\,\beta_{\overline{R}} + (\beta_1 - \alpha_1)\}.$$

Here, $\{\alpha_R, \beta_R\}$ and $\{\alpha_{\overline{R}}, \beta_{\overline{R}}\}$ refer to PrBound pairs associated with the rules $R:\; A \Longrightarrow B$ and $\overline{R}:\; \overline{A} \Longrightarrow B$, respectively, and $\{\alpha_1, \beta_1\}$ is a PrBound pair associated with the antecedent A. □

Proof: Note that

$$\begin{aligned}
P(B) = P(A)P(B|A) + P(\overline{A})\,P(B|\overline{A}) \qquad &\geq \mathbb{L}(A)\,\mathbb{L}(B|A) + \mathbb{L}(\overline{A})\,\mathbb{L}(B|\overline{A}) \\
= \mathbb{L}(A)\,\mathbb{L}(B|A) + [1-\mathbb{U}(A)]\,\mathbb{L}(B|\overline{A}) &= \alpha_1\alpha_R + (1-\beta_1)\,\alpha_{\overline{R}},
\end{aligned}$$

and

$$\begin{aligned}
P(\overline{B}) = P(A)\,P(\overline{B}|A) + P(\overline{A})\,P(\overline{B}|\overline{A}) &\geq \mathbb{L}(A)\,\mathbb{L}(\overline{B}|A) + \mathbb{L}(\overline{A})\,\mathbb{L}(\overline{B}|\overline{A}) \\
&= \mathbb{L}(A)\,[1-\mathbb{U}(B|A)] + [1-\mathbb{U}(A)]\,[1-\mathbb{U}(B|\overline{A})] \\
&= \alpha_1(1-\beta_R) + (1-\beta_1)\,(1-\beta_{\overline{R}}).
\end{aligned}$$

So,

$$1 - P(B) \geq \alpha_1(1 - \beta_R) + (1 - \beta_1)\,(1 - \beta_{\overline{R}}),$$

which yields

$$P(B) \leq 1 - [\alpha_1(1 - \beta_R) + (1 - \beta_1)\,(1 - \beta_{\overline{R}})] \leq \alpha_1\beta_R + (1 - \beta_1)\,\beta_{\overline{R}} + (\beta_1 - \alpha_1)$$

Therefore

$$P(B) \lhd \{\alpha_1\alpha_R + (1 - \beta_1)\,\alpha_{\overline{R}},\; \alpha_1\beta_R + (1 - \beta_1)\,\beta_{\overline{R}} + (\beta_1 - \alpha_1)\}. \qquad \blacksquare$$

Remark. Without the LC choice, we can write

$$\mathbb{L}(B) = \alpha_2 = \alpha_1\alpha_R + (1 - \beta_1)\,\alpha_{\overline{R}};$$

$$\mathbb{U}(B) = \beta_2 = \alpha_1\beta_R + (1 - \beta_1)\,\beta_{\overline{R}} + (\beta_1 - \alpha_1)$$

4.4.2 PrBounds-Based Statement of Lemma 1

The PrBounds-based statement of Lemma 1 of Chapter 3 is as follows:

Lemma 4 *A PrBound pair corresponding to the consequent B of rule $R:\ A \Longrightarrow B$ is*

$$P(B) \lhd \{\min\{\alpha_R, \alpha_{\overline{R}}\}, \max\{\beta_R, \beta_{\overline{R}}\}\}. \qquad \square$$

Proof: Note that

$$\begin{aligned} P(B) &= P(B|A)\,P(A) + P(B|\overline{A})\,P(\overline{A}) \\ &\geq \min\{P(B|A), P(B|\overline{A})\}P(A) + \min\{P(B|A), P(B|\overline{A})\}P(\overline{A}) \\ &= \min\{P(B|A), P(B|\overline{A})\}[P(A) + P(\overline{A})] = \min\{P(B|A), P(B|\overline{A})\} \\ &\geq \min\{\mathbb{L}(B|A), \mathbb{L}(B|\overline{A})\}. \end{aligned}$$

Similarly,

$$P(B) \leq \max\{P(B|A), P(B|\overline{A})\} \leq \max\{\mathbb{U}(B|A), \mathbb{U}(B|\overline{A})\}.$$

Therefore

$$P(B) \lhd \{\min\{\alpha_R, \alpha_{\overline{R}}\}, \max\{\beta_R, \beta_{\overline{R}}\}\}. \qquad \blacksquare$$

4.5 Summary

In this work we have proposed a framework for reasoning with i.v. probabilities. The lower/upper bounds, which we refer to as PrBound pairs, are not required to satisfy a monotonicity or other constraint. This enables us to develop i.v. versions of graphical networks which operate much like their conventional counterparts.

We have also developed PrBound pairs for conditionals in terms of the PrBound pairs of the underlying probability distribution. We also take a fresh look at how independence and conditional independence between variables are viewed, viz., we take these notions to be governed by the underlying probability distribution and not by any set of lower/upper bounds.

CHAPTER 5

Learning Parameters From Imperfect Data

5.1 Overview

In this chapter, we explore how the parameters that are needed for ML algorithms, in particular, for the implication rules in Chapter 3 and the probability bound-based framework in Chapter 4, can be extracted from a datasets which may potentially be imperfect.

We pay special attention to a type of uncertainty that is most commonly encountered in realistic datasets: attribute values that are unknown/missing or that are known to lie within a set of values but otherwise cannot be discerned further (Definition 7).

5.2 Model

Hereafter, we assume that the data record R is comprised of attribute variables A_j, $j = 1, \ldots, N_R$.

(a) We assume that A_j can take a value or state from the sample space $\Theta_j = \{\theta_{1j}, \cdots, \theta_{N_j j}\}$. The (possibly unknown) p.m.f. associated with A_j is $P_j(\cdot)$ (defined over Θ_j).

(b) We use Θ_R to denote the cross-product $\bigotimes_{j=1}^{N_R} \Theta_j$ and the short-hand notation 2^{Θ_R} to denote the set of all subsets of the cross-product Θ_R, i.e., $\prod_{j=1}^{N_R} 2^{\Theta_j}$.

(c) By $\langle A_j = a_j \rangle$, or by simply a_j, we denote the (potentially uncertain) state $a_j \in 2^{\Theta_j}$, or equivalently $a_j \subseteq \Theta_j$, that the attribute A_j assumes. Note that $a_j = \emptyset$ denotes that the attribute A_j is "not applicable". As in [Anand et al., 1996, Hewawasam et al., 2007], we assume that an attribute vector whose attributes are all "not applicable" (i.e., the "null set" of Θ_R) is nonexistent.

5.2.1 AttBounds and RecBounds

DST models have been successfully utilized to capture various types of imperfections in data (e.g., unknown/missing and incomplete/ambiguous values, probabilistic uncertainty, etc.) [Anand et al., 1996,Hewawasam et al., 2007,Wickramarathne et al., 2011a]. To capture the uncertainty associated with the value an attribute assumes, one can employ a DST *AttBBA* $m_j : 2^{\Theta_j} \mapsto [0, 1]$ defined over the FoD Θ_j; the corresponding *AttBoE* is $\{\Theta_j, \mathfrak{F}_{A_j}, m_j\}$. To capture the uncertainty associated with a complete record, one can then employ a DST *RecBBA* $m_R : 2^{\Theta_R} \mapsto [0, 1]$ defined over the cross-product FoD Θ_R; the corresponding *RecBoE* is $\{\Theta_R, \mathfrak{F}_R, m_R\}$.

Instead, in our work, we utilize PrBound pairs to capture imperfections in data.

Definition 6 *Consider the data record R.*

Table 5.1: Generating AttBound Pairs From AttBoEs

Attribute A_j	**Type**	**Value**	**Prob.**	**DS-Mass**	**AttBound Pair**
[Sample Space]		$\langle A_j = a_j \rangle$	$P_j(a_j)$	$m_{A_j}(a_j)$	$[\mathbb{L}(\mathbf{a}), \mathbb{U}(\mathbf{a})]$
A_1	Probabilistic	θ_{11}	0.6	0.6	$P_1(\theta_{11}) \in$ [0.6, 0.6]
$[\Theta_1 = \{\theta_{11}, \theta_{21}\}]$		θ_{21}	0.4	0.4	$P_1(\theta_{21}) \in$ [0.4, 0.4]
A_2	Unknown/Missing	$(\theta_{12}, \theta_{22})$	1.0	1.0	$P_2(\theta_{12}) \in$ [0.0, 1.0]
$[\Theta_2 = \{\theta_{12}, \theta_{22}\}]$					$P_2(\theta_{22}) \in$ [0.0, 1.0]
A_3	Known exactly	θ_{13}	1.0	1.0	$P_3(\theta_{13}) \in$ [1.0, 1.0]
$[\Theta_3 = \{\theta_{13}, \theta_{23}\}]$					$P_3(\theta_{23}) \in$ [0.0, 0.0]
A_4	Known to lie	$(\theta_{14}, \theta_{24})$	1.0	1.0	$P_4(\mathbf{a}) \in$[1.0, 1.0],
	within a set				$\forall \mathbf{a} \supseteq (\theta_{14}, \theta_{24})$;
$[\Theta_4 = \{\theta_{14}, \theta_{24}, \theta_{34}\}]$					$P_4(\mathbf{a}) \in [0.0, 1.0]$,
					$\forall \mathbf{a} \cap (\theta_{14}, \theta_{24}) \neq \emptyset$;
					$P_4(\mathbf{a}) \in [0.0, 0.0]$,
					$\forall \mathbf{a} \cap (\theta_{14}, \theta_{24}) = \emptyset$;
A_5	General	θ_{15}	[0.6, 0.7]	0.6	$P_5(\theta_{15}) \in$ [0.6, 0.7]
$[\Theta_5 = \{\theta_{15}, \theta_{25}\}]$		θ_{25}	[0.3, 0.4]	0.3	$P_5(\theta_{25}) \in$ [0.3, 0.4]
		$(\theta_{15}, \theta_{25})$	1.0	0.1	$P_5(\theta_{15}, \theta_{25}) \in$ [1.0,1.0]

Note. $P_j(\Theta_j) \in [1.0, 1.0], \forall j = 1, \ldots, 5.$

(i) An AttBound *pair* $\{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ *is a PrBound pair for the p.m.f.* $P_j(\cdot)$ *defined over* Θ_j*. We denote this as* $P_j(\cdot) \lhd \{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$*.*

(ii) A RecBound *pair* $\{\mathbb{L}_R(\cdot), \mathbb{U}_R(\cdot)\}$ *is a PrBound pair for the joint p.m.f.* $P_R(\cdot)$ *defined over* Θ_R*. We denote this as* $P_R(\cdot) \lhd \{\mathbb{L}_R(\cdot), \mathbb{U}_R(\cdot)\}$*.* ■

If an AttBoE or a RecBoE is available (say, from the work in [Hewawasam et al., 2007]), then an AttBound pair or a RecBound pair can be directly generated from the corresponding belief/plausibility functions, respectively.

As Example 7 below illustrates, AttBound pairs are more directly associated with, and hence are more easily generated from, the confidence or probability one can place on an attribute taking a particular state or a set of states.

Example 7 *Table 5.1 depicts a data record* $R = [\langle A_1 = a_1 \rangle, \langle A_2 = a_2 \rangle, \langle A_3 = a_3 \rangle, \langle A_4 = a_4 \rangle, \langle A_5 = a_5 \rangle]$*, where* A_1 *exhibits probabilistic uncertainty,* A_2 *is un-*

known/missing, A_3 is known exactly, and A_4 is known to lie within a set of states; A_5's form of uncertainty is more general, viz., $a_5 = \theta_{15}$ with a $[60\%, 70\%]$ *confidence and $a_5 = \theta_{25}$ with a* $[30\%, 40\%]$ *confidence.* ■

5.2.2 From AttBound Pairs to RecBound Pairs

The i.v. BN parameters are learnt from the RecBound pairs associated with the data records (see Section 5.3). How do we generate a RecordsBound pair from the AttBound pairs?

For DST models, [Hewawasam et al., 2007] employs the following scheme for this purpose:

(a) "extend" the AttBoE defined over Θ_j to the cross-product space Θ_R via the so called "cylindrical extension";

(b) repeat this for each AttBoE to generate N_R cylindrical extensions; and

(c) fuse these using the DCR to get a DST BoE for the complete data record. But by doing so, we are conceding that the AttBoEs are "independent" although evidence to support such an assumption is absent.

5.2.2.1 Independent Attributes

As mentioned earlier, our independence notion emerges, as we believe it should, from the underlying joint p.m.f. Indeed, with independent attributes $A_j(\cdot)$, the joint p.m.f. $P_R(\cdot)$ associated with the data record R can be expressed as

$$P_R(\mathbf{a}) = \prod_{j=1}^{N_R} P_j(a_j), \text{ for } \mathbf{a} = (a_1, \ldots, a_{N_R}) \subseteq \Theta_R. \tag{5.1}$$

Theorem 3 now immediately yields

Corollary 2 *Suppose* $P_j(\cdot) \lhd \{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ *for the attribute p.m.f.* $P_j(\cdot)$ *defined over* Θ_j. *With independent attributes* A_j, $j = 1, \ldots, N_R$, *for the record p.m.f.* $P_R(\cdot)$ *defined over the cross-product* Θ_R, $P_R(\cdot) \lhd \{\mathbb{L}_R(\cdot), \mathbb{U}_R(\cdot)\}$, *where*

$$\mathbb{L}_R(\mathbf{a}) = \prod_{j=1}^{N_R} \mathbb{L}_j(a_j); \;\; \mathbb{U}_R(\mathbf{a}) = \prod_{j=1}^{N_R} \mathbb{U}_j(a_j),$$

for $\mathbf{a} = (a_1, \ldots, a_{N_R})$, $a_j \subseteq \Theta_j$. ■

5.2.2.2 Attributes With Dependencies

With dependent attributes, one may proceed as follows.

General Case: For $a_j \subseteq \Theta_j$ and $2 \le K \le N_R$, let

$$\mathbf{a}^{[K]} = (a_1, \ldots, a_K); \; \mathbf{a}_j^{[K]} = (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_K). \tag{5.2}$$

So, $\mathbf{a}_j^{[K]}$ is essentially $\mathbf{a}^{[K]}$ but with its j-th entry removed. Note that, for $K = 1$, $\mathbf{a}^{[K]}$ yields a single attribute; for $K = N_R$, $\mathbf{a}^{[K]}$ yields the complete record.

The result below allows one to generate a RecBound pair from the AttBound pairs of each attribute p.m.f.:

Lemma 5 *With the initial conditions* $\mathbb{L}(\mathbf{a}_1^{[2]}) = \mathbb{L}_2(a_2)$ *and* $\mathbb{L}(\mathbf{a}_2^{[2]}) = \mathbb{L}_1(a_1)$, *for* $2 \le K \le N_R$, *consider the recursion*

$$\begin{aligned} \mathbb{L}(\mathbf{a}^{[K]}) \quad &= \max\left\{0, \max_{j=1,\ldots,K}\left\{\mathbb{L}(\mathbf{a}_j^{[K]}) + \mathbb{L}(a_j) - 1\right\}\right\}; \\ \mathbb{U}(\mathbf{a}^{[K]}) \quad &= \min_{j=1,\ldots,K}\left\{\mathbb{U}(\mathbf{a}_j^{[K]})\right\}; \\ &\textit{Use } (4.1) \quad \textit{to narrow the bound pair } \{\mathbb{L}(\mathbf{a}^{[K]}), \mathbb{U}(\mathbf{a}^{[K]})\}. \end{aligned}$$

Then, $P_R(\cdot) \lhd \left\{\mathbb{L}(\mathbf{a}^{[N_R]}), \mathbb{U}(\mathbf{a}^{[N_R]})\right\}$ *for the p.m.f.* $P_R(\cdot)$ *defined over the cross-product space* Θ_R. □

Proof: For $K \geq 2$, we note that

$$1 = P(\mathbf{a}_j^{[K]}) + P(a_j) - P(\mathbf{a}^{[K]}) + P(\overline{a}_j) - P(\mathbf{a}_j^{[K]}, \overline{a}_j),\ K \geq 2.$$

(a) We have

$$P(\mathbf{a}^{[K]}) \geq \max\Big\{0, P(\mathbf{a}_j^{[K]}) + P(a_j) - 1\Big\} \geq \max\Big\{0, \mathbb{L}(\mathbf{a}_j^{[K]}) + \mathbb{L}(a_j) - 1\Big\}$$

since $P(\overline{a}_j) - P(\mathbf{a}_j^{[K]}, \overline{a}_j) \geq 0$. This is true for arbitrary $j = 1, \ldots, K$, yielding the lower bound.

(b) Obviously, $P(\mathbf{a}^{[K]}) \leq P(\mathbf{a}_j^{[K]}) \leq \mathbb{U}(\mathbf{a}_j^{[K]})$. This is true for arbitrary $j = 1, \ldots, K$, yielding the upper bound. ■

5.2.2.2.1 Single Focal Element (SFE) Attributes A special case which offers both analytical convenience and significantly reduced computational complexity while providing adequate flexibility for capturing the types of data imperfections that one would typically encounter is the following:

Definition 7 *The attribute A_j with the p.m.f. $P_j(\cdot)$ is said to be a* single focal element (SFE) attribute *if, for some $\emptyset \neq B_j \subseteq \Theta_j$, $P_j(\cdot) \lhd \{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$, where*

$$\begin{aligned} &\mathbb{L}_j(a_j) = 1,\ \textit{if } a_j \supseteq B_j; && \mathbb{L}_j(a_j) = 0,\ \textit{otherwise};\\ &\mathbb{U}_j(a_j) = 1,\ \textit{if } a_j \cap B_j \neq \emptyset; && \mathbb{U}_j(a_j) = 0,\ \textit{otherwise}. \end{aligned}$$

A data record R comprised of only SFE attributes is called an SFE data record; *a dataset $\mathfrak{D}$ comprised of only SFE data records is called an* SFE dataset. ■

Remarks:

1. Note that the SFE attribute in Definition 7 is guaranteed to assume a state from the set B_j. Within the context of DST models, the AttBoE of an SFE attribute is of the form $m_j(B_j) = 1$ and $m_j(a_j) = 0$, $\forall a_j \neq B_j$, i.e., it has only one focal element (hence the term *SFE attribute*).

2. The AttBound pair $\{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ of an SFE attribute can only assume the values $\{0, 0\}$, $\{0, 1\}$, or $\{1, 1\}$, and $\{\mathbb{L}_j(B_j), \mathbb{U}_j(B_j)\} = \{1, 1\}$.

3. The attributes A_2, A_3, and A_4 of the data record in Table 5.1 are all SFE attributes. Note that, an unknown/missing attribute corresponds to the case when $B_j = \Theta_j$.

4. Note that, $\{\mathbb{L}_j(\Theta_j), \mathbb{U}_j(\Theta_j)\} = \{1, 1\}$, for any B_j.

One can now easily establish the following important property regarding SFE attributes:

Corollary 3 *Consider $P_j(\cdot) \lhd \{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ as in Definition 7. Suppose $\{a_j^{(\ell)}\}$, $\ell = 1, \ldots, M$, is a partition of $a_j \subseteq \Theta_j$, i.e., $a_j^{(\ell)} \subseteq \Theta_j$ and $a_j^{(\ell_1)} \cap a_j^{(\ell_2)} = \emptyset$, $\forall \ell_1, \ell_2 = 1, \ldots, M$, and $a_j = \bigcup_{\ell=1}^{M} a_j^{(\ell)}$. Then,*

$$\mathbb{L}_j(a_j) \geq \sum_{\ell=1}^{M} \mathbb{L}_j(a_j^{(\ell)}); \quad \mathbb{U}_j(a_j) \leq \sum_{\ell=1}^{M} \mathbb{U}_j(a_j^{(\ell)}). \qquad \square$$

Proof: First, suppose that $a_j^{(\ell)} \supseteq B_j$, for some ℓ. Then, $a_j = \bigcup a_j^{(\ell)} \supseteq B_j$ and $a_j^{(k)} \not\supseteq B_j$, $\forall k \neq \ell$ (since $a_j^{(\ell)}$ are mutually disjoint). So, from Definition 7, $\mathbb{L}_j(a_j^{(\ell)}) = \mathbb{L}_j(a_j) = 1$.

Next, suppose that $a_j^{(\ell)} \not\supseteq B_j$, for any ℓ. But it is still possible that $a_j = \bigcup a_j^{(\ell)} \supseteq B_j$. In such a case, we have $\mathbb{L}_j(a_j^{(\ell)}) = 0$, $\forall \ell$, and $\mathbb{L}_j(a_j) = 1$.

So, $\mathbb{L}_j(a_j) \geq \sum \mathbb{L}_j(a_j^{(\ell)})$.

The claim regarding $\mathbb{U}_j(a_j)$ can be proven similarly. ■

It turns out that Corollary 3 applies to the general evidential case too:

Corollary 4 *Consider $P_j(\cdot) \lhd \{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ as in Definition 7. Suppose $\{a_j^{(\ell)}\}$, $\ell = 1, \ldots, M$, is a partition of $a_j \subseteq \Theta_j$, i.e., $a_j^{(\ell)} \subseteq \Theta_j$ and $a_j^{(\ell_1)} \cap a_j^{(\ell_2)} = \emptyset$, $\forall \ell_1, \ell_2 = 1, \ldots, M$, and $a_j = \bigcup_{\ell=1}^{M} a_j^{(\ell)}$. Then for the general evidential data case,*

$$\mathbb{L}_j(a_j) \geq \sum_{\ell=1}^{M} \mathbb{L}_j(a_j^{(\ell)}); \;\; \mathbb{U}_j(a_j) \leq \sum_{\ell=1}^{M} \mathbb{U}_j(a_j^{(\ell)}). \qquad \square$$

Proof: Suppose $m(.)$ denotes the associated DST mass function. Then for the general evidential data case,

$$\mathbb{L}_j(a_j) = \sum_{b_j \subseteq a_j} m(b_j) \text{ and } \mathbb{L}_j(a_j^{(\ell)}) = \sum_{b_j \subseteq a_j^{(\ell)}} m(b_j).$$

So,

$$\mathbb{L}_j(a_j) = \sum_{b_j \subseteq a_j} m(b_j) \geq \sum_{\ell=1}^{M} \sum_{b_j \subseteq a_j^{(\ell)}} m(b_j) = \sum_{\ell=1}^{M} \mathbb{L}_j(a_j^{(\ell)}).$$

Note that there can exist some b_j for which $m(b_j) > 0$ and $b_j \subseteq a_j$, but $b_j \not\subseteq a_j^{(\ell)}$, for any ℓ.

The claim regarding $\mathbb{U}_j(a_j)$ can be proven similarly. ■

In essence, when the PrBound pair $\{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ of an SFE attribute is constructed as in Definition 7, it is super-additive (and hence 1-monotone). Also, $\left\{\sum_{\ell=1}^{M} \mathbb{L}_j(a_j^{(\ell)}), \sum_{\ell=1}^{M} \mathbb{U}_j(a_j^{(\ell)})\right\}$ constitutes a wider PrBound pair for $P_j(\cdot)$ than $\{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$. So when it is required to compute the PrBound pairs $\{\mathbb{L}_j(a_j), \mathbb{U}_j(a_j)\}$ *at all elements of the powerset of* Θ_j, a computationally more tractable, but admittedly more conservative, alternative is to compute $\{\mathbb{L}_j(a_j), \mathbb{U}_j(a_j)\}$ *only at the singletons of* Θ_j. The i.v. naïve Bayes' classifier in Section 5.5 exploits this property.

We now have the following interesting result which allows one to very easily obtain a RecBound pair for an SFE data record:

Lemma 6 *A PrBound pair for the record p.m.f.* $P_R(\cdot)$ *of the SFE data record* R *is* $\{\mathbb{L}_R(\cdot), \mathbb{U}_R(\cdot)\}$, *where*

$$\mathbb{L}_R(\mathbf{a}) = \prod_{j=1}^{N_R} \mathbb{L}_j(a_j); \quad \mathbb{U}_R(\mathbf{a}) = \prod_{j=1}^{N_R} \mathbb{U}_j(a_j),$$

for $\mathbf{a} = (a_1, \ldots, a_{N_R}) \subseteq \Theta_R$. □

Proof: We show the result for the $K = 2$ case; $K > 2$ cases are similar. For $K = 2$, for $\mathbf{a} = (a_1, a_2)$, use Definition 7 to get

$$\begin{aligned}
\mathbb{L}_R(\mathbf{a}) &= \max\{0, \mathbb{L}(a_1) + \mathbb{L}(a_2) - 1\} \\
&= \begin{cases} 1, & \text{if } \mathbb{L}(a_1) = \mathbb{L}(a_2) = 1; \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & \text{if } \mathbf{a} \supseteq B; \\ 0, & \text{if } \mathbf{a} \not\supseteq B; \end{cases} \\
\mathbb{U}_R(\mathbf{a}) &= \min\{\mathbb{U}(a_1), \mathbb{U}(a_2)\} \\
&= \begin{cases} 1, & \text{if } \mathbb{U}(a_1) = \mathbb{U}(a_2) = 1; \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & \text{if } \mathbf{a} \cap B \neq \emptyset; \\ 0, & \text{if } \mathbf{a} \cap B = \emptyset. \end{cases}
\end{aligned}$$

So, $\mathbb{L}_R(\mathbf{a}) = \mathbb{L}(a_1)\,\mathbb{L}(a_2)$ and $\mathbb{U}_R(\mathbf{a}) = \mathbb{U}(a_1)\,\mathbb{U}(a_2)$. ■

Noting Remark 2 above (i.e., the AttBound pair of an SFE attribute can only assume the values $\{0, 0\}$, $\{0, 1\}$, or $\{1, 1\}$), a computationally efficient method to get the RecBound pair of an SFE data record is

$$\begin{aligned}
&\mathbb{L}_R(\mathbf{a}) = 1, \text{ if } \mathbb{L}_j(a_j) = 1, \forall j; \ \mathbb{L}_R(\mathbf{a}) = 0, \text{ otherwise;} \\
&\mathbb{U}_R(\mathbf{a}) = 1, \text{ if } \mathbb{U}_j(a_j) = 1, \forall j; \ \mathbb{U}_j(a_j) = 0, \text{ otherwise.}
\end{aligned} \tag{5.3}$$

5.3 Parameter Mining From Uncertain Data

Consider a dataset $\mathfrak{D}$ of N_D records R_i, $i = 1, \ldots, N_D$, where the extra subscript i is used to differentiate between data records. As before, A_j, $j = 1, \ldots, N_R$, denote the attribute variables and $\Theta_j = \{\theta_{1j}, \ldots, \theta_{N_j j}\}$ denotes the corresponding sample space. By $\langle A_j = a_{ji} \rangle$, or simply by a_{ji}, we denote the (potentially uncertain) state $a_{ji} \subseteq \Theta_j$ that the attribute A_j of the record R_i assumes.

5.3.1 General Case

Definition 8 *Suppose* $P_{R_i}(\cdot) \lhd \{\mathbb{L}_{R_i}(\cdot), \mathbb{U}_{R_i}(\cdot)\}$ *for the record p.m.f.* $P_{R_i}(\cdot)$ *defined over the cross-product* Θ_R.

(i) The pair $\{\mathbb{L}_{\mathfrak{D}}(\cdot), \mathbb{U}_{\mathfrak{D}}(\cdot)\}$, *where*

$$\mathbb{L}_{\mathfrak{D}}(\mathbf{a}) = \sum_{i=1}^{N_D} \mathbb{L}_{R_i}(\mathbf{a})/N_D; \;\; \mathbb{U}_{\mathfrak{D}}(\mathbf{a}) = \sum_{i=1}^{N_D} \mathbb{U}_{R_i}(\mathbf{a})/N_D,$$

for $\mathbf{a} \subseteq \Theta_R$, *is referred to as a* DatasetBound pair *associated with the dataset* $\mathfrak{D}$.

(ii) For $\mathbf{a}, \mathbf{a}' \subseteq \Theta_R$ *s.t.* $\Delta_{\mathbb{L}(\mathbf{a}|\mathbf{a}')} + \Delta_{\mathbb{U}(\mathbf{a}|\mathbf{a}')} > 0$, *the pair* $\{\mathbb{L}_{\mathfrak{D}}(\mathbf{a}|\mathbf{a}'), \mathbb{U}_{\mathfrak{D}}(\mathbf{a}|\mathbf{a}')\}$ *constructed as specified in Theorem 2 is referred to as the* conditional pair at $\mathbf{a}$ given $\mathbf{a}'$ *associated with the dataset* $\mathfrak{D}$. ■

Note that, $\Delta_{\mathbb{L}(\mathbf{a}|\mathbf{a}')} = \mathbb{L}_{\mathfrak{D}}(\mathbf{a} \cap \mathbf{a}') + \mathbb{U}_{\mathfrak{D}}(\overline{\mathbf{a}} \cap \mathbf{a}')$ and $\Delta_{\mathbb{U}(\mathbf{a}|\mathbf{a}')} = \mathbb{U}_{\mathfrak{D}}(\mathbf{a} \cap \mathbf{a}') + \mathbb{L}_{\mathfrak{D}}(\overline{\mathbf{a}} \cap \mathbf{a}')$ (see Theorem 2).

5.3.2 SFE Dataset Case

A justification for Definition 8 emerges as an intuitive frequency counting method for extracting the DatasetBound pair from an SFE dataset. Consider the SFE data record

$$R_i = [\langle A_1 = \alpha_{1i} \rangle, \ldots, \langle A_{N_R} = \alpha_{N_R i} \rangle],\ \alpha_{ji} \subseteq \Theta_j. \tag{5.4}$$

For convenience, we will use $R_i = \langle \mathbf{A} = \boldsymbol{\alpha}_i \rangle$, where $\boldsymbol{\alpha}_i = (\alpha_{1i}, \ldots, \alpha_{N_R i})$, to denote the data record in (5.4); $\mathbf{a} = (a_1, \ldots, a_{N_R})$, $a_j \subseteq \Theta_j$, denotes an arbitrary vector. Notice how the computation may proceed:

5.3.2.1 Computation of Bound Pairs

Let us compute the following:

5.3.2.1.1 AttBound Pair of Attribute A_j of R_i From Definition 7,

$$\begin{aligned} &\mathbb{L}_{jR_i}(a_j) = 1 \text{ if } a_j \supseteq \alpha_{ji}; \qquad \text{else } \mathbb{L}_{jR_i}(a_j) = 0; \\ &\mathbb{U}_{jR_i}(a_j) = 1 \text{ if } a_j \cap \alpha_{ji} \neq \emptyset; \text{ else } \mathbb{U}_{jR_i}(a_j) = 0. \end{aligned} \tag{5.5}$$

5.3.2.1.2 RecBound Pair of R_i Use (5.3) to get these. We also note that $\{\mathbb{L}_{jR_i}(\Theta_j), \mathbb{U}_{jR_i}(\Theta_j)\} = \{1, 1\}$. So, one can ignore AttBound pairs for which $a_j = \Theta_j$ in computing (5.3).

5.3.2.1.3 DatasetBound Pair of $\mathfrak{D}$ From Definition 8,

$$\begin{aligned} &\mathbb{L}_{\mathfrak{D}}(\mathbf{a}) = (\#\text{ of records where } \mathbf{a} \supseteq \boldsymbol{\alpha}_i)/N_D; \\ &\mathbb{U}_{\mathfrak{D}}(\mathbf{a}) = (\#\text{ of records where } \mathbf{a} \cap \boldsymbol{\alpha}_i \neq \emptyset)/N_D. \end{aligned} \tag{5.6}$$

Table 5.2: Dataset $\mathfrak{D}$ in Example 8.
$\mathfrak{D}$ contains 3 data records $\{R_1, R_2, R_3\}$; $\mathbf{a} = (\theta_{22}, \theta_{23})$, $\mathbf{a}' = (\theta_{22}, (\theta_{13}, \theta_{33}))$, $\mathbf{c} = \theta_{23}$, $\mathbf{c}' = \theta_{22}$.

	$\langle A_1 = \alpha_{1i} \rangle$	$\langle A_2 = \alpha_{2i} \rangle$	$\langle A_3 = \alpha_{3i} \rangle$	$\langle A_4 = \alpha_{4i} \rangle$	$\langle A_5 = \alpha_{5i} \rangle$	
Data Record	$\Theta_1 =$	$\Theta_2 =$	$\Theta_3 =$	$\Theta_4 =$	$\Theta_5 =$	**RecBound**
	$\{\theta_{11}, \theta_{21}, \theta_{31}\}$	$\{\theta_{12}, \theta_{22}\}$	$\{\theta_{13}, \theta_{23}, \theta_{33}\}$	$\{\theta_{14}, \theta_{24}\}$	$\{\theta_{15}, \theta_{25}, \theta_{35}, \theta_{45}, \theta_{55}\}$	
$R_1 = \langle \mathbf{A} = \boldsymbol{\alpha}_1 \rangle$	θ_{21}	θ_{22}	θ_{23}	Θ_4	Θ_5	
AttBounds @ $\mathbf{a}$		[1, 1]	[1, 1]			[1, 1]
AttBounds @ $\mathbf{a}'$		[1, 1]	[0, 0]			[0, 0]
$R_2 = \langle \mathbf{A} = \boldsymbol{\alpha}_2 \rangle$	$(\theta_{21}, \theta_{31})$	Θ_2	$(\theta_{23}, \theta_{33})$	θ_{14}	θ_{15}	
AttBounds @ $\mathbf{a}$		[0, 1]	[0, 1]			[0, 1]
AttBounds @ $\mathbf{a}'$		[0, 1]	[0, 1]			[0, 1]
$R_3 = \langle \mathbf{A} = \boldsymbol{\alpha}_3 \rangle$	Θ_1	Θ_2	θ_{23}	Θ_4	$(\theta_{15}, \theta_{25})$	
AttBounds @ $\mathbf{a}$		[0, 1]	[1, 1]			[0, 1]
AttBounds @ $\mathbf{a}'$		[0, 1]	[0, 0]			[0, 0]

Example 8 *Consider a dataset $\mathfrak{D}$ of $N_D = 3$ data records $\{R_1, R_2, R_3\}$, each having $N_R = 5$ attribute variables $\{A_1, A_2, A_3, A_4, A_5\}$ with $\{\Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5\}$ being the corresponding sample spaces. Table 5.2 shows the 3 data records $R_i = \langle \mathbf{A} = \boldsymbol{\alpha}_i \rangle$, $i = 1, \ldots, 3$.*

Consider the bound pairs at

$$\mathbf{a} = (\Theta_1, \theta_{22}, \theta_{23}, \Theta_4, \Theta_5); \quad \mathbf{a}' = (\Theta_1, \theta_{22}, (\theta_{13}, \theta_{33}), \Theta_4, \Theta_5),$$

or, more compactly written without the unknown/missing values,

$$\mathbf{a} = (\theta_{22}, \theta_{23}); \;\; \mathbf{a}' = (\theta_{22}, (\theta_{13}, \theta_{33})).$$

(i) AttBound Pairs: *Apply* (5.5). *We need to compute AttBounds for attributes A_2 and A_3 only. All other AttBound pairs (Table 5.2 does not show these) are [1, 1].*

Consider R_1: for attribute A_2, $a_2 = \theta_{22} \supseteq \alpha_{21} = \theta_{22}$ so that $\mathbb{L}_{2R_1}(\mathbf{a}) = 1$ which implies that $\mathbb{U}_{2R_1}(\mathbf{a}) = 1$, i.e, $\{\mathbb{L}_{2R_1}(\mathbf{a}), \mathbb{U}_{2R_1}(\mathbf{a})\} = \{1, 1\}$; for attribute A_3, $a_3 = \theta_{23} \supseteq \alpha_{31} = \theta_{23}$ and so $\{\mathbb{L}_{3R_1}(\mathbf{a}), \mathbb{U}_{3R_1}(\mathbf{a})\} = \{1, 1\}$.

Consider R_2*: for attribute* A_2*,* $a_2 = \theta_{22} \not\supseteq \alpha_{22} = \Theta_2$ *so that* $\mathbb{L}_{2R_2}(\mathbf{a}) = 0$*, but* $a_2 \cap \alpha_{22} = \theta_{22} \cap \Theta_2 \neq \emptyset$ *so that* $\mathbb{U}_{2R_2}(\mathbf{a}) = 1$*, i.e.,* $\{\mathbb{L}_{2R_2}(\mathbf{a}), \mathbb{U}_{2R_2}(\mathbf{a})\} = \{0,1\}$*; for attribute* A_3*,* $a_3 = \theta_{23} \not\supseteq \alpha_{32} = (\theta_{23}, \theta_{33})$ *so that* $\mathbb{L}_{3R_2}(\mathbf{a}) = 0$*, but* $a_3 \cap \alpha_{32} = \theta_{23} \cap (\theta_{23}, \theta_{33}) \neq \emptyset$ *so that* $\mathbb{U}_{3R_2}(\mathbf{a}) = 1$*, i.e.,* $\{\mathbb{L}_{3R_2}(\mathbf{a}), \mathbb{U}_{3R_2}(\mathbf{a})\} = \{0,1\}$*.*

Consider R_3*: for attribute* A_2*,* $a_2 = \theta_{22} \not\supseteq \alpha_{23} = \Theta_2$ *so that* $\mathbb{L}_{2R_3}(\mathbf{a}) = 0$*, but* $a_2 \cap \alpha_{23} = \theta_{22} \cap \Theta_2 \neq \emptyset$ *so that* $\mathbb{U}_{2R_3}(\mathbf{a}) = 1$*, i.e.,* $\{\mathbb{L}_{2R_3}(\mathbf{a}), \mathbb{U}_{2R_3}(\mathbf{a})\} = \{0,1\}$*; for attribute* A_3*,* $a_3 = \theta_{23} \supseteq \alpha_{33} = \theta_{23}$ *so that* $\mathbb{L}_{3R_3}(\mathbf{a}) = 1$ *which implies that* $\mathbb{U}_{3R_3}(\mathbf{a}) = 1$*, i.e.,* $\{\mathbb{L}_{3R_3}(\mathbf{a}), \mathbb{U}_{3R_3}(\mathbf{a})\} = \{1,1\}$*.*

These AttBounds at $\mathbf{a}$*, and those at* $\mathbf{a}'$ *which are computed similarly, appear in Table 5.2.*

(ii) RecBound Pairs: *Apply* (5.3) *to get the RecBounds. These appear in the last column of Table 5.2.*

(iii) DatasetBound Pairs: *Simply apply Definition 8(i):*

$$\begin{aligned}
\{\mathbb{L}_{\mathfrak{D}}(\mathbf{a}), \mathbb{U}_{\mathfrak{D}}(\mathbf{a})\} &= ([1,1] + [0,1] + [0,1])/3 = \{1/3, 1\};\\
\{\mathbb{L}_{\mathfrak{D}}(\mathbf{a}'), \mathbb{U}_{\mathfrak{D}}(\mathbf{a}')\} &= ([0,0] + [0,1] + [0,0])/3 = \{0, 1/3\}.
\end{aligned}$$

(iv) Conditional Pairs: *We compute* $\{\mathbb{L}_{\mathfrak{D}}(\mathbf{c}|\mathbf{c}'), \mathbb{U}_{\mathfrak{D}}(\mathbf{c}|\mathbf{c}')\}$*, for* $\mathbf{c} = \theta_{23}$ *and* $\mathbf{c}' = \theta_{22}$*, which we have deliberately selected s.t.* $\mathbf{c} \cap \mathbf{c}' = \mathbf{a}$ *and* $\overline{\mathbf{c}} \cap \mathbf{c}' = \mathbf{a}'$*. Using Theorem 2 in Definition 8(ii) we get* $\{\mathbb{L}_{\mathfrak{D}}(\mathbf{c}|\mathbf{c}'), \mathbb{U}_{\mathfrak{D}}(\mathbf{c}|\mathbf{c}')\} = \{1/2, 1\}$*.* ■

Example 9 (Example 1 Revisited) *Let* $\mathbf{x} = (a_1, b_1, c_1)$*.*

(i) RecBound Pairs @ $\mathbf{x}$: *Applying* (5.3) *at* $\mathbf{x}$*, we get* $\{\mathbb{L}_i(\mathbf{x}), \mathbb{U}_i(\mathbf{x})\}$ *as* $\{1,1\}$ *for* $i = 3$*,* $\{0,1\}$ *for* $i = 1, 5$*, and* $\{0,0\}$ *for* $i = 2, 4, 6, 7$*.*

(ii) DatasetBound Pair @ $\mathbf{x}$: *Applying Definition 8(i) at* $\mathbf{x}$, *we get* $\{\mathbb{L}_{\mathfrak{D}}(\mathbf{x}), \mathbb{U}_{\mathfrak{D}}(\mathbf{x})\} = \{1/7, 3/7\}$.

(iii) Conditional Pair @ a_1 Given c_1: *Here,* $a_1 = (a_1, \Theta_B, \Theta_C)$ *and* $c_1 = (\Theta_A, \Theta_B, c_1)$, *where* $\Theta_A = (a_1, a_2)$, $\Theta_B = (b_1, b_2)$ *and* $\Theta_C = (c_1, c_2)$.

So, $a_1 \cap c_1 = (a_1, \Theta_B, c_1)$ *and* $\overline{a}_1 \cap c_1 = (a_2, \Theta_B, c_1)$. *Apply Definition 8 to get*

$$\begin{aligned} \{\mathbb{L}_{\mathfrak{D}}(a_1 \cap c_1), \mathbb{U}_{\mathfrak{D}}(a_1 \cap c_1)\} &= \{1/7, 3/7\}, \\ \{\mathbb{L}_{\mathfrak{D}}(\overline{a}_1 \cap c_1), \mathbb{U}_{\mathfrak{D}}(\overline{a}_1 \cap c_1)\} &= \{0, 1/7\}, \\ \text{and then } \{\mathbb{L}_{\mathfrak{D}}(a_1|c_1), \mathbb{U}_{\mathfrak{D}}(a_1|c_1)\} &= \{1/2, 1\}. \end{aligned}$$ ■

5.3.2.2 Goodness of DatasetBound Pair

How good is the dataset bound pair $\{\mathbb{L}_{\mathfrak{D}}(\cdot), \mathbb{U}_{\mathfrak{D}}(\cdot)\}$ when compared with the parameters one could have learnt if the dataset had no uncertainties? The following result helps us answer this question:

Lemma 7 *Consider two SFE datasets* $\mathfrak{D}^{(1)}$ *and* $\mathfrak{D}^{(2)}$ *each comprised of the data records* R_i, $i = 1, \ldots, N_D$. *For* $\ell = 1, 2$, *suppose the attribute value vector in* $\mathfrak{D}^{(\ell)}$ *associated with* R_i *is* $\boldsymbol{\alpha}_i^{(\ell)} \subseteq \Theta_R$. *Then the following are true:*

(i) If $\boldsymbol{\alpha}_i^{(1)} \subseteq \boldsymbol{\alpha}_i^{(2)}$, $\forall i$, *then* $\{\mathbb{L}_{\mathfrak{D}^{(1)}}(\mathbf{a}), \mathbb{U}_{\mathfrak{D}^{(1)}}(\mathbf{a})\}$ *is narrower than* $\{\mathbb{L}_{\mathfrak{D}^{(2)}}(\mathbf{a}), \mathbb{U}_{\mathfrak{D}^{(2)}}(\mathbf{a})\}$, $\forall \mathbf{a} \subseteq \Theta_R$.

(ii) If, in addition, $\boldsymbol{\alpha}_k^{(1)} \subset \boldsymbol{\alpha}_k^{(2)}$, *for some* $k = 1, \ldots, N_D$, *then there exist* $\mathbf{a}_1, \mathbf{a}_2 \subseteq \Theta_R$ *s.t.* $\mathbb{L}_{\mathfrak{D}^{(2)}}(\mathbf{a}_1) < \mathbb{L}_{\mathfrak{D}^{(1)}}(\mathbf{a}_1)$ *and* $\mathbb{U}_{\mathfrak{D}^{(1)}}(\mathbf{a}_2) < \mathbb{U}_{\mathfrak{D}^{(2)}}(\mathbf{a}_2)$. □

Proof:

(i) Suppose $\mathbf{b}_i^{(1)} \subseteq \mathbf{b}_i^{(2)}$: Clearly, a record where $\mathbf{a} \supseteq \mathbf{b}_i^{(2)}$ must also have $\mathbf{a} \supseteq \mathbf{b}_i^{(1)}$, meaning that $\mathbb{L}_{\mathfrak{D}^{(2)}}(\cdot) \leq \mathbb{L}_{\mathfrak{D}^{(1)}}(\cdot)$; a record where $\mathbf{a} \cap \mathbf{b}_i^{(1)} \neq \emptyset$ must also have $\mathbf{a} \cap \mathbf{b}_i^{(2)} \neq \emptyset$, meaning that $\mathbb{U}_{\mathfrak{D}^{(1)}}(\cdot) \leq \mathbb{U}_{\mathfrak{D}^{(2)}}(\cdot)$.

(ii) Suppose, in addition, $\mathbf{b}_k^{(1)} \subset \mathbf{b}_k^{(2)}$: Then, it is easy to show that $\mathbb{L}_{\mathfrak{D}^{(2)}}(\mathbf{a}_1) < \mathbb{L}_{\mathfrak{D}^{(1)}}(\mathbf{a}_1)$ at $\mathbf{a}_1 = \mathbf{b}_k^{(1)}$, and $\mathbb{U}_{\mathfrak{D}^{(1)}}(\mathbf{a}_2) < \mathbb{U}_{\mathfrak{D}^{(2)}}(\mathbf{a}_2)$ at $\mathbf{a}_2 = \overline{\mathbf{b}}_k^{(1)} \cap \mathbf{b}_k^{(2)}$. ■

This immediately yields:

Corollary 5 *Consider the SFE dataset $\mathfrak{D}$ comprised of the data records R_i, $i = 1, \ldots, N_D$. Suppose $\mathfrak{D}^*$ denotes the underlying "clean" dataset created by replacing all the SFE attributes in $\mathfrak{D}$ by their "true" attribute values, i.e., to create $\mathfrak{D}^*$, the attribute value vector $\boldsymbol{\alpha}_i \subseteq \Theta_R$ in $\mathfrak{D}$ is replaced by the clean attribute value vector $\boldsymbol{\alpha}_i^* \in \Theta_R$ so that $\boldsymbol{\alpha}_i^* \in \boldsymbol{\alpha}_i$, $\forall i = 1, \ldots, N_D$. Then,*

$$\mathbb{L}_{\mathfrak{D}}(\mathbf{a}) \leq \mathbb{L}_{\mathfrak{D}^*}(\mathbf{a}) = \mathbb{U}_{\mathfrak{D}^*}(\mathbf{a}) \leq \mathbb{U}_{\mathfrak{D}}(\mathbf{a}), \ \forall \mathbf{a} \subseteq \Theta_R. \qquad \square$$

Proof: From Lemma 7, it follows that $\{\mathbb{L}_{\mathfrak{D}^*}(\mathbf{a}), \mathbb{U}_{\mathfrak{D}^*}(\mathbf{a})\}$ is narrower than $\{\mathbb{L}_{\mathfrak{D}}(\mathbf{a}), \mathbb{U}_{\mathfrak{D}}(\mathbf{a})\}$, for all $\mathbf{a} \subseteq \Theta_R$. For the clean dataset $\mathfrak{D}^*$, (5.6) yields the claim, viz.,

$$\mathbb{L}_{\mathfrak{D}^*}(\mathbf{a}) = \mathbb{U}_{\mathfrak{D}^*}(\mathbf{a}) = (\text{\# of records where } \mathbf{a} = \mathbf{b}_i^*)/N_D. \qquad ■$$

Corollary 5 essentially states that, when the frequency counting strategy in Definition 8 is employed for parameter learning, the underlying probabilities are guaranteed to be constrained within the corresponding DatasetBound pair.

5.4 Experiment 1: Bound Propagation in an I.V. BN

Our first experiment illustrates how bounds of parameters of an i.v. BN could be learnt from an imperfect dataset and how they could be propagated within the i.v. BN.

5.4.1 Dataset

We use the breast cancer (BC) dataset in [Velikova et al., 2012] which consists of 20,000 records (each representing a patient) with 16 feature attribute variables, each attribute having multiple states. This dataset, which we refer to as the *ground truth (GTR) dataset,* is clean in that it has no data uncertainties.

5.4.2 Model

Out of the 16 attributes, we selected the 5 attributes $\{A, C, M, E, S\}$ which appear in Table 5.3. The corresponding sample spaces are $\Theta_A = \{\theta_{1A}, \theta_{2A}, \theta_{3A}, \theta_{4A}\}$, $\Theta_C = \{\theta_{1C}, \theta_{2C}, \theta_{3C}\}$, $\Theta_M = \{\theta_{1M}, \theta_{2M}, \theta_{3M}\}$, $\Theta_E = \{\theta_{1E}, \theta_{2E}, \theta_{3E}\}$, and $\Theta_S = \{\theta_{1S}, \theta_{2S}, \theta_{3S}, \theta_{4S}\}$.

Table 5.3: Breast Cancer Dataset: Details of the 5 Selected Attributes

Age	**Breast Cancer**	**Mass**	**Extent/Size**	**Shape**
A (Yrs)	C	M	E (cm)	S
States:				
θ_{1A} = <35	θ_{1C} = No	θ_{1M} = No	θ_{1E} = <1	θ_{1S} = Oval
θ_{2A} = 35-49	θ_{2C} = In-Situ	θ_{2M} = Benign	θ_{2E} = 1-3	θ_{2S} = Round
θ_{3A} = 50-74	θ_{3C} = Invasive	θ_{3M} = Malignant	θ_{3E} = >3	θ_{3S} = Irregular
θ_{4A} = >74	—	—	—	θ_{4S} = Other

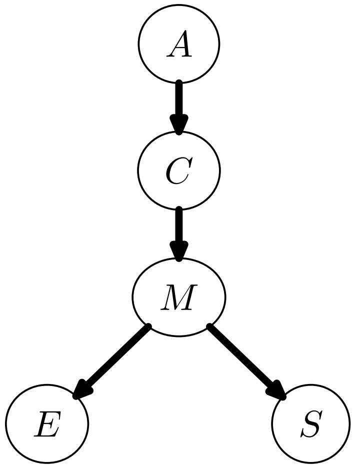


Figure 5.1: BN for the 5 Selected Attributes of the BC Dataset

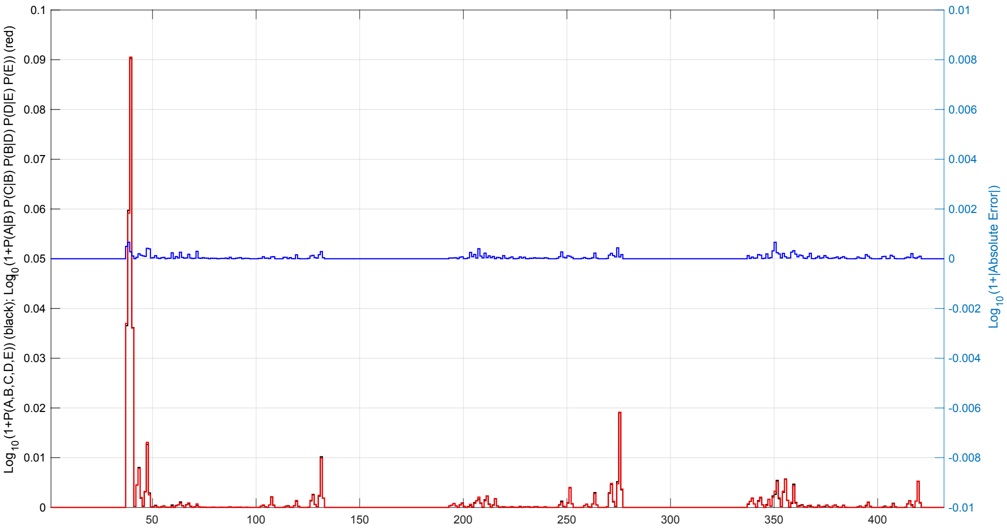
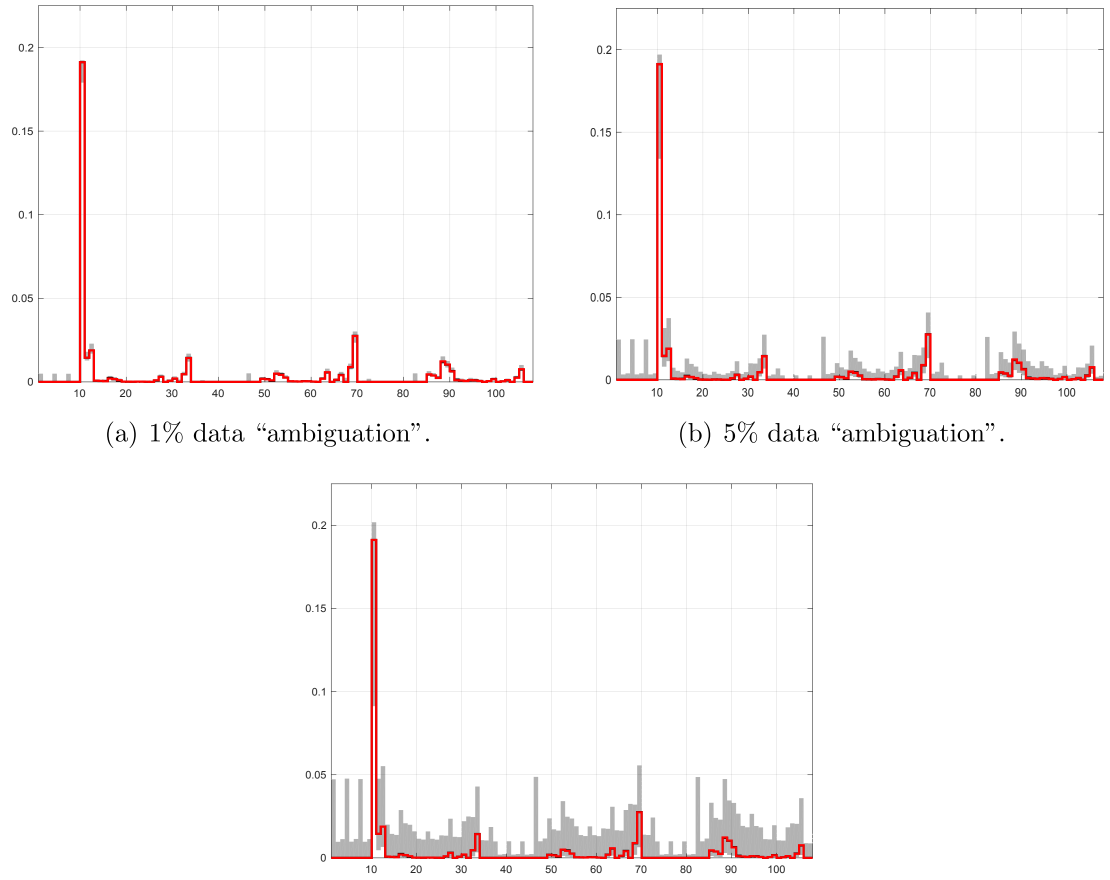


Figure 5.2: Comparison of $P(A, C, M, E, S)$ and $P(E|M)\,P(S|M)\,P(M|C)\,P(C|A)\,P(A)$. Both $P(A, C, M, E, S)$ (black) and $P(E|M)P(S|M)P(M|C)P(C|A)P(A)$ (red) are estimated directly from the GTR dataset. The absolute value of the difference $|P(A, C, M, E, S)$-$P(E|M)P(S|M)P(M|C)P(C|A)P(A)|$ (blue, right-hand axis @ 1/10th the scale) confirms the validity of the BN in Fig. 5.1. *Note:* For clarity, plots show variable x as $\log_{10}(1 + x)$.



(a) 1% data “ambiguation”.

(b) 5% data “ambiguation”.

(c) 10% data “ambiguation”.

Figure 5.3: Estimation of the PrBound Pairs $[\widetilde{\mathbb{L}}(C, M, E, S), \widetilde{\mathbb{U}}(C, M, E, S)]$.
Note. The BN in Fig. 5.1 is used to estimate the bounds $[\widetilde{\mathbb{L}}(C, M, E, S), \widetilde{\mathbb{U}}(C, M, E, S)]$ (gray) and $\widetilde{P}(C, M, E, S)$ (red) from the GTR dataset; the BN is not used to estimate $P(C, M, E, S)$ (black). The difference between the red and black plots is indistinguishable. Figs 5.3(a), 5.3(b), and 5.3(c) correspond to 1%, 5%, and 10% levels of data “ambiguation”, respectively. *Note:* For clarity, plots show variable x as $\log_{10}(1 + x)$.

Relying on intuition, the BN model in Fig. 5.1 is taken to represent the interrelationships among these 5 attributes. To test its validity, we used frequency counting to get GTR dataset-based estimates of the joint p.m.f. $P(A, C, M, E, S)$ and the product $P(E|M)\, P(S|M)\, P(M|C)\, P(C|A)\, P(A)$. These two quantities and the error between them are plotted in Fig. 5.2. The horizontal axis spans 1 to $4 \times 3 \times 3 \times 3 \times 4 = 432$ possible states that the 5 variables may assume. The two plots agree with each other extremely well confirming that the BN in Fig. 5.1 is a reasonable model for the 5 attributes.

Introducing Uncertainty. We employ a principled approach to artificially muddle (or "ambiguate") the GTR dataset to generate an *imperfect (IMP) dataset* possessing only SFE records. The 5 selected attributes provide $5 \times 20,000 = 100,000$ data points. We first randomly selected 1% of these data points. The value of a selected data point was then "ambiguated" by replacing it with a set of values which includes the original value. For example, if a selected data point has value θ_{1M}, then it was replaced by $(\theta_{1M}, \theta_{2M})$, $(\theta_{1M}, \theta_{3M})$, or $\Theta_M = (\theta_{1M}, \theta_{2M}, \theta_{3M})$; the cardinality of the new set of values was randomly selected.

5.4.3 Method

We can now compute any probability of interest, or more precisely, its corresponding PrBound pair, from the IMP dataset-based i.v. BN and compare it with the result that the GTR dataset-based BN gives.

5.4.4 Results

Let us see how we can use the i.v. BN in Fig. 5.1 to get the answers to some questions.

Question 1. What is the Marginal $P(C, M, E, S)$?

Note that

$$\begin{aligned} \widetilde{P}(A, C, M, E, S) &= \widetilde{P}(E|M)\, \widetilde{P}(S|M)\, \widetilde{P}(M|C)\, \widetilde{P}(C|A) \widetilde{P}(A); \\ \widetilde{P}(C, M, E, S) \quad &= \widetilde{P}(E|M)\, \widetilde{P}(S|M)\, \widetilde{P}(M|C)\, \widetilde{P}(C). \end{aligned} \tag{5.7}$$

Here $\widetilde{P}(\cdot)$ identifies quantities that are based on the validity of the BN in Fig. 5.1. So we have $\widetilde{P}(C, M, E, S) \lhd \{\widetilde{\mathbb{L}}(C, M, E, S), \widetilde{\mathbb{U}}(C, M, E, S)\}$, where

$$\begin{aligned} \widetilde{\mathbb{L}}(C, M, E, S) &= \widetilde{\mathbb{L}}(E|M)\, \widetilde{\mathbb{L}}(S|M)\, \widetilde{\mathbb{L}}(M|C)\, \widetilde{\mathbb{L}}(C); \\ \widetilde{\mathbb{U}}(C, M, E, S) &= \widetilde{\mathbb{U}}(E|M)\, \widetilde{\mathbb{U}}S|M)\, \widetilde{\mathbb{U}}(M|C)\, \widetilde{\mathbb{U}}(C). \end{aligned} \tag{5.8}$$

IMP dataset-based estimates of the right-hand side terms in (5.8) are now obtained using the methods in Section 5.3. Fig. 5.3(a) depicts the PrBound intervals $[\widetilde{\mathbb{L}}(C, M, E, S), \widetilde{\mathbb{U}}(C, M, E, S)]$ so obtained and the GTR dataset-based estimates of $\widetilde{P}(C, M, E, S)$ and $P(C, M, E, S)$. Note that, $\widetilde{P}(C, M, E, S)$ is based on the BN in Fig. 5.1 and hence is computed from the right-hand side of (5.7); $P(C, M, E, S)$ does not make use of the BN.

We get similar results when the experiment is repeated with 5% and 10% levels of data "ambiguation" (see Figs 5.3(b) and 5.3(c), respectively), except that the PrBound intervals get increasingly wider. Importantly, note that $\widetilde{P}(C, M, E, S)$ and $P(C, M, E, S)$ are guaranteed to lie within their corresponding PrBound intervals $[\widetilde{\mathbb{L}}(C, M, E, S), \widetilde{\mathbb{U}}(C, M, E, S)]$ and $[\mathbb{L}(C, M, E, S), \mathbb{U}(C, M, E, S)]$, respectively. The fact that $\widetilde{P}(C, M, E, S)$ and $P(C, M, E, S)$ agree very well (see Fig. 5.3) further confirms the BN model's validity.

Question 2. Given that our data is imperfect, can we determine the probability that a patient's tumor is in-situ and benign if her age is 63 (yrs) and the extent and shape of the tumor is over 3 (cm) and round, respectively? Use the BN in Fig. 5.1:

$$P(C, M|A, E, S) = \frac{P(A, C, M, E, S)}{P(A, E, S)} \quad = \frac{P(E|M)\, P(S|M)\, P(M|C)\, P(C|A)}{\sum\limits_{M} P(E|M)\, P(S|M)\, P(M|A)}. \tag{5.9}$$

So, a PrBound pair for $P(C, M|A, E, S)$ is

$$\left\{ \frac{\mathbb{L}(E|M)\, \mathbb{L}(S|M)\, \mathbb{L}(M|C)\, \mathbb{L}(C|A)}{\sum\limits_{M} \mathbb{U}(E|M)\, \mathbb{U}(S|M)\, \mathbb{U}(M|A)}, \; \frac{\mathbb{U}(E|M)\, \mathbb{U}(S|M)\, \mathbb{U}(M|C)\, \mathbb{U}(C|A)}{\sum\limits_{M} \mathbb{L}(E|M)\, \mathbb{L}(S|M)\, \mathbb{L}(M|A)} \right\}. \tag{5.10}$$

With the proposed framework, we get $P(C, M|A, E, S) \lhd [0.3130, 0.3570]$ for $C =$ In-Situ, $M =$ Benign, $50 \le A \le 74$ (yrs), $E > 3$ (cm), and $S =$ Round.

Question 3. Attributes themselves could be ambiguous as well, e.g., we get $P(C, M|A, E, S) \lhd [0.4297, 0.5911]$ for $C = (\text{In-Situ}, \text{Invasive})$, $M =$ Malignant, $A \ge 50$ (yrs), $E < 1$ (cm), and $S = (\text{Oval}, \text{Round})$.

5.5 Experiment 2: I.V. Naïve Bayes Classifier

We now illustrate the use of an i.v. naïve Bayes' classifier.

5.5.1 Dataset

This experiment is based on data gathered in the *AstroML (Machine Learning and Data Mining for Astronomy)* project and the *Sloan Digital Sky Survey* [VanderPlas et al., 2012, Ivezic et al., 2014], where the naïve Bayes' classifier is used to classify objects as stars and quasars based on four photometric variables. A quasar is an active galactic nucleus of very high luminosity. On Earth quasars appear as stars making it difficult to distinguish between stars and quasars. The classification is mainly done via their spectra where quasars exhibit greater redshift compare to stars due to the distance and the expansion of the universe (Fig. 5.4).

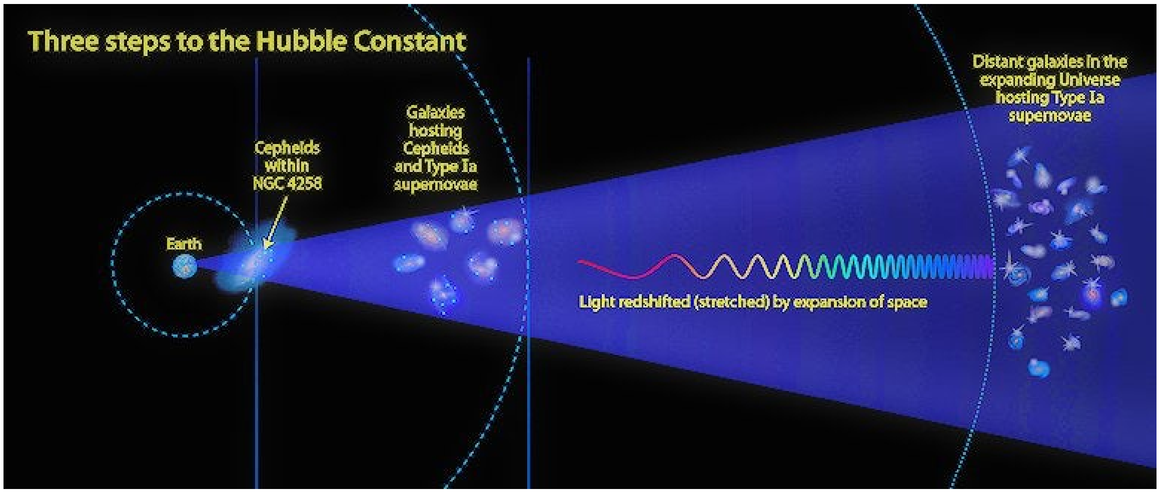

Three steps to the Hubble Constant
Earth
Cepheids within NGC 4258
Galaxies hosting Cepheids and Type Ia supernovae
Light redshifted (stretched) by expansion of space
Distant galaxies in the expanding Universe hosting Type Ia supernovae

Figure 5.4: Redshift

In the dataset, each spectrum is divided into five regions, u (ultraviolet), g (green), r (red), i (i-infrared), and z (z-infrared) (Fig 5.5). Different factors (e.g., atmospheric conditions, time of the measurement, etc.) may affect flux values within each region, and to compensate for these changes, only the relative photometric values $u-g$, $g-r$, $r-i$, and $i-z$ are used for classification purposes. This dataset, which is treated as the *ground truth (GTR) dataset,* is also clean. It contains 705,290 data records, each record consisting of the four relative photometric values (each spanning the range $[-21, +20]$) and the corresponding red-shift value (which determines the object label). The percentages of records of stars and quasars are 87.56% and 12.44%, respectively.

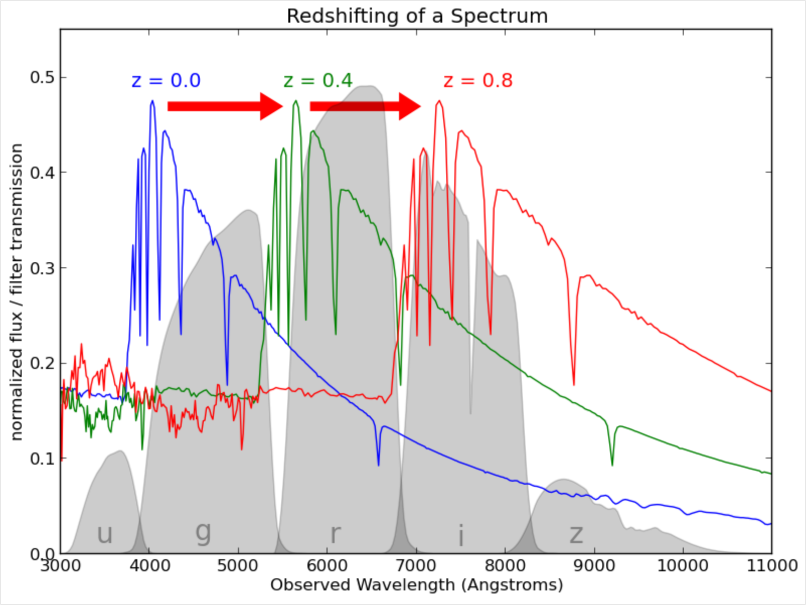


Figure 5.5: Redshfting of a Spectrum

5.5.2 Model

We use A to identify the classification or “hidden” variable which has two states, STAR and QUASAR, i.e., $\Theta_A = \{\text{STAR}, \text{QUASAR}\}$. The observed variables $\{B_i\}$, $i = \overline{1,4}$, which are taken to be conditionally independent given A, correspond to the 4 photometric variables, $u{-}g$, $g{-}r$, $r{-}i$, and $i{-}z$, respectively. We discretize the observation values to create a finite sample space for each observed variable. Let Θ_j denote this sample space associated with the observed variable B_j so that $b_j \subseteq \Theta_j$, where $\langle B_j = b_j \rangle$. The classification decision is made from $P(A|B)$ where $\langle B = \mathbf{b} \rangle$, $\mathbf{b} = (b_1, b_2, b_3, b_4)$. So, we essentially have the naïve Bayes' graphical model in Fig. 4.1 (with $n = 4$).

Introducing Uncertainty. We generate an *imperfect (IMP) dataset* having an SFE structure by first randomly selecting $r\%$ of data points and then “ambiguating” each selected data point by replacing its *value,* say x_i, by an *interval* (of width Δ)

which contains x_i. We refer to Δ as the *"ambiguation" width.* The location of x_i within this interval is randomly selected. For our experiments, we "ambiguated" only the attributes and not the class label.

5.5.3 Method

For a selected resolution q (the width of a discretized state), each observed value x_i is rounded and allocated its corresponding discretized state; an "ambiguated" i.v. entry is allocated all discretized states falling within its interval.

We use the frequency count of each discretized state and construct a frequency histogram. The resolution q affects the classification performance: with a smaller q, the number of samples available to estimate the frequency of each discretized state becomes insufficient; with a larger q, the frequency histograms lose their variability in shape.

We employed 5-fold cross validation with a $\{80\%, 20\%\}$ random split of data records for training and testing.

5.5.3.1 Training

Note that, Lemma 3 with $n = 4$ yields

$$\begin{aligned}\mathbb{L}(a_k \downarrow \mathbf{b}) &= \frac{\mathbb{L}(a_k)\prod_{j=1}^{n}\mathbb{L}(b_j|a_k)}{\mathbb{L}(a_k)\prod_{j=1}^{n}\mathbb{L}(b_j|a_k) + \mathbb{U}(\overline{a}_k)\prod_{j=1}^{n}\mathbb{U}(b_j|\overline{a}_k)};\\ \mathbb{U}(a_k \downarrow \mathbf{b}) &= \frac{\mathbb{U}(a_k)\prod_{j=1}^{n}\mathbb{U}(b_j|a_k)}{\mathbb{U}(a_k)\prod_{j=1}^{n}\mathbb{U}(b_j|a_k) + \mathbb{L}(\overline{a}_k)\prod_{j=1}^{n}\mathbb{L}(b_j|\overline{a}_k)},\end{aligned} \tag{5.11}$$

where $a_k \in \{\text{STAR}, \text{QUASAR}\}$. We employ the methods in Section 5.3 to mine the terms in the right-hand side of (5.11) from the imperfect training dataset.

5.5.3.1.1 Accommodating Uncertain Observations It is often the case that attributes within the observation vector $\mathbf{b}$ assume non-singleton values. With such uncertain observations, the calculation of the posterior pairs (left-hand sides of (5.11)) require the computation of likelihood pairs $\{\mathbb{L}(b_j|a_k), \mathbb{U}(b_j|a_k)\}$ (right-hand side of (5.11)) where some b_j s are now non-singletons.

If we restrict the data uncertainty of each attribute to be of the SFE type, we may proceed by adopting one of the following approaches:

- *Method (1):* Mine all the parameters $\{\mathbb{L}(b_j|a_k), \mathbb{U}(b_j|a_k)\}$ from the training set (where b_j could be a non-singleton). The main disadvantage of this method is the computational complexity because, in the worst case, one would have to find $2^{|\Theta_j|} - 1$ number of parameters for each $\mathbb{L}(b_j|a_k)$ and $\mathbb{U}(b_j|a_k)$. This is the method that we used in Section 5.5 with the breast cancer dataset because it has less number of parameters.

- *Method (2):* Compute $\{\mathbb{L}(b_{ij}|a_k), \mathbb{U}(b_{ij}|a_k)\}$, for all singleton $b_{ij} \in b_j$ only and make use of Corollary 3. This offers significant computational savings but the PrBounds are more conservative. This is the method we employed for this star/quasar classification task. Note that if the result for upper PrBound from Corollary 3 grows bigger than 1, then 1 can be used as the upper bound.

- *Method (3):* Use the PrBound pair $\{\mathbb{L}(b_j|a_k), \mathbb{U}(b_j|a_k)\}$ with

$$\mathbb{L}(b_j|a_k) = \min_{b_{ij} \in b_j} \{\mathbb{L}(b_{ij}|a_k)\}; \quad \mathbb{U}(b_j|a_k) = \max_{b_{ij} \in b_j} \{\mathbb{U}(b_{ij}|a_k)\}. \tag{5.12}$$

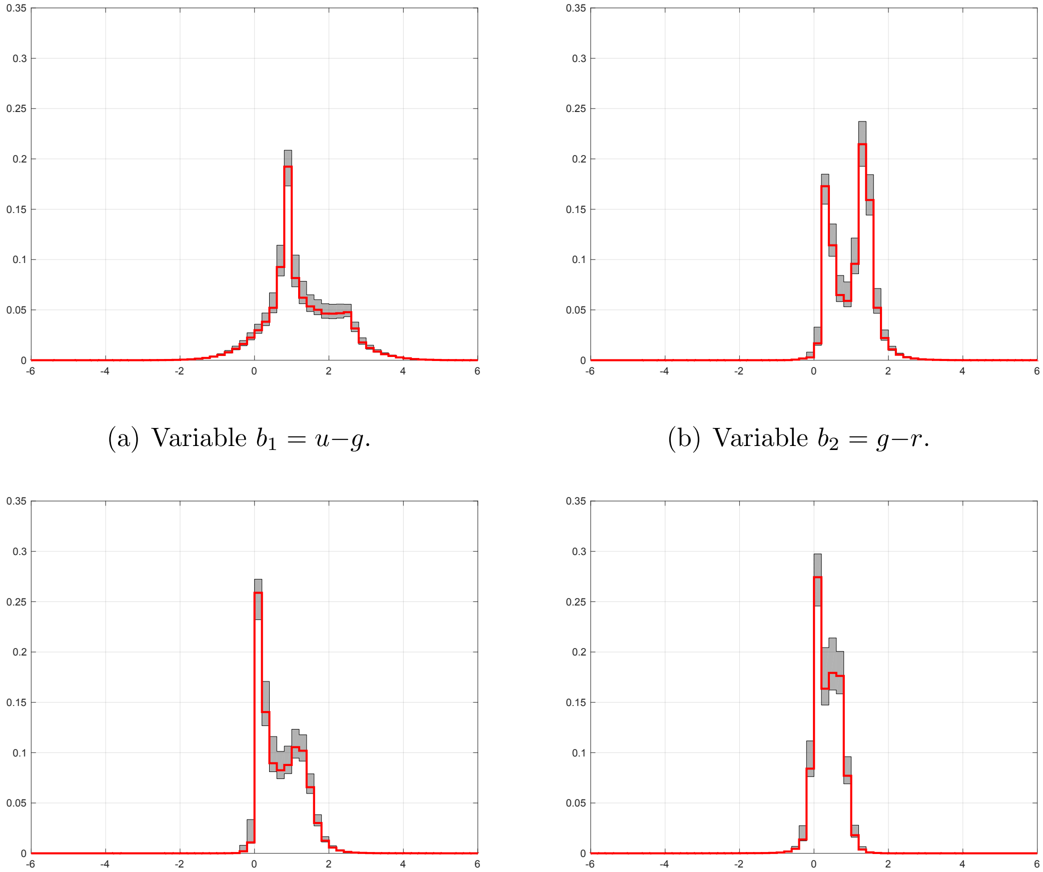
  This method, which is quite similar to the *conservative inference rule* in [Augustin et al., 2014], can be considered the most conservative method.

- *Method (4):* Ignore uncertain observations and use only clean observations for classification purposes. This is equivalent to using $\{\mathbb{L}(b_j|a_k), \mathbb{U}(b_j|a_k)\} = \{1, 1\}$ in (5.11) whenever b_j is a non-singleton. Of course, as Example 1 illustrates, this can generate flawed results.

Remark: For all of the above methods, when $b_j = \Theta_j$, $\mathbb{L}(\Theta_j|a_k) = \mathbb{U}(\Theta_j|a_k) = 1$ can be employed.

For the GTR dataset, we can represent the mined data as 8 probability histograms, 4 each for the observed variables of STARs and QUASARs. For the IMP dataset, we get 16 such histograms, 8 each for the lower and upper PrBounds. Fig. 5.6 shows the lower/upper histograms of $\{\mathbb{L}(b_j|\text{STAR}), \mathbb{U}(b_j|\text{STAR})\}$ for the 4 photometric variables (b_j s are singletons in this case). Fig. 5.6 also shows the GTR dataset-based estimates (red) and, as claimed, these lie within their corresponding lower/upper histograms.

5.5.3.1.2 Computational Complexity One can employ alternate methods to arrive at similar bounds. For example, an enumeration-optimization method would be more exhaustive and, in some cases, may even produce tighter bounds, but at higher computational complexity. In fact the worst case complexity of such a method is $\mathcal{O}(N_D \times N_\otimes)$, where N_D is the number of data records in the dataset and $N_\otimes$ is total number of distinct data records that the cross-product space of feature variables can generate [Zaffalon, 2002a]. In our example, for instance, the number of states associated with a single feature variable is $\{206, 412, 4120\}$ corresponding to the three different resolutions $\{0.2, 0.1, 0.01\}$, respectively. Noting that N_D=705, 290, we get $N_D \times N_\otimes = 705,290 \times \{206^4, 412^4, 4120^4\}$, for the resolutions $\{0.2, 0.1, 0.01\}$, respectively. This computational complexity quickly becomes prohibitive with the number of variables and the number of states for each variable. While one may employ other methods (e.g., see [Zaffalon, 2002b, Augustin et al., 2014]) to reduce this computational burden, these would still incur additional computations.



(a) Variable $b_1 = u{-}g$.

(b) Variable $b_2 = g{-}r$.

(c) Variable $b_3 = r{-}i$.

(d) Variable $b_4 = i{-}z$.

Figure 5.6: Lower/Upper Histograms (gray) of $\{\mathbb{L}(b_j|\text{STAR}), \mathbb{U}(b_j|\text{STAR})\}$ for the 4 Observed Photometric Variables of STARs.
Note. "Ambiguation" level $(r) = 10\%$, resolution $(q) = 0.2$, "ambiguation" width $(\Delta) = 0.4$. The GTR dataset-based estimates (red) lie within their corresponding lower/upper histograms.

5.5.3.2 Classification

With the IMP dataset, we must determine the label A from $\{\mathbb{L}(A|B), \mathbb{U}(A|B)\}$ and not $P(A|B)$.

Suppose we receive a new observation $\mathbf{b}$. We apply the parameters learnt from the training set into (5.11) to compute the PrBound pairs $\{\mathbb{L}(\text{STAR}|\mathbf{b}), \mathbb{U}(\text{STAR}|\mathbf{b})\}$ and $\{\mathbb{L}(\text{QUASAR}|\mathbf{b}), \mathbb{U}(\text{QUASAR}|\mathbf{b})\}$ for this new observation $\mathbf{b}$. Finally, we assign a "winning" class label (STAR or QUASAR) to $\mathbf{b}$ by employing a decision criterion in Table 5.4. As an aside, we must mention that similar decision criteria have appeared elsewhere, e.g., Criterion (A) in Table 5.4 is similar to the notions of *interval dominance* in [Zaffalon, 2002b, Augustin et al., 2014] and *strong dominance* in [Luce and Raiffa, 1957].

Table 5.4: Decision Criteria for Selecting Class Label

Criterion	**Select Class Label** a_ℓ
(A)	if $\mathbb{L}(a_\ell|\mathbf{b}) \geq \max_{k\neq\ell} \mathbb{U}(a_k|\mathbf{b})$
(B)	if $\ell = \underset{k}{\text{argmax}}\, \mathbb{L}(a_k|\mathbf{b}) = \underset{k}{\text{argmax}}\, \mathbb{U}(a_k|\mathbf{b})$
(C)	if $\ell = \underset{k}{\text{argmax}}\, \mathbb{L}(a_k|\mathbf{b})$
(D)	if $\ell = \underset{k}{\text{argmax}}\, \mathbb{U}(a_k|\mathbf{b})$

5.5.4 Results

We conducted the experiment with different "ambiguation" levels ($r\%$), different resolution values (q), and different "ambiguation" widths (Δ). We hold $r = 10\%$ and $q = 0.2$ constant and report the results for the three values of $\Delta = 0.4, 1.0$, and ∞, in Tables 5.5, 5.6, 5.7 respectively. Here, ∞ refers to the case when the entry is replaced by an interval of "full" $[-21, +20]$ width.

Table 5.8 shows the corresponding results for the GTR dataset, i.e., the perfect dataset for which $\Delta = 0$. Note that it gives same values for all the criteria since with-

Table 5.5: Classification Performance of Different Class Label Selection Criteria With "Ambiguation" Width (Δ) = 0.4.
Parameters: "ambiguation" level ($r\%$) = 10%, resolution (q) = 0.2, TS = Star classified as STAR, TQ= Quasar classified as QUASAR, FS = Quasar classified as STAR, FQ = Star classified as QUASAR.

Dataset	**Criterion**	**"Ambiguation" width (Δ) = 0.4**		
		Classified (%)		**Unclassified (%)**
		Correctly	**Incorrectly**	
		Total **(TS+TQ)**	**Total** **(FS+FQ)**	
IMP	A	93.34 (83.44+9.89)	1.91 (1.14+0.77)	4.75
	B	96.18 (85.50+10.68)	3.38 (1.58+1.80)	0.44
	C	96.39 (85.61+10.78)	3.61 (1.66+1.95)	0.00
	D	96.42 (85.65+10.77)	3.58 (1.67+1.91)	0.00

Table 5.6: Classification Performance of Different Class Label Selection Criteria With "Ambiguation" Width (Δ)=1.0.
Parameters: "ambiguation" level ($r\%$) = 10%, resolution (q) = 0.2, TS = Star classified as STAR, TQ= Quasar classified as QUASAR, FS = Quasar classified as STAR, FQ = Star classified as QUASAR.

Dataset	**Criterion**	**"Ambiguation" width (Δ) = 1.0**		
		Classified (%)		**Unclassified (%)**
		Correctly	**Incorrectly**	
		Total **(TS+TQ)**	**Total** **(FS+FQ)**	
IMP	A	87.90 (79.14+8.76)	1.20 (0.93+0.28)	10.90
	B	95.55 (85.02+10.54)	3.62 (1.61+2.01)	0.83
	C	95.95 (85.22+10.73)	4.05 (1.71+2.34)	0.00
	D	95.98 (85.35+10.63)	4.02 (1.81+2.21)	0.00

Table 5.7: Classification Performance of Different Class Label Selection Criteria With "Ambiguation" Width $(\Delta)=\infty$.
Parameters: "ambiguation" level $(r\%) = 10\%$, resolution $(q) = 0.2$, TS = Star classified as STAR, TQ= Quasar classified as QUASAR, FS = Quasar classified as STAR, FQ = Star classified as QUASAR.

		"Ambiguation" width $(\Delta) = \infty$		
		Classified (%)		**Unclassified (%)**
Dataset	**Criterion**	**Correctly**	**Incorrectly**	
		Total **(TS+TQ)**	**Total** **(FS+FQ)**	
IMP	A	46.72 (43.44+3.28)	0.40 (0.35+0.05)	52.88
	B	93.65 (85.41+8.24)	2.62 (2.02+0.61)	3.73
	C	95.84 (85.42+10.42)	4.16 (2.02+2.14)	0.00
	D	95.18 (86.94+8.24)	4.82 (4.20+0.62)	0.00

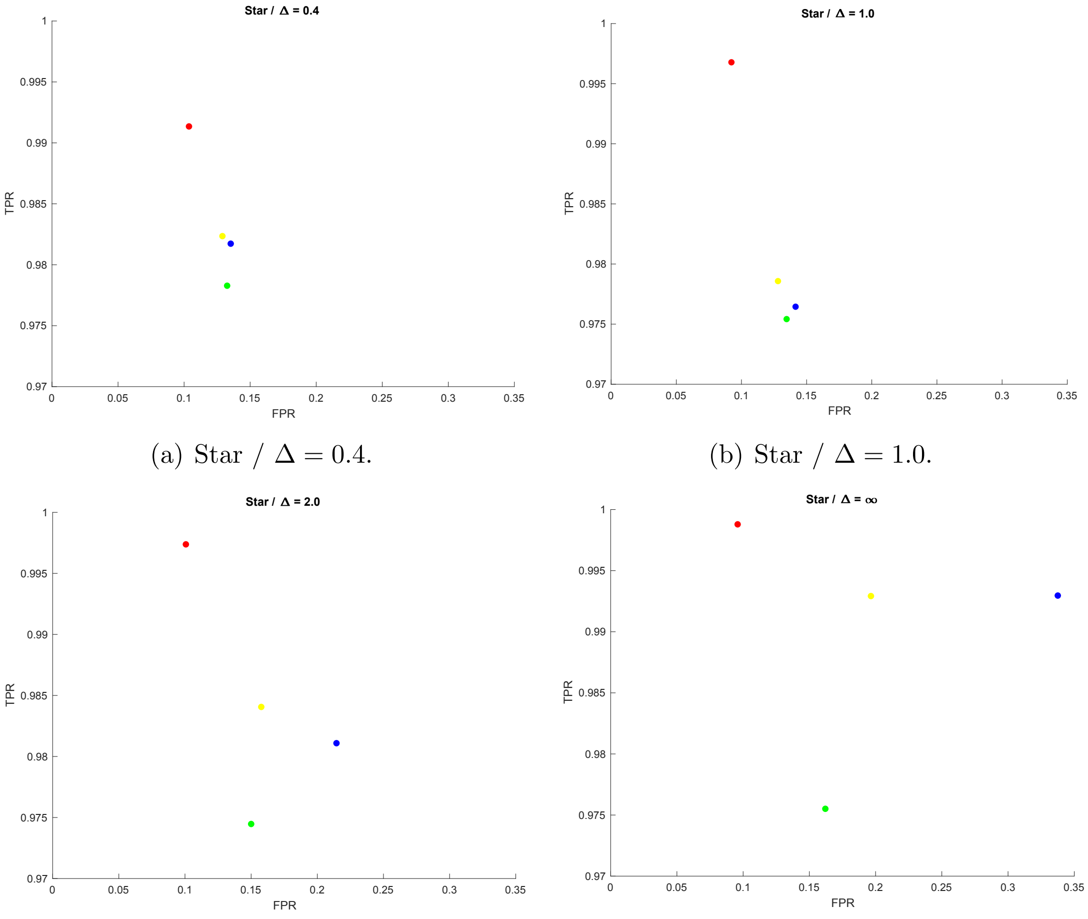
out uncertainty upper and lower probability values converge to a single probability value.

The following observations are noteworthy:

(a) While criterion (A) is able to classify a decreasing fraction of data records with increasing "ambiguation" width, it consistently yields the highest precision, and recall (which are based on only the classified cases) and lowest error. In this sense, criterion (A) can be considered the most conservative in that it renders the most guarded or safest decision. Thus it tends to leave undecided a winner for a large proportion of observation vectors. Criteria (B) and the pair (C) and (D) are increasingly less conservative. This is apparent in True Positive Rate (TPR) vs False Positive Rate (FPR) graphs for both stars and quasars, Fig. 5.7

Table 5.8: Ground Truth (GTR) Dataset Sorresponds to the Perfect Dataset Where Data are not "Ambiguated", i.e., $\Delta = 0$.
Parameters: "ambiguation" level ($r\%$) = 10%, resolution (q) = 0.2, TS = Star classified as STAR, TQ= Quasar classified as QUASAR, FS = Quasar classified as STAR, FQ = Star classified as QUASAR.

		"Ambiguation" width (Δ) = 0		
		Classified (%)		**Unclassified (%)**
Dataset	**Criterion**	**Correctly**	**Incorrectly**	
		Total	**Total**	
		(TS+TQ)	**(FS+FQ)**	
GTR	All	96.40	3.60	0.00
		(85.60+10.80)	(1.64+1.96)	



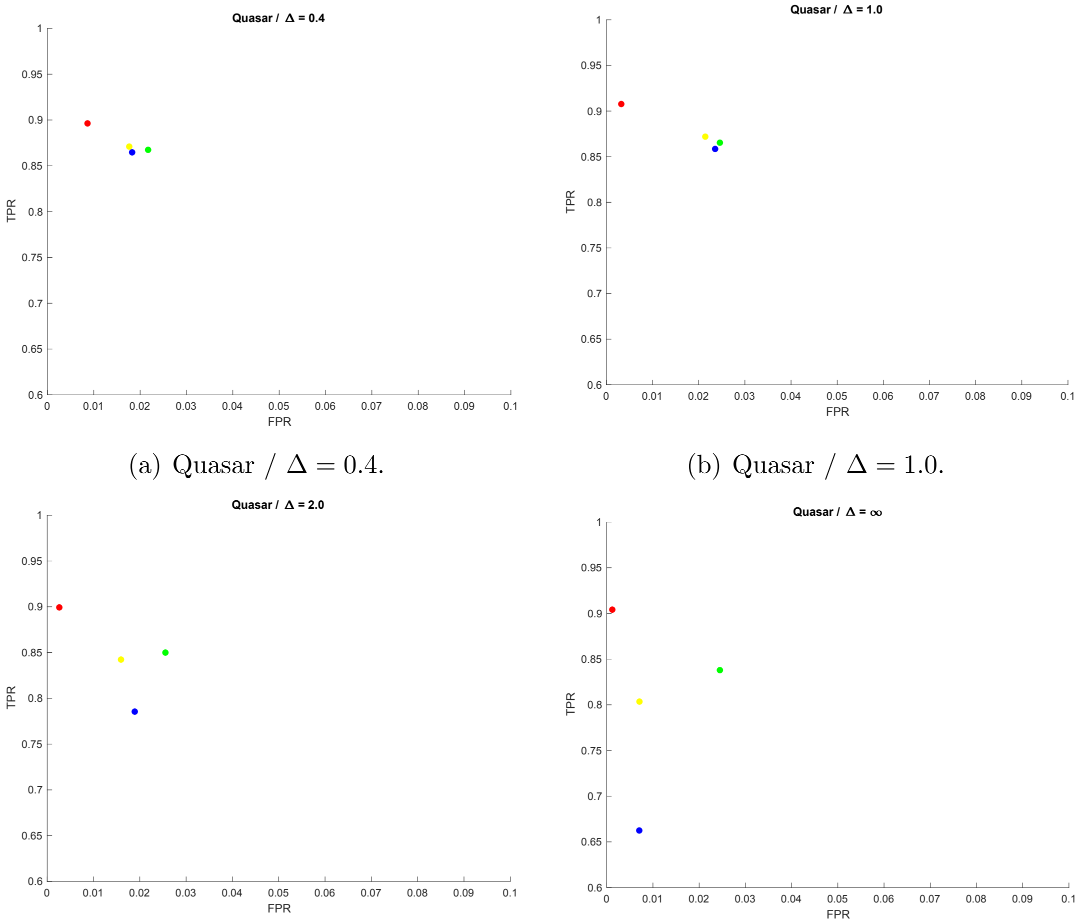
(a) Star / $\Delta = 0.4$.

(b) Star / $\Delta = 1.0$.

(c) Star / $\Delta = 2.0$.

(d) Star / $\Delta = \infty$.

Figure 5.7: TPR vs FPR for Different "Ambiguation" Widths (Δ) for Stars. Color code: red=criterion A, yellow=criterion B, green=criterion C, blue=criterion D



(a) Quasar / $\Delta = 0.4$.

(b) Quasar / $\Delta = 1.0$.

(c) Quasar / $\Delta = 2.0$.

(d) Quasar / $\Delta = \infty$.

Figure 5.8: TPR vs FPR for Different “Ambiguation” Widths (Δ) for Quasars. Color code: red=criterion A, yellow=criterion B, green=criterion C, blue=criterion D

and Fig. 5.8, where the classifier is considered better when the corresponding point is closer to top left corner but that is based on only the classified cases.

(b) Interestingly, with narrower values of "ambiguation" width, one may get comparable performance from the IMP and GTR datasets (especially with the decision criteria (C) and (D)). We believe that this is due to the lower risk posed by the IMP dataset regarding over-fitting during training.

5.6 Summary

The work in this chapter looks at how the parameters needed for ML algorithms could be extracted from an imperfect dataset. In particular, we give several important results pertaining to datasets where attribute values could be unknown/missing or are known to lie within a set of values. These include computationally more tractable and efficient strategies for computing the probability bounds and an intuitive frequency counting method for learning the lower/upper bounds of probability and conditional probability parameters. The underlying probabilities are guaranteed to be constrained within these bounds. We also extend the techniques to more general evidential data representation.

Prior to concluding this chapter, it is noteworthy that the classification scheme we have employed in the experiment in Section 5.5 applies to an SFE observation vector. In the general case of evidential data where one can have several focal elements, learning can still be carried out with the help of Lemma 5 and generalized versions of Methods (1)-(3) above (Method (4) remains the same because it ignores uncertain observations).

To explain, consider the observed attribute variable b_j of the observation vector $\mathbf{b}$. Suppose its AttBoE is $\{\Theta_j, \mathfrak{F}_{b_j}, m_{b_j}\}$. Both SFE and probabilistic cases are special cases of this more general case. Let

$$b_j^{\cup} = \bigcup_{b_j^{(\ell)} \in \mathfrak{F}_{b_j}} b_j^{(\ell)}. \tag{5.13}$$

Then, in place of the PrBounds used within Methods (1)-(3) in Section 5.5.3.1.1, we may use the following more general evidential counterparts:

- *Evidential Counterpart to Method (1):* Use

$$\mathbb{L}(b_j|a_k) = \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)})\, \mathbb{L}(b_j^{(\ell)}|a_k);\ \mathbb{U}(b_j|a_k) = \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)})\, \mathbb{U}(b_j^{(\ell)}|a_k). \tag{5.14}$$

- *Evidential Counterpart to Method (2):* Use

$$\begin{aligned} \mathbb{L}(b_j|a_k) &= \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)}) \sum_{b_j^i \in b_j^{(\ell)}} \mathbb{L}(b_j^i|a_k); \\ \mathbb{U}(b_j|a_k) &= \min\left\{ \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)}) \sum_{b_j^i \in b_j^{(\ell)}} \mathbb{U}(b_j^i|a_k), 1\right\}. \end{aligned} \tag{5.15}$$

- *Evidential Counterpart to Method (3):* Here we may utilize two methods.
  - *Method (3.1):* Use

$$\begin{aligned} \mathbb{L}(b_j|a_k) &= \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)}) \min_{b_j^i \in b_j^{(\ell)}} \{\mathbb{L}(b_j^i|a_k)\}; \\ \mathbb{U}(b_j|a_k) &= \sum_{b_j^{(\ell)} \subseteq b_j^{\cup}} m(b_j^{(\ell)}) \max_{b_j^i \in b_j^{(\ell)}} \{\mathbb{U}(b_j^i|a_k)\}. \end{aligned} \tag{5.16}$$

  - *Method (3.2):* Use

$$\mathbb{L}(b_j|a_k) = \min_{b_j^i \in b_j^{\cup}} \{\mathbb{L}(b_j^i|a_k)\};\ \mathbb{U}(b_j|a_k) = \max_{b_j^i \in b_j^{\cup}} \{\mathbb{U}(b_j^i|a_k)\}. \tag{5.17}$$

Note that Method (3.2) is more conservative than Method (3.1).

Here we employ Corollary 4, the generalized version, instead of Corollary 3 in contrast to Section 5.5.3.1.1.

CHAPTER 6

Deep Fusion Networks (DFNs)

6.1 Overview

In this chapter we take inspiration from deep learning neural network (NN) architectures and develop a PrBounds-based architecture — we refer to this as a *deep fusion network (DFN)* — which allows one to automate fusion of input data streams, fusion parameter selection, and classification of potentially uncertain data that are generated from multi-modal sensors.

For convenience of reference, Table 6.1 summarizes the notation that we have used so far. Additionally, when working in an environment with multiple sensors, we use N_S to denote the number of sensors. Moreover, without loss of generality, we assume that each sensor generates the same number N_D of synchronous data records and possesses the capability to generate data associated with the same number N_R of attributes and the same attribute types $\{A_1, \ldots, A_{N_R}\}$.

6.2 Deep Fusion Network (DFN) Architecture

Figure 6.1 shows the deep learning architecture which we envision for the proposed DFN. We now provide the details of the different layers of this architecture, where

Table 6.1: Notation

N_D; R_i	Number of data records (or data instances); i-th data record, $i = 1, \ldots, N_D$.
N_R; A_j	Number of data attributes; j-th attribute, $j = 1, \ldots, N_R$.
N_j	Size (cardinality) of the state space associated with attribute A_j.
$\Theta_j = \{\theta_{1j}, \ldots, \theta_{N_j j}\}$	State space associated with attribute A_j.
$\Theta_R = \bigotimes_{j=1}^{N_R} \Theta_j$	Cross-product state space associated with a data record.
$\langle A_j = a_{ji} \rangle$ or a_{ji}	Potentially uncertain state of attribute A_j of record R_i, $a_{ji} \subseteq \Theta_j$.
N_S; S_k	Number of sensors; k-th sensor, $k = 1, \ldots, N_S$.
$\{\Theta_j, \mathfrak{F}_{ji,k}, m_{ji,k}\}$	DST AttBoE associated with attribute A_j of record R_i of sensor S_k.
$P_{ji,k}(\cdot) \triangleleft \{\mathbb{L}_{ji,k}(\cdot), \mathbb{U}_{ji,k}(\cdot)\}$	PrBound pair for the p.m.f. $P_{ji,k}(\cdot)$ associated with attribute A_j of record R_i of sensor S_k.

the word *layer* is meant to refer to the inputs, the neuron layers, as well as those layers that do not possess neurons.

6.2.1 Input Processing Layer

When the data can have only probabilistic data uncertainties, one may channel each data record or data instance (say, R_i) generated by a sensor (say, S_k) as a vector of length not more than $\sum_{j=1}^{N_R}(N_j - 1)$. Here, a "block" of size $(N_j - 1)$ corresponds to the probabilities associated with individual singletons of the state space Θ_j. Of course, this "one channel/singleton" representation is inadequate when it comes to data records possessing i.v. uncertainties.

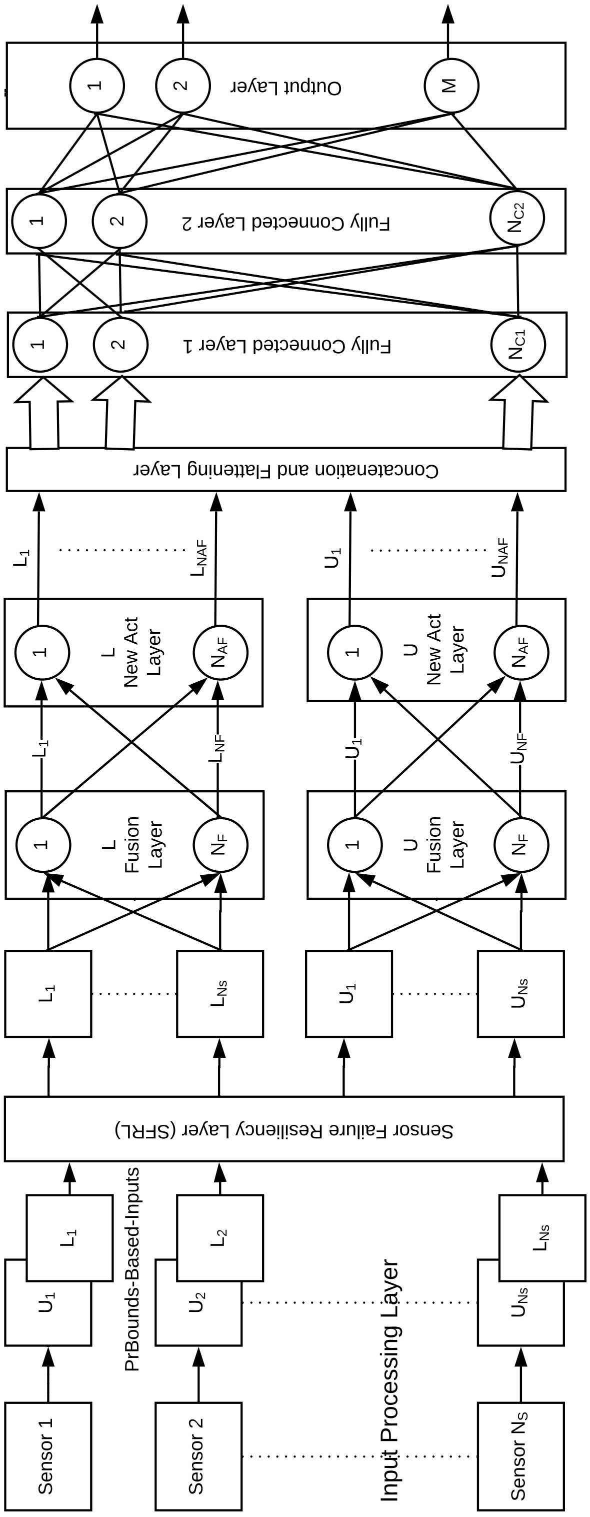


Figure 6.1: DFN Architecture.

Note. Sensor Failure Resiliency (SFR) Layer is active only during the training phase.

6.2.1.1 Inputs With DS Theory-Based Data Uncertainties

Previous work on DS theory-based generalizations of neural networks [Denoeux and Bjanger, 2000,Soua et al., 2016,Wang et al., 2016,Itkina and Kochenderfer, 2017] have essentially channeled each data record (say, R_i) generated by a sensor (say, S_k) as a vector of length not more than $\sum_{j=1}^{N_R}(2^{N_j}-1)$. Here, the "block" of size $(2^{N_j}-1)$ corresponds to the maximum number of focal elements that the attribute FoD Θ_j can generate. This "one channel/each subset of the sample space" representation clearly necessitates the learning of a prohibitively large number of parameters.

However, in practice, the number of focal elements of an AttBoE is significantly smaller than the maximum possible value it could have (viz., $|\mathfrak{F}_{ji,k}|$ versus $(2^{N_j}-1)$ for the AttBoE $\{\Theta_j, \mathfrak{F}_{ji,k}, m_{ji,k}\}$). This "one channel/focal element" representation significantly reduces the number of parameters that need to be learned. However, the burden can remain rather high even for moderately-sized problems.

This "one channel/focal element" representation also poses an additional challenge: the trained neural network will not have a channel available to dedicate to a new focal element which it has not encountered during the training phase. One could conceive of at least two solutions to address this problem:

(a) Associate the new focal element with the channel corresponding to the next available superset or the channel corresponding to the FoD.

(b) Restrict the number of channels to only certain focal elements (e.g., restricting the BoEs to be Dirichlet).

In either case, we must ensure that all channels associated with one AttBoE have DST masses summing to unity (so that it is a valid DST mass function).

6.2.1.2 Inputs With PrBounds-Based Data Uncertainties

With data records whose uncertainties are captured via PrBounds, we employ a "one channel/singleton" representation, meaning that each data record (say, R_i) generated by a sensor (say, S_k) is channeled as two vectors, each of length $\sum_{j=1}^{N_R} N_j$. Here, a "block" of size N_j corresponds to the singletons of the attribute FoD Θ_j; two vectors are needed because PrBounds occur in pairs. We note the following:

1. The computational burden associated with learning parameters is just twice that what is required with probabilistic data uncertainties.

2. PrBounds generalize DST belief/plausibility functions in that they do not impose any monotonicity condition, and therefore they offer more flexibility.

While this "one channel/singleton" representation (instead of the "one channel/each subset of the sample space" representation) cannot capture the full information regarding the data uncertainties, we believe that it constitutes a reasonable compromise between computational feasibility and representative power. Indeed, in the SFE data case, it turns out that this "one channel/singleton" representation is adequate to capture all the PrBounds-based uncertainty information. The reason for this is that the PrBound pairs of singletons completely determine the PrBound pairs of all propositions.

Lemma 8 *Consider the SFE attribute A_j with the p.m.f. $P_j(\cdot)$ (over the state space $\Theta_j = \{\theta_{1j}, \ldots, \theta_{N_j j}\}$). Suppose $P_j(\theta_{\ell,j}) \lhd \{\mathbb{L}_j(\theta_{\ell,j}), \mathbb{U}_j(\theta_{\ell,j})\}$, $\theta_{\ell,j} \in \Theta_j$, $\ell = 1, \ldots, N_j$, i.e., $\{\mathbb{L}_j(\theta_{\ell,j}), \mathbb{U}_j(\theta_{\ell,j})\}$ constitute PrBound pairs for the singletons $\theta_{\ell,j} \in \Theta_j$, $\ell =$*

$1, \ldots, N_j$. Let

$$\Theta_j^{[0,0]} = \left\{ \bigcup_{\ell=1,\ldots,N_j} \theta_{\ell,j} \mid \theta_{\ell,j} \in \Theta_j \textit{ and } \{\mathbb{L}_j(\theta_{\ell,j}), \mathbb{U}_j(\theta_{\ell,j})\} = [0,0] \right\};$$

$$\Theta_j^{[0,1]} = \left\{ \bigcup_{\ell=1,\ldots,N_j} \theta_{\ell,j} \mid \theta_{\ell,j} \in \Theta_j \textit{ and } \{\mathbb{L}_j(\theta_{\ell,j}), \mathbb{U}_j(\theta_{\ell,j})\} = [0,1] \right\};$$

$$\Theta_j^{[1,1]} = \left\{ \bigcup_{\ell=1,\ldots,N_j} \theta_{\ell,j} \mid \theta_{\ell,j} \in \Theta_j \textit{ and } \{\mathbb{L}_j(\theta_{\ell,j}), \mathbb{U}_j(\theta_{\ell,j})\} = [1,1] \right\}.$$

Then, the following are true:

(a) $\{\Theta_j^{[0,0]}, \Theta_j^{[0,1]}, \Theta_j^{[1,1]}\}$ forms a partition of Θ_j, i.e., they are mutually disjoint and their union yields Θ_j.

(b) $|\Theta_j^{[0,1]}| > 0$ iff $|\Theta_j^{[1,1]}| = 0$. Moreover, $|\Theta_j^{[1,1]}|$ can take values 0 or 1 only.

(c) The focal element associated with the SFE attribute A_j is $B_j = \Theta_j^{[0,1]} \cup \Theta_j^{[1,1]}$. Moreover, the PrBound pairs of <u>all propositions</u> can be constructed as

$$\{\mathbb{L}_j(A), \mathbb{U}_j(A)\} = \begin{cases} \{0,0\}, & \textit{if } A \cap B_j = \emptyset; \\ \{0,1\}, & \textit{if } B_j \cap A \neq \emptyset \textit{ and } \overline{B}_j \cap A \neq \emptyset; \\ \{1,1\}, & \textit{if } B_j \subseteq A. \end{cases}$$

□

Proof:

(a) The fact that $\{\Theta_j^{[0,0]}, \Theta_j^{[0,1]}, \Theta_j^{[1,1]}\}$ forms a partition of Θ_j follows directly from Definition 7 because, for an SFE attribute, $\{\mathbb{L}_j(\cdot), \mathbb{U}_j(\cdot)\}$ can only be $\{0,0\}$, $\{0,1\}$, or $\{1,1\}$ (see Remark 2 immediately after Definition 7).

(b) We clearly cannot have $|\Theta_j^{[0,1]}| = |\Theta_j^{[1,1]}| = 0$ because, by Definition 7, an SFE attribute must have one and exactly one non-empty focal element.

(b.1) If this focal element is a singleton, it will have $\{1,\ 1\}$ as its PrBound pair; the other singletons (if any) will have $\{0,0\}$ as their PrBound pairs. This means that $|\Theta_j^{[0,1]}| = 0$ and $|\Theta_j^{[1,1]}| = 1$.

(b.2) If this focal element is not a singleton, then no singleton will have $\{1,1\}$ as its PrBound pair; all singletons that constitute this focal element will have $\{0,1\}$ as their PrBound pairs; the other singletons (if any) will have $\{0,0\}$ as their PrBound pairs. This means that $|\Theta_j^{[0,1]}| > 0$ and $|\Theta_j^{[1,1]}| = 0$.

The fact that $|\Theta_j^{[1,1]}|$ can take values 0 or 1 only is already clear.

(c) This part is now easy to establish. ■

6.2.1.3 PrBounds-Based Input Representation

One can also place these blocks of size N_j side-by-side and view the lower or upper PrBound of each data record of a sensor as a "matrix" with its columns having the lengths $\{N_1, N_2, \ldots, N_{N_R}\}$. While, for convenience, we will refer to this structure as a *matrix,* it is understood that its columns are not necessarily of equal length. See Figure 6.2 and Table 6.2. Notice that the j-th columns of this pair of matrices constitute a PrBound pair for the attribute A_j of the i-th data record of the k-th sensor S_k, i.e.,

$$P_{ji,k}(\cdot) \lhd \{\mathbb{L}_{ji,k}(\cdot), \mathbb{U}_{ji,k}(\cdot)\},\ j = 1, \ldots, N_j,\ i = 1, \ldots, N_D,\ k = 1, \ldots, N_S. \qquad (6.1)$$

Therefore at the Input Processing Layer, potentially uncertain sensor data are converted to PrBounds-based inputs (unless of course the sensor data are already given as PrBounds-based inputs). This conversion employs the methods described in Chapter 5.

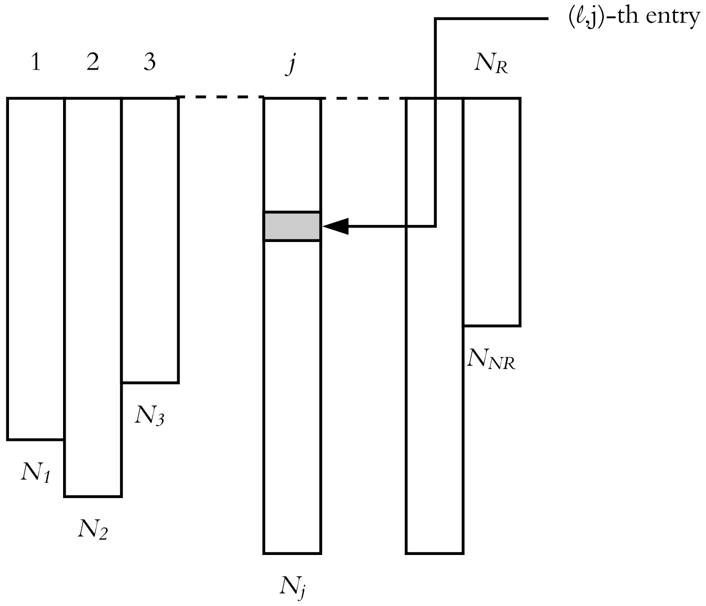


Figure 6.2: PrBounds-Based Input of the i-th Data Record R_i and k-th Sensor S_k. The j-th column has the set of PrBounds $\{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\}$ associated with p.m.f. $P_{ji,k}$ of the attribute A_j, i.e., $P_{ji,k} \lhd \{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\}$; the (ℓ, j)-th entry, or ℓ-th entry of the j-th column, has the PrBound pair of the singleton $\theta_{\ell,j} \in \Theta_j$.

Table 6.2: PrBound-Based Input Matrix of the i-th Data Record and k-th Sensor. See Figure 6.2.

(ℓ, j)-th entry pair	PrBound pair for the singleton $\theta_{\ell,j} \in \Theta_j$.
j-th column pair	Set of PrBounds $\{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\}$ associated with p.m.f. $P_{ji,k}$ of the attribute A_j, i.e., $P_{ji,k} \lhd \{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\}$.

6.2.2 Intermediate Layers

6.2.2.1 Sensor Failure Resiliency (SFR) Layer

The Sensor Failure Resiliency (SFR) Layer is purely an operational layer without any neurons. During the training phase, it deliberately makes randomly selected sensors fail in order to ensure that the system is more robust to real-time sensor failures. This is done by simply making the selected senors' PrBounds vacuous, i.e., by setting $\{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\} = \{0, 1\}$ for all (ℓ, j)-th entry pairs of randomly selected data records R_i and sensors S_k. The number of data records and the number of sensors so affected are parameters that can be adjusted within the algorithm. This strategy indeed made the system more robust to real-time sensor failures during the testing phase. See Section 6.3.

6.2.2.2 Fusion Layer

The Fusion Layer is charged with the task of fusing the different sensor input channels which so far has been treated separately. In other words, within this layer, $\{\mathbb{L}_{ji,k}, \mathbb{U}_{ji,k}\},\ k = 1, \ldots, N_S$, which gives a set of PrBounds for the p.m.f.s $P_{ji,k}(\cdot),\ k = 1, \ldots, N_S$, of attribute A_j of record R_i that the N_S sensors provide are fused together. The immediate question is, what is an appropriate fusion strategy to be used?

While the DCR [Shafer, 1976] is perhaps the most popular strategy of fusing DST evidence, its drawback of irreconcilability with probability is now well established [Smets, 1992, Smets, 1994, Smets, 1999, Heendeni et al., 2016, Núñez et al., 2018]. The *Conditional Fusion Equation (CFE)* is a more recent DST fusion strategy which has been shown to possess several attractive properties when compared with the DCR [Wickramarathne et al., 2012].

6.2.2.2.1 CFE: DST Version Expressed within the current context, the CFE fuses the DST BoEs provided by all the sensors and generates the following fused DST BoE [Wickramarathne et al., 2012]:

$$Bl_{ji}(A) = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \mathfrak{F}_{ji,k}} \beta_{ji,k}(B)\, Bl_{ji,k}(A|B), \tag{6.2}$$

where the non-negative real-valued parameters $\{\alpha_{i,k}, \beta_{ji,k}(\cdot)\}$ satisfy

$$1 = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \mathfrak{F}_{ji,k}} \beta_{ji,k}(B). \tag{6.3}$$

Here, $\{\Theta_j, \mathfrak{F}_{ji,k}, m_{ji,k}\},\ k = 1, \ldots, N_S$, is the DST BoE that the sensor S_k provides for attribute A_j of the i-th data record.

Several strategies for the selection of these CFE parameters appear in [Wickramarathne et al., 2012]. One particularly interesting strategy, referred to as the *receptive strategy,* suggests

$$\beta_{ji,k}(B) = m_{ji,k}(B),\ B \in \mathfrak{F}_{ji,k}. \tag{6.4}$$

6.2.2.2.2 CFE: Probabilistic Version With probabilistic BoEs, (6.2) and (6.3) yield

$$P_{ji}(A) = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \Theta_j} \beta_{i,k}(B)\, P_{ji,k}(A|B), \tag{6.5}$$

where the non-negative real-valued parameters $\{\alpha_k, \beta_{i,k}(\cdot)\}$ satisfy

$$1 = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \Theta_j} \beta_{ji,k}(B). \tag{6.6}$$

If we were to select the parameters to be receptive, we have

$$\beta_{ji,k}(B) = P_{ji,k}(B),\ B \in \Theta_j. \tag{6.7}$$

Substitute these paramenters in (6.5) to get

$$P_{ji}(A) = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \Theta_j} P_{ji,k}(B)\, P_{ji,k}(A|B) = \sum_{k=1}^{N_S} \alpha_{i,k} P_{ji,k}(A), \tag{6.8}$$

where

$$1 = \sum_{k=1}^{N_S} \alpha_{i,k} \sum_{B \in \Theta_j} P_{ji,k}(B) = \sum_{k=1}^{N_S} \alpha_{i,k}. \tag{6.9}$$

Essentially, the CFE strategy, when the parameters are selected to be receptive, reduces to simply the weighted sum of the p.m.f.s that the sensors provide for attribute A_j of the i-th data record.

6.2.2.2.3 CFE: PrBounds-Based Version With (6.8) and (6.9), it is now quite straight-froward to obtain the following PrBounds-based version:

$$P_{ji}(\cdot) \lhd \{\mathbb{L}_{ji}(\cdot), \mathbb{U}_{ji}(\cdot)\} = \left\{ \sum_{k=1}^{N_S} \alpha_{i,k} \mathbb{L}_{ji,k}(A), \sum_{k=1}^{N_S} \alpha_{i,k} \mathbb{U}_{ji,k}(A) \right\}, \tag{6.10}$$

where

$$1 = \sum_{k=1}^{N_S} \alpha_{i,k} \; ; \; \alpha_{i,k} \geq 0. \tag{6.11}$$

We use a pair of parallel sets of neurons to implement this PrBounds-based CFE strategy, one set for the $\mathbb{L}$-matrices and the other for the $\mathbb{U}$-matrices. We will refer to these two sets of neurons as the $\mathbb{L}$-*sublayer* and $\mathbb{U}$-*sublayer,* respectively. The architectures of the $\mathbb{L}$- and $\mathbb{U}$-sublayers being identical, let us consider the $\mathbb{L}$-sublayer.

- The $\mathbb{L}$-sublayer has N_F neurons which is an algorithmic parameter.

- Each neuron has S_N input weights and these are the non-negative real-valued CFE parameters $\alpha_{i,k}$, $k = 1, \ldots, N_S$, in (6.11). N_F sets of such parameters are chosen from a random uniform distribution (with the required conditions in (6.11) satisfied) and each set is fed into one neuron.

- At each neuron, CFE-based fusion in (6.10) takes place. Since the input weights are allowed to take the value 0, during training, the network essentially collates sensor groups for optimum performance. Therefore each neuron can be considered a fusion-cum-collation filter. No bias is used for the neurons.

- The output of each neuron is once again a matrix of the type in Figure 6.2. We use $\{\mathbb{L}_{ji,m}\},\ m = 1, \ldots, N_F$, to denote these N_F outputs. The outputs do not go through any activation function.

The $\mathbb{U}$-sublayer, where the same sets of input weights are utilized, operates in the same manner and generates the outputs $\{\mathbb{U}_{ji,m}\},\ m = 1, \ldots, N_F$.

Henceforth, when no confusion can arise, we will drop the subscripts i (which identifies the data record) and j (which identifies the attribute), and retain only the subscript m (which identifies the output count).

6.2.2.3 Activation Function Layer

As with the Fusion Layer, the Activation Function Layer also has two sublayers, the $\mathbb{L}$-sublayer and the $\mathbb{U}$-sublayer. Each sublayer possesses the following features:

- Each sublayer has N_{AF} neurons which is an algorithmic parameter.

- Each neuron in the $\mathbb{L}$-sublayer takes the N_F output matrices $\mathbb{L}_m,\ m = 1, \ldots, N_F$, of the $\mathbb{L}$-sublayer of the preceding Fusion Layer. Similarly, each neuron in the $\mathbb{U}$-sublayer takes the N_F output matrices $\mathbb{U}_m,\ m = 1, \ldots, N_F$, of the $\mathbb{U}$-sublayer of the preceding Fusion Layer. Both weights and biases are employed in the neurons of the Activation Function Layer. For each neuron of the $\mathbb{L}$-sublayer, the weights as well as the biases are chosen from a random uniform distribution. The same set of weights and biases are used for the corresponding $\mathbb{U}$-sublayer.

- The output of each $\mathbb{L}$- and $\mathbb{U}$-sublayer neurons are fed to the following new activation function AFG:

$$\begin{aligned} \text{AFG:} \quad L_{out,n} &= [2\,\mathrm{sig}(\phi\, L_n) - 1] \times [1 - (\mathrm{sig}(U_n) - \mathrm{sig}(L_n))]; \\ U_{out,n} &= [2\,\mathrm{sig}(\phi\, U_n) - 1] \times [1 - (\mathrm{sig}(U_n) - \mathrm{sig}(L_n))]. \end{aligned} \tag{6.12}$$

Here,

$$\mathrm{sig}(x) = \frac{1}{1+e^{-x}}; \quad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\,\mathrm{sig}(2x) - 1. \tag{6.13}$$

The novel feature of this activation function AFG, which is essentially a pruning mechanism, is that it takes two inputs which are the corresponding outputs of each parallel neurons $\{L_n, U_n\}$, $n = 1, \ldots, N_{AF}$, of the $\mathbb{L}$- and $\mathbb{U}$-sublayers, and "fades" the neuron output when the level of uncertainty exhibited via the PrBound pair is higher. More precisely, the higher the uncertainty, the lower the weight, and vice versa. The activation function operating on the n-th neuron of the $\mathbb{L}$-sublayer generates an output matrix which is determined by L_n and the associated uncertainty interval $\mathrm{sig}(U_n) - \mathrm{sig}(L_n)$. These output matrices, which once again are of the type in Figure 6.2, are faded with increasing uncertainty intervals. The $\mathbb{U}$-sublayer operates in a similar manner. Note the activation functions AF1 and AF2 obtained as special cases of AFG with $\phi = 1$ and

$\phi = 2$, respectively:

$$\text{AF1:} \quad L_{out,n} = [2\,\text{sig}(L_n) - 1] \times [1 - (\text{sig}(U_n) - \text{sig}(L_n))];$$

$$U_{out,n} = [2\,\text{sig}(U_n) - 1] \times [1 - (\text{sig}(U_n) - \text{sig}(L_n))], \tag{6.14}$$

and

$$\text{AF2:} \quad L_{out,n} = \tanh(L_n) \times [1 - (\text{sig}(U_n) - \text{sig}(L_n))];$$

$$U_{out,n} = \tanh(U_n) \times [1 - (\text{sig}(U_n) - \text{sig}(L_n))]. \tag{6.15}$$

- There are N_{AF} outputs generated from each of the $\mathbb{L}$ and $\mathbb{U}$-sublayers of the Activation Function Layer. Note that the outputs generated from the activation functions are bounded to within $[-1, +1]$. The reason for using -1 as the lower bound (instead of 0) is to distinguish it from a value that fades to 0 because of high uncertainty.

In our experiments, AF1 performed slightly better than AF2. We speculated that this is due to the fact that AF1 has a wider input range before the gradient of the function become low compare to AF2. This was confirmed with the results we obtained with $\phi < 1$ which performed even better than AF1. The value of ϕ can also be made to learn during the training phase instead of fixing its value beforehand.

6.2.2.4 Concatenation and Flattening Layer

The Concatenation and Flattening Layer does not have neurons. It simply concatenates the output matrices of the preceding $\mathbb{L}$- and $\mathbb{U}$-sublayers resulting in a sequence of $2\,N_{AF}$ matrices, each being of the type in Figure 6.2. The first N_{AF} matrices of this sequence are $\{L_{out,n}\},\ n = 1, \ldots, N_{AF}$, and the second N_{AF} matrices of the sequence are $\{U_{out,n}\},\ n = 1, \ldots, N_{AF}$. These matrices are "flattened" by reshaping them to one column vector of length of N_C, where

$$N_C = 2\, N_{AF} \sum_{j=1}^{N_R} N_j. \tag{6.16}$$

6.2.2.5 Fully Connected Layers

Next, we have two Fully Connected Layers. These Fully Connected Layers possess the following features:

- Both Fully Connected Layers operate in a similar manner except that the first layer has N_{C_1} number of neurons and the second layer has N_{C_2} number of neurons. Both N_{C_1} and N_{C_2} are algorithmic parameters.

- Each neuron has N_C number of inputs from each entry of the column vector generated from the preceding Concatenation and Flattening Layer. The input weights are chosen from a random distribution; biases are optional in this layer.

- At each neuron, the weighted sum of the inputs is taken and a bias, which is optional, is added.

- Each neuron outputs a single value. This output passes through batch normalization (if batch learning is used), sigmoid activation, and dropout operations, respectively.

6.2.3 Output Layer

The Output Layer has the following features:

- As usual, the number of neurons the Output Layer has is equal to the number of class labels (M).

- Each neuron has N_{C_2} number of inputs from each output of the second Fully Connected Layer. The input weights and biases are chosen from a random distribution.

- No activation function is used.

6.2.3.1 Dealing with Imperfect Labels

The class labels of the training data records may themselves be imperfect.

(a) Perfect class labels: During the training phase, the DFN may be presented with data records where the attributes values can be imperfect but the associated class labels are perfect. Even if the class labels are imperfect, one may opt to train the DFN with only those data records that have perfect class labels.

(b) Imperfect class labels: In this scenario, the DFN may be presented with data records whose class labels are imperfect. We did not explore this case further, mainly due to the difficulty in obtaining such a dataset. Generating an appropriate synthetic dataset presented a significant challenge because inserting or incorporating uncertainties to the class labels because one would have to maintain the underlying relationship, which is of course unknown, between the attributes and the class label during this process. We leave this task for future development.

6.2.3.1.1 Loss Function The loss function should be selected depending on the type of label that the DFN is anticipated to encounter. Given that we only considered perfect class labels, the loss function that we employed is not new. It simply takes the *softmax* of the outputs and use cross-entropy with one-hot encoding of the target label vector. To explain, suppose the number of labels is M. So the output layer has M number of neurons. For $i = 1, \ldots, M$, let x_i denote the output of the i-th neuron. Then the *softmax* x_i is taken as

$$y_i = \frac{e^{x_i}}{\sum\limits_{j=1}^{M} e^{x_j}}. \tag{6.17}$$

The corresponding loss function is defined as

$$\text{loss} = -\sum_{i=1}^{M} t_i \log\, y_i, \tag{6.18}$$

where the t_i s denote the 1-hot encoding of the labels.

We use back-propagation to train the parameters. We also employ mini-batch learning to achieve higher performance.

6.3 Experiment and Results

6.3.1 Dataset

We used the Letter Recognition Dataset of the UCI Machine Learning Repository [Dheeru and Taniskidou, 2017]. The dataset is perfect in that it does not contain any missing values or uncertainties. The relevant parameters are summarized in Table 6.3.

To mimic a multi-sensor dataset, we then employed the following strategy on the given "single-sensor" data set:

- Select a value for the number of sensors N_S.

- Group N_S number of single-sensor data records having the same label together.

Table 6.3: Letter Recognition Dataset.
(from the UCI Machine Learning Repository [Dheeru and Taniskidou, 2017]).

Parameter	**Symbol**	**Value**
Number of data records	N_D	20,000
Number of attribute variables	N_R	16
Number of states in each attribute	N_j	16
State space of each attribute	$\Theta_j = \{\theta_{1j}, \ldots, \theta_{N_j j}\}$	$\{0, \ldots, 15\}$
Class labels		$\{A, B, \cdots, Z\}$
Number of labels in label variable	M	26

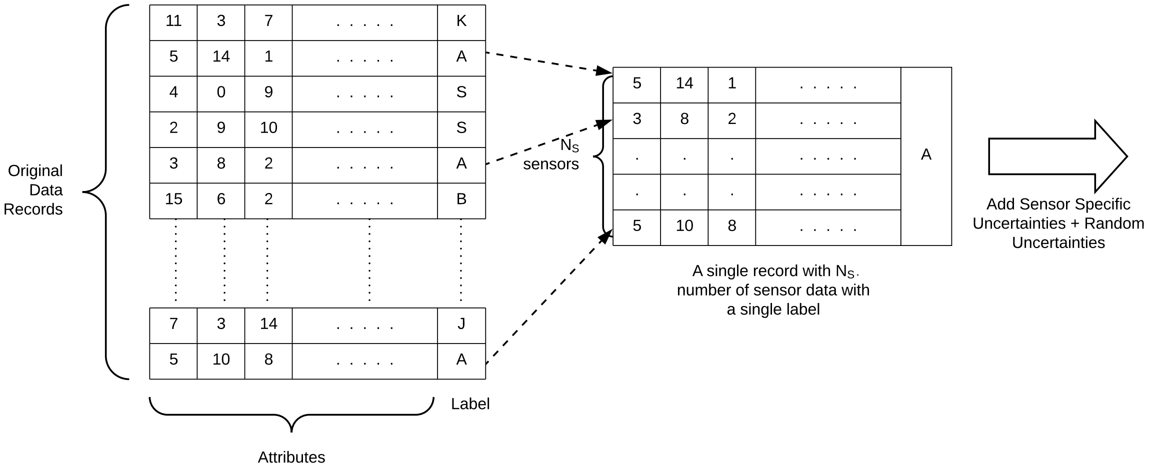


Figure 6.3: Mimicking Multi-Sensor Data Records

- Treat each such group of N_S single-sensor data records as one *multi-sensor data record* generated from N_S sensors and having the common label as its class label.

Figure 6.3 illustrates this strategy with a class label A. Let N_{MD} denote the number of such multi-sensor data records. Given that each multi-sensor data record is constituted of N_S single-sensor data records, we must have $N_{MD}N_S \leq N_D$, i.e., $N_{MD} \leq N_D/N_S$. The chosen value of N_{MD} takes the role of N_D in Tables 6.1 and 6.3.

For our experiment with the Letter Recognition Dataset, we used $N_S = 10$, $N_D = 10,000$ by randomly selecting 10,000 records out of the 20,000 total (10,000 was selected due to the time constraints of the experiments), and $N_{MD} = 855$. The same 10,000 records were employed for all the comparisons.

6.3.1.1 Introducing Uncertainties

We introduced two types of uncertainties into the attribute values (as mentioned above in Section 6.2.3.1, no uncertainties were introduced into the class labels):

(a) ***Sensor-Specific Uncertainties:*** Here, we assume that the sensor determines the attribute of which the value may be uncertain. In other words, in a given multi-sensor data record, each sensor S_k determines which attribute/attributes $A_{j,k},\ j \in \{1, \ldots, N_R\}$, should have uncertainties. At each selected attribute $A_{j,k}$, an SFE type uncertainty is introduced. The focal element $B_j \subseteq \Theta_j$ associated with this SFE uncertainty is selected so that it contains the true state of the attribute.

In the experiments, we allowed the uncertainty to occur in all the data records $i = 1, \ldots, N_{MD}$. We took the relationship between k and j as $j = k$, i.e., for the k-th sensor, the uncertainty occurs in its k-th attribute. Note that, since $N_S = 10 \leq N_R = 16$, the use of $j = k$ means that no uncertainties will be introduced into the last 6 attributes. With the size $|B_j|$ of the focal element B_j of the SFE uncertainty randomly picked to be an integer lying in $[2, |\Theta_j|]$, the elements of B_j are chosen s.t. it contains the true attribute state a_j while the remaining $|B_j| - 1$ states are selected randomly from $\Theta_j \setminus \{a_j\}$. An illustration of this scheme appears in Figure 6.4.

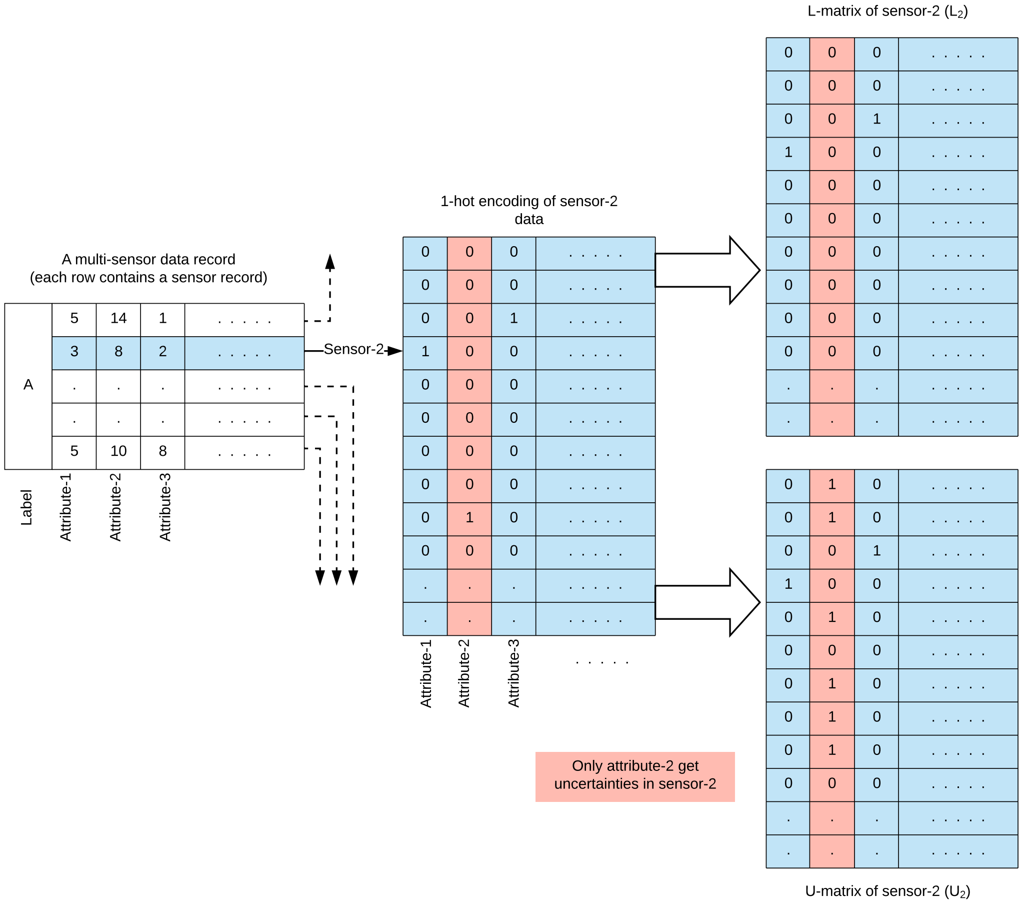
(b) ***Random Uncertainties:*** These uncertainties are not sensor-specific. A set percent $r\%$ of records are selected from the N_{MD} number of multi-sensor data records. For each such selected record, all the sensors are assumed to introduce uncertainties. At each sensor, another random set of attributes is selected. Let N_{AG} be the number of such selected attributes. At each selected attribute, an SFE uncertainty is introduced. So in contrast to the Figure 6.4, now we make uncertain the randomly selected set of attributes instead of only the k^{th} attribute for the k^{th} sensor. As in the previous case, the focal element $B_j \subseteq \Theta_j$ associated with this SFE uncertainty is selected so that it contains the true state of the attribute. In the experiments, we used $r = 10\%$, $1 \leq N_{AG} \leq 4$, and $1 \leq |B_j| \leq 5$.

6.3.2 Algorithm Implementation

The algorithm was implemented in Python with the TensorFlow library [Abadi et al., 2016]. The parameters employed for our experiment are summarized in Table 6.4.

All experiments were conducted with 5-fold cross-validation. With $N_{MD} = 855$, at each phase, we had 684 training records and 171 testing records. We also employed mini-batch learning.

As for the SFR Layer, during each training phase, we randomly selected 30% of multi-sensor data records. From each such selected record, 7 sensors were randomly selected (with repetition allowed) and allowed to fail in the sense described in Section 6.2.2.1. During each testing phase, 20% of the records are selected to do the same thing (this is to mimic the real sensor failures in testing).


L-matrix of sensor-2 (L_2)
1-hot encoding of sensor-2 data
A multi-sensor data record (each row contains a sensor record)
Sensor-2
Only attribute-2 get uncertainties in sensor-2
U-matrix of sensor-2 (U_2)

Label	Attribute-1	Attribute-2	Attribute-3	
A	5	14	1	. . . . .
	3	8	2	. . . . .
	.	.	.	. . . . .
	.	.	.	. . . . .
	5	10	8	. . . . .

Attribute-1	Attribute-2	Attribute-3	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	1	. . . . .
1	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	1	0	. . . . .
0	0	0	. . . . .
.	.	.	. . . . .
.	.	.	. . . . .

0	0	0	. . . . .
0	0	0	. . . . .
0	0	1	. . . . .
1	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
0	0	0	. . . . .
.	.	.	. . . . .
.	.	.	. . . . .

0	1	0	. . . . .
0	1	0	. . . . .
0	0	1	. . . . .
1	0	0	. . . . .
0	1	0	. . . . .
0	0	0	. . . . .
0	1	0	. . . . .
0	1	0	. . . . .
0	1	0	. . . . .
0	0	0	. . . . .
.	.	.	. . . . .
.	.	.	. . . . .

Figure 6.4: Sensor-Specific Uncertainties in Sensor S_2.
Note. Sensor S_2 corresponds to $k = 2$. Sensor-specific uncertainty is introduced into the second attribute, i.e., $j = k = 2$ (pink column in the illustration showing 1-hot encoding of sensor data). This attribute's true state is $a_2 = 8$ and its state-space has 16 states, i.e., $\Theta_2 = \{0, 1, \cdots, 15\}$. The size of the focal element B_2 of the SFE uncertainty is randomly picked from the integers lying in $[2, |\Theta_2|] = [2, 16]$ as $|B_2| = 6$. The elements of B_2 are chosen s. t. it contains the true value $a_2 = 8$ while the remaining $|B_2| - 1 = 5$ states are randomly selected from $\Theta_2 \setminus \{a_2\} = \{0, \ldots, 7, 9, \ldots, 15\}$. In this figure, the focal element randomly selected in this manner is $B_2 = \{0, 1, 4, 6, 7, 8\}$. The PrBound pair [0,1] is introduced into each of the corresponding entry (pink columns in the L- and U-matrices).

Table 6.4: Parameters

	Description	**Symbol**	**Value**
DFN:			
Input Processing Layer:	# of sensors	N_S	10
	# of data records	N_D	10,000
	# of multi-sensor data records	N_{MD}	855
	[training/testing]		[684/171]
	# of data attributes	N_R	16
	Size of state space of each attribute	N_j	16
SFR Layer:	# of affected sensors		7
	# of affected data records		
	[training/testing]		[30/20]%
Fusion Layer:	# of neurons in each sublayer	N_F	256
Activation Function Layer:	# of neurons in each sublayer	N_{AF}	256
Concatenation and Flattening Layer:	Size of the output vector	N_C	131,072
Fully Connected Layer #1:	# of neurons	N_{C_1}	256
Fully Connected Layer #2:	# of neurons	N_{C_2}	256
Output Layer:	# of neurons	N_M	26
TensorFlow:			
Drop-outs	keep-prob		0.95
Optimizer	Adam (during training)		default
Epochs			1
Other:			
Cross-validation			5-fold
Mini-batch learning	Mini-batch size		171

We also used following TensorFlow parameters: dropouts with 0.95 keep-prob; Adam optimizer with its default parameters for parameter training. Number of epochs was set to 1 due to time constraints.

6.3.3 Results

The results below are grouped into four sections: Section 6.3.3.1 explores the effect of different DFN configurations in the system performance; Section 6.3.3.2 explores the impact of using different activation functions; Section 6.3.3.3 explores the performance of the system when pruning is incorporated within the Fusion Layer where the pruning is done at the inputs of the layer; and Section 6.3.3.4 explores the performance when the input data are differently "formatted" prior to insertion into the DFN.

We employed AF1 as the "default" activation function (except within Section 6.3.3.2, where different activation functions are utilized). All the results are presented in the form of macro average ROC curves. Within each plot, the area under the ROC curve is indicated as "area", and the accuracy is indicated as "Acc".

6.3.3.1 Effect of Different DFN Configurations

Figure 6.5 compares the macro average ROC curves for four different DFN configurations: the "complete" DFN (as described in Secrtion 6.2 and Figure 6.1), the DFN without normalization of CFE coefficients (without CFE Fusion), the DFN without the SFR Layer, and the DFN with the parallel coefficient constraints (relaxing the constraint of identical coefficients for L and U sublayers in both Fusion and Activation Function layers at onece).

It is clear that both CFE fusion (with CFE normalization coefficients) and SFR Layer increase the performance of the system significantly. On the other hand, as the zoomed-in version in Figure 6.5(b) shows, relaxing the parallel constraints improves the performance only modestly. The SFR Layer appear to contribute the most to the performance.

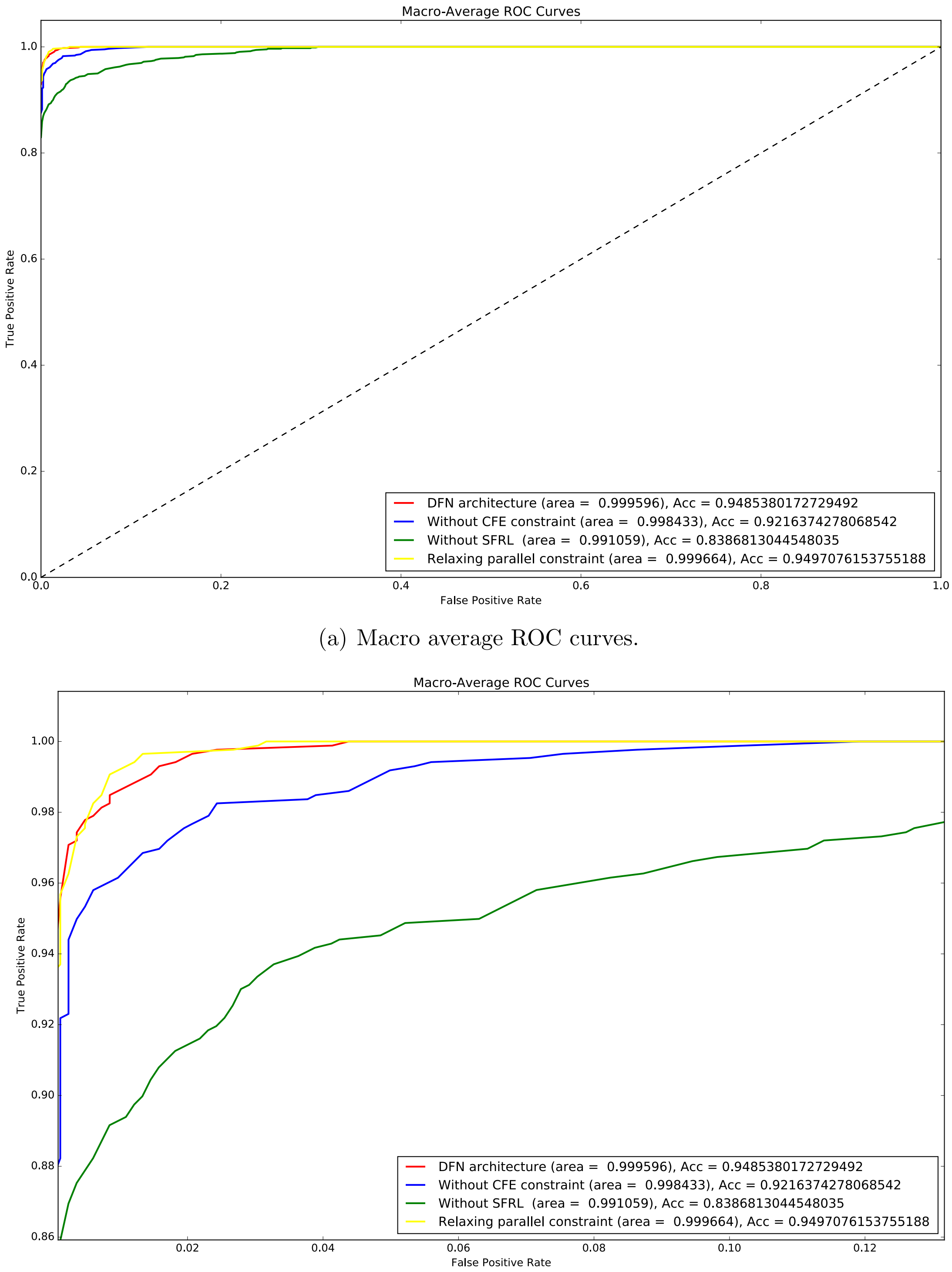
6.3.3.2 Effect of Different Activation Functions

Here we compare the system performances associated with some popular activation functions with our new activation functions. Figure 6.6 shows the results. The proposed new activation function AFG has the disadvantage of the vanishing gradient problem [Hochreiter, 1991]. However, for appropriate values of ϕ, it outperforms all the existing activation functions that we used in this experiment. This includes the the Rectified Linear Unit (ReLU) and Leaky Rectified Linear Unit (Leaky-ReLU):

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0; \\ 0, & \text{otherwise}; \end{cases} \quad ; \quad \text{Leaky-ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0; \\ \alpha\, x, & \text{otherwise}, \end{cases} \tag{6.19}$$

where we used $\alpha = 0.2$.

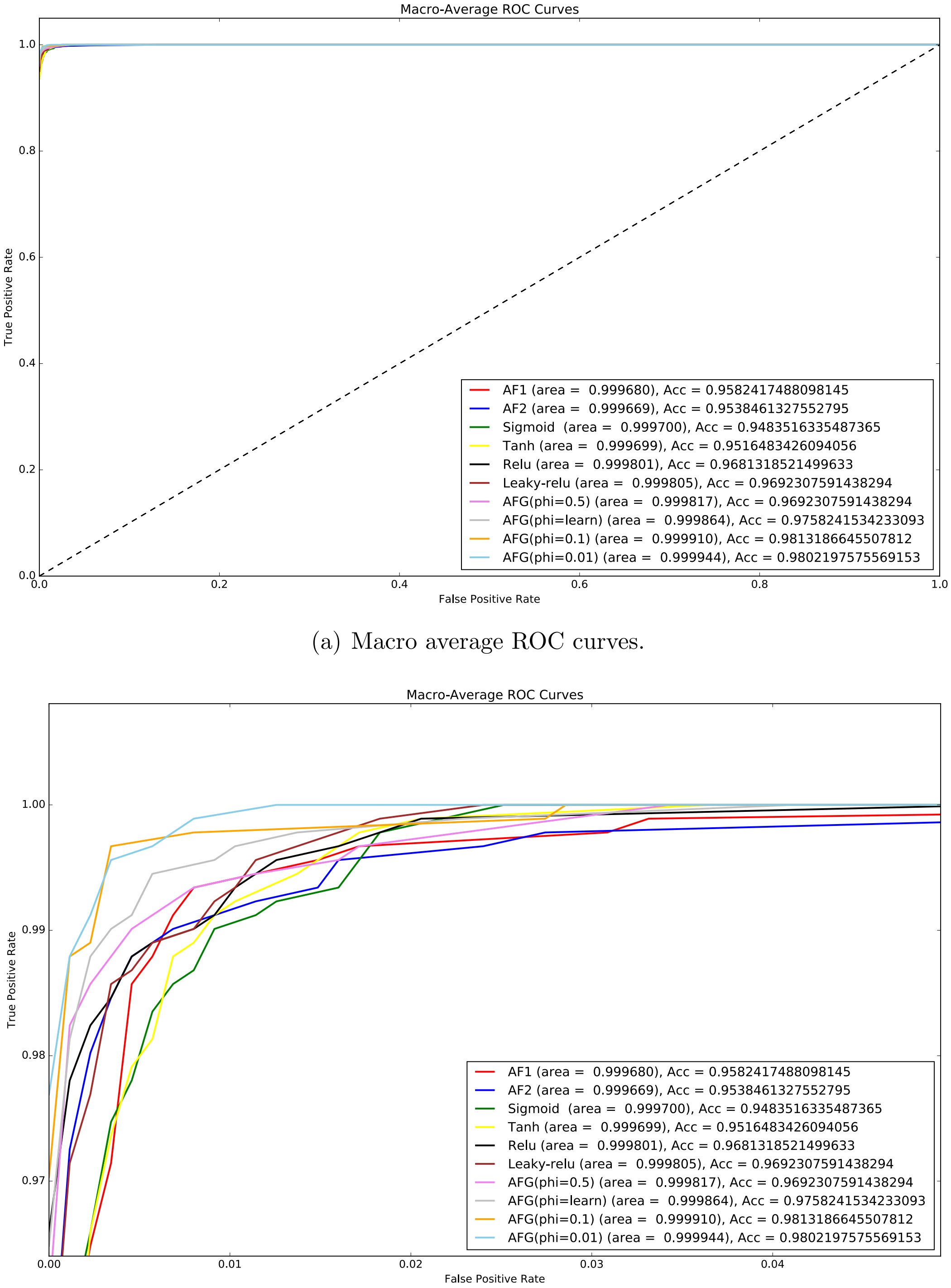
The AFG with $\phi = 0.1$ gave the highest accuracy while $\phi = 0.01$ gave the highest area under the ROC curve. We also explored learning ϕ values, meaning that ϕ is learned during the training phase (denoted by "AFG(phi=learn)" in Figure 6.6). It also outperformed existing activation functions as well as AF1, AF2 and AFG with $\phi = 0.5$. It might outperform $\phi = 0.1$ and $\phi = 0.01$ cases too provided that learning is conducted over a large dataset.



(a) Macro average ROC curves.

(b) A zoomed-in version of (a).

Figure 6.5: Macro Average ROC Curves of Different DFN Configurations



(a) Macro average ROC curves.

(b) A zoomed-in version of (a).

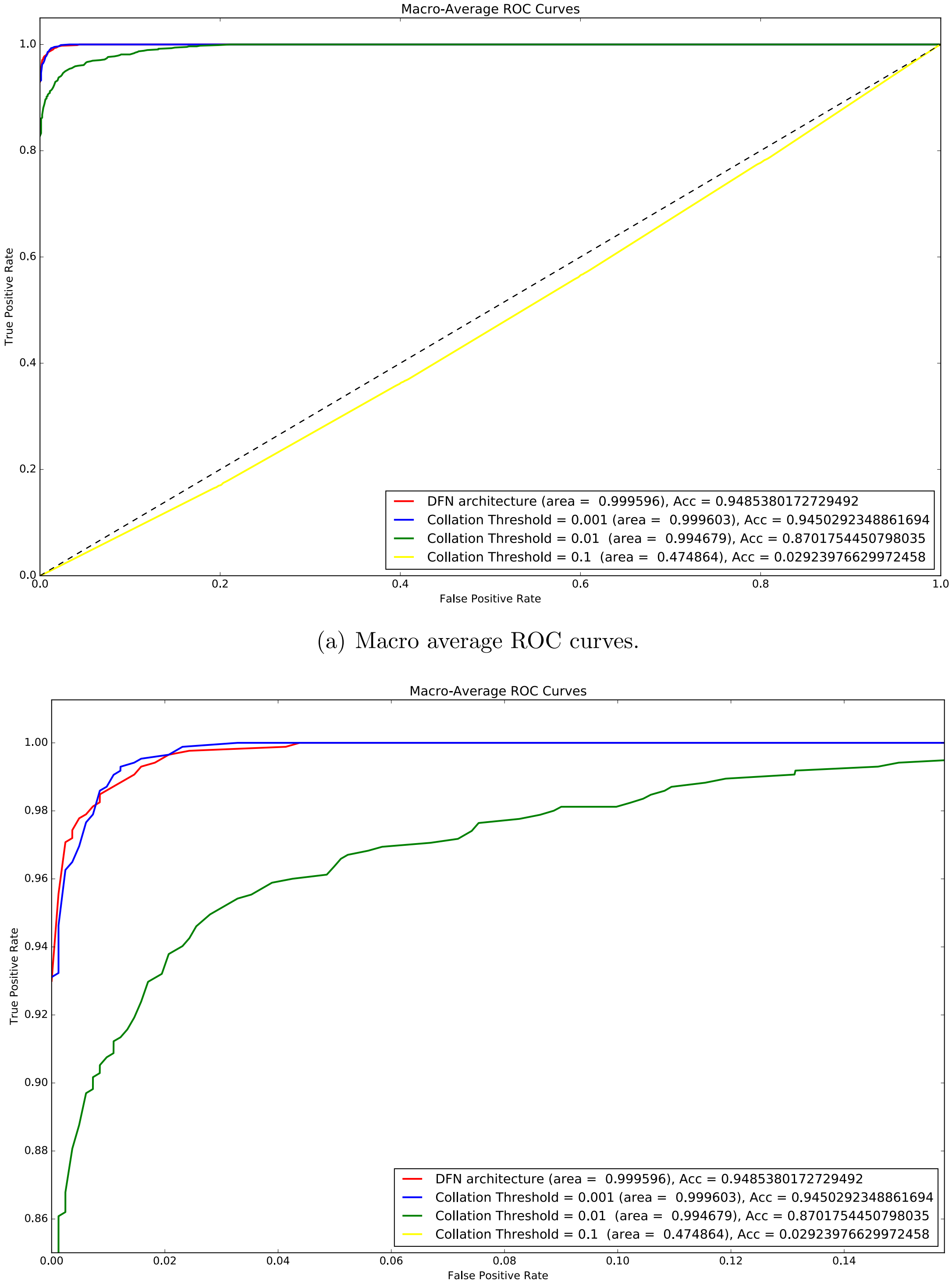
Figure 6.6: Macro Average ROC Curves of Different Activation Functions

6.3.3.3 Effect of Collation

Here the inputs to the Fusion Layer are "collated" in the sense that groups of sensors that should be fused together are identified by pruning. To be more specific, input weights are made zero if their values are below than a certain threshold. The other parameters are kept unchanged. Figure 6.7 depicts the results. Although the accuracies are lower when collation — we used the threshold values $0.001, 0.01, 0.1$ — is present, the area under the ROC curve for collation threshold 0.001 is higher than when collation is absent.

6.3.3.4 Effect of Input Data Format

Here we carried out the experiment with the input data being formatted differently. To be specific, instead of viewing N_S different data records having the same label as data coming from N_S different sensors, we select one record, replicate it 10 times as depicted in Figure 6.8, and then introduce both sensor-specific and random uncertainties (as described in Section 6.3.1.1) so that each sensor gets a sensor specific uncertainty as well as a more general uncertainty. Due to time and complexity restrictions, N_{MD} had to be limited to 3,420. The results are depicted in Figure 6.9. The accuracies are lower compared to the previous case due to the lower number of training records: 3,420 versus 8,550 in the previous case. Interestingly, in contrast to the previous case, now CFE-based fusion does not appear to improve the performance. These results seem to indicate that CFE-based fusion is more useful when the sensors have more variability.



(a) Macro average ROC curves.

(b) A zoomed-in version of (a).

Figure 6.7: Macro Average ROC Curves With Different Collation Thresholds

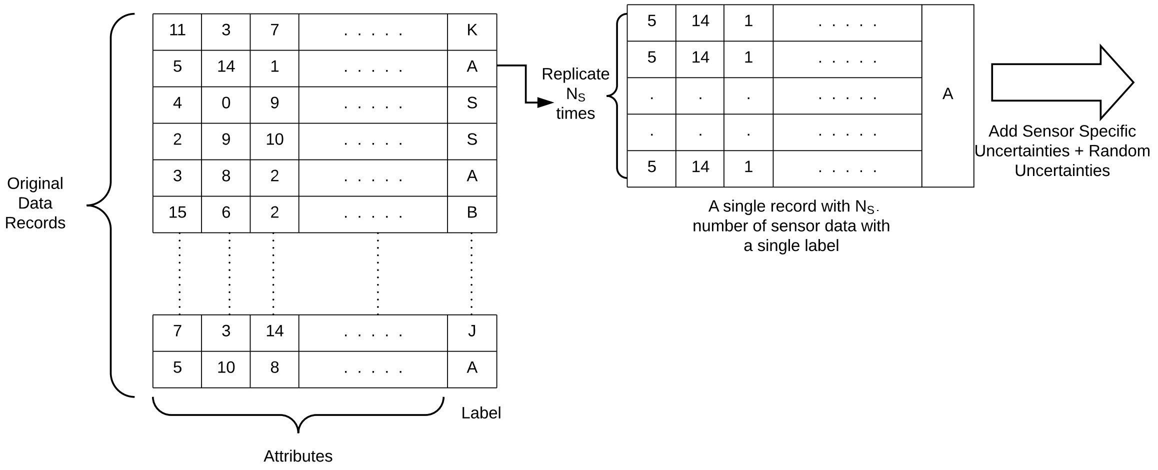
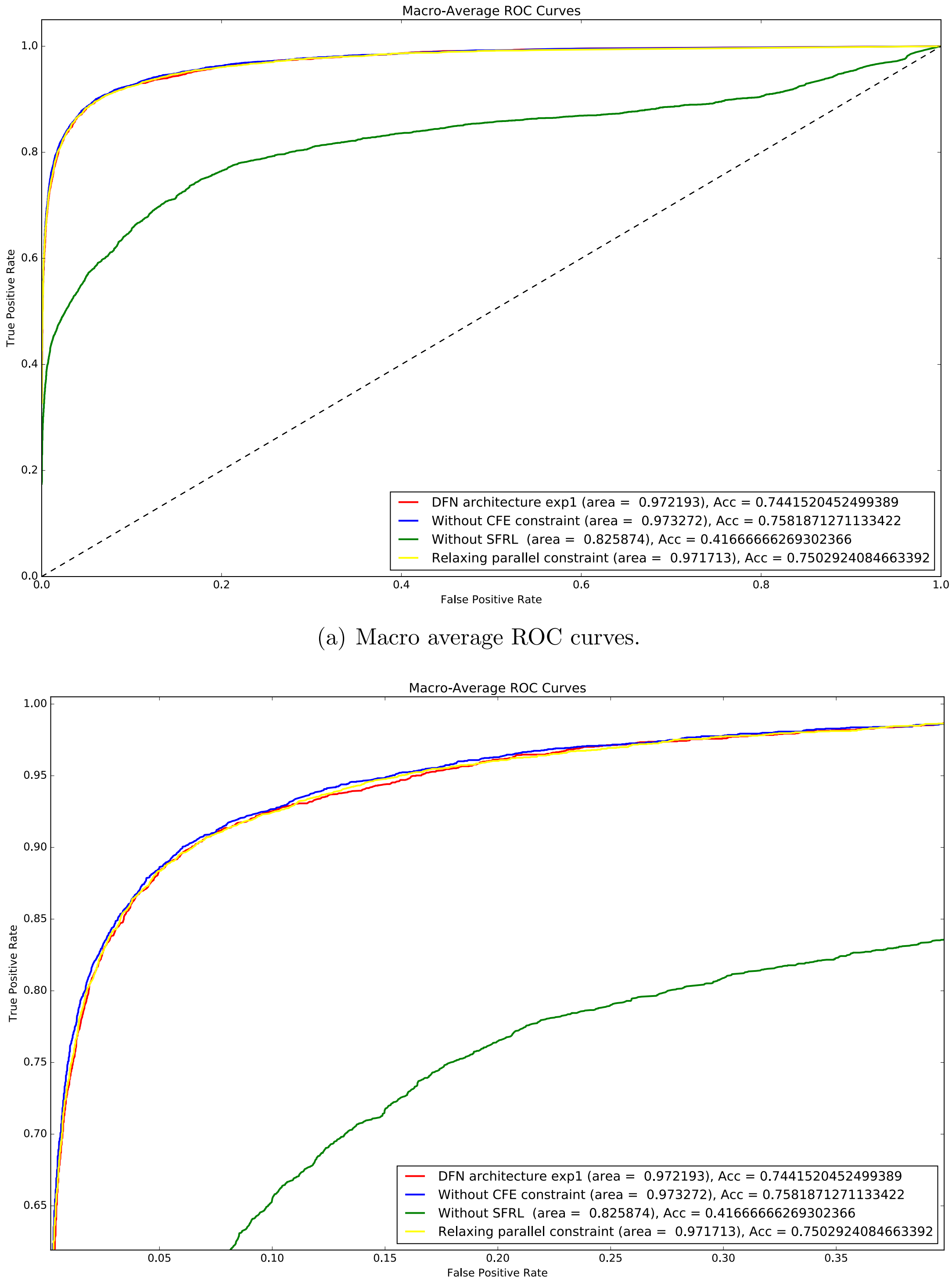


Figure 6.8: Mimicking Multi-Sensor Data Records With Input Data Formatted Differently



(a) Macro average ROC curves.

(b) A zoomed-in version of (a).

Figure 6.9: Macro Average ROC Curves With Input Data Formatted Differently

CHAPTER 7

Conclusion

7.1 Contributions

To summarize the major contribution of this dissertation, we have proposed and developed a new framework for learning and reasoning with i.v. probabilities where the latter are interpreted as having been generated from a single underlying true p.m.f. These i.v. probabilities, which we refer to as *PrBound pairs,* can be viewed as how an agent quantifies the underlying p.m.f. when it does not have full access to it.

We embarked on this work as a result of certain difficulties that we encountered in attempting to develop the notion of an imperfect implication rule as a counterpart to the classical logic-based implication rule. For this purpose, we at first employed the FH conditional from DS theory and proceeded to derive mathematical expressions for the rule consequent, given an imperfect antecedent and an imperfect rule. The results we obtained turned out to be more general and more flexible than what has so far appeared in previous works in that the probabilistic and classical logic relationships emerge as special cases of this model. However, it was apparent that, when the imperfect antecedent and rule are captured via DST belief functions, the resulting imperfect consequent does not necessarily retain the same property, viz., the ∞-

monotonic property which is equivalent to being a valid DST belief function. It is this recognition that monotonicity can be too restrictive and too unwieldy a property to maintain that eventually led us to the development of a more general, and hence more flexible, framework based on PrBounds.

While PrBounds turn out to belong to the genre of i.v. probabilities, it differs from the countless i.v. probability notions that exist in the literature. To be specific, PrBounds do not impose any monotonic condition and they are viewed as emerging from a single underlying probability distribution. This viewpoint enabled us to take a fresh look at the notions of conditioning and independence applicable to i.v. probabilities, thus allowing us to resolve several issues in i.v. probability notions that were discordant with probability. In turn, we were then able to utilize PrBounds in probabilistic graphical models so that the associated computations could be carried out with a complexity that is comparable to what is required in probabilistic settings.

Then we developed new data mining methods to learn parameters from imperfect datasets, where special attention was given to the SFE-type data imperfection which is perhaps the most commonly encountered type of uncertainty in practice. Finally, the validity of the proposed PrBounds-based framework was demonstrated by a PrBounds-based version of a BN and a PrBounds-based version of a naïve Bayes classifier.

This PrBounds-based work then inspired to build a new deep learning architecture that improves classification accuracy of imperfect data. In this architecture, the utilization of "one channel/singleton" input method reduced the number of inputs from exponential to linear, which in tern reduced the number of input parameters to be trained. Hence, the number of required training data is lesser. PrBounds-based inputs also provided a better representation of imperfect data while relaxing the constraints of the inputs, meaning that it does not require monotonicity constraints, and the architecture facilitated multi-sensor data fusion and provided robustness to real-time sensor failures while improving classification accuracy.

7.2 Future Research Directions

We believe that the PrBounds-based framework and the DFN architecture that we have proposed can be extended so that it could be utilized in various application scenarios. In this section, we identify some of these avenues of potentially significant research that one can embark on.

7.2.1 PrBounds-Based Framework

7.2.1.1 General Types of Uncertainty

Some of the the work developed in this dissertation applies to the SFE-type of uncertainty. While this SFE-type of uncertainty is quite commonly encountered in practical situations (e.g., when an attribute/class label value is missing or ambiguous), it cannot adequately well capture more general types of uncertainty (e.g., when the confidence one places on an attribute/class label value or a set of values is below 100%, when different confidence values are placed on different attribute/class label values, etc.). Modeling such general types of uncertainty usually results in wider PrBound pairs which can become useless (e.g., when the PrBound pair is closer to $\{0, 1\}$). One interesting avenue of research is to develop methods and efficient algorithms to deal with general type of evidential data uncertainty, with tighter PrBound pairs.

Another type of imperfection that one often encounters in practice are errors. These imperfections cannot be directly captured via evidential uncertainty. Therefore another stream of research is to explore how erroneous data entries could be captured [Brazdil and Clark, 1990, Brodley and Friedl, 1999].

7.2.1.2 Imperfect Logic Processing

Implication rules play a significant role in machine learning, in particular, in rule-based systems. However, how one may learn and manipulate imperfect implication rules is an issue that has attracted much less attention. The recent work in [Núñez et al., 2018] deals with this exact issue where the authors continue on to develop a complete framework for imperfect logic processing. However this work in [Núñez et al., 2018] assumes that the uncertainties are captured via DST functions.

Our PrBounds-based framework allows one to jettison the monotonicity requirement that DST functions are burdened with. So a PrBounds-based view of imperfect implication rules, and how they can be cascaded to arrive at "fused" imperfect rules, will be able to entertain more general classes of uncertainty while being consistent with probability. We also expect this work to reveal the conditions under which it makes "sense" to cascade imperfect implication rules. In other words, when the uncertainty associated with the rule itself and/or the antecedent is too high, cascading rules may not be feasible, and we might be able to determine exactly what levels of uncertainty pushes one to this realm. It is our belief that one may then continue on to develop an imperfect logic processing framework based completely upon the notion of PrBounds.

7.2.1.3 Decision Trees and Random Forests

Decision trees are well known for their accuracy and efficiency. Generalizations of decision trees to imperfect data are already available in the literature [Denoeux and Bjanger, 2000, Elouedi et al., 2001, Hady et al., 2008]. The work in [Denoeux and Bjanger, 2000] is based on the transferable belief model (TBM) developed by Smets [Smets, 1994] and it considers label uncertainties only. In addition, its emphasis is different: it considers the algorithms used in decision trees to determine splitting order such as entropy and provides TBM-based models. The work in [Elouedi et al., 2001] also uses the TBM. It does not consider uncertainties in the attribute variables at training; rather, it only considers uncertainties in the class labels. The model described in [Hady et al., 2008] does not use uncertain data at all, but uses DS theory to combine the decisions of individual decision trees of their multi-view forest. Apart from DS theory, possibility theory has also been used to handle uncertainty and imprecision in decision trees [Jenhani et al., 2008].

In contrast to the methods available in the literature, we suggest a method that uses PrBounds of focal elements in the decision tree, which branches along all the focal elements of attribute variables. This would allow uncertainties in the attribute variables. At each leaf of the tree, a table can be introduced to keep track of focal elements of label variables, which would allow the decision tree to handle label uncertainties as well. To make classifications based on PrBounds, different decision criteria similar to those elaborated upon in Chapter 5 can be employed. Such a work also can incorporate other algorithms that are already available in the literature [Denoeux and Bjanger, 2000], as well as new algorithms for splitting order, pruning etc. Then the work can be carried out further to develop imperfect information based random forests.

7.2.2 DFN Architecture

7.2.2.1 Imperfect Class Labels

As mentioned in Section 6.2.3.1, the main hindrance we have had to extend the proposed DFN architecture to deal with this scenario has been the unavailability of appropriate datasets. A tempting solution is to introduce uncertainties artificially into the labels. However, if this task were to be carried out independently from the attribute values, then there would have no relationship between the attributes and the uncertain labels that the DFN could learn. The DFN in turn can be misled resulting in poor performance.

What this essentially means is that the underlying relationship between the attributes and the class label has to be maintained when introducing uncertainties into the dataset. However, this relationship itself is unknown. If one were to utilize derive or develop a mathematical relationship between attributes and their corresponding uncertain labels and use that relationship to artificially muddle the dataset, the result will be a DFN that simply mimics the same mathematical relationship. One is then left with the question of why the DFN is required because the same mathematical relationship could be employed for classification in the first place. Another solution might be to introduce a form of "sensor-specific" uncertainty into the class label, a strategy that is somewhat similar to what we employed in Section 6.3.1.1 to introduce uncertainties into the attribute values.

7.2.2.2 DFN Parameters

Within the Fusion Layer of our DFN, we used the receptive strategy for parameter selection for the PrBound-based version of the CFE, which we have shown to yield the weighted sum of individual PrBounds (see (6.10)). But a fair amount of work has been conducted on different ways of parameter selection, the receptive strategy being only one among many others [Kulasekere et al., 2004, Premaratne et al., 2007, Premaratne et al., 2009, Wickramarathne et al., 2010, Wickramarathne et al., 2012]. An interesting area for further exploration is how these and other strategies could be incorporated and what impact they would have on the associated fusion mechanism and the DFN performance.

7.2.2.3 Deliberate Use of Data Uncertainty

While carrying out the experiment in Section 5.5, we noticed that the introduction of low levels of data uncertainty in photometric data had the effect of a slight increase in accuracy over a setting where data were perfect (see Section 5.5.4). This leads us to postulate that i.v. data and the methods developed within this dissertation may in fact be capable of improving the performance of machine learning algorithms over what could be achieved with no imperfect data. We believe that low levels of data uncertainty during training help alleviate issues related to over-fitting and increase the robustness against variations in data. Much work needs to be directed toward exploring this issue in depth.

Bibliography

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: a system for large-scale machine learning. In *Proc. USENIX Conference on Operating Systems Design and Implementation (OSDI)*, pages 265–283, Savannah, GA.

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proc. International Conference on Very Large Data Bases (VLDB)*, pages 487–499, Santiago de Chile, Chile.

[Anand et al., 1996] Anand, S. S., Bell, D. A., and Hughes, J. G. (1996). EDM: A general framework for data mining based on evidence theory. *Data and Knowledge Engineering*, 18:189–223.

[Augustin et al., 2014] Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M. (2014). *Introduction to Imprecise Probabilities.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, West Sussex, UK.

[Bauer, 1997] Bauer, M. (1997). Approximation algorithm and decision making in the Dempster-Shafer theory of evidence—an empirical study. *International Journal of Approximate Reasoning*, 17(2/3):217–237.

[Benavoli et al., 2008] Benavoli, A., Chisci, L., Farina, A., and Ristic, B. (2008). Modelling uncertain implication rules in evidence theory. In *Proc. International Conference on Information Fusion (FUSION)*, pages 1–7, Cologne, Germany.

[Berger, 1982] Berger, J. O. (1982). The robust Bayesian viewpoint. Technical Report 82-9, Purdue University, Purdue, IN.

[Blackman and Popoli, 1999] Blackman, S. and Popoli, R. (1999). *Design and Analyis of Modern Tracking Systems.* Artech House, Norwood, MA.

[Bonissone et al., 1990] Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors (1990). *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, OR. AUAI Press.

[Brazdil and Clark, 1990] Brazdil, P. and Clark, P. (1990). Learning from imperfect data. In Brazdil, P. B. and Konolige, K., editors, *Machine Learning, Meta-Reasoning and Logics*, volume 82 of *The Kluwer International Series in Engineering and Computer Science (SECS)*, pages 207–232. Kluwer Academic Publishers.

[Brodley and Friedl, 1999] Brodley, C. E. and Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11(1):131–167.

[Castillo et al., 1997] Castillo, E., Gutierrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models.* Monographs in Computer Science. Springer-Verlag, New York, NY.

[Chateauneuf and Jaffray, 1989] Chateauneuf, A. and Jaffray, J.-Y. (1989). Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, 17(3):263–283.

[Choquet, 1954] Choquet, G. (1954). Theory of capacities. *Annales de l'institut Fourier*, 5:131–295.

[Chrisman, 1992] Chrisman, L. (1992). Abstract probabilistic modeling of action. In *Proc. International Conference on Artificial Intelligence Planning Systems*, pages 28–36, College Park, MA.

[Chrisman, 1996a] Chrisman, L. (1996a). Independence with lower and upper probabilities. In [Horvitz and Jensen, 1996], pages 169–177.

[Chrisman, 1996b] Chrisman, L. (1996b). Propagation of 2-monotone lower probabilities on an undirected graph. In [Horvitz and Jensen, 1996], pages 178–185.

[Coden et al., 2016] Coden, A., Lin, W. S., Houck, K., Tanenblatt, M., Boston, J., MacNaught, J. E., Soroker, D., Weisz, J. D., Pan, S., Lai, J.-H., Lu, J., Wood, S., Xia, Y., and Lin, C.-Y. (2016). Uncovering insider threats from the digital footprints of individuals. *IBM Journal of Research and Development*, 60(4):8:1–8:11.

[Cowell, 1999] Cowell, R. G. (1999). Parameter learning from incomplete data using maximum entropy II: Application to Bayesian networks. Technical Report Statistical Research Paper 21, Cass Business School, City University, London, UK.

[Cozman and Krotkov, 1996] Cozman, F. and Krotkov, R. (1996). Quasi-Bayesian strategies for efficient plan generation: Application to the planning to observe problem. In [Horvitz and Jensen, 1996], pages 186–193.

[Cozman, 1997] Cozman, F. G. (1997). An informal introduction to quasi-Bayesian theory (and lower probability, lower expectations, Choquet capacities, robust Bayesian methods, etc...) for AI. Technical Report CMU-RI-TR 97-24, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[Cozman, 2000] Cozman, F. G. (2000). Credal networks. *Artificial Intelligence*, 120(2):199–233.

[Dabarera et al., 2016] Dabarera, R., Premaratne, K., Murthi, M. N., and Sarkar, D. (2016). Consensus in the presence of multiple opinion leaders: Effect of bounded confidence. *IEEE Transaction on Signal and Information Processing over Networks*, 2(3):336–349.

[D'Addabbo et al., 2016] D'Addabbo, A., Refice, A., Pasquariello, G., Lovergine, F. P., Capolongo, D., and Manfreda, S. (2016). A Bayesian network for flood detection combining SAR imagery and ancillary data. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3612–3625.

[Daly et al., 2011] Daly, R., Shen, Q., and Aitken, S. (2011). Learning Bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, 26(2):99–157.

[Das et al., 2017] Das, M., Ghosh, S. K., Gupta, P., Chowdary, V. M., Nagaraja, R., and Dadhwal, V. K. (2017). FORWARD: A model for FOrecasting Reservoir WAteR Dynamics using spatial Bayesian network (SpaBN). *IEEE Transactions on Knowledge and Data Engineering*, 29(4):842–855.

[de Campos and Moral, 1995] de Campos, L. M. and Moral, S. (1995). Independence concepts for convex sets of probabilities. In Besnard, P. and Hanks, S., editors, *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 108–115, San Francisco, CA. Morgan Kaufmann.

[de Cooman et al., 2010] de Cooman, G., Hermans, F., Antonucci, A., and Zaffalon, M. (2010). Epistemic irrelevance in credal nets: The case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052.

[de Cooman et al., 2011] de Cooman, G., Miranda, E., and Zaffalon, M. (2011). Independent natural extension. *Artificial Intelligence*, 175(12/13):1911–1950.

[Dempster, 1967] Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339.

[Dempster, 1968] Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247.

[Denoeux and Bjanger, 2000] Denoeux, T. and Bjanger, M. (2000). Induction of decision trees from partially classified data using belief functions. In *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, volume 4, pages 2923–2928, Nashville, TN.

[Dheeru and Taniskidou, 2017] Dheeru, D. and Taniskidou, E. K. (2017). UCI machine learning repository.

[Dubois et al., 2001] Dubois, D., Hullermeier, E., and Prade, H. (2001). Toward the representation of implication-based fuzzy rules in terms of crisp rules. In *Joint IFSA World Congress/NAFIPS International Conference*, pages 1592–1597, Vancouver, BC, Canada.

[Elouedi et al., 2001] Elouedi, Z., Mellouli, K., and Smets, P. (2001). Belief decision trees: Theoretical foundations. *International Journal of Approximate Reasoning*, 28(2/3):91–124.

[Fagin and Halpern, 1990] Fagin, R. and Halpern, J. Y. (1990). A new approach to updating beliefs. In [Bonissone et al., 1990], pages 317–325.

[Ferdous et al., 2012] Ferdous, R., Khan, F., Sadiq, R., Amyotte, P., and Veitch, B. (2012). Handling and updating uncertain information in bow-tie analysis. *Journal of Loss Prevention in the Process Industries*, 25:8–19.

[Fertig and Breese, 1993] Fertig, K. W. and Breese, J. (1993). Probability intervals over influence diagrams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):280–286.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

[Ginsberg, 1984] Ginsberg, M. L. (1984). Nonmonotonic reasoning using dempster's rule. In *National Conference on Artificial Intelligence (AAAI)*, pages 126–129, Austin, TX.

[Giron and Rios, 1980] Giron, F. J. and Rios, S. (1980). Quasi-Bayesian behaviour: A more realistic approach to decision making? *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1):17–38.

[Good, 1962] Good, I. J. (1962). Subjective probability as the measure of a non-measurable set. In Nagel, E., Suppes, P., and Tarski, A., editors, *Proc. International Congress on Logic, Methodology and Philosophy of Science*, pages 319–329. Stanford University Press, Stanford, CA.

[Grosof, 1985] Grosof, B. N. (1985). An inequality paradigm for probabilistic knowledge the augmented logic of conditional probability intervals. In *Proceedings of the First Conference on Uncertainty in Artificial Intelligence*, UAI'85, pages 1–8, Arlington, Virginia, United States. AUAI Press.

[Haddawy and Suwandi, 1994] Haddawy, P. and Suwandi, M. (1994). Decision-theoretic refinement planning using inheritance. In *Proc. International Conference on Artificial Intelligence Planning Systems*, pages 266–271, Chicago, IL.

[Hady et al., 2008] Hady, M. F. A., Schwenker, F., and Palm, G. (2008). Multi-view forests based on Dempster-Shafer evidence theory: A new classifier ensemble method. In *Proc. IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, pages 18–23, Innsbruck, Austria.

[Halpern and Fagin, 1992] Halpern, J. Y. and Fagin, R. (1992). Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54(3):275–317.

[Halpern and Leung, 2016] Halpern, J. Y. and Leung, S. (2016). Weighted expected regret: New approaches for representing uncertainty and making decisions. arXiv:1302.5681v1 [cs.GT].

[Hau and Kashyap, 1990] Hau, H. Y. and Kashyap, R. L. (1990). Belief combination and propagation in a lattice-structured interference network. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(1):45–57.

[Heendeni et al., 2014] Heendeni, J. N., Premaratne, K., Murthi, M. N., and Scheutz, M. (2014). Modelling and fusion of imperfect implication rules. In Cuzzolin, F., editor, *Proc. International Conference on Belief Functions (BELIEF)*, volume 8764 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 313–320. Springer, Oxford, UK.

[Heendeni et al., 2016] Heendeni, J. N., Premaratne, K., Murthi, M. N., Uscinski, J., and M, S. (2016). A generalization of Bayesian inference in the Dempster-Shafer belief theoretic framework. In *Proc. International Conference on Information Fusion (FUSION)*, pages 798–804, Heidelberg, Germany.

[Hewawasam et al., 2007] Hewawasam, K. K. R. G. K., Premaratne, K., and Shyu, M.-L. (2007). Rule mining and classification in a situation assessment application: A belief theoretic approach for handling data imperfections. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 37(6):1446–1459.

[Hewawasam et al., 2005] Hewawasam, K. K. R. G. K., Premaratne, K., Subasingha, S. P., and Shyu, M. L. (2005). Rule mining and classification in imperfect databases. In *International Conference on Information Fusion (FUSION)*, volume 1, pages 661–668, Philadelphia, PA.

[Hochreiter, 1991] Hochreiter, J. (1991). *Untersuchungen zu dynamischen neuronalen Netzen*. PhD thesis, Institut für Informatik, Technische Universität München, München, Germany.

[Horvitz and Jensen, 1996] Horvitz, E. and Jensen, F., editors (1996). *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, San Francisco, CA. Morgan Kaufmann.

[Huber and Ronchetti, 2009] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statisics*. John Wiley & Sons, Inc.

[Itkina and Kochenderfer, 2017] Itkina, M. and Kochenderfer, M. J. (2017). Convolutional neural network information fusion based on Dempster-Shafer theory for urban scene understanding. Technical report, Stanford University, Stanford, CA.

[Ivezic et al., 2014] Ivezic, Z., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2014). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton Series in Modern Observational Astronomy. Princeton University Press, Princeton, NJ.

[Jaeger, 2006] Jaeger, M. (2006). The AI&M procedure for learning from incomplete data. In Dechter, R. and Richardson, T., editors, *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 225–232, Arlington, VA. AUAI Press.

[Jaffray, 1992] Jaffray, J.-Y. (1992). Bayesian updating and belief functions. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):1144–1152.

[Janez and Appriou, 1998] Janez, F. and Appriou, A. (1998). Theory of evidence and non-exhaustive frames of discernment: Plausibilities correction methods. *International Journal of Approximate Reasoning*, 18(1/2):1–19.

[Jenhani et al., 2008] Jenhani, I., Ben-Amor, N., and Elouedi, Z. (2008). Decision trees as possibilistic classifiers. *International Journal of Approximate Reasoning*, 48(3):784–807.

[Josang, 2010] Josang, A. (2010). Cumulative and averaging fusion of beliefs. *Information Fusion*, 11(2):192–200.

[Khaleghi et al., 2013] Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14:28–44.

[Klawonn and Smets, 1992] Klawonn, F. and Smets, P. (1992). The dynamic of belief in the transferable belief model and specialization-generalization matrices. In Dubois, D., Wellman, M., D'Ambrosio, B., and Smets, P., editors, *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 130–137, San Mateo, CA. Morgan Kaufmann.

[Kohavi and Provost, 1998] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning – Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*, 30(2/3):271–274.

[Kourou et al., 2017] Kourou, K., Papaloukas, C., and Fotiadis, D. I. (2017). Integration of pathway knowledge and dynamic Bayesian networks for the prediction of oral cancer recurrence. *IEEE Journal of Biomedical and Health Informatics*, 21(2):320–327.

[Kschischang et al., 2001] Kschischang, F. R., Frey, B. J., and Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

[Kulasekere et al., 2004] Kulasekere, E. C., Premaratne, K., Dewasurendra, D. A., Shyu, M.-L., and Bauer, P. H. (2004). Conditioning and updating evidence. *International Journal of Approximate Reasoning*, 36(1):75–108.

[Lauritzen, 1995] Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19(2):191–201.

[Lauritzen and Wermuth, 1989] Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17(1):31–57.

[Leicester et al., 2016] Leicester, P. A., Goodier, C. I., and Rowley, P. N. (2016). Probabilistic analysis of solar photovoltaic self-consumption using Bayesian network models. *IET Renewable Power Generation*, 10(4):448–455.

[Levi, 1990] Levi, I. (1990). Compromising Bayesianism: A plea for indeterminacy. *Journal of Statistical Planning and Inference*, 25:347–362.

[Lewis, 1976] Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, LXXXV(3):297–315.

[Li et al., 2001] Li, W., Han, J., and Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 369–376, San Jose, CA.

[Liao and Ji, 2009] Liao, W. and Ji, Q. (2009). Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 42:3046–3056.

[Liu et al., 1998] Liu, B., Hsu, W., and Ma, Y. M. (1998). Integrating classification and association rule mining. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 80–86, New York, NY.

[Luce and Raiffa, 1957] Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey.* John Wiley and Sons, New York, NY.

[Miranda, 2008] Miranda, E. (2008). A survey of the theory of coherent lower previsions. *International Journal Approximate Reasoning*, 48(2):628–658.

[Motro and Smets, 1997] Motro, A. and Smets, P., editors (1997). *Uncertainty Management in Information Systems: From Needs to Solutions.* Kluwer Academic Publishers, Boston, MA.

[Nanavati et al., 2001] Nanavati, A. A., Chitrapura, K. P., Joshi, S., and Krishnapuram, R. (2001). Mining generalized disjunctive association rules. In *Proc. International Conference on Information and Knowledge Management (CIKM)*, pages 482–489, Atlanta, GA.

[Nguyen et al., 2002] Nguyen, H. T., Mukaidono, M., and Kreinovich, V. (2002). Probability of implication, logical version of Bayes theorem, and fuzzy logic operations. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 530–535, Honolulu, HI.

[Nguyen and Smets, 1993] Nguyen, H. T. and Smets, P. (1993). On dynamics of cautious belief and conditional objects. *International Journal of Approximate Reasoning*, 8(2):89–104.

[Nilsson, 1986] Nilsson, N. J. (1986). Probabilistic logic. *Artificial Intelligence*, 28(1):71–88.

[Núñez et al., 2013] Núñez, R. C., Dabarera, R., Scheutz, M., Briggs, G., Bueno, O., Premaratne, K., and Murthi, M. N. (2013). DS-based uncertain implication rules for inference and fusion applications. In *Proc. International Conference on Information Fusion (FUSION)*, pages 1934–1941, Istanbul, Turkey.

[Núñez et al., 2018] Núñez, R. C., Murthi, M. N., Premaratne, K., Scheutz, M., and Bueno, O. (2018). Uncertain logic processing: Logic-based inference and reasoning using Dempster-Shafer models. *International Journal of Approximate Reasoning*, 95:1–21.

[Nunez et al., 2013] Nunez, R. C., Scheutz, M., Premaratne, K., and Murthi, M. N. (2013). Modeling uncertainty in first-order logic: A Dempster-Shafer theoretic approach. In *Proc. of the International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, pages 1934–1941, Compiegne, France.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Francisco, CA.

[Pearl, 1990] Pearl, J. (1990). Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5/6):363–389.

[Polpitiya et al., 2016] Polpitiya, L. G., Premaratne, K., Murthi, M. N., and Sarkar, D. (2016). A framework for efficient computation of belief theoretic operations. In *Proc. of the International Conference on Information Fusion (FUSION)*, pages 1570–1577, Heildelberg, Germany.

[Polpitiya et al., 2017] Polpitiya, L. G., Premaratne, K., Murthi, M. N., and Sarkar, D. (2017). Efficient computation of belief theoretic conditionals. In *Proc. International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, volume 62 of *Proceedings of Machine Learning Research (PMLR)*, pages 265–276. Lugano, Switzerland.

[Premaratne et al., 2007] Premaratne, K., Dewasurendra, D. A., and Bauer, P. H. (2007). Evidence combination in an environment with heterogeneous sources. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 37(3):298–309.

[Premaratne et al., 2009] Premaratne, K., Murthi, M. N., Zhang, J., Scheutz, M., and Bauer, P. H. (2009). A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. In *Proc. International Conference on Information Fusion (FUSION)*, pages 2122–2129, Seattle, WA.

[Quinlan, 1983] Quinlan, J. R. (1983). Inferno: A cautious approach to uncertain inference. *The Computer Journal*, 26(3):255269.

[Ramoni and Sebastiani, 2001] Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data. *Machine Learning*, 45(2):147–170.

[Roudposhti et al., 2016] Roudposhti, K. K., Nunes, U., and Dias, J. (2016). Probabilistic social behavior analysis by exploring body motion-based patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1679–1691.

[Sambhoos et al., 2008] Sambhoos, K., Llinas, J., and Little, E. (2008). Graphical methods for real-time fusion and estimation with soft message data. In *Proc. International Conference on Information Fusion (FUSION)*, pages 1–8, Cologne, Germany.

[Sarathy et al., 2017] Sarathy, V., Scheutz, M., and Malle, B. F. (2017). Learning behavioral norms in uncertain and changing contexts. In *Proc. IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen, Hungary.

[Seidenfeld et al., 1989] Seidenfeld, T., Kadane, J. B., and Schervish, M. J. (1989). On the shared preferences of two bayesian decision makers. *The Journal of Philosophy*, 86(5):225–244.

[Shafer, 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

[Shi et al., 2017] Shi, P., Fan, X., Ni, J., Khan, Z., and Li, M. (2017). A novel underwater dam crack detection and classification approach based on sonar images. *PLoS ONE*, 12(6):1–17.

[Smets, 1992] Smets, P. (1992). Resolving misunderstanding about belief functions. *International Journal of Approximate Reasoning*, 6(3):321–344.

[Smets, 1994] Smets, P. (1994). What is Dempster-Shafer's model? In Yager, R. R., Fedrizzi, M., and Kacprzyk, J., editors, *Advances in the Dempster-Shafer Theory of Evidence*, pages 5–34. John Wiley and Sons, New York, NY.

[Smets, 1999] Smets, P. (1999). Practical uses of belief functions. In Laskey, K. B. and Prade, H., editors, *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 612–621, San Francisco, CA. Morgan Kaufmann.

[Smets, 2002] Smets, P. (2002). The application of the matrix calculus to belief functions. *International Journal of Approximate Reasoning*, 31(1/2):1–30.

[Soua et al., 2016] Soua, R., Koesdwiady, A., and Karray, F. (2016). Big-data-generated traffic flow prediction using deep learning and Dempster-Shafer theory. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pages 3195–3202, Vancouver, BC, Canada.

[Tassem, 1992] Tassem, B. (1992). Interval probability propagation. *International Journal of Approximate Reasoning*, 7(3/4):95–120.

[van der Gaag, 1990] van der Gaag, L. (1990). Computing probability intervals under independency constraints. In [Bonissone et al., 1990], pages 457–466.

[VanderPlas et al., 2012] VanderPlas, J., Connolly, A. J., Ivezic, Z., and Gray, A. (2012). Introduction to astroML: Machine learning for astrophysics. In *Conference on Intelligent Data Understanding (CIDU)*, pages 47–54, Boulder, CO.

[Vannoorenberghe, 2004] Vannoorenberghe, P. (2004). On aggregating belief decision trees. *Information Fusion*, 5(3):179–188.

[Velikova et al., 2012] Velikova, M., Lucas, P. J. F., Samulski, M., and Karssemeijer, N. (2012). A probabilistic framework for image information fusion with an application to mammographic analysis. *Medical Image Analysis*, 16(4):865–875.

[Villalba et al., 2016] Villalba, J., Miguel, A., Ortega, A., and Lleida, E. (2016). Bayesian networks to model the variability of speaker verification scores in adverse environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2327–2340.

[Walley, 1991] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, UK.

[Walley, 1996] Walley, P. (1996). Measures of uncertainty in expert systems. *Artificial Intelligence*, 83:1–58.

[Wang et al., 2016] Wang, X., Wang, Y., and Sun, H. (2016). Exploring the combination of Dempster-Shafer theory and neural network for predicting trust and distrust. *Computational Intelligence and Neuroscience*, 2016.

[Wasserman and Kadane, 1992] Wasserman, L. and Kadane, J. B. (1992). Symmetric upper probabilities. *The Annals of Statistics*, 20(4):1720–1736.

[Whittaker, 1990] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons, Ltd., Chichester, West Sussex, England.

[Wickramarathne et al., 2011a] Wickramarathne, T. L., Premaratne, K., Kubat, M., and Jayaweera, D. T. (2011a). CoFiDS: A belief-theoretic approach for automated collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 23(2):175–189.

[Wickramarathne et al., 2012] Wickramarathne, T. L., Premaratne, K., and Murthi, M. N. (2012). Consensus-based credibility estimation of soft evidence for robust data fusion. In Denoeux, T. and Masson, M.-H., editors, *Proc. International Conference on Belief Functions (BELIEF)*, volume 164 of *Advances in Intelligent and Soft Computing (AISC)*, pages 301–309. Springer-Verlag, Compiegne, France.

[Wickramarathne et al., 2013] Wickramarathne, T. L., Premaratne, K., and Murthi, M. N. (2013). Towards efficient computation of the Dempster-Shafer belief theoretic conditionals. *IEEE Transactions on Cybernetics*, 43(2):712–724.

[Wickramarathne et al., 2014] Wickramarathne, T. L., Premaratne, K., Murthi, M. N., and Chawla, N. V. (2014). Convergence analysis of iterated belief revision in complex fusion environments. *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Signal Processing for Social Networks*, 8(4):598–612.

[Wickramarathne et al., 2010] Wickramarathne, T. L., Premaratne, K., Murthi, M. N., and Scheutz, M. (2010). A Dempster-Shafer theoretic evidence updating strategy for non-identical frames of discernment. In *Proc. Workshop on the Theory of Belief Functions (BELIEF)*, Brest, France.

[Wickramarathne et al., 2011b] Wickramarathne, T. L., Premaratne, K., Murthi, M. N., Scheutz, M., Kübler, S., and Pravia, M. (2011b). Belief theoretic methods for soft and hard data fusion. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2388–2391, Prague, Czech Republic.

[Wilson, 2001] Wilson, N. (2001). Algorithms for Dempster-Shafer theory. In Gabbay, D. M. and Smets, P., editors, *Algorithms for Uncertainty and Defeasible Reasoning*, volume 5 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 421–475. Kluwer.

[Xu and Smets, 1996] Xu, H. and Smets, P. (1996). Reasoning in evidential networks with conditional belief functions. *International Journal of Approximate Reasoning*, 14(2/3):155–185.

[Yager et al., 1994] Yager, R. R., Fedrizzi, M., and Kacprzyk, J., editors (1994). *Advances in the Dempster-Shafer Theory of Evidence*. John Wiley and Sons, New York, NY.

[Zaffalon, 2002a] Zaffalon, M. (2002a). Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122.

[Zaffalon, 2002b] Zaffalon, M. (2002b). The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21.

[Zhang et al., 2017] Zhang, Z., She, Z., and Zhang, A. (2017). A semi-supervision fault diagnosis method based on attitude information for a satellite. *IEEE Access*, 5:20303–20312.