Belief rule-based classification system: extension of FRBCS in belief functions framework

Lianmeng Jiao[a,b], Quan Pan[a,*], Thierry Denœux[b], Yan Liang[a], Xiaoxue Feng[a,c]

[a] School of Automation, Northwestern Polytechnical University, 710072 Xi'an, PR China
[b] UMR CNRS 7253, Heudiasyc, Université de Technologie de Compiègne, 60205 Compiègne, France
[c] UMR CNRS 6279, ICD-LM2S, Université de Technologie de Troyes, 10010 Troyes, France

Abstract

Among the computational intelligence techniques employed to solve classification problems, the fuzzy rule-based classification system (FRBCS) is a popular tool capable of building a linguistic model interpretable to users. However, it may face lack of accuracy in some complex applications, by the fact that the inflexibility of the concept of the linguistic variable imposes hard restrictions on the fuzzy rule structure. In this paper, we extend the fuzzy rule in FRBCS with a belief rule structure and develop a belief rule-based classification system (BRBCS) to address imprecise or incomplete information in complex classification problems. The two components of the proposed BRBCS, i.e., the belief rule base (BRB) and the belief reasoning method (BRM), are designed specifically by taking into account the pattern noise that existes in many real-world data sets. Four experiments based on benchmark data sets are carried out to evaluate the classification accuracy, robustness, interpretability and time complexity of the proposed method.

Keywords: Pattern classification, Fuzzy rule-based classification system, Belief rule-based classification system, Belief functions theory, Pattern noise

1. Introduction

The fuzzy rule-based classification system (FRBCS) [10, 24, 50] is a useful tool to address classification problems, and it has become a popular framework for classifier design. It is widely employed due to its capability of building a linguistic model interpretable to users and addressing both quantitative and qualitative information coming from expert knowledge, mathematical models or empirical measures [45]. The FRBCS has been successfully applied to different real world classification tasks, including, but not limited to, image processing [49], intrusion detection [52], fault classification [44], target recognition [55], and medical applications [2, 53].

However, on one hand, the FRBCS may face lack of accuracy when dealing with some complex applications, due to the inflexibility of the concept of the linguistic variable, which imposes hard restrictions on the fuzzy rule structure

*Corresponding author at: School of Automation, Northwestern Polytechnical University, Xi'an, 710072, PR China. Tel.: +86 29 88431307; fax: +86 29 88431306.

Email addresses: jiaolianmeng@mail.nwpu.edu.cn (Lianmeng Jiao), panquannpu@gmail.com (Quan Pan), thierry.denoeux@hds.utc.fr (Thierry Denœux), liangyan@nwpu.edu.cn (Yan Liang), fengxiaoxue@mail.nwpu.edu.cn (Xiaoxue Feng)

[4]. For example, when the input-output mapping varies in complexity within the space, homogeneous partitioning using linguistic variables for the input and output spaces becomes inefficient. Moreover, as the size of the rule base directly depends on the number of fuzzy partitions for each feature, the derivation of an accurate FRBCS causes the number of rules to rise significantly, which may make the system lose the capability of being interpretable to the users. As both the interpretability (depends on several factors, mainly the rule structure, the number of rules, the number of features, the number of fuzzy partitions, and the shape of the fuzzy sets [21]) and the accuracy are important in system modeling, two ways of improving the accuracy of an interpretable FRBCS have been developed. One is learning an optimized rule base with sophisticated methods, such as FURIA [23], PDFC [9], or FARC-HD [3]. The other, as our main focus in this paper, is changing the rule structure to make it more flexible to characterize the input-output mapping, for example through introducing probability distribution [11] or interval-valued fuzzy sets [45].

In fact, different types of uncertainty, such as fuzziness, imprecision and incompleteness, may coexist in real-world complex systems. The FRBCS, which is based on fuzzy sets theory [60], cannot effectively address imprecise or incomplete information in the modeling and reasoning processes. The belief functions theory, also called Dempster-Shafer theory, proposed and developed by Dempster [13] and Shafer [46] et al., has become one of the most powerful frameworks for uncertain modeling and reasoning. As fuzzy sets theory is well suited to dealing with fuzziness, and belief functions theory provides an ideal framework for handling imprecision and incompleteness, many researchers have investigated the relationship between fuzzy sets theory and belief functions theory and suggested different methods of integrating them [6, 7, 34, 56, 57]. Among these methods, Yang et al. [57] extended the fuzzy rule in belief functions theory and proposed a new knowledge representation scheme in a belief rule structure, which is capable of capturing fuzzy, imprecise, and incomplete causal relationships. The belief rule structure has been successfully applied in clinical risk assessment [31], inventory control [32], and new product development [51, 58].

This paper aims to extend the fuzzy rule in FRBCS with the belief rule structure developed in [57] for classification applications. Compared with the fuzzy rule, the consequence part of the belief rule is in a belief distribution form, which is more informative and can characterize the uncertain information (i.e., fuzziness, imprecision, and incompleteness) existing in the training set. In addition, feature weights are introduced in the belief rule to characterize the different degrees of importance of features to the consequence. Therefore, the belief rule is more suitable for modeling those complex classification problems with great uncertainty. Based on the belief rule structure, a belief rule-based classification system (BRBCS) is developed as an extension of FRBCS in belief functions framework. In the proposed BRBCS, a data-driven belief rule base (BRB) generation method is developed to establish the uncertain association between the feature space and the class space. This BRB generation method enables the automatic generation of belief rules from the training data without the requirement of a priori expert knowledge. Then, to classify a query pattern based on the BRB, a belief reasoning method (BRM) is developed based on belief functions theory. This BRM can well address the uncertainty existing in the consequences of activated belief rules for a query pattern.

To handle the pattern noise commonly existing in many real-world data sets, two techniques are developed in the BRB generation and BRM design processes. First, the consequence part of each belief rule in BRB is generated by fusing the information coming from all of the training patterns assigned to the corresponding antecedent fuzzy region. In this way, the adverse effects of the noisy training patterns on the consequence of the belief rule can be reduced to some extent. Furthermore, in BRM, the final consequent class of a query pattern is obtained by combining the consequence parts of all of the belief rules activated by the query pattern. Thus, even if some unreliable belief rules are generated in noisy conditions, this procedure can further reduce the risk of misclassification.

Using belief functions theory to solve classification problems is not new, and some proposals have been introduced [15, 16, 18, 36, 37]. In [15], an evidential K-nearest neighbor (EK-NN) classification rule was proposed by extending the classical K-NN rule within the framework of belief functions theory. Liu et al. [36, 37] further extended the EK-NN method considering the belief assigned to meta-classes. These methods mainly focus on the uncertainty modeling and reasoning in K-nearest neighbor classification, whereas the BRBCS proposed in this paper aims to use the belief functions theory in the rule-based classification system, so these two schemes are different in nature. Recently, Chen et al. [8] started to study the possibility of using belief functions theory in the rule-based classification system, but although the belief rule structure was introduced, the authors did not provide a method to derive the belief rule base from the data sets, and only the traditional fuzzy rule base was used in the experiment for illustration. Later, a belief rule base representation, generation and inference method was proposed in [33], where each training sample was developed as a rule with belief degrees embedded in the antecedent terms to model the input-output relationship. Although it has a similar theoretical foundation, the BRBCS proposed in our paper is developed in a quite different way. First, we use the rule structure considering belief distribution in the consequence part to well characterize the incompleteness induced by the limited training patterns. Second, we construct a belief rule associated with each activated fuzzy region by fusing all of the training samples assigned to this fuzzy region. In this manner, the constructed belief rules are more reliable, and the number of rules can be reduced greatly.

The rest of this paper is organized as follows. In Section 2, the basics of the fuzzy rule-based classification system and belief functions theory are introduced. The belief rule-based classification system (BRBCS) is developed in Section 3, and then four experiments are performed to evaluate the accuracy, robustness, interpretability and time complexity of the proposed BRBCS in Section 4. At last, Section 5 concludes the paper.

2. Preliminaries: fuzzy rule-based classification system and belief functions theory

In this section, we provide some preliminaries of the fuzzy rule-based classification system (FRBCS) and the belief functions theory. In Section 2.1, we describe the classical fuzzy rule structure followed by one of the most popular FRBCSs developed by Chi et al. [10]. In Section 2.2, we introduce some basic concepts of belief functions theory that are necessary for developing the belief rule-based classification system.

2.1. Fuzzy rule-based classification system

A fuzzy rule-based classification system (FRBCS) is composed of two main conceptual components, the fuzzy rule base (FRB) and the fuzzy reasoning method (FRM). The FRB establishes an association between the space of pattern features and the space of consequent classes. The FRM provides a mechanism to classify a query pattern based on the FRB.

The fuzzy rule in the FRB for an M-class (denoted as $\Omega \triangleq \{\omega_1, \omega_2, \cdots, \omega_M\}$) pattern classification problem with P features has the following structure [10]:

$$\begin{aligned} \text{Fuzzy Rule } R^q: \quad & \text{If } x_1 \text{ is } A_1^q \text{ and } \cdots \text{ and } x_P \text{ is } A_P^q, \text{ then the consequence is } C^q \\ & \text{with rule weight } \theta^q, \quad q = 1, 2, \cdots, Q, \end{aligned} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_P)$ is the pattern feature vector and $\mathbf{A}^q = (A_1^q, \cdots, A_P^q)$ is the antecedent part, with each A_p^q belonging to fuzzy partitions $\{A_{p,1}, A_{p,2}, \cdots, A_{p,n_p}\}$ associated with the p-th feature. $C^q \in \Omega$ is the label of the consequent class, and Q is the number of fuzzy rules in the FBR. The rule weight θ^q, characterizing the certainty grade of the fuzzy rule R^q, is used as the strength of R^q in fuzzy reasoning.

Based on the above fuzzy rule structure, several FRB generation methods have been proposed [9, 10, 25]. Here, we introduce the one proposed by Chi et al. [10] because it is one of the most widely used algorithms. To generate the FRB, this method uses the following steps:

1. *Establishment of the fuzzy regions.* Usually, the partition of the pattern space is related to the specific classification problem. If no prior knowledge is available, the method based on simple fuzzy grids is usually employed [24]. Fig. 1 shows an example of the fuzzy partition of a two-dimensional pattern space with triangular fuzzy sets. Based on this method, once the domain interval and the partition number for each feature are determined, the fuzzy regions are easily computed.

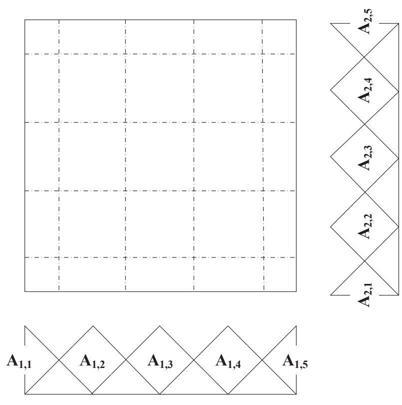

Figure 1: An example of the fuzzy partition of a two-dimensional pattern space by simple fuzzy grids

2. *Generation of a fuzzy rule for each training pattern.* Assume that N labeled P-dimensional training patterns $\mathbf{x}_i = (x_{i1}, \cdots, x_{iP})$, $i = 1, 2, \cdots, N$ are available. For each training pattern $\mathbf{x}_i$, the following steps are necessary
   (a) To calculate its matching degree $\mu(\mathbf{x}_i)$ with different fuzzy regions using the product operator

$$\mu_{\mathbf{A}^q}(\mathbf{x}_i) = \prod_{p=1}^{P} \mu_{A_p^q}(x_{ip}), \tag{2}$$

where $\mu_{A_p^q}(\cdot)$ is the membership function of the fuzzy set A_p^q,

(b) To assign the training pattern $\mathbf{x}_i$ to the fuzzy region with the greatest matching degree,

(c) To generate a rule for this training pattern, with the antecedent part determined by the selected fuzzy region and the consequent class equal to the class label of the training pattern, and

(d) To compute the rule weight θ^q.

Rules with the same antecedent part may be generated during the learning process. In this case, only the one having the maximum rule weight is maintained, whereas the remainder are removed.

Once the FRB is constructed, the query patterns can be classified by the following single winner FRM. Let S be the set of Q constructed fuzzy rules. A query pattern $\mathbf{y} = (y_1, y_2, \cdots, y_P)$ is classified by a single winner rule R^w, which is chosen from rule set S as

$$R^w = \arg\max_{R^q \in S} \{\mu_{\mathbf{A}^q}(\mathbf{y}) \cdot \theta^q\}. \tag{3}$$

That is, the winner rule R^w has the maximum product of the matching degree $\mu_{\mathbf{A}^q}(\mathbf{y})$ and the rule weight θ^q in S. Because it is limited by the number of training patterns, the query pattern may not always be covered by any rule in the FRB, in which case the classification is rejected. To avoid the non-covering problem, several techniques have been proposed, such as using bell-shaped fuzzy sets instead of the triangular fuzzy sets [39] or stretching a rule by deleting one or more of its antecedent terms [23].

2.2. *Basics of belief functions theory*

In belief functions theory, a problem domain is represented by a finite set $\Theta = \{\theta_1, \theta_2, \cdots, \theta_n\}$ of mutually exclusive and exhaustive hypotheses called the *frame of discernment*. A *mass function* or *basic belief assignment* (BBA) expressing the belief committed to the elements of 2^Θ by a given source of evidence is a mapping function $\mathrm{m}(\cdot)$: $2^\Theta \rightarrow [0, 1]$, such that

$$\mathrm{m}(\emptyset) = 0 \text{ and } \sum_{A \in 2^\Theta} \mathrm{m}(A) = 1. \tag{4}$$

Elements $A \in 2^\Theta$ having $\mathrm{m}(A) > 0$ are called the *focal elements* of the BBA $\mathrm{m}(\cdot)$. The BBA $\mathrm{m}(A)$ measures the degree of belief exactly assigned to a proposition A and represents how strongly the proposition is supported by evidence. The belief assigned to Θ, or $\mathrm{m}(\Theta)$, is referred to as the degree of *global ignorance*. The probability assigned to any subset of 2^Θ, except for any individual proposition θ_i ($i = 1, \cdots, n$) and Θ, is referred to as the degree of *local ignorance*. If there is no local or global ignorance, a belief function reduces to a conventional probability function, and the BBA $\mathrm{m}(\cdot)$ is said to be *Bayesian*.

Shafer [46] also defines the *belief function* and *plausibility function* of $A \in 2^\Theta$ as follows

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} \mathrm{m}(B) \text{ and } \mathrm{Pl}(A) = \sum_{B \cap A \neq \emptyset} \mathrm{m}(B). \tag{5}$$

Bel(A) represents the exact support to A and its subsets, and Pl(A) represents the total possible support to A and its subsets. The interval $[\mathrm{Bel}(A), \mathrm{Pl}(A)]$ can be seen as the lower and upper bounds of support to A. The belief functions m(·), Bel(·) and Pl(·) are in one-to-one correspondence.

For decision-making support, Smets [47] proposed the *pignistic probability* BetP(A) (from the Latin word *pignus*, meaning a bet) to approximate the unknown probability in $[\mathrm{Bel}(A), \mathrm{Pl}(A)]$ as follows

$$\mathrm{BetP}(A) = \sum_{\substack{B\subseteq\Theta \\ A\cap B\neq\emptyset}} \frac{|A\cap B|}{|B|}\mathrm{m}(B), \tag{6}$$

where $|X|$ is the cardinality of set X.

Two useful operations in the manipulation of belief functions are *Shafer's discounting operation* and *Dempster's rule of combination*. The discounting operation is used in the case that a source of evidence provides a BBA m(·), but one knows that this source has probability $\alpha \in [0, 1]$ of reliability. Then, one may adopt $(1 - \alpha)$ as the discount rate, which results in a new BBA $\mathrm{m}^{\alpha}(\cdot)$ defined by

$$\mathrm{m}^{\alpha}(A) = \begin{cases} \alpha\mathrm{m}(A), & \text{for } A \neq \Theta \\ \alpha\mathrm{m}(\Theta) + (1-\alpha), & \text{for } A = \Theta. \end{cases} \tag{7}$$

Several distinct bodies of evidence characterized by different BBAs can be combined using Dempster's rule. Mathematically, Dempster's rule of combination of two BBAs $\mathrm{m}_1(\cdot)$ and $\mathrm{m}_2(\cdot)$ defined on the same frame of discernment Θ is

$$\mathrm{m}(A) = \begin{cases} 0, & \text{for } A = \emptyset \\ \frac{\sum_{B\cap C=A} \mathrm{m}_1(B)\mathrm{m}_2(C)}{1-\sum_{B\cap C=\emptyset} \mathrm{m}_1(B)\mathrm{m}_2(C)}, & \text{for } A \in 2^{\Theta} \text{ and } A \neq \emptyset. \end{cases} \tag{8}$$

In Dempster's rule, the total conflicting belief mass $\sum_{B\cap C=\emptyset} \mathrm{m}_1(B)\mathrm{m}_2(C)$ is redistributed back to all of the focal elements through normalization.

Sometimes, the dissimilarity measure between two sources of evidence is needed, which is characterized by the distance between two BBAs. The choice for a well-adapted distance is not easy, and many distances have been defined, as shown in [30]. Here, we present the definition of Jousselme's distance d_J [29], which is one of the most commonly used distances. The d_J between two BBAs $\mathrm{m}_1(\cdot) \triangleq \mathbf{m}_1$ and $\mathrm{m}_2(\cdot) \triangleq \mathbf{m}_2$ defined on the same frame of discernment Θ is

$$d_J(\mathbf{m}_1, \mathbf{m}_2) = \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)\mathbf{D}(\mathbf{m}_1 - \mathbf{m}_2)^T}, \tag{9}$$

where $\mathbf{D}$ is a $2^{|\Theta|} \times 2^{|\Theta|}$ matrix with elements given by $D_{i,j} = \frac{|A_i\cap B_j|}{|A_i\cup B_j|}$, $A_i, B_j \in 2^{\Theta}$.

To interpret the belief functions theory in a more obvious way and to demonstrate its capacity for representing and reasoning with uncertain information, we provide the following murder example [17] for illustration.

Example 1. Suppose a murder has been committed. There are three suspects: *Peter*, *John*, and *Mary*. In the belief functions framework, the set $\Theta = \{Peter, John, Mary\}$ can be seen as the frame of discernment of the considered problem.

Suppose a witness saw the murderer going away, but he is short-sighted, and he only saw that it was a man. Based on this evidence, the following BBA can be constructed for the murderer

$$\mathrm{m}_1(\{Peter, John\}) = 1. \tag{10}$$

Further, we know that the witness is drunk 20% of the time. This means that the above evidence holds with probability $\alpha = 0.8$. Thus, Shafer's discounting operation in Eq. (7) can be used to obtain the corresponding BBA as

$$\mathrm{m}_1^{\alpha}(\{Peter, John\}) = 0.8, \quad \mathrm{m}_1^{\alpha}(\Theta) = 0.2. \tag{11}$$

Suppose when investigating the scene of the crime, a blond hair has been found, and we know there is a probability $\alpha = 0.6$ that the room has been cleaned before the crime. In a similar way, we can obtain the second BBA as

$$\mathrm{m}_2^{\alpha}(\{John, Mary\}) = 0.6, \quad \mathrm{m}_2^{\alpha}(\Theta) = 0.4. \tag{12}$$

With the above two distinct pieces of evidence, Dempster's rule of combination can be used to fuse them into a final BBA as

$$\mathrm{m}(\{John\}) = 0.48, \quad \mathrm{m}(\{Peter, John\}) = 0.32, \quad \mathrm{m}(\{John, Mary\}) = 0.12, \quad \mathrm{m}(\Theta) = 0.08. \tag{13}$$

To judge who is most likely to be the murderer based on the above final BBA, we compute the belief $\mathrm{Bel}(\cdot)$, the plausibility $\mathrm{Pl}(\cdot)$ and the pignistic probability $\mathrm{BetP}(\cdot)$ for each suspect using Eqs.(5,6), as shown in Table 1. The belief $\mathrm{Bel}(\cdot)$ and the plausibility $\mathrm{Pl}(\cdot)$ for each suspect provide its lower and upper probabilities to be the murderer, respectively. The pignistic probability $\mathrm{BetP}(\cdot)$ provides an approximate estimation between the lower and upper probabilities. For this example the same decision is made by maximizing $\mathrm{Bel}(\cdot)$, $\mathrm{Pl}(\cdot)$ and $\mathrm{BetP}(\cdot)$: John is most likely to be the murderer.

Table 1: The belief, plausibility and pignistic probability with regard to each suspect

Suspects	Bel	Pl	BetP
Peter	0	0.4	0.19
John	0.48	1	0.73
Mary	0	0.2	0.09

From the above example, it can be seen that the belief functions theory can adequately deal with uncertainty induced by partial evidence. As a generalized probability distribution, the mass function can represent local ignorance (such as the belief assigned to $\{John, Mary\}$ and $\{Peter, John\}$) as well as global ignorance (the belief assigned to the frame of discernment Θ). In addition, belief functions theory provides effective tools (such as Dempster's rule of combination) for uncertain reasoning at the credal level using the formalism of mass function. At last, the final mass function is transformed into a probability measure (pignistic probability) to support the decision making at the pignistic level.

3. Belief rule-based classification system (BRBCS)

Considering the advantages of belief functions theory for representing and reasoning with uncertain information, in this section we extend the classical FRBCS in belief functions framework and develop the belief rule-based classification system (BRBCS). As in Fig. 2, the proposed BRBCS is composed of two components: the belief rule base (BRB) that establishes an association between the feature space and the class space and the belief reasoning method (BRM) that provides a mechanism to classify a query pattern based on the BRB. In Section 3.1, we first describe the belief rule structure for classification applications, which extends the traditional fuzzy rule structure in belief functions framework. Based on the belief rule structure, we learn the belief rule base from the training patterns in Section 3.2, and then the belief reasoning method is developed in Section 3.3.

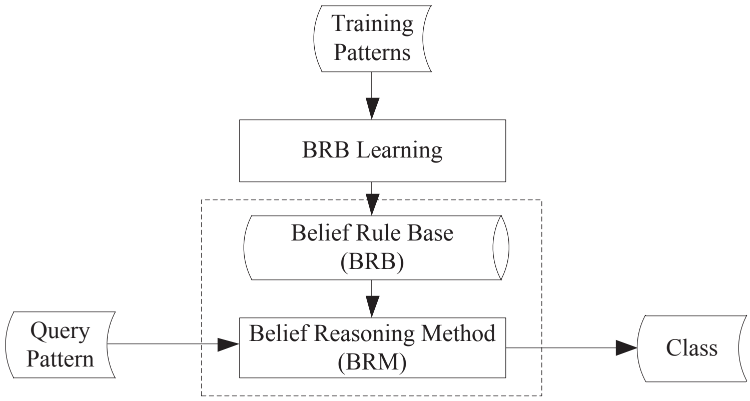

Figure 2: Belief rule-based classification system

3.1. Belief rule structure

The fuzzy rule structure expressed in Eq. (1) is relatively simple in that it does not consider the distribution of consequence and the relative importance of each feature. To take the above aspects into consideration, two concepts are introduced [57]:

- *Belief degrees of consequence*. For a complex classification problem, it is likely that the consequence of a rule may take a few values with different belief degrees. Suppose the consequence may have M different classes, $\omega_1, \omega_2, \cdots, \omega_M$, and the corresponding belief degrees are represented by $\beta_i (i = 1, 2, \cdots, M)$, then the consequence with a belief structure can be represented by $\{(\omega_1, \beta_1), (\omega_2, \beta_2), \cdots, (\omega_M, \beta_M)\}$.

- *Feature weights*. As indicated in [26, 27], in real-world classification problems, different features may behave distinctly in determining the consequent class. Thus, there is a need to assign a weight to each feature to describe such degrees of importance.

To take into account the belief degrees of consequence and the feature weights, the fuzzy rule structure expressed in Eq. (1) can be extended to the following belief rule structure for classification purposes:

$$\begin{aligned}\text{Belief Rule } R^q: \quad & \text{If } x_1 \text{ is } A_1^q \text{ and } \cdots \text{ and } x_P \text{ is } A_P^q, \text{ then the consequence is } \mathbf{C}^q = \{(\omega_1, \beta_1^q), \cdots, (\omega_M, \beta_M^q)\} \\ & \text{with rule weight } \theta^q \text{ and feature weights } \delta_1, \cdots, \delta_P, \quad q = 1, 2, \cdots, Q,\end{aligned} \tag{14}$$

where β_m^q is the belief degree to which ω_m is believed to be the consequent class for the q-th belief rule. In the belief structure, the consequence may be incomplete, i.e., $\sum_{m=1}^{M} \beta_m^q < 1$, and the left belief $1 - \sum_{m=1}^{M} \beta_m^q$ denotes the degree of global ignorance about the class label. The rule weight θ^q with $0 \leq \theta^q \leq 1$ characterizes the certainty grade of the belief rule R^q and the feature weights $\delta_1, \cdots, \delta_P$ with $0 \leq \delta_1, \cdots, \delta_P \leq 1$ describe the importance of different features in determining the consequent class.

Remark 1. Compared with the fuzzy rule structure, the belief rule structure has some advantages for classification problems as follows. a) In the belief rule structure, the consequence is in a belief distribution form. On the one hand, with the distribution form, any difference in the antecedent part can be clearly reflected in the consequence, whereas in a traditional fuzzy rule, different antecedents may lead to the same consequence. On the other hand, by introducing belief functions theory, the belief structure makes the rule more appropriate to characterize the uncertain information. In generating each rule, only limited training patterns are available, and each training pattern only provides partial evidence about the consequence of this rule. Thus, the corresponding consequence of this rule should not be complete. The belief rule structure can well characterize this incompleteness, with the remained belief $1 - \sum_{m=1}^{M} \beta_m^q$ denoting the degree of global ignorance about the class label induced by the limited training patterns. b) With the introduction of feature weights, the importance of different features to the consequence can be well characterized, which is closer to reality. In summary, compared with the traditional fuzzy rule, the belief rule is more informative, more flexible and thus more suitable for modeling those complex classification problems.

3.2. Belief rule base (BRB) generation

To make a classification with BRBCS, the first step is to generate a BRB from the training set. The FRB generation method given by Chi et al. [10] is used as a base model in this section to develop the BRB generation method in belief functions framework. As displayed in Eq. (14), each belief rule is composed of four components, namely, the antecedent part, the belief degrees of consequence, the rule weight and the feature weights. Because the antecedent part in the belief rule is the same as that in the fuzzy rule, here we only focus on the generation of the latter three components, i.e., the belief degrees of consequence, the rule weight and the feature weights.

3.2.1. Generation of belief degrees of consequence

In this part, with the symbols defined in Section 2.1 and Section 3.1, we develop an algorithm to generate the belief degrees of consequence in BRB.

Similar to the generation of the consequent class in FRB in the first step, we also need to calculate the matching degree $\mu_{\mathbf{A}^q}(\mathbf{x}_i)$ of each training pattern $\mathbf{x}_i$ with the antecedent part $\mathbf{A}^q$ using Eq. (2). In FRB, the consequent class is directly specified as the class label of the training pattern having the greatest matching degree with the antecedent part $\mathbf{A}^q$. However, this procedure may entail great risk, especially when class noise exists in the training set. In BRB, we fuse the class information of all of the training patterns assigned to the corresponding antecedent fuzzy region to get the consequence in a belief distribution form.

Denote as T^q the set of training patterns assigned to the antecedent fuzzy region $\mathbf{A}^q$. From the view of belief functions theory, the class set $\Omega = \{\omega_1, \cdots, \omega_M\}$ can be regarded as the frame of discernment of the problem. For any training pattern $\mathbf{x}_i \in T^q$, the class label $\text{Class}(\mathbf{x}_i) = \omega_m$ can be regarded as a piece of evidence that increases the belief that the consequent class belongs to ω_m. However, this piece of evidence does not by itself provide 100% certainty. In belief functions theory, this can be expressed by saying that only some part of the belief (measured by the matching degree $\mu_{\mathbf{A}^q}(\mathbf{x}_i)$) is committed to ω_m. Because $\text{Class}(\mathbf{x}_i) = \omega_m$ does not point to any other particular class, the rest of the belief should be assigned to the frame of discernment Ω representing global ignorance. Therefore, this item of evidence can be represented by a BBA $\mathrm{m}_i^q(\cdot)$ verifying:

$$\begin{cases} \mathrm{m}_i^q(\{\omega_m\}) &= \mu_{\mathbf{A}^q}(\mathbf{x}_i) \\ \mathrm{m}_i^q(\Omega) &= 1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i) \\ \mathrm{m}_i^q(A) &= 0, \quad \forall A \in 2^\Omega \setminus \{\Omega, \{\omega_m\}\}, \end{cases} \tag{15}$$

with $0 < \mu_{\mathbf{A}^q}(\mathbf{x}_i) \leq 1$.

For each $\mathbf{x}_i \in T^q$, a BBA depending on both its class label and its matching degree with the antecedent part can therefore be defined. To obtain the consequence associated with the antecedent part $\mathbf{A}^q$ in a belief distribution form, these BBAs can be combined using Dempster's rule. As shown in Eq. (15), only two focal elements are involved in each BBA. Because of the particular structure of the BBA, the computational burden of Dempster's rule can be greatly reduced, and the analytical formulas can be derived as

$$\begin{aligned} \mathrm{m}^q(\{\omega_m\}) &= \frac{1}{1-K^q}\left(1 - \prod_{\mathbf{x}_i \in T_m^q}(1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i))\right) \cdot \prod_{r \neq m} \prod_{\mathbf{x}_i \in T_r^q}(1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i)), \\ & \quad m = 1, 2, \cdots, M, \\ \mathrm{m}^q(\Omega) &= \frac{1}{1-K^q} \prod_{r=1}^{M} \prod_{\mathbf{x}_i \in T_r^q}(1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i)), \end{aligned} \tag{16}$$

where T_m^q is a subset of T^q, corresponding to those training patterns belong to class ω_m, and K^q is the total conflicting belief mass

$$K^q = 1 + (M-1) \prod_{r=1}^{M} \prod_{\mathbf{x}_i \in T_r^q}(1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i)) - \sum_{m=1}^{M} \prod_{r \neq m} \prod_{\mathbf{x}_i \in T_r^q}(1 - \mu_{\mathbf{A}^q}(\mathbf{x}_i)). \tag{17}$$

Therefore, the belief degrees of consequence of rule R^q can be given by

$$\begin{aligned} \beta_m^q &= \mathrm{m}^q(\{\omega_m\}), \quad m = 1, 2, \cdots, M, \\ \beta_\Omega^q &= \mathrm{m}^q(\Omega), \end{aligned} \tag{18}$$

where β_Ω^q is the belief degree unassigned to any individual class.

Remark 2. In the classical FRBCS reviewed in Section 2.1, the consequence of each rule is only determined by the class label of the training pattern having the greatest matching degree with the antecedent part, whereas the consequence in the belief rule fuses information that comes from all of the training patterns assigned to the corresponding antecedent fuzzy region. Thus, it can effectively reduce the adverse effects of some noisy training patterns. The method to generate the consequence is similar to some data-cleaning approaches [5, 12]. The difference is that the data-cleaning approaches remove the unreliable training samples, whereas our method retains all training samples and generates the consequence in a belief distribution form, which can be considered as soft labels. Compared with the data cleaning approaches, the belief distribution form maintains more information from the training samples and can be further combined with the consequences of other rules in later processing.

Remark 3. The idea to generate the belief degrees of consequence in this paper is inspired by the EK-NN classification method developed by Denœux [15], in which each of the k nearest neighbors of the query sample is considered as an item of evidence that supports certain hypotheses regarding the class membership of that pattern. The corresponding relations of the two methods are as follows: a) the training patterns assigned to the corresponding antecedent fuzzy region correspond to the k nearest neighbors of the query pattern in EK-NN, and b) the importance of each training pattern is measured by the matching degree with the antecedent part, and in EK-NN, that is measured by the distance from the query pattern. With the above relationship, it can be further deduced that the consequence generation method in FRB revisited in Section 2.1 has a similar idea to the voting K-NN classification method [20]. As illustrated in [15], the EK-NN classifier can obtain much better performance than the voting K-NN classifier, especially in noisy conditions. Thus, it is expected that the belief distribution form of consequence in the BRB can handle the class noise more effectively than the single class form of consequence in the FRB.

3.2.2. *Generation of rule weights*

In the area of data mining, two measures called *confidence* and *support* have often been used for evaluating association rules [1]. Our belief rule R^q in Eq. (14) can be viewed as a type of association rule of the form $\mathbf{A}^q \Rightarrow \mathbf{C}^q$. The main difference from the standard formulation of the association rule is that in our belief if-then rule, the input variable is in fuzzy form and the output variable is in belief distribution form. In this part, we will draw the rule weight θ^q from the concepts of confidence and support.

The confidence is defined as a measure of the validity of one association rule [1]. For our belief if-then rule, the consequence part $\mathbf{C}^q$ is obtained by combining the items of evidence coming from all of the training patterns assigned to the antecedent fuzzy region $\mathbf{A}^q$. It is believed that if the items of evidence involved are in conflict with each other (for example, if the items of evidence assign different classes with the highest belief), then the consequence has low validity. In belief functions theory, several models are proposed to measure the conflict among different items of evidence [35, 48]. The conflict factor $\sum_{B \cap C = \emptyset} \mathrm{m}_1(B)\mathrm{m}_2(C)$ derived in Dempster's rule is employed here for its simplicity and convenience. The confidence of the belief rule R^q is hence defined as

$$c(R^q) = 1 - \overline{K^q}, \tag{19}$$

with the average conflict factor $0 \leq \overline{K^q} \leq 1$ calculated by

$$\overline{K^q} = \begin{cases} 0, & \text{if } |T^q| = 1, \\ \frac{1}{|T^q|(|T^q|-1)} \sum_{\mathbf{x}_i, \mathbf{x}_j \in T^q; i<j} \sum_{B \cap C = \emptyset} \mathrm{m}_i^q(B)\mathrm{m}_j^q(C), & \text{otherwise.} \end{cases} \tag{20}$$

where $|T^q|$ is the number of training patterns assigned to the fuzzy region $\mathbf{A}^q$.

On the other hand, as described in [1], the support indicates the grade of the coverage by one association rule. For our belief if-then rule, N training patterns are available for rule generation, while only those assigned to the corresponding antecedent fuzzy region are used to generate the consequence. Therefore, the support of the belief rule R^q is defined as

$$s(R^q) = \frac{|T^q|}{N}. \tag{21}$$

As defined above, the confidence and support characterize the weight of the belief rule in two distinct aspects and should therefore be considered jointly. On the one hand, if the belief rule R^q has high confidence but low support (for example, if only one training pattern is assigned to the antecedent fuzzy region $\mathbf{A}^q$), the belief rule weight should be decreased, as the consequence may be easily affected by the class noise. On the other hand, if the belief rule R^q has high support but low confidence (for example, if a large number of training patterns are contained in T^q but with great divergence in the class label), the belief rule weight should also be decreased, considering the great conflicts. The product of the confidence $c(R^q)$ and the support $s(R^q)$ is used to characterize the weight of the belief rule R^q as

$$\theta^q \propto c(R^q) \cdot s(R^q). \tag{22}$$

Following a normalization process, we obtain the weights of all of the belief rules as

$$\theta^q = \frac{c(R^q) \cdot s(R^q)}{\max_q \{c(R^q) \cdot s(R^q), q = 1, \cdots, Q\}}, \quad q = 1, 2, \cdots, Q. \tag{23}$$

3.2.3. Generation of feature weights

In the belief rule displayed as Eq. (14), the feature weights reflect the relative importance of the antecedent features with respect to their influence on the consequence of the rule. In other words, an antecedent feature with a higher weight is more influential on the consequence. Therefore, to determine a feature weight is to find a way to measure the relative intensity of the influence that this antecedent feature imposes on the consequence in comparison with others. In this part, such a measurement is quantified by a so-called correlation factor (CF) between each feature and the consequence.

Suppose n_p fuzzy partitions $\{A_{p,1}, A_{p,2}, \cdots, A_{p,n_p}\}$ are established for feature A_p. Now, we will derive the weight of feature A_p through correlation analysis with the corresponding consequence. Specifically, we use the relationship between the changes of different fuzzy partitions that A_p takes and the changes of the consequence to determine the correlation between A_p and the consequence.

As Q belief rules with different antecedent parts are available in the BRB, in the first place, for feature A_p, according to its n_p fuzzy partitions $\{A_{p,1}, A_{p,2}, \cdots, A_{p,n_p}\}$, we divide the BRB into n_p sub-BRBs B_k, $k = 1, 2, \cdots, n_p$, with each sub-BRB B_k containing all of the belief rules using the fuzzy partition $A_{p,k}$ for feature A_p:

$$B_k = \left\{R^q \mid A_p^q = A_{p,k}, q = 1, 2, \cdots, Q\right\}, \quad k = 1, 2, \cdots, n_p. \tag{24}$$

Then, for each sub-BRB B_k, the consequence parts of all of the contained belief rules are combined to obtain the integrated consequence $\mathrm{m}_k(\cdot)$ with the weighted averaging operation:

$$\begin{aligned} \mathrm{m}_k(\{\omega_m\}) &= \frac{1}{\sum\limits_{R^q \in B_k} \theta^q} \sum_{R^q \in B_k} \theta^q \cdot \beta_m^q, \quad m = 1, 2, \cdots, M, \\ \mathrm{m}_k(\{\Omega\}) &= \frac{1}{\sum\limits_{R^q \in B_k} \theta^q} \sum_{R^q \in B_k} \theta^q \cdot \beta_\Omega^q, \end{aligned} \tag{25}$$

where β_m^q, $m = 1, 2, \cdots, M$, and β_Ω^q are the belief degrees of consequence of rule R^q generated in Section 3.2.1, and θ^q is the rule weight of R^q generated in Section 3.2.2.

Thus, when A_p changes its fuzzy partition from $A_{p,k}$ to $A_{p,k+1}$, $k = 1, 2, \cdots, n_p - 1$, the change of the consequence is

$$\Delta C_{p,k} = d_J(\mathbf{m}_k, \mathbf{m}_{k+1}), \tag{26}$$

where $\mathbf{m}_k = \mathrm{m}_k(\cdot)$, and d_J is Jousselme's distance, as displayed in Eq. (9).

Then, the average change of the consequence for k changing from 1 to $n_p - 1$ is given by

$$\Delta C_p = \frac{\sum\limits_{k=1}^{n_p - 1} \Delta C_{p,k}}{n_p - 1}. \tag{27}$$

In this paper, we define ΔC_p as the correlation factor (CF) between feature A_p and the consequence, i.e.,

$$CF_p = \Delta C_p. \tag{28}$$

In a similar way, we obtain the correlation factors $CF_p, p = 1, 2, \cdots, P$ for all features. Further, δ_p, the weight of feature A_p, can be generated from the normalized CF_p as follows

$$\delta_p = \frac{CF_p}{\max\limits_p \{CF_p, p = 1, 2, \cdots, P\}}, \quad p = 1, 2, \cdots, P. \tag{29}$$

3.3. Belief reasoning method (BRM)

As revisited in Section 2.1, in FRBCS, the single winner FRM is used to classify a new query pattern. However, when excessive noise exists in the training set, this method may have a great risk of misclassification. In this section, we will fuse the consequences of all of the rules activated by the query pattern within the framework of belief functions theory to get a more robust classification. The main idea is firstly calculating the association degrees of the query pattern with the consequences of the activated belief rules and then combining these consequences with respect to their reliability (characterized by the association degrees) based on belief functions theory.

3.3.1. Association degree with the consequence of a belief rule

Denote $\mathbf{y} = (y_1, y_2, \cdots, y_P)$ as a query pattern to be classified. In the first place, the matching degree of the query pattern with the antecedent part of each rule is calculated. As the feature weights are complemented in the belief rule, we use the following simple weighted multiplicative aggregation function to calculate the matching degree

$$\mu_{\mathbf{A}^q}(\mathbf{y}) = \left[\prod_{p=1}^{P} \left[\mu_{A_p^q}(y_p)\right]^{\delta_p}\right]^{1/P}, \tag{30}$$

where $\mu_{A_p^q}(\cdot)$ is the membership function of the antecedent fuzzy set A_p^q, and δ_p is the weight of the p-th feature given in Eq. (29).

Remark 4. In Eq. (30), the contribution of a feature towards the matching degree is positively related to the weight of the feature. In other words, a more important feature plays a greater role in determining the matching degree. Particularly, if $\delta_p = 0$, then $\left[\mu_{A_p^q}(y_p)\right]^{\delta_p} = 1$, which shows that a feature with zero importance does not have any impact on the matching degree; if $\delta_p = 1$, then $\left[\mu_{A_p^q}(y_p)\right]^{\delta_p} = \mu_{A_p^q}(y_p)$, which shows that the most important feature has the largest impact on the matching degree.

Let S be the set of Q constructed belief rules in the BRB. Denote as $S' \subseteq S$ the set of belief rules activated by query pattern $\mathbf{y}$:

$$S' = \{R^q \mid \mu_{\mathbf{A}^q}(\mathbf{y}) \neq 0, q = 1, 2, \cdots, Q\}. \tag{31}$$

The association degree of query pattern $\mathbf{y}$ with the consequence of one activated belief rule $R^q \in S'$ is determined by two factors, the matching degree and the rule weight. The matching degree reflects the similarity between the query pattern and the antecedent part of the belief rule, while the rule weight characterizes the reliability of the belief rule. Thus, the association degree is defined as

$$\alpha^q = \mu_{\mathbf{A}^q}(\mathbf{y}) \cdot \theta^q, \quad \text{for } R^q \in S'. \tag{32}$$

Remark 5. As a result of limitation due to the number of training patterns, in some applications, there may be no rule activated by query pattern $\mathbf{y}$. In such a case, we classify the non-covered query pattern based on the generated rule which has the nearest distance with it. For a non-covered query pattern $\mathbf{y}$, we first find the fuzzy region $\mathbf{A}^* = (A_1^*, A_2^*, \cdots, A_P^*)$ that has the greatest matching degree with it. The distance between a non-covered query pattern $\mathbf{y}$ and one generated rule R^q is defined as $d(\mathbf{y}, R^q) = \|\mathbf{A}^* - \mathbf{A}^q\|_2$, where $\mathbf{A}^q$ is the antecedent fuzzy region of rule R^q. Further, if several generated rules have the same nearest distance to the non-covered query pattern, the consequences of these rules are combined using belief functions theory in the following section, considering the rule weight as the reliability factor α.

3.3.2. Reasoning using belief functions theory

In the previous part, the association degrees of query pattern $\mathbf{y}$ with the consequences of the activated belief rules are calculated. In [57], an evidential reasoning (ER) method was used to combine different rules, considering their association degrees (or weights). As the ER method was first proposed for multi-attribute decision making problems [59], it considers the weights as the importance factors representing the decision maker's subjective preference in the fusion process. However, for pattern classification problems, the weights should be considered as the reliability factors representing the rules' capability to provide a correct assessment of the input query pattern [28]. Therefore, the ER method is not reasonable for the pattern classification problems, and in the following, a belief reasoning method (BRM) is developed to combine the belief rules activated by the query pattern.

In belief functions theory, Shafer's discounting operation, displayed as Eq. (7), is usually used to discount the unreliable evidence before combination. Regarding association degree α in Eq. (32) as the reliability factor, the consequence of one activated belief rule in Eq. (18) is discounted using Shafer's discounting operation as

$$\begin{aligned} \mathrm{m}^{\alpha}(\{\omega_m\}) &= \alpha \cdot \beta_m, \quad m = 1, 2, \cdots, M, \\ \mathrm{m}^{\alpha}(\Omega) &= \alpha \cdot \beta_{\Omega} + (1 - \alpha) \end{aligned} \tag{33}$$

For all of the $|S'| = L$ activated belief rules, through the above formula, we can get the corresponding discounted consequences $\mathrm{m}_i^{\alpha}(\cdot)$, $i = 1, 2, \cdots, L$.

To make a decision regarding the discounted consequences of activated belief rules, the corresponding BBAs can be combined using Dempster's rule. However, as indicated in [19, 54], the direct use of Dempster's rule will result in an exponential increase in computational complexity for the reason of enumerating all subsets or supersets of a given subset A of Ω, and the operation becomes impractical when the frame of discernment has more than 15 to 20 elements. The following part is intended to develop an operational algorithm for evidence combination with linear computational complexity, considering the fact that the focal elements of each associated BBA are all singletons except the ignorance set Ω.

Define $I(i)$ as the index set of the former i BBAs. Let $\mathrm{m}_{I(i)}(\cdot)$ be the BBA after combining all of the former i BBAs associated with $I(i)$. Given the above definitions, a recursive evidence combination algorithm can be developed as follows

$$\begin{aligned} \mathrm{m}_{I(i+1)}(\{\omega_q\}) &= K_{I(i+1)} \Big[\mathrm{m}_{I(i)}(\{\omega_q\}) \cdot \mathrm{m}_{i+1}^{\alpha}(\{\omega_q\}) + \mathrm{m}_{I(i)}(\Omega) \cdot \mathrm{m}_{i+1}^{\alpha}(\{\omega_q\}) \\ &\quad + \mathrm{m}_{I(i)}(\{\omega_q\}) \cdot \mathrm{m}_{i+1}^{\alpha}(\Omega) \Big], \quad q = 1, 2, \cdots, M \\ \mathrm{m}_{I(i+1)}(\Omega) &= K_{I(i+1)} \Big[\mathrm{m}_{I(i)}(\Omega) \cdot \mathrm{m}_{i+1}^{\alpha}(\Omega) \Big] \\ K_{I(i+1)} &= \left[1 - \sum_{j=1}^{M} \sum_{p=1, p \neq j}^{M} \mathrm{m}_{I(i)}(\{\omega_j\}) \cdot \mathrm{m}_{i+1}^{\alpha}(\{\omega_p\}) \right]^{-1} \\ & \quad i = 1, 2, \cdots, L-1, \end{aligned} \tag{34}$$

where $K_{I(i+1)}$ is a normalizing factor, so that $\sum_{q=1}^{M} \mathrm{m}_{I(i+1)}(\{\omega_q\}) + \mathrm{m}_{I(i+1)}(\Omega) = 1$.

Note that $\mathrm{m}_{I(1)}(\{\omega_q\}) = \mathrm{m}_1^{\alpha}(\{\omega_q\})$ for $q = 1, 2, \cdots, M$ and $\mathrm{m}_{I(1)}(\Omega) = \mathrm{m}_1^{\alpha}(\Omega)$. Thus, this recursive evidence combination algorithm can initiate with the first BBA. Accordingly, as the recursive index i reaches $L-1$, the final results $\mathrm{m}_{I(L)}(\{\omega_q\})$ and $\mathrm{m}_{I(L)}(\Omega)$ ($\mathrm{m}(\{\omega_q\})$ and $\mathrm{m}(\Omega)$ for short, respectively) are obtained by combining all of the L BBAs. This combination result is the basis for the later decision process.

For decision making based on the combined BBA $\mathrm{m}(\cdot)$ calculated with Eq. (34), the belief function $\mathrm{Bel}(\cdot)$, plausibility function Pl(·) and pignistic probability BetP(·) are common alternatives. As the focal elements of the combined BBA m(·) are all singletons except the ignorance set Ω, the credibility, plausibility and pignistic probability of each class ω_q are calculated as follows

$$\begin{array}{rcl} \mathrm{Bel}(\{\omega_q\}) & = & \mathrm{m}(\{\omega_q\}) \\ \mathrm{Pl}(\{\omega_q\}) & = & \mathrm{m}(\{\omega_q\}) + \mathrm{m}(\Omega) \\ \mathrm{BetP}(\{\omega_q\}) & = & \mathrm{m}(\{\omega_q\}) + \frac{\mathrm{m}(\Omega)}{M} \\ & & q = 1, 2, \cdots, M. \end{array} \tag{35}$$

It is supposed that based on this evidential body, a decision has to be made in assigning query pattern **y** to one of the classes in Ω. Because of the particular structure of the combined BBA (i.e., the focal elements are either singletons or the whole frame Ω), it can be easily discovered that

$$\begin{array}{rcl} \omega & = & \arg\max\limits_{\omega_q \in \Omega} \mathrm{Bel}(\{\omega_q\}) \\ & = & \arg\max\limits_{\omega_q \in \Omega} \mathrm{Pl}(\{\omega_q\}) \\ & = & \arg\max\limits_{\omega_q \in \Omega} \mathrm{BetP}(\{\omega_q\}) \\ & = & \arg\max\limits_{\omega_q \in \Omega} \mathrm{m}(\{\omega_q\}). \end{array} \tag{36}$$

That is, the strategies maximizing the three criteria Bel(·), Pl(·), and BetP(·) in Eq. (35) lead to the same decision: the pattern is assigned to the class with maximum basic belief assignment m(·).

Remark 6. For some classification applications under harsh working conditions (e.g., battlefield target recognition), significant noise may exist in the training set. Though the consequence generation method proposed in Section 3.2.1 can reduce the adverse effects from pattern noise, the consequence of one rule may still be unreliable in excessively noisy conditions. The BRM developed within the framework of belief functions theory combines the consequences of all of the activated rules to obtain the final consequent class. Therefore, compared with the single winner FRM, the BRM can further reduce the risk of misclassification.

4. Experiments

The performance of the proposed BRBCS will be empirically assessed through three different experiments with 20 real-world classification problems from the well-known UCI Repository of Machine Learning Databases [38]. In the first experiment, the original data sets are used to evaluate the classification accuracy of the proposed BRBCS. In the second one, the noise is added to the data sets artificially in controlled settings to evaluate the classification robustness of the proposed BRBCS in noisy training set conditions. In the latter two experiments, we will provide an analysis of its interpretability and time complexity, respectively.

4.1. Data sets and experiment conditions

Twenty well-known benchmark data sets from the UCI repository are selected to evaluate the performance of the BRBCS. The main characteristics of the 20 data sets are summarized in Table 2, where "# Instances" is the number of instances in the data set, “# Features” is the number of features, and “# Classes” is the number of classes. Notice that for the data sets *Cancer*, *Diabetes* and *Pima*, we have removed the instances with missing feature values.

Table 2: Description of the benchmark data sets employed in the study

Data set	# Instances	# Features	# Classes
Banknote	1,372	4	2
Breast	106	9	6
Cancer[a]	683	9	2
Diabetes[a]	393	8	2
Ecoli	336	7	8
Glass	214	9	6
Haberman	306	3	2
Iris	150	4	3
Knowledge	403	5	4
Letter	20,000	16	26
Liver	345	6	2
Magic	19,020	10	2
Pageblocks	5,473	10	4
Pima[a]	336	8	2
Satimage	6,435	36	6
Seeds	210	7	3
Transfusion	748	4	2
Vehicle	846	18	4
Vertebral	310	6	3
Yeast	1,484	8	10

[a]For the data sets containing missing values, instances with missing feature values are removed.

To develop the different experiments, we consider the *B-Fold Cross-Validation* (B-CV) model [41]. Each data set is divided into B blocks, with $B-1$ blocks as a training set and the remaining block as a test set. Therefore, each block is used exactly once as a test set. We use the 5-CV here, i.e., five random partitions of the original data set, with four of them (80%) as the training set and the remainder (20%) as the test set. For each data set, we consider the average results of the five partitions.

For the first, third and fourth experiments, the original data sets described above are used directly, while for the second, some additional processes are needed. As discussed in [42, 43, 61], the pattern noise in the data set can be distinguished into two categories: class noise and feature noise. The class noise, also known as labeling error, occurs when a sample is assigned to an incorrect class. It can be attributed to several causes, including subjectivity during the labeling process, data entry errors, or limitations of the equipped measure instrument. In contrast, the feature noise is used to refer to corruptions in the values of one or more features of samples in a data set, which is often encountered in harsh working conditions. With the above consideration, in the second experiment, we will manage the robustness evaluation under two types of noise scenarios, class noise and feature noise.

As the initial amount of noise present in the original data sets is unknown, we use manual mechanisms to independently add noise to each data set to control the noise level for comparison. Additionally, to observe how noise affects the accuracy of the classifiers, the noise is only added in the training sets, while the test sets remain unchanged. Based on the type of noise, as in [42], different schemes of noise introduction are designed as follows.

- *Introduction of class noise.* In this scheme, a class noise level of $x\%$ indicates that $x\%$ of the samples in the training set are mislabeled. The class labels of these samples are randomly changed to different ones within the domain of the class.
- *Introduction of feature noise.* In this scheme, a feature noise level of $x\%$ indicates that $x\%$ of the feature values in the training set are erroneous. The corrupted feature is assigned a random value between the minimum and maximum of the domain of that feature, following a uniform distribution.

To evaluate the performance of the difference methods, in the first experiment, the classification accuracy rate criterion is utilized. In the second experiment, apart from the classification accuracy rate under each level of induced noise, we also take into account the following *relative loss of accuracy* (RLA) to observe the form in which the accuracy of one algorithm is affected when increasing the level of noise with respect to the case without noise.

$$RLA_{x\%} = \frac{Acc_{0\%} - Acc_{x\%}}{Acc_{0\%}}, \tag{37}$$

where $RLA_{x\%}$ is the relative loss of accuracy at noise level $x\%$, $Acc_{0\%}$ is the classification accuracy in the test with the original data set, and $Acc_{x\%}$ is the classification accuracy when testing the data set with noise level $x\%$.

To assess whether significant differences exist among different methods, we adopt a nonparametric statistical analysis. For conducting multiple statistical comparisons over multiple data sets, as suggested in [14, 22], the Friedman test and the corresponding *post hoc* Bonferroni-Dunn test are employed. For performing multiple comparisons, it is necessary to check whether the results obtained by different methods present any significant difference (Friedman test), and in the case of finding one, we can find out by using a *post hoc* test to compare the control method with the remaining methods (Bonferroni-Dunn test). We use $\alpha = 0.05$ as the level of significance in all cases. For a detailed description of these tests, one can refer to [14, 22].

4.2. Classification accuracy evaluation

In the first experiment, we aim to compare the classification accuracy of our proposed BRBCS with the classical FRBCS proposed by Chi et al. [10], as reviewed in Section 2.1, and the following two methods improved through changing the rule structure to make it more flexible to characterize the input-output mapping.

1. *FRBCS extended by Cordón et al. [11] (EFRBCS)*: Compared with the basic fuzzy rule structure in Eq. (1), in this method, a fuzzy rule with certainty degrees for all classes in the consequent part is considered.

$$R^q: \text{ If } x_1 \text{ is } A_1^q \text{ and } \cdots \text{ and } x_P \text{ is } A_P^q, \text{ then } (r_1^q, \cdots, r_M^q), \text{ for } q = 1, 2, \cdots, Q, \tag{38}$$

where r_j^q is the certainty degree for rule R^q to predict class ω_j for a pattern belonging to the fuzzy region represented by the antecedent part of the rule. Different from the belief rule structure employed in our proposed BRBCS, the consequence of this rule is in probability distribution, i.e., the sum of certainty degrees for all classes equals one. Accordingly the additive combination reasoning method is employed to classify a query pattern.

2. *EBRB method proposed by Liu et al. [33]*: In this method, a rule base is designed with belief degrees embedded in the consequence part as well as in all of the antecedent terms of each rule for expert inference. However, when using for classification, the consequence part of each rule degenerates into a signal class label.

$$R^q: \quad \text{If } x_1 \text{ is } \left\{(A_{1,j}, \alpha_{1,j}^q), j = 1, \cdots, n_1\right\} \text{ and } \cdots \text{ and } x_P \text{ is } \left\{(A_{P,j}, \alpha_{P,j}^q), j = 1, \cdots, n_P\right\}, \text{ then the consequence is } C^q \text{ with rule weight } \theta^q, \quad q = 1, 2, \cdots, Q, \tag{39}$$

where the antecedent term for each feature $\left\{(A_{p,j}, \alpha_{p,j}^q), j = 1, \cdots, n_p\right\}$ is in belief distribution. Based on the above rule structure, each training sample is developed as a rule to model the input-output relationship, and accordingly, the query pattern is classified by the additive combination of the weighted consequences of all of the rules.

The settings of the considered methods are summarized in Table 3. As for the considered data sets no prior knowledge about the establishment of the fuzzy regions is available, and the simple fuzzy grids are used to partition the feature space. We normalize each feature value into a real number in the unit interval $[0, 1]$. Once the number of partitions for each feature is determined, the fuzzy partitions can be easily computed. Here, different numbers of partitions ($C = 3, 5, 7$) are employed to make the comparison.

Table 3: Settings of considered methods for classification accuracy evaluation

Method	**Setting**			
	Rule structure	Reasoning method	Membership function	Partition number
FRBCS	Rule structure in Eq. (1)	Single winner	Triangular	C = 3,5,7
EFRBCS	Rule structure in Eq. (38)	Additive combination	Triangular	C = 3,5,7
EBRB	Rule structure in Eq. (39)	Additive combination	Triangular	C = 3,5,7
BRBCS	Rule structure in Eq. (14)	Belief reasoning	Triangular	C = 3,5,7

Table 4 and Table 5 show the classification accuracy rates of our proposed BRBCS in comparison with other rule-based methods over the training data and the test data, respectively. The numbers in brackets represent the rank of each method. It can be seen that in both cases, the proposed BRBCS outperforms other methods for most of the data sets. To compare the results statistically, we use nonparametric tests for multiple comparisons to find the best method, considering the average ranks obtained over the test data. First, we use the Friedman test to determine whether significant differences exist among all of the mean values. Table 6 shows the Friedman statistic $\mathcal{F}_F$ for each number of partitions, and it relates them to the corresponding critical values by using a level of significance of $\alpha = 0.05$. Given that the Friedman statistics are clearly greater than their associated critical values, there are significant differences among the observed results with a level of significance $\alpha = 0.05$ for all of the three partition numbers. Then, we apply the Bonferroni-Dunn test to compare the best ranking method (BRBCS) with the remaining methods. Table 7 presents these results. We can see that the Bonferroni-Dunn test rejects all of the hypotheses of equality with the rest of the methods with $p < \alpha/(k-1)$. Therefore, by the analysis of the statistical study shown in Tables 6 and 7, we conclude that our BRBCS is a solid model for classifier design, as it has shown itself to be the best accuracy method when compared with the other rule-based methods applied in this study.

Table 4: Classification accuracy rate (in %) of our proposed BRBCS in comparison with other rule-based methods for different numbers of partitions (over the training data)

	C = 3				C = 5				C = 7			
	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS
Banknote	95.63(3)	96.27(2)	93.40(4)	96.54(1)	98.36(3)	99.27(2)	97.45(4)	99.64(1)	99.64(3)	99.73(2)	97.08(4)	99.82(1)
Breast	70.59(4)	76.47(3)	80.00(2)	89.41(1)	71.76(4)	77.65(3)	89.41(2)	94.12(1)	63.53(4)	75.29(3)	81.18(2)	90.59(1)
Cancer	93.24(4)	96.89(2)	95.98(3)	97.44(1)	96.53(3)	96.89(2)	95.06(4)	99.82(1)	93.78(3)	94.88(2)	93.59(4)	98.35(1)
Diabetes	88.25(4)	91.11(2)	88.89(3)	92.06(1)	95.24(3)	93.65(4)	95.87(2)	96.83(1)	91.11(4)	95.56(3)	99.05(1)	97.46(2)
Ecoli	89.22(4)	92.94(1)	89.59(3)	91.08(2)	94.42(2)	92.19(3)	89.96(4)	98.88(1)	92.94(2)	89.59(3)	88.10(4)	94.80(1)
Glass	73.84(3)	73.26(4)	75.58(2)	83.14(1)	93.02(2)	87.21(3)	85.47(4)	94.77(1)	85.47(3)	86.05(2)	81.40(4)	90.12(1)
Haberman	72.65(4)	76.73(1)	73.47(3)	75.51(2)	80.82(1)	74.29(3)	73.47(4)	78.78(2)	78.78(3)	80.00(2)	73.47(4)	81.22(1)
Iris	95.00(2)	90.00(4)	95.83(1)	94.17(3)	93.33(4)	96.67(1)	95.83(2)	95.00(3)	95.83(4)	96.00(3)	99.17(1)	96.67(2)
Knowledge	89.16(4)	90.40(2)	89.78(3)	91.64(1)	92.88(2)	92.26(3)	91.33(4)	94.74(1)	96.59(3)	97.21(2)	95.05(4)	99.07(1)
Letter	93.75(3)	96.88(2)	93.06(4)	97.50(1)	92.19(4)	95.94(2)	92.81(3)	96.56(1)	91.19(3)	93.44(1)	89.32(4)	93.06(2)
Liver	63.77(2)	58.70(4)	62.68(3)	67.03(1)	71.74(2)	67.75(4)	70.29(3)	76.81(1)	86.96(2)	77.17(4)	79.71(3)	88.04(1)
Magic	82.81(3)	82.94(2)	82.28(4)	83.60(1)	82.81(2)	82.15(3)	81.49(4)	85.44(1)	81.49(4)	82.02(3)	82.68(2)	84.25(1)
Pageblocks	91.35(4)	95.46(2)	93.17(3)	96.37(1)	92.72(4)	95.00(2)	93.63(3)	96.83(1)	89.29(4)	91.57(3)	93.40(2)	94.09(1)
Pima	78.05(4)	79.67(3)	81.30(2)	86.18(1)	82.93(4)	91.71(2)	86.18(3)	94.15(1)	78.54(4)	81.14(3)	84.55(1)	82.76(2)
Satimage	88.75(3)	87.98(4)	90.52(2)	92.64(1)	88.58(4)	91.30(3)	92.46(2)	93.24(1)	89.72(4)	91.28(2)	90.89(3)	91.86(1)
Seeds	89.29(4)	89.88(3)	92.86(2)	94.05(1)	92.86(4)	94.05(3)	97.62(1)	97.02(2)	88.10(4)	90.48(3)	94.05(1)	93.45(2)
Transfusion	77.13(3)	77.30(2)	75.57(4)	78.46(1)	76.96(2)	76.46(3)	73.46(4)	81.80(1)	79.13(3)	79.47(2)	74.79(4)	84.97(1)
Vehicle	79.91(4)	81.24(3)	82.72(2)	84.49(1)	78.43(4)	79.03(3)	79.91(2)	80.85(1)	74.59(4)	77.55(2)	78.43(1)	75.04(3)
Vertebral	74.60(4)	80.65(2)	76.61(3)	81.45(1)	82.26(3)	87.10(2)	79.44(4)	89.11(1)	82.66(2)	81.05(3)	79.71(4)	85.89(1)
Yeast	55.56(3)	52.02(4)	59.85(2)	65.74(1)	58.92(4)	59.68(3)	60.69(2)	69.63(1)	58.08(3)	53.73(4)	66.67(1)	64.31(2)
Av. Rank	3.45	2.60	2.75	1.20	3.05	2.70	3.05	1.20	3.30	2.60	2.70	1.40

Table 5: Classification accuracy rate (in %) of our proposed BRBCS in comparison with other rule-based methods for different numbers of partitions (over the test data)

	C = 3				C = 5				C = 7			
	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS
Banknote	94.23(4)	95.33(3)	98.16(1)	96.42(2)	94.53(4)	97.59(1)	96.13(3)	97.15(2)	99.05(3)	99.64(2)	97.08(4)	99.71(1)
Breast	58.57(4)	66.10(2)	65.33(3)	68.33(1)	62.38(4)	67.38(3)	69.24(2)	73.57(1)	59.52(4)	63.33(3)	69.24(2)	70.38(1)
Cancer	90.00(4)	90.44(3)	92.44(2)	95.82(1)	91.82(3)	92.00(2)	91.47(4)	96.74(1)	89.32(4)	91.82(3)	93.59(1)	92.47(2)
Diabetes	67.95(4)	68.56(2)	68.21(3)	69.67(1)	73.08(3)	71.54(4)	73.21(2)	77.82(1)	75.28(2)	75.44(1)	73.67(4)	74.05(3)
Ecoli	76.12(3)	77.79(2)	71.49(4)	78.34(1)	86.57(2)	82.39(3)	81.19(4)	88.06(1)	85.16(2)	84.00(4)	84.49(3)	86.57(1)
Glass	66.05(3)	61.38(4)	66.67(2)	69.04(1)	72.94(1)	68.57(3)	67.00(4)	71.84(2)	67.14(1)	64.57(2)	63.29(4)	64.29(3)
Haberman	67.13(4)	71.80(2)	72.79(1)	68.85(3)	69.18(3)	71.80(2)	62.95(4)	72.46(1)	70.16(3)	71.80(2)	63.77(4)	73.44(1)
Iris	92.67(4)	93.00(3)	95.33(1)	93.67(2)	95.33(2)	94.67(3)	92.00(4)	96.33(1)	96.33(2)	95.67(3)	90.00(4)	96.67(1)
Knowledge	83.25(3)	80.25(4)	86.75(2)	87.25(1)	91.00(3)	82.75(4)	92.75(2)	93.75(1)	83.50(3)	85.00(1)	80.25(4)	84.75(2)
Letter	92.05(3)	94.44(2)	91.92(4)	95.60(1)	90.50(4)	94.12(2)	92.86(3)	95.15(1)	89.32(4)	91.46(3)	93.68(1)	93.00(2)
Liver	56.52(3)	55.65(4)	60.29(1)	59.42(2)	63.07(3)	64.87(2)	58.84(4)	66.52(1)	66.39(3)	66.84(2)	60.00(4)	68.22(1)
Magic	79.55(4)	82.75(2)	82.12(3)	84.96(1)	82.44(2)	81.38(3)	81.06(4)	85.32(1)	76.28(4)	79.45(3)	81.55(2)	82.14(1)
Pageblocks	89.34(4)	95.58(2)	91.63(3)	96.03(1)	90.41(4)	92.67(3)	94.87(2)	95.42(1)	87.89(4)	88.34(3)	90.37(2)	91.67(1)
Pima	64.05(3)	68.10(2)	61.18(4)	69.93(1)	65.36(4)	72.16(2)	71.10(3)	74.71(1)	63.40(4)	66.33(1)	64.10(3)	64.75(2)
Satimage	86.45(4)	87.78(3)	90.65(2)	91.15(1)	82.83(4)	84.36(3)	91.48(1)	89.04(2)	79.38(4)	82.65(3)	89.97(1)	84.56(2)
Seeds	79.52(4)	82.38(3)	85.90(2)	87.00(1)	88.57(3)	90.00(2)	84.76(4)	90.48(1)	86.67(3)	87.30(2)	81.90(4)	88.57(1)
Transfusion	71.81(4)	76.24(2)	75.57(3)	76.51(1)	77.84(2)	76.24(3)	71.68(4)	80.84(1)	78.52(2)	77.72(3)	73.47(4)	83.89(1)
Vehicle	60.36(4)	64.50(3)	69.64(2)	70.91(1)	60.36(4)	66.44(3)	67.45(2)	68.25(1)	57.99(4)	62.72(3)	65.99(2)	66.30(1)
Vertebral	67.42(4)	72.90(2)	69.03(3)	73.87(1)	82.26(3)	83.74(2)	77.29(4)	86.77(1)	81.29(2)	79.71(3)	78.06(4)	84.84(1)
Yeast	48.51(3)	47.30(4)	49.32(2)	56.62(1)	56.32(3)	57.77(2)	52.70(4)	58.53(1)	55.81(1)	53.73(4)	53.95(3)	54.05(2)
Av. Rank	3.65	2.70	2.40	1.25	3.05	2.60	3.20	1.15	2.95	2.55	3.00	1.50

Table 6: Friedman test of the accuracy rate for the considered methods ($\alpha = 0.05$)

Partition number	Statistic $\mathcal{F}_F$	Critical value	Hypothesis
C = 3	27.005	2.490	Rejected
C = 5	21.000	2.490	Rejected
C = 7	7.798	2.490	Rejected

Table 7: Bonferroni-Dunn test of the accuracy rate for comparing BRBCS with other methods ($\alpha = 0.05$)

Partition number	Method	z value	p value	Critical value $\alpha/(k-1)^a$	Hypothesis
	FRBCS	5.88	4.13E-9	0.0167	Rejected
C = 3	EFRBCS	3.55	3.83E-4	0.0167	Rejected
	EBRB	2.82	0.0048	0.0167	Rejected
	FRBCS	4.65	3.26E-6	0.0167	Rejected
C = 5	EFRBCS	3.55	3.83E-4	0.0167	Rejected
	EBRB	5.02	5.13E-7	0.0167	Rejected
	FRBCS	3.55	3.83E-4	0.0167	Rejected
C = 7	EFRBCS	2.57	0.0101	0.0167	Rejected
	EBRB	3.67	2.39E-4	0.0167	Rejected

[a] k is the number of considered methods.

To analyze the effect of partition numbers on classification performance, in Table 4 and Table 5, the best accuracy rate for each data set is underlined. It can be seen that the classification accuracy is not always ideally improving according to the increase of partition number, especially for those data sets with relatively more features, which is caused by the limited number of training samples. Additionally, as will be shown in Section 4.5, a larger partition number usually means a greater computation burden. Therefore, in practice, for those data sets with fewer features ($M < 10$), we suggest using a partition number $C = 5$; otherwise, a partition number $C = 3$ is suggested to get a better trade-off between accuracy and complexity.

4.3. Classification robustness evaluation

In the second experiment, we aim to analyze the classification robustness of our proposed BRBCS when noise is present in the training sets. Apart from the classical FRBCS introduced in Section 2.1, the following two robust classifiers are also considered for comparison.

1. *C4.5* [40]: C4.5 is considered to be a robust learner tolerant to noisy data. It iteratively builds a decision tree that correctly classifies the largest number of examples. Additionally, a pruning strategy is used to reduce the chances of the classifier being affected by noisy data from the training set.
2. *BagC4.5* [42]: This is a multiple classifier system that considers C4.5 as the base classifier. In this method, the bagging technique is used to resample the original training set, and then the base classifier is trained with different data sets. As experimentally analyzed in [42], BagC4.5 is a good noise-robust multiple classifier system.

The settings of the considered methods are summarized in Table 8. In the following, different types of noise (class noise and feature noise) with different noise levels ($NL = 10\%, 20\%, 30\%, 40\%, 50\%$) are tested for comparison.

Table 8: Settings of considered methods for classification robustness evaluation

Method	Setting
FRBCS	• C = 5 for feature number $M < 10$, otherwise C = 3
C4.5	• Confidence level $c = 0.25$; • Minimal instances per leaf $i = 2$; • Prune after the tree building
BagC4.5	• Replicate number $T = 10$; • Majority vote combination
BRBCS	• C = 5 for feature number $M < 10$, otherwise C = 3

Fig. 3 shows the classification accuracy rate of each data set at different class noise levels. It may be observed that for most data sets, the proposed BRBCS outperforms the other methods at any class noise level. To verify the robustness of the proposed method more specifically, Table 9 gives the RLA of our proposed BRBCS in comparison with other robust methods at different class noise levels. The numbers in brackets represents the rank of each method. For nonparametric statistical analysis, firstly, based on the average ranks of the different methods in Table 9, the Friedman test is conducted to evaluate whether significant differences exist among the different methods. Table 10 shows the Friedman test result of RLA for the considered methods at different class noise levels. Given that the Friedman statistics are clearly greater than their associated critical values, there are significant differences among the observed results with a level of significance of $\alpha = 0.05$ at all class noise levels. Then, we use the Bonferroni-Dunn test to compare the best ranking method (BRBCS) with the remaining methods. Table 11 presents these results. We can see that the Bonferroni-Dunn test rejects all of the hypotheses of equality with the rest of the methods with $p < \alpha/(k-1)$, except for the BagC4.5 method at noise level $NL = 10\%$. This means that there is no significant difference only between BRBCS and BagC4.5 at noise level $NL = 10\%$ with significance level $\alpha = 0.05$. With the increase of the noise level, the p value associated with each of the remaining methods becomes much lower. Thus, the RLA differences are more significant at higher noise levels, showing the superior robustness of the proposed BRBCS in disruptive class noise conditions.

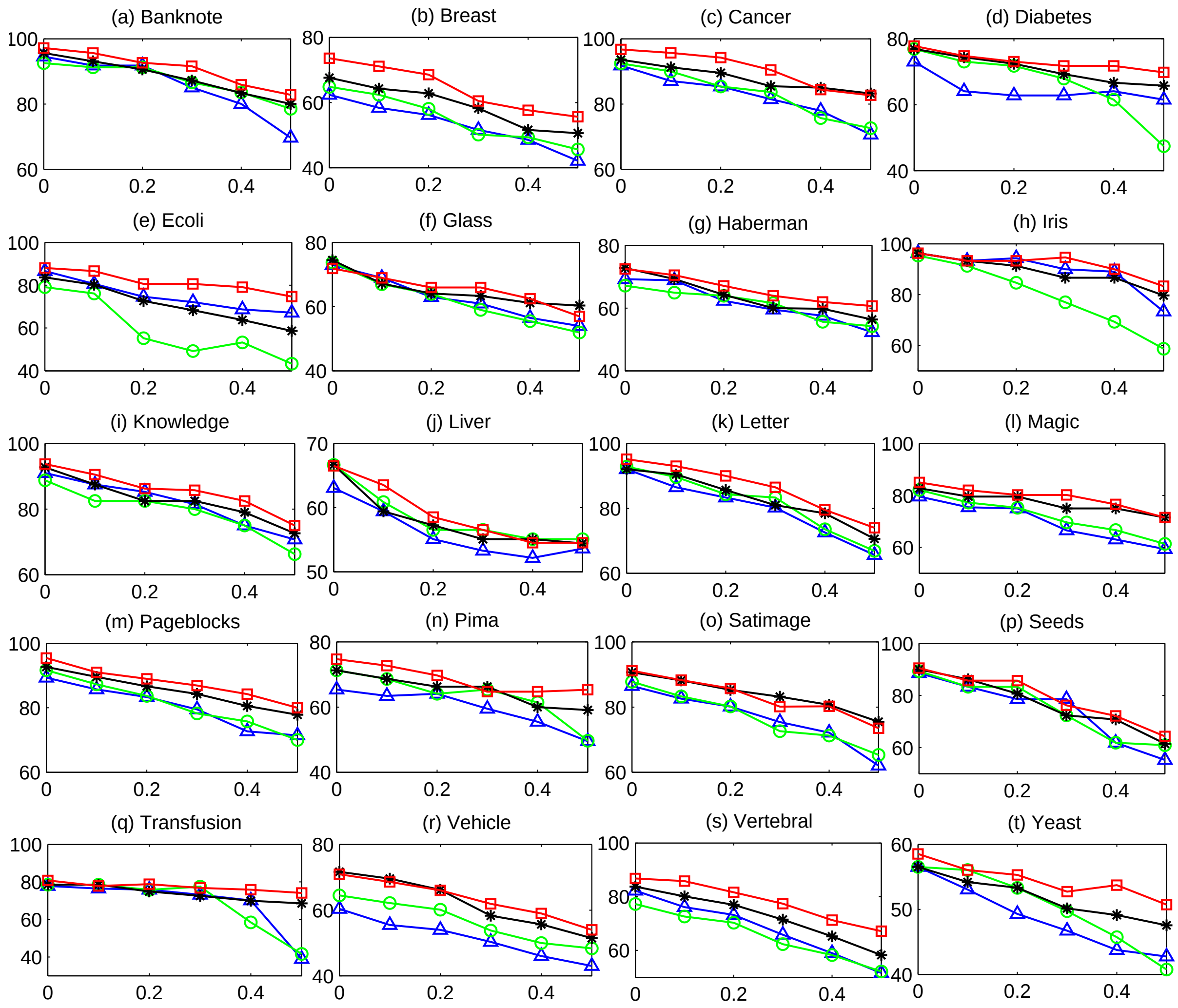

Figure 3: Classification accuracy rate (in %) of our proposed BRBCS in comparison with other methods at different class noise levels. The symbol '△' denotes the FRBCS, '○' denotes the C4.5, '∗' denotes the BagC4.5, and '□' denotes the BRBCS.

Table 9: RLA (in %) of our proposed BRBCS in comparison with other robust methods at different class noise levels.

	NL = 10%				NL = 30%				NL = 50%			
	FRBCS	C4.5	BagC4.5	BRBCS	FRBCS	C4.5	BagC4.5	BRBCS	FRBCS	C4.5	BagC4.5	BRBCS
Banknote	2.89(4)	1.40(1)	2.67(3)	1.58(2)	10.04(4)	6.40(2)	8.78(3)	5.71(1)	26.34(4)	15.13(2)	16.32(3)	14.74(1)
Breast	6.41(4)	3.75(2)	4.93(3)	3.34(1)	17.33(2)	22.63(4)	13.87(1)	17.80(3)	32.52(4)	29.61(3)	25.14(2)	24.40(1)
Cancer	5.19(4)	2.64(3)	2.50(2)	1.07(1)	11.27(4)	9.47(3)	8.69(2)	6.48(1)	23.12(4)	21.50(3)	11.10(1)	14.57(2)
Diabetes	12.28(4)	5.00(3)	3.27(1)	3.86(2)	14.04(4)	11.67(3)	9.97(2)	7.74(1)	15.79(3)	38.33(4)	14.50(2)	10.28(1)
Ecoli	6.93(4)	3.77(2)	4.07(3)	1.69(1)	16.71(2)	37.74(4)	18.34(3)	8.47(1)	22.41(2)	45.28(4)	29.83(3)	15.24(1)
Glass	5.48(2)	9.22(3)	9.85(4)	4.08(1)	16.45(3)	20.07(4)	14.98(2)	8.26(1)	26.05(3)	29.56(4)	18.99(2)	17.78(1)
Haberman	1.47(1)	3.34(3)	4.85(4)	2.72(2)	13.98(3)	8.27(1)	17.57(4)	11.77(2)	24.41(4)	19.25(2)	22.61(3)	16.29(1)
Iris	3.11(2)	4.20(4)	3.53(3)	2.89(1)	6.57(2)	19.23(4)	10.03(3)	1.73(1)	23.88(3)	38.46(4)	17.30(2)	13.50(1)
Knowledge	3.85(2)	7.04(4)	5.66(3)	3.47(1)	10.45(3)	9.86(2)	11.05(4)	8.53(1)	22.25(3)	25.35(4)	21.70(2)	20.00(1)
Letter	6.06(4)	3.46(3)	1.74(1)	2.26(2)	12.86(4)	10.26(2)	12.03(3)	9.09(1)	28.69(4)	27.97(3)	23.31(2)	22.23(1)
Liver	5.79(2)	8.70(3)	10.87(4)	4.51(1)	15.53(3)	15.22(2)	17.40(4)	15.03(1)	14.98(1)	17.39(3)	18.37(4)	17.04(2)
Magic	5.22(3)	5.97(4)	3.88(2)	3.49(1)	16.40(4)	15.27(3)	9.41(1)	9.94(2)	25.46(4)	25.38(3)	13.64(1)	15.89(2)
Pageblocks	4.09(2)	4.78(4)	3.33(1)	4.63(3)	10.99(3)	14.60(4)	9.05(2)	8.93(1)	20.04(2)	23.61(4)	20.99(3)	16.16(1)
Pima	3.00(2)	3.67(4)	3.25(3)	2.68(1)	9.00(3)	8.26(2)	6.96(1)	12.10(4)	24.30(3)	30.28(4)	17.06(2)	12.60(1)
Satimage	4.49(3)	5.13(4)	2.76(1)	3.16(2)	12.79(3)	17.30(4)	8.27(1)	12.10(2)	28.28(4)	25.62(3)	16.73(1)	19.40(2)
Seeds	5.91(3)	6.72(4)	4.23(1)	5.26(2)	11.29(1)	19.03(3)	19.58(4)	15.29(2)	37.63(4)	32.77(3)	31.67(2)	28.95(1)
Transfusion	1.71(3)	0.59(2)	0.25(1)	3.71(4)	6.00(2)	8.89(3)	10.85(4)	4.95(1)	49.91(4)	47.01(3)	12.68(2)	8.24(1)
Vehicle	8.10(4)	3.66(3)	2.88(1)	3.26(2)	17.67(3)	16.52(2)	18.62(4)	12.70(1)	28.76(4)	25.07(2)	28.20(3)	23.85(1)
Vertebral	7.45(4)	6.09(3)	4.32(2)	1.11(1)	20.00(4)	19.34(3)	14.68(2)	10.78(1)	37.26(4)	32.54(3)	30.39(2)	22.60(1)
Yeast	6.14(4)	0.83(1)	4.10(2)	4.22(3)	17.34(4)	12.03(3)	11.30(2)	9.91(1)	24.41(3)	27.95(4)	15.89(2)	13.33(1)
Av. rank	3.05	3.00	2.25	1.70	3.05	2.90	2.60	1.45	3.35	3.25	2.20	1.20

Table 10: Friedman test of RLA for considered methods at different class noise levels ($\alpha = 0.05$)

Noise level	Statistic $\mathcal{F}_F$	Critical value	Hypothesis
NL = 10%	6.3672	2.490	Rejected
NL = 30%	8.7372	2.490	Rejected
NL = 50%	30.096	2.490	Rejected

Table 11: Bonferroni-Dunn test of RLA for comparing BRBCS with other methods at different class noise levels ($\alpha = 0.05$)

Noise level	Method	z value	p value	Critical value $\alpha/(k-1)$	Hypothesis
	FRBCS	3.31	9.44E-4	0.0167	Rejected
NL = 10%	C4.5	3.18	0.0015	0.0167	Rejected
	BagC4.5	1.35	0.1779	0.0167	**Accepted**
	FRBCS	3.92	8.88E-5	0.0167	Rejected
NL = 30%	C4.5	3.55	3.83E-4	0.0167	Rejected
	BagC4.5	2.82	0.0048	0.0167	Rejected
	FRBCS	5.27	1.39E-7	0.0167	Rejected
NL = 50%	C4.5	5.02	5.13E-7	0.0167	Rejected
	BagC4.5	2.45	0.0143	0.0167	Rejected

Fig. 4 shows the classification accuracy rate of each data set with different feature noise levels. Similar to the results under class noise conditions, for most data sets, the test accuracy is always higher for BRBCS than for the other robust methods under the feature noise scheme. Table 12 shows the RLA of our proposed BRBCS in comparison with other robust methods at different feature noise levels. In a similar manner, we first use the Friedman test to evaluate whether significant differences exist among the different methods. Table 13 shows the Friedman test result of RLA for the considered methods at different feature noise levels. It can be seen that the Friedman statistics are clearly greater than their associated critical values at all feature noise levels, which means that there are significant differences among the observed results with a level of significance of $\alpha = 0.05$. Then, we use the Bonferroni-Dunn test to compare the best ranking method (BRBCS) with the remaining methods. As shown in Table 14, the Bonferroni-Dunn test rejects all of the hypotheses of equality with the rest of the methods with $p < \alpha/(k-1)$, except for the BagC4.5 method at noise levels $NL = 10\%$ and $NL = 30\%$. In other words, there is no significant difference between BRBCS and BagC4.5 at noise levels $NL = 10\%$ and $NL = 30\%$ with significance level $\alpha = 0.05$. This is mainly because the feature noise is not very disruptive at relatively lower noise levels. With the increase of the noise level, the p value associated with each of the remaining methods becomes much lower. Thus, the RLA differences are more significant at higher noise levels, which shows the superior robustness of the proposed BRBCS in disruptive feature noise conditions.

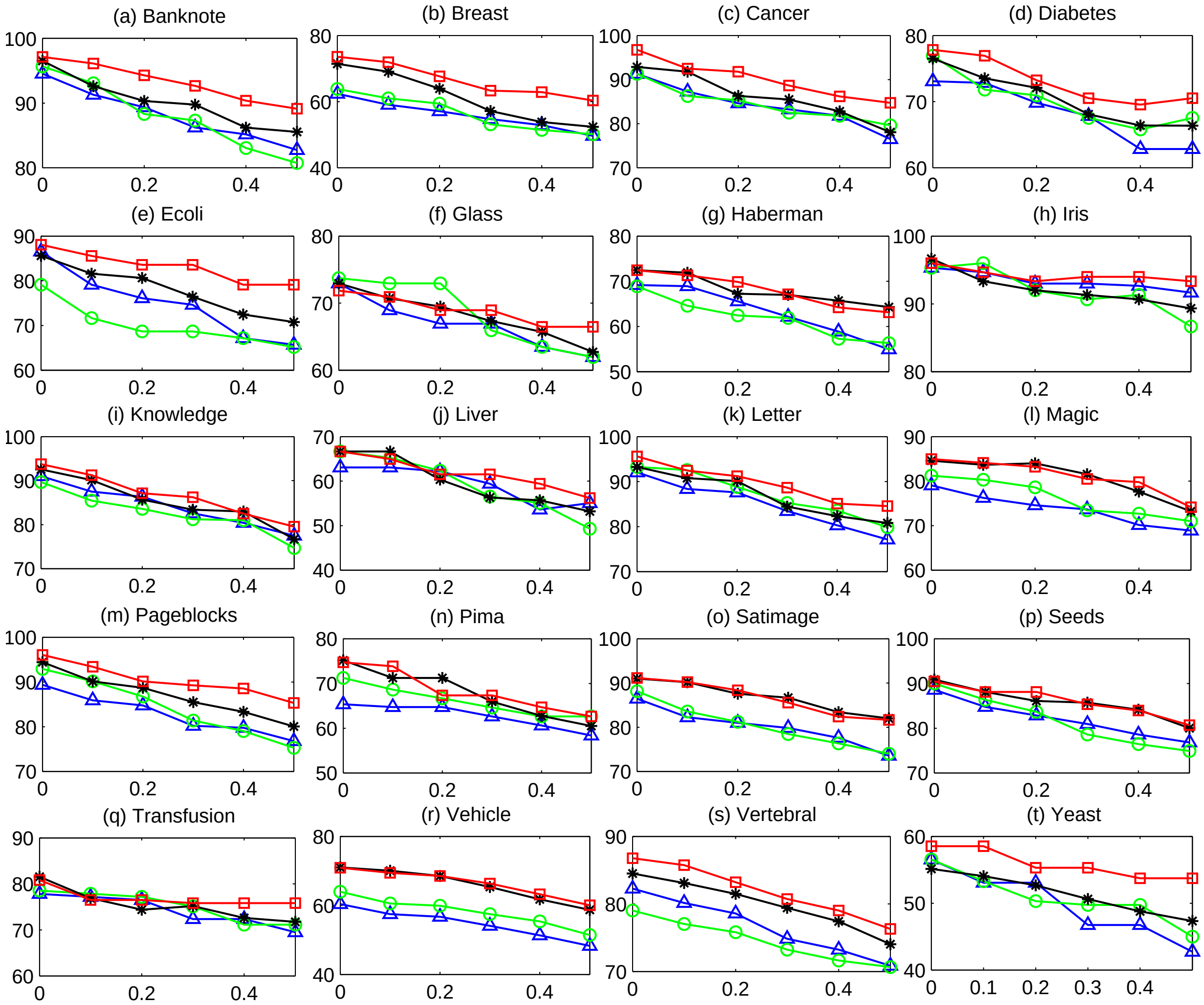

Figure 4: Classification accuracy rate (in %) of our proposed BRBCS in comparison with other methods at different feature noise levels. The symbol '$\triangle$' denotes the FRBCS, '$\circ$' denotes the C4.5, '$*$' denotes the BagC4.5, and '$\square$' denotes the BRBCS.

Table 12: RLA (in %) of our proposed BRBCS in comparison with other robust methods at different feature noise levels.

	NL = 10%				NL = 30%				NL = 50%			
	FRBCS	C4.5	BagC4.5	BRBCS	FRBCS	C4.5	BagC4.5	BRBCS	FRBCS	C4.5	BagC4.5	BRBCS
Banknote	3.44(3)	2.71(2)	4.06(4)	1.05(1)	8.79(4)	8.53(3)	7.02(2)	4.62(1)	12.48(3)	15.60(4)	11.40(2)	8.27(1)
Breast	5.34(4)	4.48(3)	3.33(2)	2.26(1)	12.34(1)	16.79(3)	20.00(4)	13.91(2)	20.46(2)	21.42(3)	26.67(4)	17.99(1)
Cancer	4.91(3)	5.52(4)	1.17(1)	4.38(2)	8.81(3)	9.63(4)	7.95(1)	8.38(2)	16.22(4)	12.78(2)	15.94(3)	12.44(1)
Diabetes	0.36(1)	6.67(4)	3.92(3)	1.15(2)	7.20(1)	12.23(4)	11.03(3)	9.39(2)	14.04(4)	12.23(2)	13.27(3)	9.39(1)
Ecoli	8.62(3)	9.44(4)	4.69(2)	2.82(1)	13.78(4)	13.21(3)	10.72(2)	5.09(1)	24.14(4)	17.63(3)	17.35(2)	10.18(1)
Glass	5.48(4)	1.08(1)	3.07(3)	1.30(2)	8.23(3)	10.58(4)	7.70(2)	4.08(1)	15.08(3)	16.00(4)	14.03(2)	7.53(1)
Haberman	0.38(1)	6.22(4)	0.79(2)	1.59(3)	11.19(4)	10.15(3)	7.56(2)	7.35(1)	20.62(4)	18.20(3)	11.28(1)	12.90(2)
Iris	0.70(1)	1.77(3)	3.45(4)	1.39(2)	2.45(2)	4.90(3)	5.52(4)	2.08(1)	3.85(2)	9.09(4)	7.59(3)	2.78(1)
Knowledge	3.85(3)	4.69(4)	2.14(1)	2.66(2)	9.26(2)	9.40(3)	9.89(4)	7.99(1)	14.84(1)	16.73(3)	17.15(4)	15.13(2)
Letter	4.00(4)	0.73(1)	2.64(2)	3.31(3)	9.34(3)	8.58(2)	9.44(4)	7.28(1)	16.21(4)	14.36(3)	13.35(2)	11.56(1)
Liver	0.33(1)	2.17(4)	1.55(3)	0.67(2)	5.79(1)	15.22(3)	15.51(4)	7.72(2)	12.68(1)	26.09(4)	20.22(3)	15.68(2)
Magic	3.55(4)	1.77(3)	1.30(2)	1.01(1)	6.82(3)	9.63(4)	3.58(1)	5.27(2)	12.88(2)	13.63(4)	13.45(3)	12.72(1)
Pageblocks	3.86(3)	2.99(2)	4.55(4)	2.73(1)	10.22(3)	12.43(4)	9.48(2)	7.07(1)	14.07(2)	18.98(4)	15.24(3)	11.15(1)
Pima	1.00(1)	3.67(3)	5.22(4)	1.14(2)	4.18(1)	9.28(2)	12.28(4)	9.89(3)	10.70(1)	12.09(2)	19.50(4)	16.17(3)
Satimage	4.85(3)	5.20(4)	0.90(1)	1.02(2)	7.63(3)	11.01(4)	4.74(1)	6.14(2)	14.93(3)	16.03(4)	9.91(1)	10.41(2)
Seeds	4.30(4)	4.08(3)	3.04(2)	2.63(1)	8.60(3)	12.70(4)	5.66(1)	5.98(2)	13.28(3)	16.79(4)	11.92(2)	10.84(1)
Transfusion	1.84(2)	0.85(1)	5.71(4)	5.35(3)	7.05(3)	4.33(1)	7.80(4)	6.19(2)	10.70(3)	9.40(2)	12.02(4)	6.19(1)
Vehicle	4.82(3)	5.29(4)	1.30(1)	2.06(2)	10.40(4)	10.12(3)	7.99(2)	6.47(1)	20.00(4)	19.58(3)	17.34(2)	15.34(1)
Vertebral	2.58(4)	2.13(3)	1.18(1)	1.68(2)	9.02(4)	7.35(3)	6.03(1)	6.95(2)	13.88(4)	10.61(1)	12.40(3)	12.08(2)
Yeast	6.14(4)	5.71(3)	1.94(2)	0.22(1)	17.34(4)	12.03(3)	8.14(2)	5.52(1)	24.41(4)	20.40(3)	14.17(2)	8.20(1)
Av. rank	2.80	3.00	2.40	1.80	2.80	3.15	2.50	1.55	2.90	3.10	2.65	1.35

Table 13: Friedman test of RLA for considered methods at different feature noise levels ($\alpha = 0.05$)

Noise level	Statistic $\mathcal{F}_F$	Critical value	Hypothesis
NL = 10%	3.8365	2.490	Rejected
NL = 30%	7.4993	2.490	Rejected
NL = 50%	11.303	2.490	Rejected

Table 14: Bonferroni-Dunn test of RLA for comparing BRBCS with other methods at different feature noise levels ($\alpha = 0.05$)

Noise level	Method	z value	p value	Critical value $\alpha/(k-1)$	Hypothesis
	FRBCS	2.45	0.0143	0.0167	Rejected
NL = 10%	C4.5	2.94	0.0033	0.0167	Rejected
	BagC4.5	1.47	0.1416	0.0167	**Accepted**
	FRBCS	3.06	0.0022	0.0167	Rejected
NL = 30%	C4.5	3.92	8.89E-5	0.0167	Rejected
	BagC4.5	2.33	0.0200	0.0167	**Accepted**
	FRBCS	3.80	1.47E-4	0.0167	Rejected
NL = 50%	C4.5	4.29	1.81E-5	0.0167	Rejected
	BagC4.5	3.18	0.0015	0.0167	Rejected

4.4. Interpretability analysis

As mentioned in the introduction, in addition to accuracy, interpretability is also an important criterion for rule-based systems. In [21], the authors presented an overview of the interpretability measures of rule-based systems. In this experiment, the following three widely accepted interpretability measures recommended in [21] are utilized to make a quantitative analysis.

- *Number of rules (# Rule)*: According to the principle of Occam's razor (the best model is the simplest one fitting the system behavior well), the set of fuzzy rules should be as small as possible under conditions in which the model performance is preserved to a satisfactory level.

- *Number of conditions (# Conditions)*: The number of conditions should be as small as possible in order to ease the readability of the rules.

- *Number of fired rules for a given input (# Fired rule)*: The number of rules used for the reasoning of a given input should be as small as possible in order to control the semantic interpretability.

Based on the above three interpretability measures, we test the four rule-based methods shown in Table 3. Twenty real-world problems shown in Table 2 are considered for evaluation, and the 5-CV model is used to calculate the average results. For those data sets with few features ($M < 10$), we use the partition number $C = 5$; otherwise the partition number $C = 3$ is selected. Table 15 shows the results of the three interpretability measures for different methods.

Table 15: Results of interpretability measures for our proposed BRBCS in comparison with other rule-based methods

	# Rule				# Conditions				# Fired rule			
	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS	FRBCS	EFRBCS	EBRB	BRBCS
Banknote	74.4	74.4	1098	74.4	4	4	4	4	1	9.8	1098	9.8
Breast	49	49	85	49	9	9	9	9	1	7.2	85	7.2
Cancer	167.2	167.2	547	167.2	9	9	9	9	1	4.2	547	4.2
Diabetes	248.8	248.8	315	248.8	8	8	8	8	1	18.5	315	18.5
Ecoli	105.8	105.8	269	105.8	7	7	7	7	1	11.3	269	11.3
Glass	80.4	80.4	172	80.4	9	9	9	9	1	7.2	172	7.2
Haberman	49	49	245	49	3	3	3	3	1	5.2	245	5.2
Iris	42.6	42.6	120	42.6	4	4	4	4	1	5.8	120	5.8
Knowledge	120.2	120.2	323	120.2	5	5	5	5	1	6.8	323	6.8
Letter	1354.6	1354.6	16000	1354.6	16	16	16	16	1	67.3	16000	67.3
Liver	112.6	112.6	276	112.6	6	6	6	6	1	16.8	276	16.8
Magic	346.2	346.2	15216	346.2	10	10	10	10	1	29.3	15216	29.3
Pageblocks	55.4	55.4	4379	55.4	10	10	10	10	1	21.7	4379	21.7
Pima	190.8	190.8	615	190.8	8	8	8	8	1	26.9	615	26.9
Satimage	1443.2	1443.2	5148	1443.2	36	36	36	36	1	87	5148	87
Seeds	96.8	96.8	168	96.8	7	7	7	7	1	9.5	168	9.5
Transfusion	28.6	28.6	599	28.6	4	4	4	4	1	5.5	599	5.5
Vehicle	288.2	288.2	677	288.2	18	18	18	18	1	61.5	677	61.5
Vertebral	93.2	93.2	248	93.2	6	6	6	6	1	18.8	248	18.8
Yeast	208	208	1188	208	8	8	8	8	1	27.4	1188	27.4

By the analysis of the results presented in Table 15, we can draw the following conclusions.

1. According to the measure *# Rule*, the proposed BRBCS and the other two fuzzy rule-based methods, FRBCS and EFRBCS, have higher interpretability than the EBRB method. As the same fuzzy partition approach of the feature space is used for the FRBCS, EFRBCS and BRBCS methods, the three methods generate the same number of rules. However, for the EBRB method, as each training pattern is used to generate a rule, the number of generated rules equals the number of training patterns.
2. According to the measure *# Conditions*, the four considered methods show the same interpretability. All of the four methods do not take into account the selection of features, and so all features are considered as conditions in the antecedent part of a rule.
3. According to the measure *# Fired rule*, the proposed BRBCS has higher interpretability than the EBRB method but a lower interpretability than the FRBCS method. This is because for the FRBCS method, the single winner rule is used to classify a query pattern, while for the proposed BRBCS method, all of the activated rules are combined to obtain the final result.

Considering the above three measures, the proposed BRBCS shows a slight decrease in interpretability compared with the classical FRBCS. As declared in [21], accuracy and interpretability are always two contradictory requirements. The accuracy improvement of the proposed BRBCS is achieved at the expense of a slight decrease in interpretability.

4.5. Time complexity analysis

In this section, a time complexity analysis of the proposed BRBCS is provided to show to what extent the runtime depends on factors such as the number of instances, the number of features and the number of partitions. Twenty real-world problems (with the numbers of training instances ranging from 85 to $16,000$ and the numbers of features ranging from 3 to 36) shown in Table 2 are considered for evaluation. Three different numbers of partitions, $C = 3, 5, 7$, are tested, and the 5-CV model is used to calculate the average runtime. The numerical experiments are executed by MATLAB 7.12 on an HP EliteBook 8570p with an Intel(R) Core(TM) i7-3540 M CPU @3.00 GHz and 8 GB memory. Table 16 shows the average runtime of the proposed BRBCS in the training and testing phases for different data sets and different partition numbers, where "# Tra." is the number of training instances in the data set, "# Fea." is the number of features, and "# Rule" is the number of generated rules. "T. Tra." and "T. Tes." are the average runtimes in the training phase and testing phase (classifying one pattern), respectively.

Table 16: Average runtime (in s) of our proposed BRBCS for different data sets and different partition numbers

			C = 3			C = 5			C = 7		
Data sets	# Tra.	# Fea.	# Rule	T. Tra.	T. Tes.	# Rule	T. Tra.	T. Tes.	# Rule	T. Tra.	T. Tes.
Banknote	1,098	4	30.6	0.166	1.3E-3	74.4	0.225	2.7E-3	151.2	0.293	5.2E-3
Breast	85	9	26	0.025	2.2E-3	49	0.031	3.3E-3	61.6	0.047	4.3E-3
Cancer	547	9	209	0.181	1.4E-2	267.2	0.209	1.8E-2	323.2	0.284	2.1E-2
Diabetes	315	8	80.2	0.091	5.4E-3	248.8	0.125	1.6E-2	297.8	0.162	1.8E-2
Ecoli	269	7	44.8	0.050	2.1E-3	105.8	0.072	4.4E-3	178.4	0.094	7.1E-3
Glass	172	9	39.8	0.053	2.9E-3	80.4	0.072	5.7E-3	115.8	0.099	8.2E-3
Haberman	245	3	17.2	0.029	6.0E-4	49	0.041	1.4E-3	82.2	0.047	2.4E-3
Iris	120	4	14.4	0.025	6.1E-4	42.6	0.029	1.6E-3	63	0.031	2.3E-3
Knowledge	323	5	61.4	0.033	2.4E-3	120.2	0.041	5.0E-3	154.2	0.059	6.3E-3
Letter	16,000	16	1354.6	7.025	8.4E-2	3593	11.963	2.0E-1	6939	17.708	3.7E-1
Liver	276	6	42.8	0.053	2.4E-3	112.6	0.078	5.7E-3	171.6	0.106	8.2E-3
Magic	15,216	10	346.2	4.034	2.8E-2	1854.8	5.123	5.7E-2	4573.2	6.933	8.2E-2
Pageblocks	4,379	10	55.4	0.642	4.7E-3	162	0.950	1.1E-2	286.6	1.542	2.2E-2
Pima	615	8	104.6	0.149	6.9E-3	190.8	0.231	2.4E-2	340.8	0.303	3.2E-2
Satimage	5,148	36	1443.2	1.857	1.6E-1	2046.6	5.997	2.5E-1	2531.6	6.730	3.1E-1
Seeds	168	7	53.2	0.041	3.2E-3	96.8	0.062	5.9E-3	121.8	0.078	7.7E-3
Transfusion	599	4	12.6	0.081	5.1E-4	28.6	0.112	1.1E-3	56	0.147	2.0E-3
Vehicle	677	18	288.2	0.374	5.2E-2	334	0.577	8.0E-2	370.8	0.633	8.5E-2
Vertebral	248	6	34	0.049	1.9E-3	93.2	0.066	4.7E-3	142.2	0.094	6.9E-3
Yeast	1,188	8	96	0.312	6.6E-3	208	0.515	1.3E-2	462	0.608	2.8E-2

By analyzing the results presented in Table 16, we can see that the runtimes for both the training and testing phases mainly depend on the number of generated rules. More rules usually means more time to train the BRB and also more time to classify a pattern based on the generated BRB. Thus, we can instead analyze how the factors affect the number of rules. First, for each data set, the number of rules always increases with increases in the partition number. However, the number of rules cannot increase indefinitely, as it is constrained by the number of training instances. Second, by comparing different data sets, we can see that a larger number of features usually results in a larger number of rules (e.g., the data sets *Letter* and *Satimage*). However, this tendency is also constrained by the number of training instances. For example, although the data set *Vehicle* has a larger number of features than *Magic*, it has a relatively smaller number of rules under any partition condition, mainly because its number of training instances is quite small. In brief, with the growth in the numbers of partitions and features, the runtime of the proposed BRBCS will increase, but this increase is constrained by the number of available training instances.

5. Conclusion

This paper extends the traditional FRBCS in belief functions framework and develops a belief rule-based classification system (BRBCS) to address imprecise or incomplete information in complex classification problems. The two components of the proposed BRBCS, i.e., the belief rule base (BRB) and the belief reasoning method (BRM), are designed specifically by taking into account the possible pattern noise in many real-world data sets. The experiments show that the proposed BRBCS achieves better classification accuracy compared with other rule-based methods. Moreover, this method can effectively address the class or feature noise in the training data set. This allows us to conclude that the introduction of the belief rule structure improves the behavior of the rule-based classification system.

Acknowledgments

The authors would like to thank the anonymous reviewers for their careful reading and in-depth criticisms and suggestions. This work is partially supported by China Natural Science Foundation (Nos. 61135001, 61374159, 61403310) and the Doctorate Foundation of Northwestern Polytechnical University (No. CX201319).

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499.

[2] M.R. Akbarzadeh-Totonchi, M. Moshtagh-Khorasani, A hierarchical fuzzy rule-based approach to aphasia diagnosis, J. Biomed. Inform. 40 (2007) 465–475.

[3] J. Alcala-Fdez, R. Alcala, F. Herrera, A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, IEEE Trans. Fuzzy Syst. 19 (2011) 857–872.

[4] J.M. Alonso, L. Magdalena, Special issue on interpretable fuzzy systems, Inform. Sci. 181 (2011) 4331–4339.

[5] F. Angiulli, Fast condensed nearest neighbor rule, in: Proceedings of the 22nd International Conference on Machine Learning, pp. 7–11.

[6] E. Binaghi, P. Madella, Fuzzy Dempster-Shafer reasoning for rule-based classifiers, Int. J. Intell. Syst. 14 (1999) 559–583.

[7] L.H. Chen, An extended rule-based inference for general decision-making problems, Inform. Sci. 102 (1997) 247–261.

[8] S. Chen, J. Liu, H. Wang, J.C. Augusto, A data driven rule-base inference approach for classification systems, in: Proceedings of the 2nd International Conference on Software Engineering and Service Science, pp. 78–81.

[9] Y. Chen, J.Z. Wang, Support vector learning for fuzzy rule-based classification systems, IEEE Trans. Fuzzy Syst. 11 (2003) 716–728.

[10] Z. Chi, H. Yan, T. Pham, Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition, Singapore: World Scientific, 1996.

[11] O. Cordón, M. José del Jesus, F. Herrera, A proposal on reasoning methods in fuzzy rule-based classification systmes, Int. J. Approx. Reason. 20 (1999) 21–45.

[12] B. Dasarathy, Noising around the neighbourhood: a new system structure and classification rule for recognition in partially exposed environments, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1980) 67–71.

[13] A. Dempster, Upper and lower probabilities induced by multivalued mapping, Ann. Math. Statist. 38 (1967) 325–339.

[14] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[15] T. Denœux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, IEEE Trans. Syst. Man Cybern. 25 (1995) 804–813.

[16] T. Denœux, A neural network classifier based on Dempster-Shafer theory, IEEE Trans. Syst. Man Cybern. Part A-Syst. 30 (2000) 131–150.

[17] T. Denœux, Introduction to belief functions, Technical Report, Université de Technologie de Compiègne, Compiègne, France, 2011.

[18] T. Denœux, P. Smets, Classification using belief functions: the relationship between the case-based and model-based approaches, IEEE Trans. Syst. Man Cybern. Part B-Cybern. 36 (2006) 1395–1406.

[19] T. Denœux, A.B. Yaghlane, Approximating the combination of belief functions using the fast möbius transform in a coarsened frame, Int. J. Approx. Reason. 31 (2002) 77–101.

[20] E. Fix, J. Hodges, Discriminatory analysis, nonparametric discrimination: consistency properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[21] M. Gacto, R. Alcalá, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Inform. Sci. 181 (2011) 4340–4360.

[22] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Inform. Sci. 180 (2010) 2044–2064.

[23] J. Hühn, E. Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction, Data Min. Knowl. Disc. 19 (2009) 293–319.

[24] H. Ishibuchi, K. Nozaki, H. Tanaka, Distributed representation of fuzzy rules and its application to pattern classification, Fuzzy Sets Syst. 52 (1992) 21–32.

[25] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy GBML approaches for pattern classification problems, IEEE Trans. Syst. Man Cybern. Part B-Cybern. 35 (2005) 359–365.

[26] L. Jiao, Q. Pan, X. Feng, Multi-hypothesis nearest-neighbor classifier based on class-conditional weighted distance metric, Neurocomputing 151 (2015) 1468–1476.

[27] L. Jiao, Q. Pan, X. Feng, F. Yang, An evidential k-nearest neighbor classification method with weighted attributes, in: Proceedings of the 16th International Conference on Information Fusion, pp. 300–306.

[28] L. Jiao, Q. Pan, Y. Liang, X. Feng, F. Yang, Combining sources of evidence with reliability and importance for decision making, Cent. Eur. J. Oper. Res. (2013) in press, doi: 10.1007/s10100–013–0334–3.

[29] A.L. Jousselme, D. Grenier, E. Bossé, A new distance between two bodies of evidence, Inform. Fusion 2 (2001) 91–101.

[30] A.L. Jousselme, P. Maupin, Distances in evidence theory: Comprehensive survey and generalizations, Int. J. Approx. Reason. 53 (2012) 118–145.

[31] G.L. Kong, D.L. Xu, R. B., A belief rule-based decision support system for clinical risk assessment of cardiac chest pain, Eur. J. Oper. Res. 219 (2012) 564–573.

[32] B. Li, H.W. Wang, J.B. Yang, M. Guo, C. Qi, A belief-rule-based inventory control method under nonstationary and uncertain demand, Expert

Syst. Appl. 38 (2011) 14997–15008.

[33] J. Liu, L. Martinez, A. Calzada, H. Wang, A novel belief rule base representation, generation and its inference methodology, Knowl.-Based Syst. 53 (2013) 129–141.

[34] J. Liu, J.B. Yang, J. Wang, H.S. Sii, Y.M. Wang, Fuzzy rule based evidential reasoning approach for safety analysis, Int. J. Gen. Syst. 33 (2004) 183–204.

[35] W. Liu, Analyzing the degree of conflict among belief functions, Artif. Intell. 170 (2006) 909–924.

[36] Z. Liu, Q. Pan, J. Dezert, A new belief-based k-nearest neighbor classification method, Pattern Recogn. 46 (2013) 834–844.

[37] Z. Liu, Q. Pan, J. Dezert, G. Mercier, Credal classification rule for uncertain data based on belief functions, Pattern Recogn. 44 (2014) 2532–2541.

[38] C.J. Merz, P.M. Murphy, D.W. Aha, UCI Repository of Machine Learning Databases, Dept. of Information and Computer Science, Univ. of California, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html, 1997.

[39] K. Nozaki, H. Ishibuchi, H. Tanaka, Adaptive fuzzy rule-based classification systems, IEEE Trans. Fuzzy Syst. 4 (1996) 238–250.

[40] J.R. Quinlan, C4.5: programs for machine learning, San Francisco, CA, USA: Morgan Kaufmann Publishers, 1993.

[41] S. Raudys, A. Jain, Small sample effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252–264.

[42] J.A. Sáez, M. Galar, J. Luengo, F. Herrera, Tackling the problem of classification with noisy data using Multiple Classifier Systems: Analysis of the performance and robustness, Inform. Sci. 247 (2013) 1–20.

[43] J.A. Sáez, J. Luengo, F. Herrera, On the suitability of fuzzy rule-based classification systems with noisy data, IEEE Trans. Fuzzy Syst. 20 (2012) 1–19.

[44] S.R. Samantaray, Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection, Appl. Soft Comput. 13 (2013) 928–938.

[45] J.A. Sanz, A. Fernández, H. Bustince, F. Herrera, Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning, Inform. Sci. 180 (2010) 3674–3685.

[46] G. Shafer, A Mathematical Theory of Evidence, Princeton, NJ: Princeton University Press, 1976.

[47] P. Smets, Decision making in the TBM: the necessity of the pignistic transformation, Int. J. Approx. Reason. 38 (2005) 133–147.

[48] P. Smets, Analyzing the combination of conflicting belief functions, Inform. Fusion 8 (2007) 387–412.

[49] D.G. Stavrakoudis, G.N. Galidaki, I.Z. Gitas, J.B. Theocharis, A genetic fuzzy-rule-based classifier for land cover classification from hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 50 (2012) 130–148.

[50] C.T. Sun, Rule-base structure identification in an adaptive-network based fuzzy inference system, IEEE Trans. Fuzzy Syst. 2 (1994) 64–73.

[51] D.W. Tang, J.B. Yang, K.S. Chin, A methodology to generate a belief rule base for customer perception risk analysis in new product development, Expert Syst. Appl. 38 (2011) 5373–5383.

[52] C. Tsang, S. Kwong, H. Wang, A systematic fuzzy rule based approach for fault classification in transmission lines, Pattern Recogn. 40 (2007) 2373–2391.

[53] S.A. Vinterbo, E.Y. Kim, L. Ohno-Machado, Small, fuzzy and interpretable gene expression based classifiers, Bioinformatics 21 (2005) 1964–1970.

[54] N. Wilson, Algorithms for Dempster-Shafer theory, in: D.M. Gabbay, P. Smets (Eds.), Handbook of Defeasible Reasoning and Uncertainty Management, volume 5, Boston, MA: Kluwer, 2000, pp. 421–475.

[55] H. Wu, J. Mendel, Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers, IEEE Trans. Fuzzy Syst. 15 (2007) 56–72.

[56] R.R. Yager, D.P. Filev, Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory, IEEE Trans. Syst. Man Cybern. 25 (1995) 1221–1230.

[57] J.B. Yang, J. Liu, J. Wang, H.S. Sii, H.W. Wang, Belief rule-based inference methodology using the evidential reasoning approach–RIMER, IEEE Trans. Syst. Man Cybern. Part A-Syst. 36 (2006) 266–285.

[58] J.B. Yang, I.M. Wang, D.L. Xu, K.S. Chin, L. Chatton, Belief rule-based methodology for mapping consumer preferences and setting product targets, Expert Syst. Appl. 39 (2012) 4749–4759.

[59] J.B. Yang, D.L. Xu, On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty, IEEE Trans. Syst. Man Cybern. Part A-Syst. 32 (2002) 289–304.

[60] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.

[61] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. 22 (2004) 177–210.