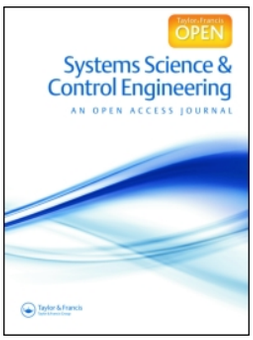

Systems Science & Control Engineering

An Open Access Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tssc20

Causes detection of unqualified bendability of hot rolled strip via improved RankBoost with multiple feature ranking algorithms

Fei He , Lidong Wang & Honglei Wang

To cite this article: Fei He , Lidong Wang & Honglei Wang (2020): Causes detection of unqualified bendability of hot rolled strip via improved RankBoost with multiple feature ranking algorithms, Systems Science & Control Engineering, DOI: 10.1080/21642583.2020.1843084

To link to this article: https://doi.org/10.1080/21642583.2020.1843084

Published online: 11 Nov 2020.

SYSTEMS SCIENCE & CONTROL ENGINEERING: AN OPEN ACCESS JOURNAL
https://doi.org/10.1080/21642583.2020.1843084

Causes detection of unqualified bendability of hot rolled strip via improved RankBoost with multiple feature ranking algorithms

Fei He[a], Lidong Wang[a] and Honglei Wang[b]

[a]Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing, People's Republic of China; [b]Tiandi Science and Technology Company Ltd., Beijing, People's Republic of China

ABSTRACT

In hot rolling process, mechanical properties of steel materials are important to steel quality. The bendability is one of the key parameters to evaluate the formability of the strip. When the bendability is unqualified, how to detect causes becomes a big challenge. In this paper, a model to find the causes of bendability of hot rolled strip based on improved RankBoost with multiple feature selection algorithms using historical data is built. Firstly, the related process variables and bendability results are collected. And then, seven feature ranking methods including Fisher score, Relief, Gini index, T-test, Kruskal–Wallis, mutual information entropy and minimum redundancy maximum relevance (MRMR), are used to rank the significance of features individually. Finally, to summarize the results of the seven methods, the total importance of every feature can be obtained using the improved RankBoost method to select the most important features as the major causes. The real field data set from hot rolling strip steel process is used to validate the model. The results demonstrate that the RankBoost method can give a more credible result.

ARTICLE HISTORY
Received 30 June 2020
Accepted 24 October 2020

KEYWORDS
Bendability; hot rolled strip; causes detection; feature selection; information entropy

Hot rolled steel strip is manufactured through various production processes. Iron ore and coke are fed to the blast furnace to make iron. The blast furnace is a huge chemical reactor where reduction reactions take place. The iron is then sent to the steel producing making process where bloom is produced. The steel making process consists of converters for removing carbon, refiners for adjusting elements, and continuous casters. Then, the blooming process resizes bloom to slab for the next rolling process. The purpose of hot rolling is to turn reheated steel slabs into strips.

Bendability is one of the deformation modes in press forming (Lester, 1973). Therefore, one type of steel strip, which is mainly used for automotive parts, is required to have better bendability. In manufacturing there are some unqualified products. The physical models and finite element methods are used to analyse the relationship between the microstructure and bendability (Eiji et al., 2014). The analysis results can be used to design and control the bendability of strip. When unqualified products happen, it is hard to find the causes.

Modern hot rolling process is highly automated and often monitored by many sensors. The large amount of sensing data provides great opportunities for effective quality control of hot rolling process. Mechanical properties are influenced by chemical content, and all kinds of process parameters in manufacturing. When the product quality cannot satisfy the need of customer, fault diagnosis methods are used to detect the major causes. Multivariate statistical approaches for process monitoring and fault diagnosis have been rapidly developed in recent decades, mainly due to the adoption of powerful latent projection techniques such as principal components analysis (PCA) and partial least squares (PLS) (Sharma et al., 2013). To cope with nonlinearity problem, the principal curve based on neural network and kernel PCA (KPCA), kernel PLS (KPLS) (Liu et al., 2011; Samuel & Cao, 2016) are proposed.

When the process data from normal process is supplied, PCA and KPCA methods can be used for process monitoring and detection causes. When the process data and continuous quality data from normal process is collected, PLS and KPLS methods can be used. Bendability value is tested offline and there is a delay of several hours. Sometime there is a batch of unqualified products, and how to determine the cause quickly becomes a big challenge. Now, there are process data and discrete quality, we propose that feature selection methods are used to judge the significance of features for classification. The features that can separate the different classes clearly are selected preferentially. If the importance of the features can be ranked, the first several features can be considered

CONTACT Fei He hefei@ustb.edu.cn

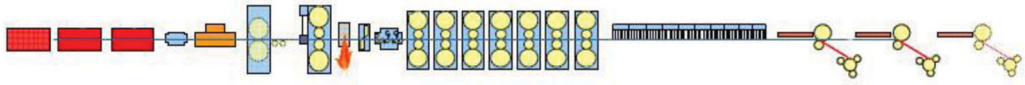

Figure 1. Layout drawing of hot rolling process.

as the causes of the unqualified products. There are a lot of feature selection methods can be used, but we do not know which one can give the correct significance analysis. In this paper, Fisher score, Relief, Gini index, T-test, Kruskal–Wallis, mutual information entropy and minimum redundancy maximum relevance (MRMR) are used to rank the significance of features individually. The contribution of this paper lies in the following two aspects: in order to select the most important features as the major causes, an improved RankBoost method is given by combining the other seven feature selection methods, and the total significance of every feature can be obtained that considers the results of every method. On the other hand, the improved method is applied to tackle the cause detection of unqualified bendability.

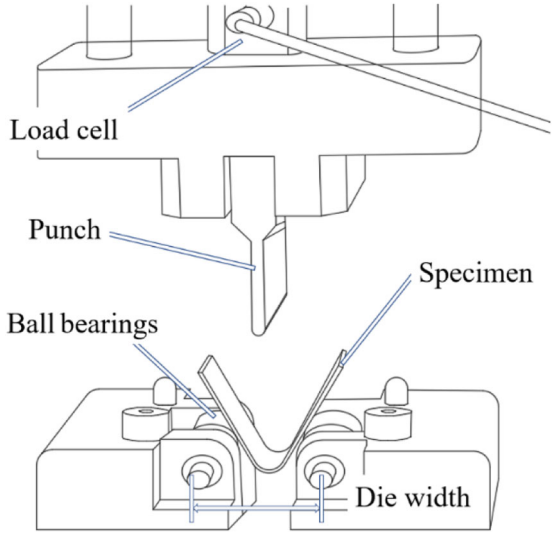

Figure 2. Bending test for small specimens.

1. Hot strip rolling process

1.1. Introduction of hot strip rolling process

The purpose of hot rolling is to turn reheated steel slabs into strips. A hot strip line is always composed of reheat furnaces, a roughing mill, several finishing mills, and two coilers. In the roughing mill, the reheated slabs are reduced to a thickness of 25–50 mm and narrowed to the desired width. The resulting sheet slab is then transported to the finishing mill, where it is further reduced to the final thickness of 1–20 mm. The resulting strip is then coiled to form the finished coil of steel strip. In the finishing rolling, to achieve the required reduction, final qualities and tolerances, several passes of rolling are executed by tandem rolling with six or seven successive stands in the presence of inter-stand tension. A simplified schematic diagram of a steel rolling mill for the production of coil plate is presented in Figure 1. It shows the route of slabs from entry at the reheat furnaces to their exit at the coiler. The process route can best be described in terms of the major items of equipment (Bissessur et al., 2000; Wang & He, 2019).

(1) Reheat furnace: The feed stock for the rolling mill are slabs produced by the continuous casting process in a steel plant. These are normally supplied at ambient temperature. The purpose of the furnace is to raise the temperature of the whole slab to around 1300°C.
(2) Roughing mill: This is a reversing mill that produces a breakdown slab (the product between the roughing mill and the finishing mill) by rolling the slab through a series of forward and reverse passes, typically reducing the slab thickness from 200 to 30 mm.
(3) Finishing mill: This is designed to reduce the gauge (thickness) of the breakdown slab to that of the finished strip, while maintaining the desired thickness. A sequential of combination of stands is used, e.g. six to seven. The mill control system is critical as constant mass flow that must be maintained in all stands to ensure continuous production.
(4) Down coiler: On exit from the mill, the hot strip typically has a velocity of up to 10 m/s and can be hundreds of metres in length. The down coiler allows the strip to be converted into a coil.

1.2. Bendability of hot strip

In hot strip rolling, mechanical properties of steel materials are important to steel quality which are detected offline and destructively. The main parameters of the mechanical properties include elongation rate, yield point, and tensile strength, which are continuous values. The bendability is one of the key parameters to evaluate the formability of the strip. Bending test is used to evaluate how easy it is to form by bending with approximate plane strain deformation and crack generation is checked after carrying out specified radius bending as shown in Figure 2 (Mertin et al., 2019). And then, the bendability can be described as qualified or unqualified.

Table 1. Variables table of bendability of hot rolled strip.

Category	Variables	Number
Rate	Thickness ratio	1
Temperature	Rough Milling exit Temp.	2
	Finishing entry Temp.	3
	Finishing exit Temp.	4
	Coiling Temp.	5
Composition	C	6
	Si	7
	Mn	8
	P	9
	S	10
Quality	Recurvation quality	

In order to detect causes using historical data, process variables that maybe impact the bendability quality as described in Table 1 according to expert knowledge.

2. Feature ranking indices

2.1. Fisher score

In the Fisher score method, given training vectors if the numbers of positive and negative instances are $n+$ and n- respectively, then the fisher score of the ith feature is explained as follows (Chen & Lin, 2006; Kemal & Volkan, 2011):

$$F(i) = \frac{(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i^{(-)})^2 + (\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i)^2}{((1)/(n_+ - 1)) \sum_{k=1}^{n_+} (\boldsymbol{x}_{ki}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (\boldsymbol{x}_{ki}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)})^2}, \quad (1)$$

where $\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{x}}_i^{(+)}$, $\bar{\boldsymbol{x}}_i^{(-)}$ are the average of the ith feature of the whole, positive and negative data sets, respectively. $\boldsymbol{x}_{ki}^{(+)}$ is the ith feature of the kth positive instance, and $\boldsymbol{x}_{ki}^{(-)}$ is the ith feature of the kth negative instance. The numerator presents the discrimination between the positive and negative sets, and the denomination explains the one within each of the two sets. The larger the Fisher score $F(i)$ is, the more likely this feature is more discriminative (Duda et al., 2001; Gunes et al., 2010).

2.2. Relief

The key idea of Relief is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. For that purpose, feature weights are iteratively estimated according to their ability to discriminate between neighbouring patterns (Aldehim & Wang, 2015). In each iteration, a data point $\boldsymbol{x}$ is randomly selected and then two nearest neighbours of $\boldsymbol{x}$ are found, one from the same class (termed the nearest *hit* or *NH*) and the other from a different class (termed the nearest *miss* or *NM*). The weight of the ith feature is then updated (Kira & Rendell, 1992):

$$\boldsymbol{w}_i = \boldsymbol{w}_i + diff(\boldsymbol{x}^{(i)}, NM^{(i)}(\boldsymbol{x}))/m - diff(\boldsymbol{x}^{(i)}, NH^{(i)}(\boldsymbol{x}))/m, \quad (2)$$

where i respects the ith feature, $\boldsymbol{x}$ is randomly selected data point, m is the sample size, and $diff()$ is the distance between samples. $HM(\boldsymbol{x})$ and $NH(\boldsymbol{x})$ are nearest neighbour sample points with the same class and different class, respectively. Then every weight is calculated through T iterations. The detail of Relief algorithm is depicted as (Aldehim & Wang, 2015; Kira & Rendell, 1992; Kononenko, 1994).

2.3. Gini index

Suppose that $\boldsymbol{x}_i$ is ith feature of data sets, its class label attribute has two different values, which defines different classes of C_j ($j = 1, 2$). According to the class label attribute value, $\boldsymbol{x}_i$can be divided into two subsets. If $\boldsymbol{x}_i^{(j)}$ is the subset of samples belongs to class C_j, and m_i is the number of the samples in the subset $\boldsymbol{x}_i^{(j)}$, then the Gini index of set $\boldsymbol{x}_i$is (Shang et al., 2006; Zhu & Lin, 2013)

$$Gini(\boldsymbol{x}_i) = 1 - \sum_{j=1}^{2} P_j^2, \quad (3)$$

where P_i is the probability of any sample of C_j, which estimated by m_i/m. When the minimum of Gini($\boldsymbol{x}_i$) is 0 that mean all records belong to the same category at this collection; it indicates the maximum useful information can be obtained. When all the samples in this collection have uniform distribution for certain category, Gini($\boldsymbol{x}_i$) reaches maximum, it indicates the minimum useful information obtained.

The form of the Gini index is used to measure the 'impurity' of attribute for categorization. The smaller its value that means lesser 'impurity', the better attribute. Then, the Gini index of every feature is computed and sorted in ascending order.

2.4. T-test

When we want to compare the difference between two set, T-test is used to test whether the mean values are different. Define the data set $\mathbf{X}$ which has two classes C_1 and C_2. We use m_1, v_1 stand for the mean and variance of features in class C_1, and m_2, v_2 for the mean and variance of features in class C_2. Then, T-test statistics are as follows (Wang et al., 2019):

$$\boldsymbol{t} = \frac{|\boldsymbol{m}_1 - \boldsymbol{m}_2|}{\sqrt{\left((\boldsymbol{v}_1)/\left(\boldsymbol{n}_1 + \frac{\boldsymbol{v}_2}{\boldsymbol{n}_2}\right)\right)}}. \quad (4)$$

In which, n_1, n_2 are the sample size of C_1 and C_2. Then, $\boldsymbol{t}$ is a vector in which t_i represents T-test results of ith Feature.

The ith is more significant when t_i is larger. Thus, T-test statistics is computed and t_i is sorted in descending order.

2.5. Kruskal–Wallis

The Kruskal–Wallis statistical test is a non-parametric test that makes no assumptions about the distribution of the data (e.g. normality) (Hollander & Wolfe, 1973). Many non-parametric test methods use data rank rather than raw values to calculate the statistic.

Let $n_1, n_2, \ldots, n_K$ represent the sample sizes for each of the K classes. The total sample size is $N = \sum_{k=1}^{K} n_k$. The combined sample is ranked and then, the sum of the ranks for the class k is computed as $R_k = \sum_{i|x_i \in \text{Class}_k} \text{rank}(x_i)$. The Kruskal–Wallis test statistic is calculated as below:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{R_k^2}{n_k} - 3(N+1). \tag{5}$$

If the null hypothesis of equal median holds, this test statistic corresponds approximately to a chi-square distribution with $K-1$ degrees of freedom. The larger the test statistic H, the weaker the null hypothesis becomes, since a strong separation of the medians indicates that the feature under consideration has a high classification power (Cor et al., 2006).

2.6. Mutual information entropy

Information theory was conceptualized by Shannon to deal with the problem of optimally transmitting message over noisy channels. In information theory, entropy is regarded as a measure of information and Hartley called it the 'amount of information' (Principe, 2010). Since it is capable of quantifying the uncertainty of random variables and scaling the amount of information shared by them effectively, it has been widely used in many fields (Principe, 2010).

Let $\boldsymbol{X}$ denote a random variable taking values in a finite set $\boldsymbol{X} = \{x_1, x_2, \ldots, x_k, \ldots x_N\}$ according to a probability distribution $p(x_k)$; its uncertainty can be measured by entropy $H(\boldsymbol{X})$, which is defined as

$$H(\boldsymbol{X}) = -\sum_{x_k \in X} p(x_k) \log p(x_k). \tag{6}$$

Note that entropy does not depend on actual values, but just the probability distribution of random variable.

The total decrease of uncertainty in $\boldsymbol{X}$ by observing $\boldsymbol{Y}$ is known as the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is defined as (Sylvain et al., 2008)

$$I(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{x_k \in X} \sum_{y_i \in Y} p(x_k, y_i) \log_2 \frac{p(x_k, y_i)}{p(x_k) p(y_i)}. \tag{7}$$

The mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ is used to quantify how much information shared by two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$.

2.7. MRMR

Minimum redundancy maximum relevance is based on mutual information method (Peng et al., 2005). We propose to expand the space covered by the feature set by requiring that features are maximally dissimilar to each other, for example, their mutual Euclidean distances are maximized, or their pairwise correlations are minimized. These minimum redundancy criteria are of course supplemented by the usual maximum relevance criteria such as maximal mutual information with the targeted phenotypes. The benefits of this approach can be realized in two ways (Ding & Peng, 2005): (1) with the same number of features, the MRMR feature set is expected to be more representative of the targeted phenotypes, therefore leading to better generalization property; (2) equivalently, a smaller MRMR feature set can be used to effectively cover the same space that a larger conventional feature set does.

Minimal redundancy will make the feature set a better representation of the entire dataset. The minimum redundancy condition is

$$\min(W_i), W_i = \frac{1}{S^2} \sum_{i,j \in S} I(x_i, x_j) \tag{8}$$

where $I(x_i, x_j)$ is used to represent mutual information between x_i and x_j. S denote the feature set. S is the number of features.

The maximum relevance condition is to maximize the total relevance of all feature in S:

$$\max(V_i), V_i = \frac{1}{S} \sum_{i \in S} I(x_i, y) \tag{9}$$

where $I(x_i, y)$ represent mutual information between x_i and y.

The MRMR feature set is obtained by optimizing the conditions in Equations (8) and (9) simultaneously. Then the simplest combination criterion is considered as

$$F = \max(V_i - W_i). \tag{10}$$

3. Improved RankBoost with multiple feature ranking algorithms

Let $\boldsymbol{X}$ be a set called the feature space. These are many feature selection methods that can give rankings. The

Table 2. The advantages and disadvantages of each feature selection method.

No.	Method	Advantages	Disadvantages
1	MRMR	Suitable for high-dimensional data observed in two or more different groups.	High sensitivity of standard correlation and redundant measurement of outliers.
2	Mutual Info	Strong theoretical basis in information theory.	No way to normalize, not very convenient to calculate continuous variables.
3	Gini	Eliminate data redundancy, combine other features to test feature correlation.	High time complexity
4	T-test	Eliminate differences between subjects, no need for large sample data sets.	Ignore the dependence and correlation between variables.
5	Fisher score	Independent calculation of feature scores and feature selection.	Ignore feature correlation, resulting in feature subsets that may be sub-optimal.
6	Kruskai–Wallis	The sample data does not have to be normally distributed.	Not very convenient to calculate discrete variables.
7	Relief	Strong versatility, low complexity, remove irrelevant features, suitable for large-scale data sets.	Independent of the specific learning algorithm.

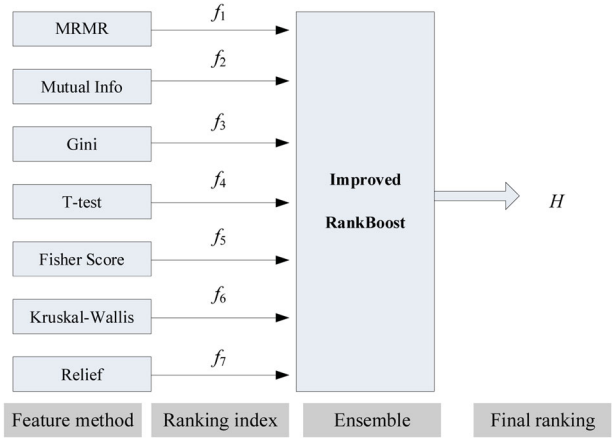

Figure 3. Basic layout of ensemble of feature selection methods via improved RankBoost.

advantages and disadvantages of each feature selection method are shown in Table 2. Our goal is to combine a given set of feature ranking. A feature ranking is nothing more than an ordering of the features from most preferred to least preferred. In the paper the improved RankBoost method is used to combine the feature ranking results got from different methods as shown in Figure 3, instead of the mean processing method.

We assume that n learning algorithms give n ranking features denoted as $f_1, f_2, \ldots, f_n$. Since each ranking feature f_i defines a linear ordering of the features, we can equivalently think of f_i as a scoring function where higher scores are assigned to more preferred feature. That is, we can represent any ranking feature as a real-valued function where $f_i(x_1) > f_i(x_0)$ means that feature x_1 is preferred to x_0 by f_i.

Note that, every feature selection method may give different ranking. So that, RankBoost method is introduced to combine all of the ranking order into a single ranking called the final ranking that can be represented by a function H. If $H(x_1) > H(x_0)$, x_1 is preferred to x_0. RankBoost is an iterative algorithm based on Adaboost to solve ranking problem. Like all boosting algorithms, RankBoost operates in rounds. The pseudo code is shown in Figure 4 (Freund et al., 2003). We assume access to a separate procedure called the weak learner that, on each round, is called to produce a weak ranking. RankBoost maintains a distribution D_t over $\boldsymbol{X} \times \boldsymbol{X}$ that is passed on round t to the weak learner. Intuitively, RankBoost chooses D_t to emphasize different parts of the training data. A high weight assigned to a pair of features indicates a great importance that the weak learner order that pair correctly. The boosting algorithm uses the weak rankings to update the distribution. The weight is decreased if h_t gives a correct ranking and increased otherwise. The final ranking H is a weighted sum of the weak rankings.

RankBoost Algorithm is used in combining all the feature selection methods as shown in Figure 5, the detail description as follows:

Firstly, we get ranking features $f_1, f_2, \ldots, f_n$ by the feature selection methods, which are used as weak learners on each round of RankBoost. Here, we can equivalently think of f_i as a scoring function where $f_i(x_0) = m + 1 - idx(x_0)$. And m is the number of features, $idx(x_0)$ represents the ranking index of x_0 in the linear ordering f_i.

Secondly, we start to combine them by RankBoost. The initial distribution D over $X \times X$ is needed here. Assume the function has the form $\Phi : X \times X \to R$. Here, $\Phi(x_0, x_1) > 0$ means that x_1 should be ranked above x_0 while $\Phi(x_0, x_1) < 0$ means the opposite; a value of zero indicates no preference between x_0 and x_1, so $\Phi(x_0, x_1) = 0$. To minimize the (weighted) number of pairs of features, let $D(x_0, x_1) = c \cdot \max\{0, \Phi(x_0, x_1)\}$. Here, c is a positive constant chosen so that

$$\sum_{x_0, x_1} D(x_0, x_1) = 1. \tag{11}$$

In this paper, we set function $\Phi(x_0, x_1) = F(x_1) - F(x_0)$, where F is the combination of the linear orderings $f_1, f_2, \ldots, f_n$ by feature selection methods and $F(x_0)$ has

Given: initial distribution D over $X \times X$.
Initialize: $D_1 = D$.
For $t = 1, \cdots, T$

- Train weak learner using distribution D_t.
- Get weak ranking $h_t: X \to \mathbb{R}$.
- Choose $\alpha_t \in \mathbb{R}$.
- Update:

$$D_{t+1}(x_0, x_1) = \frac{D_t(x_0, x_1) exp\left(\alpha_t\left(h_t(x_0) - h_t(x_1)\right)\right)}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution.).
Output the final ranking:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

Figure 4. RankBoost algorithm.

the same formation as $f_i(x_0)$, that is, $F(x_0) = m + 1 - idx(x_0)$, where m is the number of features, $idx(x_0)$ represents the ranking index of x_0 in the linear ordering f_i. Here we use the average of ranking indices of each feature in $f_1, f_2, \ldots, f_n$ to restart sorting, then we get F.

Thirdly, the distribution D_t is passed on round t to weak learner. On each round, we need to find the best one among all the weak leaners according to D_t, which can minimize the ranking loss defined to be

$$rloss_D(H) = \sum_{x_0, x_1} D(x_0, x_1) H(x_1) \le H(x_0), \tag{12}$$

where $*$ is 1 when $*$ is true, and it is 0 when $*$ is false. Then, we have $rloss_D(H) \le \prod_{t=1}^{T} Z_t$, where $Z_t = \sum_{x_0, x_1} D_t(x_0, x_1) exp(\alpha_t(h_t(x_0) - h_t(x_1)))$. So, we can minimize the Z_t in each round to reduce the ranking loss. At the same time, the parameter α_t is chosen.

Suppose in the current round, the weak learner is $h(x) = f_i(x)$, when $h(x) \in [-1, +1]$, we have

$$\begin{aligned} Z &\le \sum_{x_0, x_1} D(x_0, x_1) \left[\left(\frac{1 + h(x_0) - h(x_1)}{2} \right) e^{\alpha} \right. \\ &\quad \left. + \left(\frac{1 - h(x_0) + h(x_1)}{2} \right) e^{-\alpha} \right] \\ &= \left(\frac{1 - r}{2} \right) e^{\alpha} + \left(\frac{1 + r}{2} \right) e^{-\alpha} \end{aligned} \tag{13}$$

where

$$r = \sum_{x_0, x_1} D(x_0, x_1)(h(x_1) - h(x_0)). \tag{14}$$

Z is minimized when

$$\alpha = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r} \right), \tag{15}$$

which, plugging into Equation (13), yields $Z \le \sqrt{1 - r^2}$.

In particular, we will use weak rankings h of the form

$$h(x) = \begin{cases} 1 & if\ f_i(x) > \theta \\ 0 & if\ f_i(x) < \theta \end{cases}, \tag{16}$$

where the threshold $\theta \in \{\theta_j\}_{j=1}^{J}$ is made up of different values in $f_i(x)$ and $-\infty$, which is ascending ordered. So that Equation (14) has another form

$$r = \sum_{x} h(x)\pi(x), \tag{17}$$

where $\pi(x) = \sum_{x'} (D(x', x) - D(x, x'))$. As the weak rankings h has the 0–1 form, we have

$$r = \sum_{x: f_i(x) > \theta} h(x)\pi(x) + \sum_{x: f_i(x) \le \theta} h(x)\pi(x) = \sum_{x: f_i(x) > \theta} \pi(x). \tag{18}$$

So, we can make the unknown parameter in $|r|$ a group as (f_i, θ). Change f_i among all the scoring function by feature selection methods and $\theta \in \{\theta_j\}_{j=1}^{J}$ mentioned above, we will get different values of $|r|$. When $|r|$ gets the most, the f_i is the best learner. Then bring the value of r into Equation (15), we get the α_t.

Finally, update the distribution D_t to D_{t+1}, which will be passed on to next round. Until the ranking loss tends to a lower stability point, the final ranking H is obtained.

4. Experiments

4.1. Data set

There are 10 production process variables collected from the real hot strip rolling field, including thickness reduction ratio, rolling process temperature information(rough

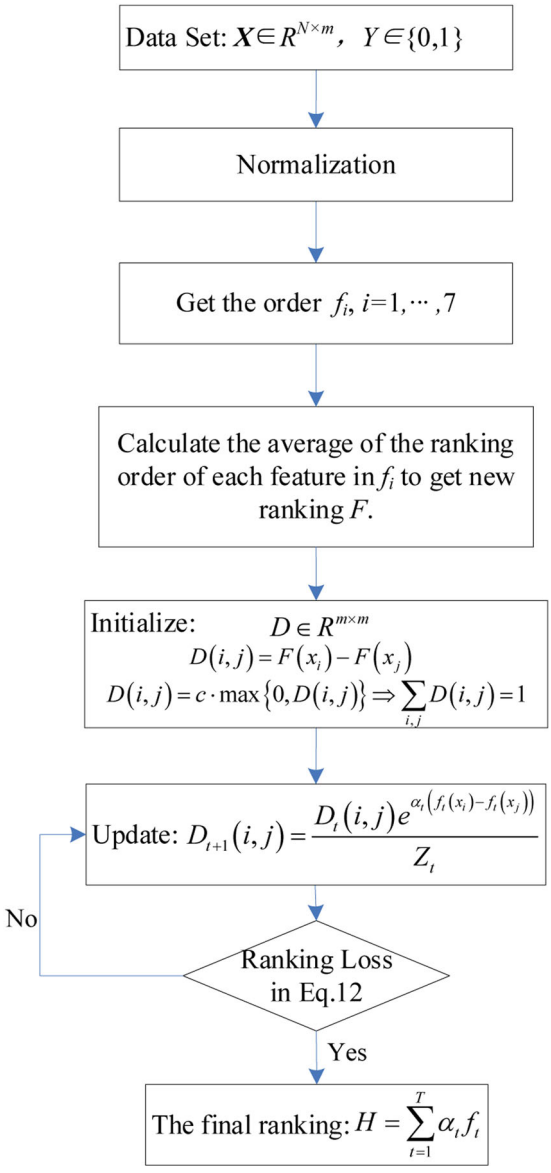

Figure 5. Flow chart of RankBoost.

milling exit temperature, finishing entry temperature, finishing exit temperature and coiling temperature), content of chemical components(C, Si, Mn, P, S). Because the bendability is detected destructively, only one sample can be chosen to test bendability that represent the entire coil quality in each steel coil. The mean values of the process parameters in each steel coil are computed to correspond to the bendability. In all, 961 samples are collected, in which 890 samples come from qualified products process and the other 71 samples come from unqualified products process.

In order to summarize the dataset, the statistics of hot rolled strip data including maximum, minimization, average and standard deviation of every variable are shown in Table 3. From Table 3, each variable has different data range. Firstly, data normalization is used to deal with the raw data to remove the impact of different data ranges. Data normalization is that the raw data minus the mean and divided by the standard deviation.

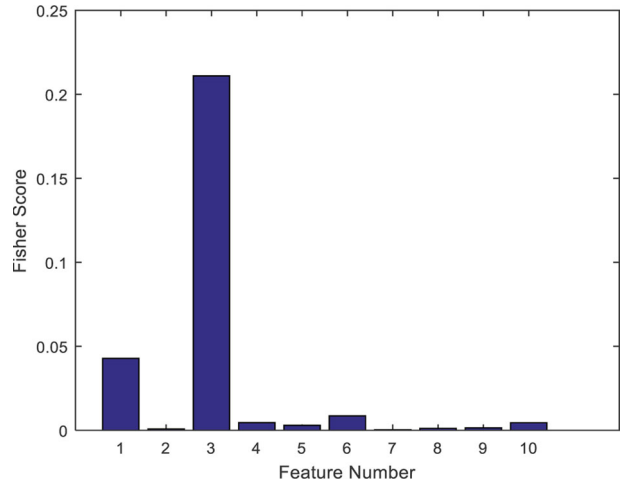

Figure 6. Feature select based on Fisher score.

4.2. Feature select process

Seven feature selection methods including Fisher score, Relief, Gini index, T-test, Kruskal–Wallis, mutual information entropy and Minimum redundancy maximum relevance, are used to rank the significance of features individually.

(1) Fisher score

In the Fisher score method, the Fisher score $F(i)$ of every feature is computed using Equation (1), and then is plotted as Figure 6. The larger the Fisher score $F(i)$ is, the more discriminative this feature is. In Figure 3 the 3rd feature has the largest value, so finishing entry temperature is the most important feature in the Fisher score method.

(2) Relief

In the Relief method, the weight of every feature is calculated using Equation (2). The larger the weight is, the more important the feature is. And then, the weights are described in Figure 7. Then we can get that the thickness reduction ratio is the most important feature in the Relief method.

(3) Gini index

Gini index reaches maximum, it indicates the minimum useful information obtained. Its value means lesser

Table 3. Statistical values.

Variables	Maximum	Minimization	Average	Standard Deviation
Thickness reduction ratio/%	0.8687	0.8421	0.8538	0.0046
Rough Milling exit Temp./℃	1119	1001	1061.01	20.9347
Finishing entry Temp. /℃	1040	915	987.50	23.2820
Finishing exit Temp. /℃	926	825	858.71	12.0632
Coiling Temp. /℃	699	593	619.02	15.7791
C/%	0.21	0.112	0.1603	0.0142
Si/%	0.36	0.10	0.2043	0.0321
Mn/%	1.43	0.399	1.2957	0.0662
P/%	0.037	0.01	0.0201	0.0042
S/%	0.028	0.003	0.0130	0.0037

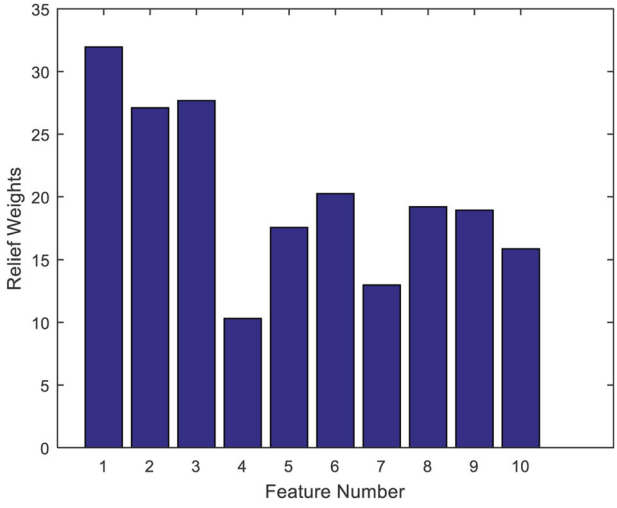

Figure 7. Feature select based on relief.

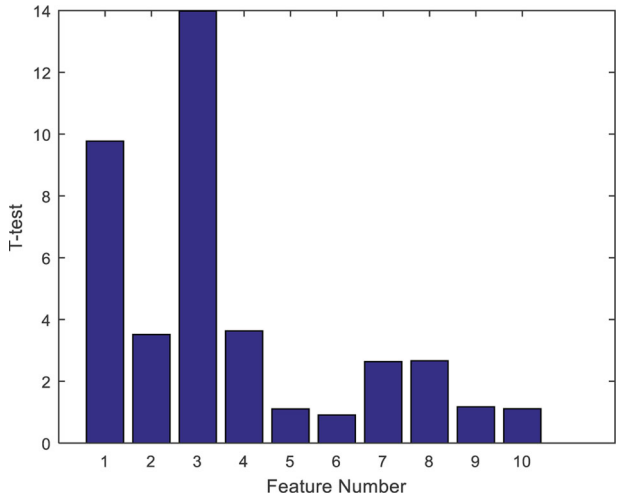

Figure 9. Feature select based on T-test.

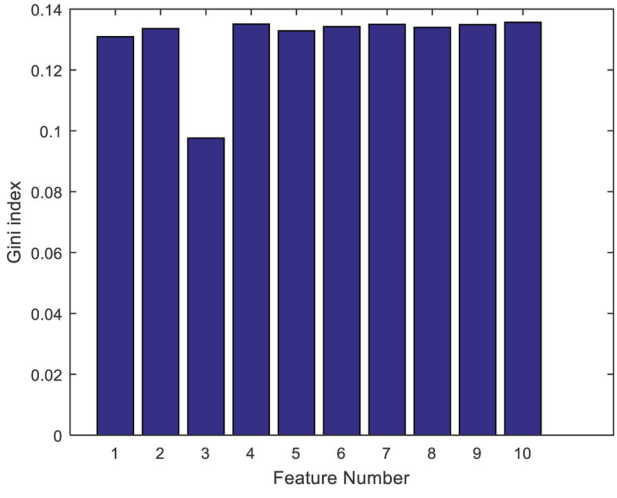

Figure 8. Feature select based on Gini index.

'impurity', the better attribute. Gini index is accounted using Equation (3) and the result is given in Figure 8. From Figure 8, the 3rd feature gets the smallest Gini index, so the finishing entry temperature is the most important feature in the Gini index method.

(4) T-test

In T-test method, the weight of every feature is calculated using Equation (4) that presents how significant difference by comparing the mean between the two classes. The larger the weight is, the more significant the feature is to separate the two classes. And then, the weights are described in Figure 9. Then we can get that the finishing entry temperature is the most important feature in the T-test method.

(5) Kruskal–Wallis

In Kruskal–Wallis method, the weight of every feature is calculated using Equation (5). The larger the weight is, the more important the feature is. And then, the weights are described in Figure 10. So finishing entry temperature is the most important feature in Kruskal–Wallis method, and the thickness reduction ratio on its heels.

(6) Mutual information entropy

The mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ is used to quantify how much information shared by two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. In

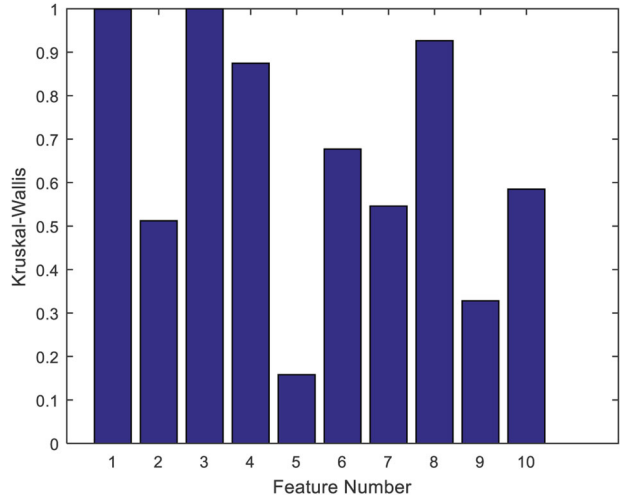

Figure 10. Feature select based on Kruskal–Wallis.

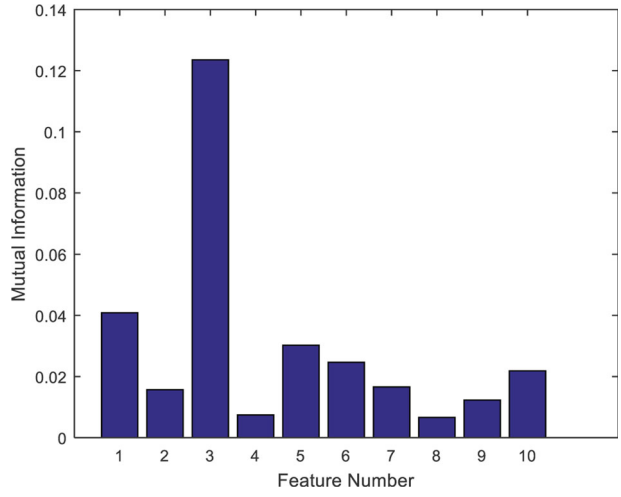

Figure 11. Feature select based on mutual information entropy.

the feature select process, $\boldsymbol{X}$ respects the feature and $\boldsymbol{Y}$ respects the quality information. The larger the mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ is, the more correlative between the feature and quality is. The mutual information between every feature and the quality is calculated using Equation (7) and shown in Figure 11. The finishing entry temperature is also considered as the most important feature in mutual information entropy method.

(7) MRMR

In MRMR method, the minimal redundancy will make the feature set a better representation of the entire dataset. The maximum relevance condition is to maximize the total relevance of all feature. To achieve minimum redundancy and maximum relevance, features ranking is computed as Equation (10). Then we can get that the finishing entry temperature is the most important feature in MRMR method.

4.3. Causes detection based on RankBoost

Finally, to summarize the results of the seven methods, the total importance of every feature can be obtained using the RankBoost method to select the most important features as the major causes.

Feature important rankings based on every method are collected in Table 4. Through Table 4, the 3rd feature is ranked as the most important feature six times in seven methods. But in the Relief method, the 3rd feature is not regarded as the first one. And the first five features in every method are not absolutely the same. If we only use one feature selection method, maybe cannot get the credible cause. To summarize the results of the seven methods, mean processing and RankBoost method are used and the result are shown in Table 5. In the mean processing method, features are sorted by the mean ranking values of each feature in all the methods. To show the result clearly, the selected features are used for classification, then fault detection rate and false alarm rate are used to evaluate the ranking results. The better ranking result is, the higher fault detection rate is and the lower false alarm rate is. The support vector machines (SVM) are introduced to classify the normal and abnormal samples. The parameters of SVM are optimized using cross validation. The results of fault detection rate and lower false alarm rate are shown in Figures 12 and 13 respectively with the increasing feature number based on every method.

As shown Figures 12 and 13, fault detection rate and false alarm rate are improving as feature number increasing in almost every method. But in mean processing and RankBoost method, both the fault detection rate and false alarm rate are improving faster and more steadily

Table 4. Feature important orders based on every method.

Methods	Feature important sort
MRMR	3 1 10 6 9 8 4 7 5 2
Mutual Info	3 1 5 6 10 7 2 9 4 8
Gini	3 1 5 2 8 6 9 7 4 10
T-test	3 1 4 2 8 7 9 10 5 6
Fisher score	3 1 6 4 10 5 9 8 2 7
Kruskal–Wallis	3 1 8 4 6 10 7 2 9 5
Relief	1 3 2 6 8 9 5 10 7 4

Table 5. Combined feature important order based on mean processing and RankBoost.

Methods	Feature important sort
Mean processing	3 1 6 8 2 10 4 5 9 7
RankBoost	3 1 6 8 9 2 5 4 7 10

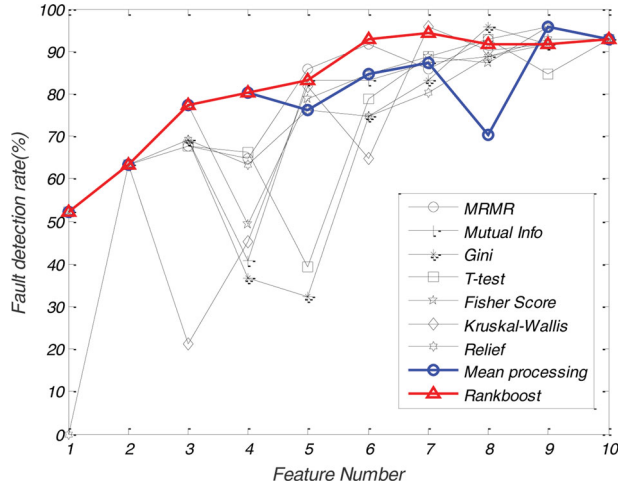

Figure 12. Fault detection rate with increasing feature number based on each method.

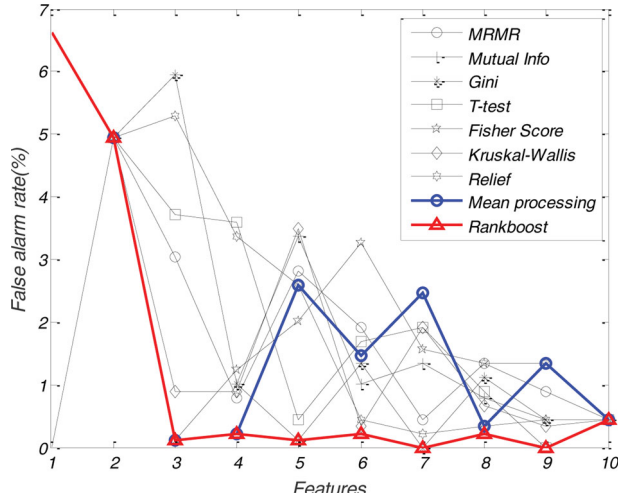

Figure 13. False alarm rate with increasing feature number based on each method.

than each feature selection method. Comparing to the mean processing method, RankBoost gets much better result. In RankBoost method, false alarm rate reaches around the lowest value when selecting 3 features and fault detection rate gets the maximum value when selecting 7 features. RankBoost algorithm can combine different feature rankings to a final feature ranking which is conducive to indicating the major causes. It is hard to select a suitable method in the real data set. If only one method is used to select the feature importance, maybe the wrong decision is done as Relief method in Table 4. In the other hand, in most cases the fault detection rate is lower than the RankBoost method especially via smaller feature number as Figure 12. As a result, there are more mistakes about quality prediction based on one single method.

As shown in Table 4, the 3rd feature corresponding to the finishing entry temperature is the most important feature based on RankBoost method, and then the 1st feature corresponding to thickness reduction ratio is the second most important feature. In the actual manufacturing process, the control accuracy of the finishing entry temperature should be improved. To compare clearly, the finishing entry temperature between the qualified and unqualified steel is shown as Figure 14. In Figure 14, the first 890 values come from the qualified bendability and the others from the unqualified bendability. When the finishing entry temperature is small, there is more probability to unqualified bendability. To improve the bendability, maybe we should increase the finishing entry temperature. In the real hot rolling process, there are many quality parameters, it is a multi-objective optimization problem. Besides, the thickness reduction ratio should be optimized in the future.

5. Discussion and conclusions

In this paper, a model to find the causes of bendability of hot rolled strip based on improved RankBoost method with multiple feature selection algorithms using historical data is built. Seven feature selection methods including

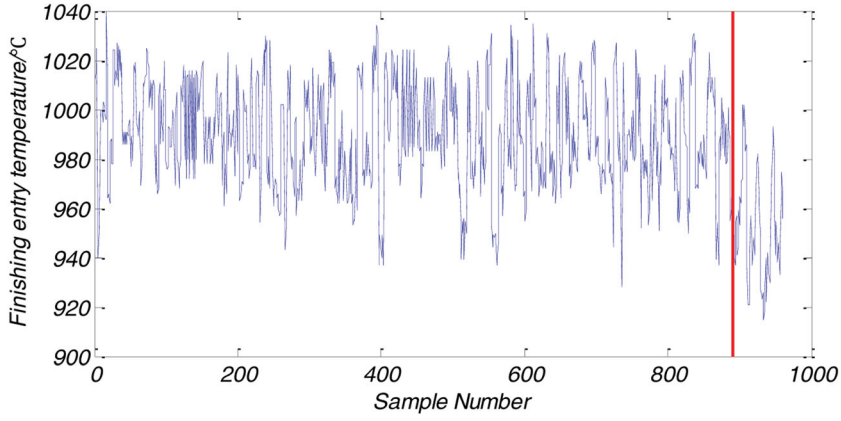

Figure 14. The values of finishing entry temperature.

Fisher score, Relief, Gini index, T-test, Kruskal–Wallis, mutual information entropy and Minimum redundancy maximum relevance, are used to rank the significance of features individually. Finally, to summarize the results of the seven methods, the total importance of every feature can be obtained using the RankBoost method to select the most important features as the major causes. Nine hundred and sixty samples including 890 qualified and 71 unqualified are collected to validate the model. The result shows that the finishing entry temperature is most important feature that causes the unqualified bendability. In the actual manufacturing process, we should to improve the control accuracy of the finishing entry temperature.

The cause detection based on feature selection me thod can be applied, when a large number of unqualified products appear. But we only can give the reason from all the unqualified products, and cannot give the cause of only one product. The number of unqualified products is always smaller than the normal products, so there is an unbalanced classification problem. In the future the feature selection method should be improved considering the unbalance.

Acknowledgements

This research is supported by the National Key Technology R&D Program of China (Grant no. 2015BAF30B01), the Open Foundation of the State Key Laboratory of rolling and automation, Northeastern University (Grant no. 2018RALKFKT003), the USTB-NTUT Joint Research Program (Grant no. TW2019013), and CCTEG Science and Technology Innovation Fund (2018-TD-ZD006, 2018-TD-QN021).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research is supported by the National Key Technology Research and Develpment Project of China (Grant no. 2015BAF30B01), the Open Foundation of the State Key Laboratory of rolling and automation, Northeastern University (Grant no. 2018RALKFKT003), the USTB-NTUT Joint Research Program (Grant no. TW2019013), and CCTEG Science and Technology Innovation Fund (2018-TD-ZD006, 2018-TD-QN021).

References

Aldehim, G., & Wang, W. (2015). Determining appropriate approaches for using data in feature selection. *International Journal of Machine Learning and Cybernetics*, *8*, 1–14. https://doi.org/10.1007/s13042-015-0469-8

Bissessur, Y., Martin, E. B., Morris, A. J., & Kitson, P. (2000). Fault detection in hot steel rolling using neural networks and multivariate statistics. *IEE Proceedings of Control Theory & Applications*, *147*(6), 633–640. https://doi.org/10.1049/ip-cta:20000763

Chen, Y. W., & Lin, C. J. (2006). Combining SVMs with various feature selection strategies. *Feature Extraction Studies in Fuzziness and Soft Computing*, *207*, 315–324. https://doi.org/10.1007/978-3-540-35488-8_13

Cor, A., Ambroise, C., & Cocquerez, J. P. (2006). Feature selection in robust clustering based on Laplace mixture. *Pattern Recognition Letters*, *27*(6), 627–635. https://doi.org/10.1016/j.patrec.2005.09.028

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics & Computational Biology*, *3*(2), 185–205. https://doi.org/10.1142/S0219720005001004

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Wiley.

Eiji, N., Takeshi, H., Hiroyuki, K., Yusuke, M., & Hideo, M. (2014). Process metallurgy analyses to design a high-bendability and high-spring back property sheet by using two-scale finite element method. *International Journal of Mechanical Sciences*, *87*, 89–101. https://doi.org/10.1016/j.ijmecsci.2014.06.001

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933–969. doi: 10.1162/jmlr.2003.4.6.933

Gunes, S., Polat, K., & Yosunkaya, S. (2010). Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Systems with Applications*, *37*(12), 7922–7928. https://doi.org/10.1016/j.eswa.2010.04.043

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. Wiley.

Kemal, P., & Volkan, K. (2011). Determining of gas type in counter flow vortex tube using pairwise fisher score attribute reduction method. *International Journal of Refrigeration*, *34*(6), 1372–1286. https://doi.org/10.1016/j.ijrefrig.2011.05.010

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *Proceedings of the Nineth International Conference on Machine Learning*, 249–256. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Proceedings of European Conference Machine Learning*, 171–182. doi:10.1.1.51.6297

Lester, E. A. (1973). *Metals Handbook ninth edition vol.4*. American Society for Metals.

Liu, X., Li, K., McAfee, M., & Irwin, G. W. (2011). Improved nonlinear PCA for process monitoring using support vector data description. *Journal of Process Control*, *21*(9), 1306–1317. https://doi.org/10.1016/j.jprocont.2011.07.003

Mertin, C., Stellmacher, T., Schmitz, T., & Hirt, G. (2019). Enhanced springback prediction for bending of high-strength spring steel using material data from an inverse modelling approach. *Procedia Manufacturing*, *29*, 153–160. https://doi.org/10.1016/j.promfg.2019.02.120

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis & Machine Intelligence IEEE Transactions on*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Verlag.

Samuel, R. T., & Cao, Y. (2016). Nonlinear process fault detection and identification using kernel PCA and kernel density estimation. *Systems Science & Control Engineering*, *4*(1), 165–174. https://doi.org/10.1080/21642583.2016.1198940

Shang, W. Q., Huang, H. K., Liu, Y. L., & Lin, Y. M. (2006). Research on the algorithm of feature selection based on Gini index for text categorization. *Computer Research and Development*, *43*(10), 1688–1694. https://doi.org/10.1360/crad20061002

Sharma, A., Paliwal, K. K., Imoto, S., & Miyano, S. (2013). Principal component analysis using QR decomposition. *International Journal of Machine Learning and Cybernetics*, *4*(6), 679–683. https://doi.org/10.1007/s13042-012-0131-7

Sylvain, V., Teodor, T., & Abdessamad, K. (2008). Fault detection and identification with a new feature selection based on mutual information. *Journal of Process Control*, *18*(5), 479–490. https://doi.org/10.1016/j.jprocont.2007.08.003

Wang, C., & He, F. (2019). State clustering of the hot strip rolling process via kernel entropy component analysis and weighted cosine distance. *Entropy*, *21*(10), 1019. https://doi.org/10.3390/e21101019

Wang, J., Xu, J., Zhao, C. G., Peng, Y., & Wang, H. P. (2019). An ensemble feature selection method for high-dimensional data based on sort aggregation. *Systems Science & Control Engineering*, *7*(2), 32–39. https://doi.org/10.1080/21642583.2019.1620658

Zhu, W. D., & Lin, Y. M. (2013). Using Gini-index for feature weighting in text categorization. *Journal of Computational Information Systems*, *14*(9), 5819–5826.