

Belief Combination of Classifiers for Incomplete Data

Zuwei Zhang, *Student Member, IEEE*, Songtao Ye, Yiru Zhang, Weiping Ding, *Senior Member, IEEE*, and Hao Wang, *Senior Member, IEEE*

Abstract—Data with missing values, or incomplete information, brings some challenges to the development of classification, as the incompleteness may significantly affect the performance of classifiers. In this paper, we handle missing values in both training and test sets with uncertainty and imprecision reasoning by proposing a new belief combination of classifier (BCC) method based on the evidence theory. The proposed BCC method aims to improve the classification performance of incomplete data by characterizing the uncertainty and imprecision brought by incompleteness. In BCC, different attributes are regarded as independent sources, and the collection of each attribute is considered as a subset. Then, multiple classifiers are trained with each subset independently and allow each observed attribute to provide a sub-classification result for the query pattern. Finally, these sub-classification results with different weights (discounting factors) are used to provide supplementary information to jointly determine the final classes of query patterns. The weights consist of two aspects: global and local. The global weight calculated by an optimization function is employed to represent the reliability of each classifier, and the local weight obtained by mining attribute distribution characteristics is used to quantify the importance of observed attributes to the pattern classification. Abundant comparative experiments including seven methods on twelve datasets are executed, demonstrating the out-performance of

BCC over all baseline methods in terms of accuracy, precision, recall, F1 measure, with pertinent computational costs.

Index Terms—Classifier fusion, classification, evidence theory, incomplete data, missing values.

I. INTRODUCTION

CLASSIFICATION is a traditional and prevalent problem in data analysis, aiming to identify objects to the categories they belong to. Incompleteness problem in data is one of critical challenges in classification applications, caused by various data collection or access mechanisms. Incomplete data, also called incomplete patterns or missing data, refers to the data with missing values, attributes, or contents¹. This phenomenon has affected classification applications with unsatisfactory results [1]–[3]. The incompleteness of data is a critical issue in risk-sensitive fields, such as industrial systems [4], [5], health management [6] and financial market [7]. Many methods have emerged to resolve the incompleteness issues around three types of missing mechanism [4], [8]–[10]: such as missing completely at random, missing at random, not missing at random. These methods can be roughly categorized into four groups:

1) Deletion methods. The pattern with missing values is simply discarded. The deletion method is only applicable to cases where the number of incomplete patterns accounts for a small proportion (less than 5%) of the whole dataset [8], [11]. It inevitably leads to a waste of patterns that are difficult (or costly) to obtain sometimes.

2) Model-based methods. The missing values are imputed based on statistical assumptions of joint distribution, then the fulfilled patterns are classified by conventional classifiers. For example, a (supervised) logistic regression algorithm is proposed in [12] to deal with incomplete datasets, where the missing values are modeled by performing analytic integration with an estimated conditional density function (conditioned on the observed data) based on the Gaussian mixture model (GMM) [13]. However, the approximated model is not robust enough, bringing over-fitting or under-fitting problems.

3) Machine learning methods. Incomplete patterns are directly used to train some specific classifiers. For decision trees, in algorithm C4.5 [14], missing values are simply ignored in gain and entropy calculations, while C5.0 [15] and CART neural network [16] employ imputative frameworks [17]. In Hybrid neural networks [18], missing values are

Manuscript received September 19, 2021; revised November 24, 2021; accepted December 30, 2021. This work was supported in part by the Center-initiated Research Project and Research Initiation Project of Zhejiang Laboratory (113012-AL2201, 113012-PI2103), the National Natural Science Foundation of China (61300167, 61976120), the Natural Science Foundation of Jiangsu Province (BK20191445), the Natural Science Key Foundation of Jiangsu Education Department (21KJA510004), and Qing Lan Project of Jiangsu Province. Recommended by Associate Editor Long Chen. (*Corresponding author: Yiru Zhang.*)

Citation: Z. W. Zhang, S. T. Ye, Y. R. Zhang, W. P. Ding, and H. Wang, “Belief combination of classifiers for incomplete data,” *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 4, pp. 652–667, Apr. 2022.

Z. W. Zhang is with the Research Center for Optical Fiber Sensing, Zhejiang Laboratory, Hangzhou 310000, and also with the School of Automation, Northwestern Polytechnical University, Xi’an 710072, China (e-mail: zhangzuwei0720@gmail.com).

S. T. Ye is with the Research Center for Optical Fiber Sensing, Zhejiang Laboratory, Hangzhou 310000, China (e-mail: songtaoye9@gmail.com).

Y. R. Zhang is with the School of Cyber Engineering, Xidian University, Xi’an 710000, China, and also with the Department of Computer Science, CY Cergy Paris Université, 2 Av. Adolphe Chauvin, 95302 Cedex, France (e-mail: zyrbruce@gmail.com).

W. P. Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).

H. Wang is with the Research Center for Optical Fiber Sensing, Zhejiang Laboratory, Hangzhou 310000, the School with of Cyber Engineering, Xidian University, Xi’an 710000, China, and also with the Department of Computer Science, Norwegian University of Science and Technology, Trondheim 7491, Norway (e-mail: hawa@ntnu.no).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JAS.2022.105458

¹ To avoid ambiguity, we apply the term incomplete data for a dataset with missing values, and incomplete pattern for a pattern with missing values.

imputed by two FCM (fuzzy *c*-means) based methods. In a support vector solution [19], the modified support vector machine generalizes the approach of mean imputation in the linear case by taking into account the uncertainty of the predicted outputs. In word embedding models [20], the missing attributes are usually valued as 0, which can be regarded as imputation. When most of the attributes are lost, however, the classification performance of these methods is often unsatisfactory.

4) Estimation methods. This is the most widely used methods for dealing with incomplete data. The missing value is replaced (imputed) with an estimation [3], and then the pattern with estimations is classified by conventional classifiers (e.g., Bayes classifier [21]). We will review some estimation-based methods separately.

There are some popular and representative methods for estimating missing values. The simplest method is the mean imputation [22], where the missing values are imputed by the mean of the observed values of the corresponding attribute. *K*-nearest neighbor imputation (KNNI) [23] is another simple idea, in which various weights depending on the distances between the neighbors and the incomplete pattern are designed to model the different effects of neighbors on the missing values. In [24], fuzzy *c*-means imputation (FCMI), the missing values are imputed according to the clustering centers generated by FCM and the distances between one object and all the centers. There are also other effective methods for dealing with incomplete data, such as the self-organizing maps (SOM) imputation [9], and the regression imputation [8], [12]. In particular, a fuzzy-based information decomposition (FID) method [25] was proposed recently to address the class imbalance and missing values problem simultaneously. In FID, the incomplete pattern is imputed and used to create synthetic patterns for the minority class to rebalance the training data. In [26], it is assumed that two batches extracted randomly in the same dataset have the same distribution. Then optimal transport distances are leveraged to quantify that criterion and turn it into a loss function to impute missing data values. Besides, practical methods are proposed to minimize these losses using end-to-end learning that can exploit parametric assumptions on the underlying distributions of values. Moreover, some works [27]–[31] are devoted to multiple imputations for missing values to model the uncertainty of the incomplete pattern caused by the lack of information. For example, in [29], a novel method, based on the generative adversarial network (GAN) framework [32], attempts to model the data distribution and then performs multiple imputations by drawing numerous times to capture the uncertainty of the interpolated values. These estimation-based methods assume reasonable correlations between missing and observed values, which are not always reliable.

In addition to the issues mentioned above in each type, most methods, such as deletion, model-based, and estimation methods, treat only missing values and do not consider the negative impact of missing values on the classification. Although machine learning-based methods can classify missing values, they do not consider the uncertainty caused by missing values in the process, therefore their performances are

sometimes not robust. Targeting on these limitations, the motivation of our work is based on the following three aspects.

1) Comparing to the deletion method, we investigate a method that does not remove any observed information since this information may be precious.

2) Comparing to model-based and estimation methods, we develop a method that does not introduce new uncertainties since statistical assumption will cause uncertainty.

3) Comparing to machine learning methods, we aim to improve the classification performance by modeling missing values with uncertainty and imprecision considered.

Based on the above analysis, we observe that these methods focus on the test set and assume that the training set is complete in most cases. When the training set is incomplete, the incomplete patterns are either imputed for completion or deleted directly. Moreover, these methods tend to directly model missing values, such as estimation strategies or model prediction. However, this would bring new uncertainties because estimated values can never replace the real world. In this case, this paper aims to answer an important question: how to improve the classification accuracy of incomplete data without losing information or introducing new uncertainty information in the presence of many missing values in both training and test sets? To drive such an answer, we design a new belief combination of classifiers (BCC) method for missing data based on evidence theory.

Evidence theory [33], [34] has been widely used in pattern classification since it is an efficient tool to characterize and combine uncertain and imprecise information, and it can well compromise (more or less) useful supplementary information provided by different sources in classifier fusion [35]–[37]. For instance, a classifier combination method depending on the concepts of internal reliability and relative reliability is proposed for classifier fusion with contextual reliability evaluation (CF-CRE) [38] based on evidence theory, where the internal reliability and relative reliability capture different aspects of the classification reliability. For non-independence classifiers, the literature [35] studies a method of combining other operators (i.e., parameterized *t*-norm) with the Dempster's rule, aiming to make their behavior between the Dempster's rule and the cautious rule. In [36], the transferable belief model (TBM) [39], an uncertain reasoning framework based on evidence theory, is employed to improve the performances of mailing address recognition systems by combining the outputs from several postal address readers (PARs). Reputably, the idea of group decision-making is introduced in [40] for reasoning with multiple pieces of evidence to identify and discount unreliable evidence automatically. The core is to construct an adaptive robust combination rule that incorporates the information contained in the consistent focal elements. These classifier fusion methods based on evidence theory have achieved satisfactory performances. However, they are designed for complete patterns. The uncertainty brought by the incompleteness can also be considered in the data process [6], [41], making one of the key ideas of this paper.

The proposed method, named belief combination of

classifiers (BCC), is able to characterize the uncertainty and imprecision caused by missing values both in training and test sets without imposing assumptive models. The main contributions of this paper cover the following aspects:

1) A novel classification method based on evidence theory is proposed, applicable for data with missing values, where incompleteness may exist in both training and query sets. To overcome the incompleteness in patterns, classifiers are trained by subsets of attributes, leading to sub-classification results.

2) Uncertainty and imprecision reasoning is used for missing values, used for training multiple classifiers. Afterwards, multiple evidential sub-classification results are combined for a final decision. Such designs with the context of uncertainty and imprecision make the classification results more robust.

3) An optimization function is designed to calculate the global weight to represent the reliability of each classifier, as well as the local weight obtained by mining attribute distribution characteristics is used to quantify the importance of observed attributes to the pattern classification. Moreover, abundant experiments are conducted, demonstrating the supremacy of BCC over many conventional methods in terms of classification results.

The remainder of this paper is organized as follows. After a brief introduction of evidence theory in Section II, the belief combination of classifiers (BCC) method for missing data is proposed in detail in Section III. Simulations results are presented in Section IV to evaluate the performance of BCC with different real datasets. Finally, Section V concludes the entire work and gives research perspectives.

II. BASICS OF EVIDENCE THEORY

Evidence theory, also known as Dempster-Shafer theory (DST) or the theory of belief functions, is firstly introduced by Dempster [33], then developed by Shafer in his *A Mathematical Theory of Evidence* [34]. Evidence theory is considered as an extended version of fuzzy set theory and has been widely used in data fusion [36], [39], [40], decision-making [37], [42], [43], clustering [44]–[46] and classification [47]–[49] applications. Evidence theory is a powerful framework for imprecise probability. It works with a discernment framework (DF) $\Omega = \{\omega_1, \dots, \omega_c\}$ consisting of c exclusive and exhaustive status of a variable, the uncertainty and imprecision degree of this attributes are expressed by subsets of the power-set 2^Ω with different basic belief assignments (BBAs), also called mass functions of belief.

In classification problems, the evidential class² of a pattern x under the power-set 2^Ω is mathematically defined as a BBA mapping $m(\cdot)$ from 2^Ω to $[0, 1]$, which satisfies the following conditions: $m(\emptyset) = 0$ and

$$\sum_{A \in 2^\Omega} m(A) = 1 \quad (1)$$

where $m(A) > 0$ represents the support degree of the object associated with the element A . In classification problems, A

may represent singleton class with $|A| = 1$, or meta-class with $|A| > 1$. Ω denotes all elements and $m(\Omega)$ is the degree of total ignorance. Total ignorance usually plays a neutral role in the fusion process because it characterizes the vacuous belief source of evidence.

In multiple classifier fusion processing, each classification result can be regarded as an evidence source represented by a BBA, and then the famous Dempster's rule is used to combine multiple BBAs, which is conjunctive, commutative and associative. The DS combination of two distinct sources of evidence characterized by the BBAs $m_1(\cdot)$ and $m_2(\cdot)$ over 2^Ω is denoted as $m = m_1 \oplus m_2$, and it is mathematically defined (assuming the denominator is not equal to zero) by $m(\emptyset) = 0$, and $\forall A \neq \emptyset \in 2^\Omega$ by

$$m(A) = (m_1 \oplus m_2)(A) = \frac{\sum_{\substack{B, C \in 2^\Omega \\ B \cap C = A}} m_1(B)m_2(C)}{1 - \sum_{\substack{B, C \in 2^\Omega \\ B \cap C = \emptyset}} m_1(B)m_2(C)} \quad (2)$$

with the whole conflicting factor smaller than 1. formally,

$$\sum_{\substack{B, C \in 2^\Omega \\ B \cap C = \emptyset}} m_1(B)m_2(C) < 1. \quad (3)$$

Evidential degree represented by a BBA m can be converted into Bayesian probabilistic values by belief function $Bel(\cdot)$, plausibility function $Pl(\cdot)$ and pignistic probability function $BetP(\cdot)$, respectively representing the upper bound, lower bound and intermediate value induced from a BBA. Herein, we introduce only the pignistic function on m , $BetP_m(\cdot)$ as follows, which is applied in the proposed BCC method.

$$BetP_m(A) = \sum_{B \in 2^\Omega, B \neq \emptyset} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)}. \quad (4)$$

Often, the probability degrees converted from BBA are applied for decision making over singletons in DF Ω .

There are also a few methods [31], [41], [50] based on evidence theory to deal with incomplete data. For example, a prototype-based credal classification (PCC) method is proposed in [31]. In PCC, the incomplete pattern is edited with c possible versions, for a c -class problem, to obtain c different classification results by a single classifier. Then the c results with different weights are fused to obtain the final classification of the incomplete pattern. Although PCC can characterize uncertainty caused by missing values, the estimation strategy also introduces new uncertainty information. Besides, it assumes that the training set is complete.

III. BELIEF COMBINATION OF CLASSIFIERS

This section presents a belief combination of classifiers (BCC) method based on evidence theory for classifying incomplete data. The BCC can faithfully make use of the observed data without imposing any assumption on missing values.

Given an S -dimensional dataset \mathcal{X} in class DF $\Omega = \{\omega_1, \dots, \omega_c\}$. The s th ($s \in \{1, \dots, S\}$) dimension of attributes, denoted

² In evidence theory, the term *evidential* refers to variables with both uncertainty and imprecision.

by \mathcal{X}^s , independently trains a classifier, denoted as C_s . For a pattern $\mathbf{x} \in \mathcal{X}$, based on the attribute knowledge in \mathcal{X}^s , classifier C_s provides a probability output of membership to classes ω_i , $i \in \{1, \dots, c\}$, denoted as $\mathbf{p}_s = \{p_s(1), \dots, p_s(c)\}$. To distinguish from the final classification result, we call the classification given from attribute subsets as sub-classification. Therefore, the final classification accuracy of the pattern can be improved by a discounted combination of the results of the S classifiers, where discounting indexes are defined by the reliable prior information in the training set. The discounting indexes depend strongly on two factors: the completeness of the data and the importance of the attribute, respectively referred to as the reliability weight and the importance weight.

For an incomplete query pattern $\mathbf{x} \in \mathcal{X}$ with H observed attributes, only H classifiers corresponding to the H observed attributes could provide reliable results among S classifiers.

The calculation of BCC mainly consists of three steps: evaluation of classifier reliability, evaluation of attributes' importance, and the global fusion of classifiers with decisions.

A. Classifier Reliability Calculation

Ideally, the output of each classifier would be perfect if it is completely consistent with the truth, i.e., the query pattern truly belongs to class ω_i , $p_s(i) = 1$ and $p_s(t) = 0$ for $t \neq i$. For some classifiers returning soft classification results, such as Bayesian classifier [51] and multi-layer perceptrons (MLP) [52], $p_s(i)$ is the returned value from the classifier, indicating the probability of the pattern belonging to the corresponding class. However, this is usually not true as all training subsets cannot train a classifier, and the classifier obtained by global optimization is not always suitable for a specific query pattern. Therefore, the reliability of each classifier C_s is not constant and should be reviewed. After that, the sub-classification outputs with different reliabilities are combined to make one final result for the query pattern. In the process of obtaining the reliability of each classifier, all patterns in the training set are employed. The reliability is estimated as follows:

Denote the reliability of classifier C_s by α_s , the reliability values of all classifiers are expressed by a $1 \times S$ matrix $\alpha = [\alpha_1, \dots, \alpha_S]$, subject to $\sum_{s=1}^S \alpha_s = 1$, $\alpha_s \in [0, 1]$. Denote the training set by $\mathcal{X}_{\text{train}}$, the output from classifier C_s on the s th attribute of pattern \mathbf{x} by $\mathbf{p}_s^x = [p_s^x(1), \dots, p_s^x(c)]$, where $p_s^x(i)$ represents the possibility of \mathbf{x} belonging to the class ω_i , and $p_s^x(i) \in [0, 1]$ with $\sum_{i=1}^c p_s^x(i) = 1$. The real class (ground truth) of \mathbf{x} is represented by $\mathbf{T}^x = [T^x(1), \dots, T^x(c)]$, satisfying

$$\mathbf{T}^x(i) = \begin{cases} 1, & \text{if } \mathcal{L}(\mathbf{x}) = \omega_i \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

For example, in a 3-class problem, \mathbf{T}^x is valued by $\mathbf{T}^x = [0, 1, 0]$ if the pattern \mathbf{x} belongs to the class ω_2 .

Based on the denotations above, an optimised reliability matrix α can be developed by

$$\arg \min_{\alpha} \left(\sum_{\exists s, s=1}^S \alpha_s \mathbf{p}_s - \mathbf{T} \right). \quad (6)$$

Therefore, a set of equations are constructed as follows:

$$\begin{cases} \left\| \sum_{\exists s, s=1}^S \alpha_s \mathbf{p}_s^1 - \mathbf{T}^1 \right\|^2 = \varphi_1 \\ \left\| \sum_{\exists s, s=1}^S \alpha_s \mathbf{p}_s^2 - \mathbf{T}^2 \right\|^2 = \varphi_2 \\ \vdots \\ \left\| \sum_{\exists s, s=1}^S \alpha_s \mathbf{p}_s^N - \mathbf{T}^N \right\|^2 = \varphi_M \end{cases} \quad (7)$$

$$\text{s.t. } \sum_{s=1}^S \alpha_s = 1, \alpha_s \in [0, 1] \quad (8)$$

where $\|\cdot\|$ represents a matrix norm. In our work, the Euclidean norm is accepted. In the calculation, the possibility of classification caused by a missing value (let's say the s th attribute of \mathbf{x} is missing) is set as 0, formally

$$\mathbf{p}_s^x = 0, \text{ if } \mathbf{x}(s) \text{ is missing.} \quad (9)$$

From the equations above, it can be induced that a small φ implies a better weight matrix α . Thus, the optimal discounting matrix α is estimated by

$$\alpha = \arg \min_{\alpha} \left(\sum_{n=1}^N \left\| \sum_{\exists s, s=1}^S \alpha_s \mathbf{p}_s^n - \mathbf{T}_n \right\|^2 \right) \quad (10)$$

where $n \in \{1, \dots, N\}$ with N denoting the number of patterns included in the training set $\mathcal{X}_{\text{train}}$. The solution of reliability matrix α is a classical constrained nonlinear least squares optimization problem, which can be estimated by the active-set algorithm [53]. In our work, we employ the sequential quadratic programming (SQP) method based on the constraints (8), with the estimation of the Lagrange-Hessen equation updated by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula (quasi-Newton method)³.

B. Attribute Importance Calculation

Many works have been devoted to exploring the importance of attributes in various data analysis domains [6]. In this work, the reliability matrix α of classifiers is estimated by a global optimal combination process, which may not apply to some specific patterns. The discounting factors of evidence are also related to the distributions of (missing) attributes of query patterns. As a simple example, Fig. 1 illustrates the distribution of the 1st and 5th attributes for the real dataset, named cloud, from the UCI (University of California, Irvine) repository (available at <http://archive.ics.uci.edu/ml>).

It can be reported from Fig. 1 that the distributions of the 5th attribute under the two classes are quite different. Therefore, the cross-entropy between the two classes in the

³ This paper focuses on the classification of incomplete data, which means that the reliability weight can be obtained by the proposed optimization strategy quickly based on the training set.

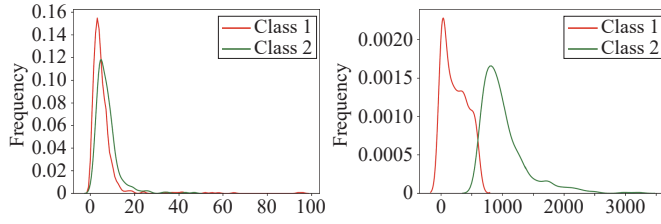


Fig. 1. The distribution of attributes in different classes with cloud dataset.

5th attribute is larger than that in the 1st attribute, indicating that the 5th attribute provides more prior information than the 1st one. Therefore, it is reasonable to assign different importance weights to different attributes so that some of the crucial attributes play more decisive roles while those who are not very useful for classification are less influential.

There are many methods to extract the correlation between attribute distribution and the class of patterns. A popular one, called Pearson product-moment correlation coefficient [54]⁴, is adopted to estimate the attribute weight β_s . Specifically, denote \mathcal{V}_s for the set of the s th observed attribute of all training patterns, and $\mathcal{V}_{s,j}$ for the value of j th pattern. To build the set \mathcal{L}_s of class labels for \mathcal{V}_s , we also define $\mathcal{L}_{s,j}$ to represent the label value of the attribute (value) $\mathcal{V}_{s,j}$, $\mathcal{L}_{s,j} \in \{1, \dots, c\}$. Obviously, \mathcal{V}_s and \mathcal{L}_s are objective. Thus, the Pearson product-moment correlation coefficient ρ_s between the s th attribute variable \mathcal{V}_s and the class variable \mathcal{L}_s is defined as follows:

$$\rho_s = \frac{\text{cov}(\mathcal{V}_s, \mathcal{L}_s)}{\sigma_{\mathcal{V}_s} \sigma_{\mathcal{L}_s}} = \frac{E[(\mathcal{V}_s - \mu_{\mathcal{V}_s})(\mathcal{L}_s - \mu_{\mathcal{L}_s})]}{\sigma_{\mathcal{V}_s} \sigma_{\mathcal{L}_s}} \quad (11)$$

where $\text{cov}(\mathcal{V}_s, \mathcal{L}_s)$ denotes the covariance between \mathcal{V}_s and \mathcal{L}_s , σ the standard deviation, μ the average value, and $|\rho_s|$ ($\rho_s \in [-1, 1]$) the degree of linear correlation between the s th attribute and the class. In other words, a smaller $|\rho_s|$ implies less prior information provided by the distribution of the s th attribute for classification, and vice versa. The importance β_s of the s th attribute is thereby defined by

$$\beta_s = \frac{|\rho_s|}{\sum_{s=1}^S |\rho_s|}. \quad (12)$$

It should be noted that the distribution of missing values for different patterns are different, thus, the local attribute weight also changes depending on the observed attributes of \mathbf{x} . This will be discussed in detail in the next part.

C. Global Combination of Classifiers and Final Decision

After determining the reliability and importance of each classifier, the S sub-classification results are globally combined for the final decision on the query pattern.

1) *Global Combination*: Due to the diversity of different classifiers' performances, some of the S classification results may strongly support different classes. In such a case, the final results merely demonstrate high conflicts if the Dempster's

rule directly combines these pieces of evidence. Thus, some discounting techniques are employed to weigh the impact of these sources of evidence in the global fusion process differently. Since the weights of different outputs are not only related to the reliability α of the classifier, but also the importance β of different attributes, the joint discounting factor γ_s of the s th output can be defined by

$$\gamma_s = \alpha_s \beta_s. \quad (13)$$

Moreover, since the number and relative position of missing attributes are different for query patterns, the discounting factor γ_s of s th output should adapt to such difference. To avoid such effects, we provide a benchmark to normalize discounting factors. For the specific query pattern \mathbf{x} with \mathcal{H} available attributes, $H \in \{1, \dots, S\}$, the relative discounting factor $\hat{\gamma}_h$ is defined by

$$\hat{\gamma}_h = \frac{\gamma_h}{\max_j \gamma_j}, \quad h, j = 1, \dots, H \quad (14a)$$

$$\max_j \gamma_j = \max\{\gamma_1, \dots, \gamma_H\}. \quad (14b)$$

With factor $\hat{\gamma}_h$ generally discounting the h th classification result, the discounting method proposed by Shafer [34] is employed for the fusion of differently discounted BBAs. According to the Shafer's method, the discounted BBA with discounting factor $\hat{\gamma}_h$, denoted by $\hat{\gamma}_h m_h$, is given by

$$\begin{cases} \hat{\gamma}_h m_h(A) = \hat{\gamma}_h m_h(A), \quad \forall A \in 2^\Omega \\ \hat{\gamma}_h m_h(\Omega) = \hat{\gamma}_h m_h(\Omega) + 1 - \hat{\gamma}_h. \end{cases} \quad (15)$$

In (15), the information is discounted by a belief degree on the total ignorance represented by all possible results, denoted by the DF Ω . Total ignorance Ω plays a neutral role in fusion process, thus, high conflicts among different classification results are well managed. With this proper mechanism of conflict management, the outputs can be fused directly by the Dempster's combination rule, defined in (2). In such a case, the Dempster's rule can offer a reasonable compromise between the specificity of the result and the level of complexity of the combination.

The combined global BBA m_x representing the final credal classification result for pattern \mathbf{x} is calculated by

$$m_x(A) = \bigoplus_{h \in \{1, \dots, H\}} m_x^h(A), \quad A \in 2^\Omega \quad (16)$$

where m_x^h denotes the discounted sub-classification results depending on the output of the s th attribute, and \bigoplus the Dempster's rule (2).

2) *Decision Making*: For a query pattern \mathbf{x} , its label $\mathcal{L}(\mathbf{x})$ is decided based on the probability in the frame of Ω , induced from the global BBA m_x with pignistic function $\text{Bet}P(m_x)$, reference (4). The decision is simple, solely the class with the highest possibility is selected, formally

$$\mathcal{L}(\mathbf{x}) = \arg \max_{A \in \Omega} (\text{Bet}P(m_x(A))). \quad (17)$$

The pseudo-code of BCC method is given in Algorithm 1 for convenience.

⁴ It is used to measure the similarity between two sets of variables. Other correlation coefficients, such as Spearman's correlation coefficient [55], Kendall's correlation [56] are also applicable.

Algorithm 1: Belief combination of classifiers

Require: Training set: $\mathcal{X}_{\text{train}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$; Test set: $\mathcal{X}_{\text{test}} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$; A basic classifier: C .

Ensure: Class decision results

- 1: **function** BCC training data $\mathcal{X}_{\text{train}}$, test data $\mathcal{X}_{\text{test}}$
- 2: Reconstruct the S training subsets, $\mathcal{X}_{\text{train}}^s$;
- 3: Train S corresponding basic classifiers, C_s ;
- 4: Construct the reliability optimization equations by (6) and (7);
- 5: Optimize classifier reliability matrix by (10) with constraint of (8);
- 6: Obtain the Pearson product-moment correlation coefficient ρ_s by (11);
- 7: Calculate local attribute importance by (12);
- 8: **for** $i = 1$ to N **do**
- 9: Obtain the classification results of \mathbf{x}_i with the trained classifiers;
- 10: Calculate the joint discounting factor γ_s by (13);
- 11: Extract the relative discounting factor $\hat{\gamma}_h$ by (14a) and (14b);
- 12: Discount these different evidence by (15);
- 13: Fuse these different evidence by (16);
- 14: Classify the query pattern \mathbf{x}_i and decide by (17).
- 15: **end for**
- 16: **return** class label
- 17: **end function**

D. Discussion

1) *Selection of Basic Classifiers:* Since we focus on improving the accuracy of classifying incomplete data rather than improving the classifier's performance, any classifier can be employed in principle. However, in the combination process, each observed attribute is considered an independent pattern. In such a case, the patterns used to train the basic classifier are 1-dimensional. Therefore, it is better to choose those general classifiers as benchmarks rather than those designed specifically for high-dimensional data. Most of the conventional classifiers are designed with the framework of probability, so the focal element A usually represents a specific class under the framework Ω in (7), thereby considering specific classes as an admissible solution of the classification. Nevertheless, there are also some classifiers based on the framework of belief functions [47], which can generate specific classes as well as the total ignorant class Ω . In such a case, $m(A)$ is the belief mass committed to the focal element (class) A in (15). Of course, $p^i = p(\omega_i)$ is equal to $m(A)$ if the focal element A represents the specific class ω_i in (7).

2) *Selection of Combination Rule:* For the selection of combination rule, it is known that numerous combination rules exist dealing with different kinds of evidence resources and conflicts. However, our goal is not to propose a new combination rule but to improve the classification performance by reasonably characterizing the uncertainty and imprecision caused by missing values based on evidence theory when classifying incomplete data. A number of

experiments prove that this is feasible. In fact, many combination rules have been proposed when dealing with conflicting evidence, such as Smet's rule [57], Yager's rule [58], Dubois-Prade (DP) rule [59] and proportional conflict redistribution (PCR) rules [60].

The property of associativity is important in our application since the fusion of multiple evidences are calculated in a sequential way in which the order makes no difference. The above rules are not associative, which makes them less attractive in applications. High conflict within evidence sources is another issue in information fusion which often makes results hardly reliable. By considering these two issues, in our method, the sources of evidence are firstly modified to prevent (possible) highly conflicts, and then combined by the Dempster's rule to determine the final class of query patterns. Based on the Shafer's discounting method, the whole conflicts are distributed to total ignorance Ω due to normalization when different evidences are highly conflicting or in some special low-conflict situations.

IV. EXPERIMENT APPLICATIONS

To validate the effectiveness of BCC method confronting missing data, a number of benchmark datasets are employed to compare with several other conventional methods based on four common criteria: 1) accuracy (AC); 2) precision (PE); 3) recall (RE); 4) F1-measure (F1) [61].

A. Methods for Comparison

The classification performance of this new BCC is evaluated by comparisons with several other conventional methods including mean imputation (MI) [22], K -nearest neighbors imputation (KNNI) [23], fuzzy c -means imputation (FCMI) [24], prototype-based credal classification (PCC) [31], fuzzy-based information decomposition (FID) [25], generative adversarial imputation nets (GAIN) [29] and batch Sinkhorn imputation (BSI) [26]. In MI, the missing values in the training set are replaced by the average values of the same class, and the missing values in the test set are imputed by the means of the observed values of the position in the training set. In KNNI, the incomplete pattern in the training set is estimated by the KNNs with different weights depending on the distance between the pattern and the neighbors in the same class, and the incomplete pattern in the test set is estimated by the global KNNs. Since the training set is complete by default in FCMI, PCC, and FID, we thereby use the average values of the class to impute the missing values in the training set, similar to MI. In FCMI, the missing values in the test set are imputed according to the clustering centers generated by FCM and the distances between the object and the centers. In PCC, the incomplete pattern in the test is imputed with c possible versions for a c -class problem, while the centers of c classes are obtained from the training set. In FID, the missing values in the test set are estimated by taking into account different contributions of the observed data. In GAIN, generative adversarial nets are trained to estimate the missing values in the test set. In BSI, the missing values in the test set are imputed by minimizing optimal transport distances between quantitative variables. For all parameters of the compared

methods, we use the default values as in the original papers.

Different from the above methods, the proposed BCC considers each attribute as an independent source, and the collection of each attribute is considered as a subset. Afterward, each subset trains a classifier independently, which allows each observed attribute to provide a sub-classification result for the query pattern. Then these sub-classification results with different weights (discounting factors) are used to provide supplementary information to jointly determine the final classes of query patterns.

B. Basic Classifiers

In our simulations, K-NN technique [62], evidential K -nearest neighbor (EK-NN) [47] and Bayesian classifier (Bayes) [21] are employed as the basic classifiers to generate pieces of evidence. In the Bayesian classifier, Gaussian distributions are assumed for each attribute. For the parameters in the classifier, we apply the default values identical to the original papers. The outputs of EK-NN are BBAs consisting of the singletons and the total ignorance, and the outputs of K-NN and Bayesian classifiers are probability values. Both BBAs and probability can be directly applied to the optimal combination in the BCC method as explained in Section III-D.

C. Benchmark Datasets

Twelve datasets from the UCI repository are used to evaluate the effectiveness of BCC with comparison to main conventional methods. The basic features of these datasets are shown in Table I, including number of classes (#Class), number of attributes (#Attr.) and number of instances (#Inst.). The size of each dataset is defined by the number of its attributes (#Attr.) and the number of instances (#Inst.).

TABLE I
BASIC INFORMATION OF THE USED DATASETS

Datasets	#Class	#Attr.	#Inst.
Iris (Ir)	3	4	150
Cleve (Cl)	2	13	296
ILPD (IL)	2	10	583
Abalone (Ab)	18	8	4139
Vehicle (Ve)	4	18	946
Heart (He)	2	13	270
White wine quality (Ww)	7	11	4898
Wine (We)	3	13	178
Cloud (Co)	2	10	1024
Parkinsons (Pa)	42	21	5875
Knowledge (Kn)	4	5	403
Red wine quality (Rw)	6	11	1599

In the experiments, each attribute independently reconstructs a subset and trains a basic classifier. In order to demonstrate the effectiveness under different incompleteness levels, in the training and test sets, we assume that each pattern has ϕ missing (unobserved) values with missing-

completely-at-random mechanism. In experiments, different values of ϕ are employed to verify the performance of the BCC method.

D. Performance Evaluation

We use the simplest 2-fold cross validation. Since the size of the training and test sets are equal, all patterns can be respectively used for training and testing on each fold. The program is randomly run 10 times⁵, and the performance of BCC is shown for various ϕ , denoting the number of missing values for each training and test pattern, as reported in Tables II–VII. Specifically, the accuracy values (AC) for the compared methods, based on K-NN, EK-NN and Bayesian basic classifiers⁶, are respectively reported in Tables II, IV and VI. Other indexes, PE, RE, F1, are recorded in Tables III, V, and VII, with K-NN, EK-NN and Bayesian basic classifiers integrated respectively. In Table III, taking the CI dataset as an example, PE is the average value with different ϕ ($\phi = 4, 6, 8$ in Table II), based on the K-NN classifier. In addition, the average histograms for ϕ with different datasets and basic classifiers are plotted in Fig. 2 to compare the effectiveness of BCC more intuitively. y -axis represents the accuracy, calculated by the average of the outputs with different ϕ in Tables II, IV and VI.

From these results, it can be observed that the BCC method generally provides better results compared to other conventional methods in most cases. Moreover, these results support us to make the following analysis.

As typical single imputation strategies, MI, KNNI, and FCMI predict possible estimations of missing values based on different mechanisms, but such estimations may not be reasonable enough. For example, in KNNI, similar patterns (neighbors) are employed to impute missing values. In this case, the selection of similarity measure norm is an essential process. If an inappropriate measure is chosen, unsatisfactory results are often obtained. Moreover, the disadvantage of direct modeling of missing values is that it is impossible to avoid bringing new uncertainties because estimation can never replace the real world. Furthermore, only modeling missing values is insufficient because the uncertainty caused by missing values can also negatively affect the classifier's performance. Therefore, we can see that the results obtained by these methods are often not satisfactory.

PCC and GAIN are multiple imputation strategies, which means that a missing value may be estimated as multiple versions to characterize the uncertainty. This is an improvement and reasonable in some ways, however, multiple estimations are not always better than single estimation strategies [63]. In particular, as an imputation strategy, GAIN models the uncertainty of missing values but still does not characterize the uncertainty and imprecision in the model and results from a classification perspective. On the other hand, as an evidence theory-based method, PCC is similar to the proposed BCC method in characterizing the uncertainty and

⁵ All results demonstrated in this paper are average values.

⁶ The differences between the chosen classifiers are beyond this paper.

TABLE II
THE ACCURACY (AC) OF DIFFERENT METHODS WITH K-NN CLASSIFIER (%)

Datasets	ϕ	MI	KNNI	FCMI	PCC	FID	GAIN	BSI	BCC
Cl	4	65.98±2.59	65.54±3.82	64.05±4.85	67.57±2.78	62.57±3.33	62.43±3.04	68.24±2.65	75.81 ±5.15
	6	65.03±2.38	64.53±5.26	64.39±2.89	68.04±3.09	57.36±4.24	58.78±3.48	64.18±3.24	73.65 ±2.66
	8	62.16±4.16	63.85±3.34	61.96±5.01	68.45±4.19	54.05±3.70	58.10±4.52	68.24±3.16	70.14 ±3.89
IL	4	68.66±1.31	60.79±6.05	64.55±2.57	67.02±1.85	66.71±2.80	66.91±1.83	64.38±2.14	71.92 ±3.17
	6	70.45±0.82	58.73±3.15	57.71±5.82	65.86±2.06	64.97±3.67	62.79±0.67	63.01±1.35	71.09 ±1.21
	8	69.38±1.42	62.16±0.73	54.97±4.56	63.05±3.71	65.79±2.70	65.34±1.33	61.98±1.92	70.80 ±1.37
Co	5	80.06±1.19	89.06±0.55	90.47±1.08	92.52±0.95	68.15±1.71	92.66±0.92	93.17±1.26	93.44 ±0.90
	7	72.99±1.38	85.55±0.07	85.20±1.82	88.12±1.56	64.97±3.67	85.70±1.95	87.89±2.03	90.27 ±1.23
	9	64.99±1.66	78.32±0.83	76.88±2.51	80.16±1.88	65.79±2.70	74.41±1.84	71.50±2.13	82.11 ±2.34
Ww	1	43.76±0.64	44.00±1.36	43.79±0.96	43.85±0.80	43.45±1.73	45.41±0.78	47.63±1.52	48.67 ±0.58
	2	42.79±0.56	42.60±0.68	41.22±0.97	42.94±1.57	41.79±0.75	43.54±0.68	45.77±0.69	48.06 ±0.51
	3	42.23±0.89	41.83±0.73	40.66±0.98	41.20±0.95	41.56±0.57	42.87±1.04	44.54±0.64	46.80 ±0.46
He	6	62.96±2.21	62.59±3.67	61.63±3.93	69.48±3.79	56.15±4.56	61.62±1.91	73.07±3.46	73.19 ±5.68
	8	59.15±3.22	62.96±2.10	63.63±2.50	68.81±4.33	54.52±3.22	60.44±2.59	70.68±4.51	71.85 ±3.19
	9	58.33±2.91	61.11±3.67	62.81±3.98	66.59±3.60	51.04±2.76	56.89±5.11	67.42±3.13	69.78 ±3.53
We	3	61.35±3.22	68.54±5.21	63.03±4.81	76.29±4.57	56.74±4.41	68.31±2.90	91.01±3.43	91.57 ±3.05
	5	55.17±3.51	71.91±4.90	64.04±6.51	75.28±4.71	49.10±5.14	68.76±4.99	88.76±4.37	90.00 ±3.57
	7	49.66±3.29	67.64±4.16	58.88±4.24	73.15±3.84	47.30±5.86	66.97±5.19	83.14±3.45	86.18 ±4.21
Ir	1	77.53±3.86	92.53±2.42	91.87±2.84	93.33±3.27	78.00±5.38	92.00±3.77	89.33±3.42	94.93 ±1.74
	2	60.20±2.53	89.07±4.75	87.47±3.34	89.87 ±2.68	61.87±6.98	89.13±3.99	82.66±3.36	89.33±3.13
	3	46.40±3.42	74.93±3.18	76.27±4.90	75.47±7.42	53.47±7.53	75.13±2.59	62.66±2.14	76.73 ±3.82
Ve	8	41.26±1.40	49.57±2.13	46.24±2.41	47.94±2.34	39.08±1.87	52.50±2.67	54.13 ±2.78	53.05±2.26
	10	36.78±1.98	46.49±2.53	40.73±2.23	45.01±2.19	35.56±2.36	51.67±3.92	48.93±2.48	51.99 ±2.57
	16	26.89±1.17	38.53±2.13	33.74±2.49	34.92±1.83	29.08±1.98	37.21±1.83	34.04±2.75	39.23 ±2.95
Kn	2	61.11±2.84	61.22±0.29	63.27±3.28	61.88±3.45	45.00±3.15	66.71±2.47	52.47±2.89	67.08 ±3.56
	3	50.54±2.55	53.63±0.57	50.99±5.36	48.66±2.15	38.56±4.40	59.70±4.08	42.07±2.21	60.25 ±3.23
	4	38.02±2.85	44.05±2.62	38.91±3.75	38.71±3.80	36.24±3.48	47.00 ±2.61	41.08±3.04	45.94±3.83
Rw	1	47.19±1.32	47.79±1.49	47.15±1.75	48.20±1.60	47.01±0.60	49.17±0.80	52.12±2.47	52.17 ±1.80
	2	46.69±1.39	45.96±1.06	46.86±1.54	47.45±1.12	45.39±1.39	48.07±1.05	51.36±1.68	52.12 ±1.19
	3	45.11±1.01	44.96±0.47	45.00±2.13	46.71±1.35	45.04±2.09	47.92±2.06	50.47±2.34	51.54 ±1.58
Ab	2	21.07±0.36	22.75±0.58	22.91±0.98	22.86±1.49	19.84±0.65	22.53±0.35	23.14±0.45	23.75 ±0.63
	4	18.60±0.56	20.95±0.33	22.38±0.69	21.72±0.84	18.18±0.65	21.37±0.61	21.78±0.26	22.62 ±0.27
	5	17.98±0.48	20.69±0.07	21.32±1.10	21.14±0.78	17.30±1.15	20.68±0.55	20.19±0.19	22.38 ±0.24
Pa	5	40.30±0.94	54.81±1.01	45.13±1.80	63.82±2.35	32.59±0.70	44.45±1.13	74.77±1.41	82.05 ±0.77
	7	31.11±0.86	48.62±0.25	36.49±0.28	61.94±0.51	24.78±0.82	36.43±1.87	66.30±2.04	75.85 ±1.45
	9	24.13±1.54	43.57±0.25	29.62±0.86	56.74±0.63	19.19±0.43	30.90±3.61	52.45±2.49	65.71 ±2.80

imprecision caused by missing values. However, in PCC, the use of class centers as a benchmark for estimating missing values is unreasonable and does not assess the reasonableness and necessity of multiple estimations. Therefore, it can characterize the imprecision in the results, but the performance is still not good enough.

As the latest work, FID, GAIN, and BSI have proposed some feasible solutions from data distribution but only partially address the problem of modeling incomplete data. For example, GAIN and BSI, as model-based methods, are dedicated to perfectly approximating the real world, which is

practically impossible. In particular, BSI assumes that two random batches from the same dataset should follow the same distribution. Still, in some scenarios, the distribution of the data itself is hard to estimate. If the data distribution is not modeled precisely enough, the reasonableness of the estimations is questionable. FID is a pioneering work that works on both incomplete and imbalanced data classification. In FID, the imbalance is also considered incomplete data, so its essence is still a classification of incomplete data. However, the classification results are also less than ideal because the process does not reasonably characterize the

TABLE III
THE PE, RE AND F1 OF DIFFERENT METHODS WITH K-NN CLASSIFIER (%)

Datasets	MI			KNNI			FCMI			PCC			FID			GAIN			BSI			BCC		
	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1
Cl	64.60	64.16	63.72	64.71	64.73	64.60	63.88	63.50	62.87	68.24	68.18	67.82	58.01	57.54	57.01	60.05	59.80	59.39	68.23	67.02	66.36	73.60	72.50	72.39
IL	84.02	65.03	62.92	83.26	83.94	83.50	83.62	84.81	83.82	86.35	86.22	86.23	54.86	53.16	52.49	64.11	53.07	52.39	57.27	51.06	50.48	88.15	87.96	88.00
Co	60.41	59.07	57.78	62.62	61.86	61.04	62.18	62.14	61.95	68.09	68.08	67.85	53.55	53.19	51.91	68.65	63.16	63.34	68.52	63.44	63.09	72.16	71.67	70.35
Ww	22.58	18.84	19.06	23.48	19.39	19.83	23.35	19.87	20.22	24.09	20.04	20.54	23.13	18.71	18.92	25.82	19.87	20.20	24.76	23.16	23.42	27.91	24.09	24.38
He	60.41	59.07	57.78	62.62	61.86	61.04	62.18	62.14	61.95	68.09	68.08	67.85	53.55	53.19	51.91	59.01	58.75	58.17	71.28	70.77	69.56	72.16	71.67	70.35
We	62.15	55.73	53.69	68.18	68.62	67.03	61.62	61.57	60.01	75.61	74.20	73.91	49.88	49.90	45.64	68.11	66.78	65.65	89.45	88.85	88.79	89.65	89.79	89.36
Ir	80.73	61.64	58.81	85.96	85.71	85.39	86.07	85.38	85.29	86.86	86.56	86.31	70.01	64.87	64.74	86.60	86.07	86.44	79.76	78.50	78.44	87.45	87.20	86.86
Ve	41.34	34.86	29.70	45.10	45.32	43.27	40.21	40.82	37.97	42.48	42.79	42.09	35.48	34.84	34.28	47.73	47.63	46.89	47.89	47.30	47.22	47.80	48.47	47.37
Kn	61.85	43.86	42.01	53.22	50.60	51.13	54.65	49.26	49.21	49.30	49.51	47.98	41.01	35.66	33.28	58.14	55.13	55.09	48.42	42.78	42.44	60.13	55.05	55.44
Rw	23.52	20.80	20.54	26.22	21.96	22.42	23.20	21.44	21.24	23.74	21.77	21.76	22.75	19.98	19.67	26.52	21.30	21.10	26.78	21.47	21.09	27.03	22.40	21.53
Ab	12.41	8.18	7.00	15.45	17.31	15.35	15.28	14.97	13.68	15.85	17.58	15.69	11.92	10.44	10.23	16.14	14.01	13.56	16.11	14.46	13.32	17.43	15.47	15.01
Pa	42.68	32.00	32.95	52.50	49.32	49.79	44.27	37.09	38.32	63.17	60.93	61.17	31.29	25.46	26.23	42.42	37.29	37.81	64.42	64.21	64.38	75.44	74.93	74.43

TABLE IV
THE ACCURACY (AC) OF DIFFERENT METHODS WITH EK-NN CLASSIFIER (%)

Datasets	ϕ	MI	KNNI	FCMI	PCC	FID	GAIN	BSI	BCC
Cl	4	63.72±1.93	62.70±3.89	66.76±2.63	67.03±3.86	57.36±4.21	60.54±3.26	70.94±4.76	75.00±1.72
	6	63.24±4.67	62.57±3.04	60.54±5.50	70.54±5.09	54.46±6.49	54.18±7.63	60.81±4.45	72.16±3.59
	8	59.32±3.28	66.22±4.35	66.22±5.47	69.59±4.25	52.57±3.64	54.72±4.16	48.86±3.28	70.14±4.04
IL	4	70.82±1.36	60.34±6.02	67.33±3.41	68.42±2.51	67.67±5.40	69.93±1.94	70.89±2.13	71.30±3.04
	6	70.31±1.14	59.45±3.29	60.48±6.72	66.64±1.89	69.28±1.42	69.86±1.47	70.27±1.48	70.62±1.77
	8	67.33±2.17	63.15±2.15	54.45±6.22	63.15±0.99	67.71±2.24	65.13±4.63	69.86±1.98	69.52±1.43
Co	5	80.51±1.01	89.10±1.56	90.94±1.16	92.97±1.32	74.10±3.40	91.40±2.45	91.60±1.63	94.39±0.33
	7	74.28±1.31	84.18±0.85	85.16±1.04	88.63±0.72	64.47±2.07	86.05±2.75	87.69±2.44	89.06±1.85
	9	63.52±3.11	76.50±1.77	77.38±1.16	79.57±0.92	61.35±2.89	75.35±2.11	66.60±2.87	79.98±1.02
Ww	1	49.24±0.86	49.45±0.76	49.20±1.15	50.09±0.80	49.14±0.83	49.96±1.14	50.67±1.23	51.57±0.54
	2	46.08±0.36	46.15±1.12	44.91±0.77	46.18±1.19	45.75±0.90	47.21±0.71	48.28±1.08	47.52±0.77
	3	43.35±0.55	42.97±1.26	40.90±1.28	42.85±1.06	44.37±0.92	44.56±1.35	44.12±0.98	45.02±0.77
He	6	61.26±2.64	61.73±1.86	60.44±5.46	72.19±4.58	54.00±4.42	58.81±3.03	66.67±3.57	72.44±2.48
	8	60.07±2.10	62.47±1.54	65.33±5.68	71.48±3.33	53.78±6.21	53.33±4.23	68.89±4.26	71.56±4.04
	9	57.85±1.24	62.96±2.96	63.11±5.77	70.81±3.50	51.70±6.04	56.88±5.50	55.56±5.43	71.26±3.12
We	3	60.34±3.19	67.42±2.25	66.74±3.86	78.20±2.59	54.49±7.31	65.84±3.85	89.88±4.17	91.69±3.33
	5	54.38±2.86	65.17±2.25	62.70±4.38	72.58±3.86	49.89±5.16	71.23±5.42	87.32±5.31	89.66±3.08
	7	50.22±2.87	65.17±4.49	56.18±5.02	81.12±3.22	44.04±2.95	65.16±4.27	86.21±3.09	86.97±3.04
Ir	1	77.47±2.28	92.00±0.00	90.93±4.04	93.33±4.81	74.53±7.38	92.53±2.42	93.33±1.53	93.87±2.02
	2	62.27±5.87	92.44±4.07	89.33±3.40	91.47±1.19	58.13±6.33	91.99±1.88	89.33±1.45	90.40±1.46
	3	48.13±4.77	75.11±2.78	75.47±3.07	74.40±5.28	51.07±6.45	91.46±1.52	74.66±1.79	65.87±9.02
Ve	8	40.71±1.52	50.43±1.93	45.15±3.29	48.37±3.14	37.64±2.78	52.60±3.29	54.48±2.46	53.14±1.06
	10	38.37±1.34	47.83±1.52	44.73±4.59	47.66±2.71	35.25±2.94	50.82±2.00	47.46±3.27	51.39±1.08
	16	28.09±1.57	35.15±0.36	34.14±0.76	35.70±2.61	27.71±2.08	38.39±2.09	31.91±2.90	39.20±1.50
Kn	2	60.50±2.69	64.19±3.30	63.27±2.20	60.69±4.12	41.34±3.28	67.29±2.69	50.00±2.98	69.21±2.63
	3	51.53±2.52	53.80±2.81	50.89±2.49	50.50±2.62	37.57±4.17	58.31±2.99	42.57±2.46	62.18±3.55
	4	37.67±1.69	45.21±5.08	38.51±4.57	37.62±1.64	33.42±1.96	39.10±3.19	38.11±3.25	46.44±5.72
Rw	1	51.89±0.89	53.04±1.09	51.90±1.14	51.15±1.88	49.61±1.66	52.92±1.98	52.76±1.87	53.25±1.69
	2	48.15±1.37	50.96±1.82	47.52±1.21	49.33±1.21	46.79±1.50	51.37±0.57	51.25±0.81	51.42±1.91
	3	45.88±1.23	47.42±3.34	46.05±1.62	48.88±1.30	45.09±1.95	49.15±1.52	50.37±1.19	50.96±2.42
Ab	5	19.55±0.91	22.37±1.17	23.24±0.50	23.12±0.85	19.17±1.05	22.32±0.85	22.27±0.98	23.46±0.99
	7	18.04±0.45	20.40±1.16	22.54±1.10	22.20±0.48	17.30±0.81	20.93±0.79	20.91±1.01	22.75±0.97
	9	17.68±0.38	20.38±0.53	21.58±1.14	21.28±0.34	16.75±1.06	19.84±1.37	19.85±0.77	21.67±1.20
Pa	5	39.67±0.83	58.29±1.27	44.70±0.91	63.86±0.93	40.99±0.93	54.32±2.18	74.33±1.74	84.01±0.95
	7	30.41±1.02	52.58±0.21	35.92±1.46	59.81±1.27	30.76±0.78	45.60±1.29	65.75±1.95	77.37±1.72
	9	24.24±1.37	47.46±0.19	30.44±1.01	56.94±1.06	23.83±0.62	38.81±4.97	54.39±2.36	68.80±2.44

TABLE V
THE PE, RE AND F1 OF DIFFERENT METHODS WITH EK-NN CLASSIFIER (%)

Datasets	MI			KNNI			FCMI			PCC			FID			GAIN			BSI			BCC		
	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1
Cl	62.64	62.01	61.23	63.99	63.71	63.51	64.85	64.62	64.08	69.17	68.77	68.60	55.23	54.38	50.25	56.98	51.12	50.17	66.19	66.02	66.02	73.03	72.00	71.75
IL	52.01	50.78	45.84	58.37	59.85	57.18	56.04	56.10	52.40	58.99	59.04	58.73	51.27	50.91	47.28	50.06	44.97	57.87	53.14	50.65	50.96	59.29	53.16	48.86
Co	84.06	65.57	63.33	82.44	82.42	82.33	83.72	85.12	84.02	86.56	86.01	86.20	65.04	64.28	62.87	83.22	83.17	82.62	80.54	80.51	80.03	86.89	86.81	86.80
Ww	33.18	21.43	22.63	24.95	23.51	23.55	29.80	21.49	22.44	38.08	23.84	25.37	32.87	20.04	20.76	21.17	22.18	22.31	25.77	25.08	24.65	39.10	24.26	25.51
He	60.00	58.54	57.25	62.60	62.65	62.26	62.98	62.94	62.57	71.61	71.53	71.27	51.02	51.57	44.51	59.36	58.94	57.28	63.05	61.24	61.76	72.56	71.66	71.39
We	60.35	54.74	51.27	65.38	63.06	61.01	61.90	61.60	59.97	77.18	76.26	76.04	48.81	48.96	44.36	66.19	63.20	63.26	88.23	88.26	87.94	90.29	90.15	89.76
Ir	81.45	61.99	59.42	86.78	86.25	86.29	86.32	85.32	85.19	86.98	86.39	86.24	69.20	61.25	60.31	88.18	88.01	87.35	85.69	85.62	85.58	84.46	84.03	83.70
Ve	42.38	35.18	29.90	44.32	44.67	43.16	41.40	42.08	39.25	43.65	44.26	43.49	36.09	33.92	30.07	47.20	47.84	47.48	47.14	47.45	47.13	47.73	48.41	47.23
Kn	59.07	44.14	42.59	54.34	52.74	52.72	53.50	48.72	49.01	49.59	50.19	48.30	37.78	34.12	30.50	57.93	57.89	57.27	41.35	41.24	41.01	59.33	56.10	56.34
Rw	24.40	22.84	22.63	24.38	24.68	24.52	23.42	23.70	23.15	23.52	24.10	23.59	22.62	21.28	20.91	23.13	22.95	22.39	25.94	23.62	23.93	26.19	24.78	24.39
Ab	11.33	7.58	5.98	13.88	15.92	14.22	14.09	14.92	13.17	15.48	18.05	15.91	11.95	9.55	8.26	13.20	11.97	12.94	12.91	11.73	12.82	15.59	15.07	14.23
Pa	42.26	31.48	32.53	57.28	51.93	54.43	44.60	36.96	38.29	62.59	60.38	60.59	35.85	31.78	31.02	46.78	45.48	45.23	64.90	64.23	64.24	77.03	76.15	75.81

TABLE VI
THE ACCURACY (AC) OF DIFFERENT METHODS WITH BAYESIAN CLASSIFIER (%)

Datasets	ϕ	MI	KNNI	FCMI	PCC	FID	GAIN	BSI	BCC
Cl	4	77.36±1.96	76.89±2.59	75.68±2.85	77.64±3.29	74.93±2.03	78.92±2.63	76.35±2.38	79.32 ±3.58
	6	75.95±1.56	75.95±3.78	75.61±1.92	75.14±3.23	72.50±3.56	75.56±3.14	75.44±3.18	76.42 ±2.56
	8	71.82±5.28	70.95±4.61	71.08±3.68	72.77±6.04	69.19±3.39	72.59±5.13	68.24±4.31	73.65 ±3.15
IL	4	70.58±1.87	59.86±3.53	61.95±3.33	65.58±2.91	47.50±4.05	52.32±3.87	51.63±3.75	70.91 ±2.09
	6	69.88±1.46	63.97±2.73	64.08±5.50	66.54±2.55	46.61±4.78	50.48±2.80	49.31±2.66	69.81 ±2.83
	8	66.03±2.56	66.16±2.50	65.17±4.46	64.49±2.61	42.57±2.95	43.56±3.59	47.26±3.13	66.42 ±3.21
Co	5	82.25±2.50	89.69±1.28	87.32±1.52	89.30±1.66	69.26±2.02	82.50±1.89	88.28±2.38	91.43 ±1.17
	7	67.16±1.87	85.59±0.79	84.45±1.97	86.70±1.44	65.68±3.83	77.97±1.98	83.98±1.57	88.18 ±1.54
	9	62.71±0.93	77.97±1.84	77.58±2.39	78.11±1.68	54.41±7.15	75.19±2.81	63.08±4.13	78.39 ±2.61
Ww	1	44.69±0.48	44.34±1.12	43.74±1.01	44.48±0.63	43.82±1.19	44.54±1.09	43.87±0.83	49.99 ±1.52
	2	44.36±0.69	43.81±1.13	44.36±0.74	43.53±1.31	43.37±0.90	44.18±0.64	44.18±0.97	47.97 ±0.27
	3	44.24±0.71	42.38±1.09	43.30±0.88	44.12±0.92	42.93±1.10	43.72±1.92	42.38±1.04	45.99 ±1.00
He	6	77.85±1.69	75.26±3.65	77.56±3.65	76.22±4.14	73.93±3.77	78.01±1.12	74.07±3.57	78.74 ±5.26
	8	73.41±3.62	73.19±6.44	65.93±8.28	74.52±3.56	69.78±6.00	74.26±2.89	72.59±4.63	74.81 ±2.70
	9	68.30±4.19	69.48±4.36	65.26±6.76	71.41±2.59	63.41±3.88	73.52 ±3.49	69.62±3.71	71.78±3.80
We	3	90.96±3.16	92.81±2.59	91.91±3.55	91.80±3.00	74.38±5.34	92.38±2.38	94.38 ±2.47	93.22±2.37
	5	82.92±3.75	87.87±5.41	86.18±4.17	89.44±2.32	65.47±6.18	88.88±3.17	90.77 ±4.33	89.66±3.74
	7	71.80±4.55	82.02±5.45	79.44±4.71	85.73±3.82	54.94±8.85	88.33±2.19	87.24±3.90	87.53 ±4.25
Ir	1	73.27±4.68	91.73±2.89	91.87±2.89	93.20±3.79	64.00±4.02	91.20±3.34	94.89±2.61	95.20 ±2.89
	2	55.47±4.17	87.47±4.07	89.20±3.69	89.47±4.38	60.13±5.98	89.19±4.75	85.33±4.47	90.27 ±3.33
	3	38.87±5.09	70.93±5.53	74.27±7.57	72.93±6.10	48.00±4.87	86.86±8.25	54.66±5.48	76.93 ±2.52
Ve	8	44.05±2.00	47.80±5.57	43.45±2.36	46.55±2.63	38.96±3.63	43.49±2.95	38.29±3.09	47.20 ±2.91
	10	40.86±2.40	44.97±0.49	43.17±3.53	45.63±2.74	36.31±2.26	44.82±3.07	41.60±2.82	45.67 ±3.12
	16	28.06±2.33	35.51±1.42	34.35±3.46	35.63±2.41	27.90±2.21	35.80±3.13	36.40±3.73	36.83 ±2.05
Kn	2	63.61±2.80	65.45±3.20	63.22±3.04	62.87±2.79	42.92±2.98	63.06±1.65	56.93±2.80	66.24 ±2.98
	3	51.93±2.32	54.84±5.62	52.18±3.20	51.63±2.84	37.87±3.04	51.78±1.54	47.52±2.72	55.40 ±2.26
	4	39.28±3.12	42.34±2.28	39.65±2.87	37.62±3.28	35.35±4.28	42.61±3.52	40.59±2.93	43.63 ±2.32
Rw	1	53.26±1.43	50.71±1.39	51.76±2.34	53.12±1.31	47.93±1.66	52.90±3.13	53.12±1.94	55.07 ±2.59
	2	51.81±1.73	47.42±1.66	51.31±1.82	51.17±2.67	48.40±1.31	50.65±3.14	48.87±3.22	55.43 ±2.03
	3	51.54±1.33	50.00±3.25	52.01±1.28	51.23±2.58	46.99±2.29	40.82±2.13	50.12±2.84	53.08 ±1.44
Ab	2	24.88±0.55	23.43±1.16	22.96±0.94	25.20±0.48	23.16±1.05	23.47±1.80	23.33±1.45	27.31 ±1.42
	4	21.96±0.71	23.17±1.78	23.30±0.41	23.12±0.90	21.33±1.42	21.81±1.38	21.88±1.07	26.25 ±1.07
	5	20.84±0.37	23.78±1.31	23.94±1.00	22.37±0.34	18.53±1.76	21.38±3.59	18.98±1.87	26.28 ±0.55
Pa	5	66.63±0.78	77.82±0.75	68.43±0.95	77.77±0.55	45.05±1.42	54.18±2.51	61.13±1.34	78.13 ±1.80
	7	55.95±0.55	69.63±0.56	57.78±0.78	71.55±0.54	37.79±1.40	45.13±2.36	51.83±1.94	72.86 ±2.32
	9	46.09±0.51	61.33±1.47	51.52±0.84	66.99±0.83	30.87±1.44	38.33±2.64	43.70±1.87	68.41 ±1.86

TABLE VII
THE PE, RE AND F1 OF DIFFERENT METHODS WITH BAYESIAN CLASSIFIER (%)

Datasets	MI			KNNI			FCMI			PCC			FID			GAIN			BSI			BCC		
	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1	PE	RE	F1
Cl	75.47	75.02	74.72	74.74	74.57	74.45	74.28	74.13	73.91	75.17	74.97	74.84	72.30	72.17	71.93	76.70	75.58	75.61	74.43	73.36	73.29	77.06	75.84	75.83
IL	58.96	54.14	49.59	60.13	61.20	58.43	60.24	59.95	56.45	60.48	61.69	60.24	56.11	55.90	45.23	56.92	54.61	48.65	55.73	55.08	48.28	61.50	60.43	59.20
Co	78.30	62.61	58.03	83.64	83.51	83.41	82.59	83.52	82.55	83.86	84.22	83.97	65.01	65.19	62.19	82.04	80.01	78.19	79.07	79.71	78.19	85.06	85.63	85.21
Ww	27.60	28.0	26.07	27.68	27.99	26.05	26.81	27.68	25.30	26.85	28.28	26.06	24.39	25.70	23.56	27.59	27.43	25.67	26.95	25.77	24.29	27.76	17.14	15.06
He	73.93	72.74	72.15	72.51	72.05	71.93	70.72	69.72	68.56	73.94	73.86	73.61	68.86	68.67	68.42	74.76	73.77	73.63	72.85	71.95	71.50	75.81	74.12	74.05
We	88.01	80.21	80.67	89.22	86.68	87.27	87.99	85.06	85.67	90.42	89.12	89.23	70.77	64.30	62.08	90.74	90.62	90.18	91.82	91.47	90.16	90.43	90.82	90.29
Ir	66.41	56.34	50.76	85.01	83.64	83.55	85.51	85.49	85.04	85.81	85.46	85.18	66.69	57.21	55.90	86.85	86.96	86.50	78.66	78.58	78.44	87.82	87.80	87.39
Ve	38.62	37.32	33.39	43.47	43.10	42.09	41.75	41.08	37.55	42.99	42.75	42.25	35.91	34.33	33.80	42.54	41.84	41.67	41.53	39.98	33.50	43.75	43.90	42.05
Kn	67.54	48.05	47.54	55.14	52.06	52.54	59.43	48.93	49.85	50.52	51.80	49.65	41.99	35.19	31.38	55.73	50.86	50.87	48.19	46.11	46.15	60.06	51.87	52.79
Rw	29.25	29.24	28.58	28.79	30.16	28.50	28.68	29.21	28.43	29.98	31.23	29.78	26.48	26.83	25.92	29.04	28.76	27.78	31.34	32.74	29.93	22.97	22.71	20.87
Ab	10.21	9.87	8.24	16.34	18.10	15.54	12.81	14.59	12.33	16.54	19.08	16.10	13.34	12.71	11.50	14.66	12.37	10.96	18.31	19.98	14.96	14.83	17.05	15.22
Pa	83.17	55.89	62.70	73.90	69.64	70.21	71.23	59.68	61.38	77.86	74.59	75.44	44.61	37.92	37.16	51.87	46.16	43.68	55.76	52.55	50.32	77.56	73.46	72.11

TABLE VIII
EXECUTION TIME WITH K-NN CLASSIFIER (S)

Data	ϕ	MI	KNNI	FCMI	PCC	FID	GAIN	BSI	BCC
Cl	6	0.0697	16.233	0.0977	0.0538	0.1276	27.458	67.101	1.0617
IL	5	0.1136	33.920	0.1443	0.0779	0.2153	27.355	46.481	1.4826
Co	6	0.2119	56.604	0.2182	0.1475	0.4678	27.426	66.956	2.1495
Ww	2	0.8237	270.36	1.1519	2.9101	1.8665	27.452	49.215	23.036
He	7	0.0648	15.750	0.0747	0.0488	0.1206	27.152	59.258	0.8247
We	8	0.0418	9.9419	0.0638	0.0448	0.1099	25.511	32.881	0.4997
Ir	2	0.0367	8.0887	0.0738	0.0369	0.0362	26.046	41.427	0.2015
Ve	9	0.2068	51.183	0.3058	0.2233	0.5046	29.746	93.118	3.8033
Kn	2	0.0628	24.494	0.1531	0.0978	0.1068	27.914	60.312	0.7250
Rw	6	0.3062	94.604	0.4018	0.8605	0.6513	28.946	51.655	4.3880
Ab	4	0.6782	248.60	5.0948	27.599	1.6579	29.041	53.208	14.340
Pa	10	1.1748	330.12	21.930	267.26	6.2400	28.203	29.803	152.97

uncertainty and imprecision caused by missing values.

The proposed BCC method avoids modeling missing values directly. Thus, it can handle well the cases where both the training and test sets contain many missing values without losing information and introducing new uncertainty information. We model each attribute independently while avoiding negative interactions between attributes. Simultaneously, different attributes are able to provide complementary information under the framework of evidence theory. By doing this, the distribution characteristics of different attributes are thoroughly mined, and each attribute can train a basic classifier independently. Moreover, the performance of the classifiers depends on the quality of the training sets constrained by missing values. In this case, the global measurement of the weight on each classifier is an important part of making decisions as cautious as possible. In addition, the proposed BCC method is end-to-end, which can characterize the uncertainty and imprecision in the data, the model, and the results simultaneously. Therefore, the proposed BCC is often able to outperform other comparison methods.

Furthermore, from Fig. 3, it can be observed that the accuracy is less affected by the data incompleteness level for BCC than other methods, indicating that BCC is more robust. Indeed, the BCC can be regarded as an intermediate of

multiple classifiers, effected by the combination process. Since the incompleteness is characterized by imprecision rather than uncertainty in evidence theory, the imputed values make less impact on the classification results than other methods, realized by the Dempster's combination rule. Such a mechanism brings robustness to the BCC method.

In addition, we can also observe that with the increase of ϕ , the classification accuracy results of different methods decreases in most cases. This is consistent with our intuitive perception, because a larger ϕ implies the loss of more attributes from a pattern. The less information the pattern contains, the more difficult it is to classify the pattern correctly, the classification accuracy thereby gradually declines. However, the performance of BCC is still better than that of other conventional methods in the same case. In addition, the increase in ϕ also reflects the increase in the proportion of missing from another perspective. For example, if $\phi = 9$, for the Cloud dataset, the missing rate of the training and testing sets is 90%.

We admit that a few issues exist in BCC, one of which is the computational cost. Since BCC is a combination mechanism of multiple classifiers, it is less efficient than a single classifier. Thus, the satisfactory results are obtained at the cost of more computational resources. Another potential issue is the combination step. The Dempster's rule is applied

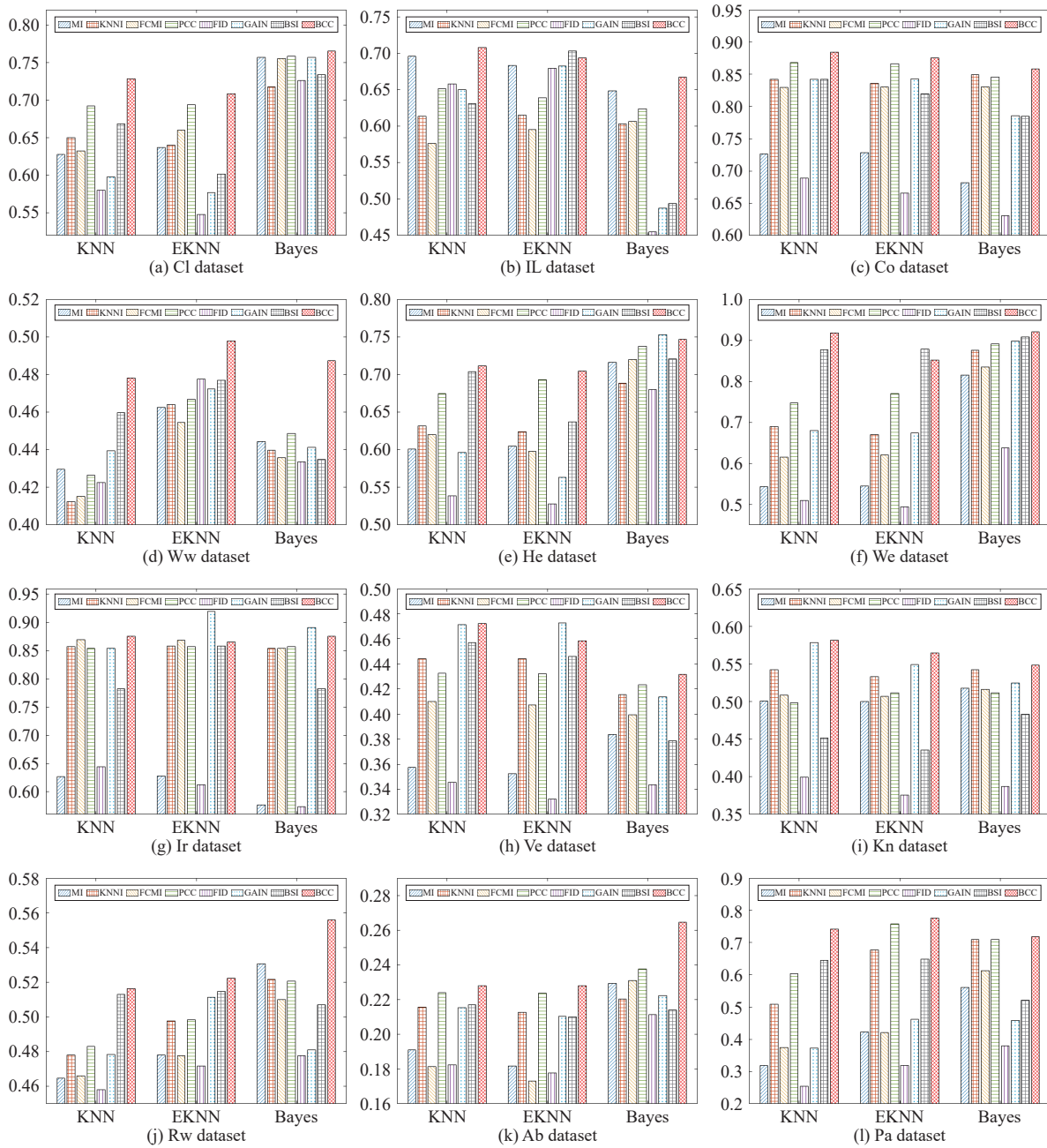


Fig. 2. The average accuracy (AC) of different methods in various datasets.

to combine multiple classification results. When high conflicts still exist between discounted evidence, the rule may return an undecidable BBA. In this case, the issue becomes non-negligible.

E. Computational Cost

The execution time of the different methods with K-NN as the basic classifier is shown in Table VIII. It shows that the BCC method is indeed more time-consuming than other methods since BCC needs to spend more time training the basic classifier and optimizing the discounting factors of pieces of evidence. However, the computational cost is much less than KNNI, GAIN and BSI methods in most cases, for example, for KNNI, because searching for neighbors is a very time-consuming task as the number of patterns increases.

Therefore, it is necessary to take a trade-off between performance and computational cost when using the BCC method. Generally speaking, BCC is more suitable for applications in which high classification accuracy is required whereas efficient computation is not a strong requirement.

V. CONCLUSION

Confronting the problem of classification over incomplete data, we proposed a new belief combination of classifiers (BCC) method in the framework of evidence theory under the fact that patterns in the training set and test set are incomplete. The BCC method characterizes the uncertainty and imprecision caused by missing values with the theory of belief functions. By doing so, BCC is able to make full use of observed data while introducing little impact in dealing with

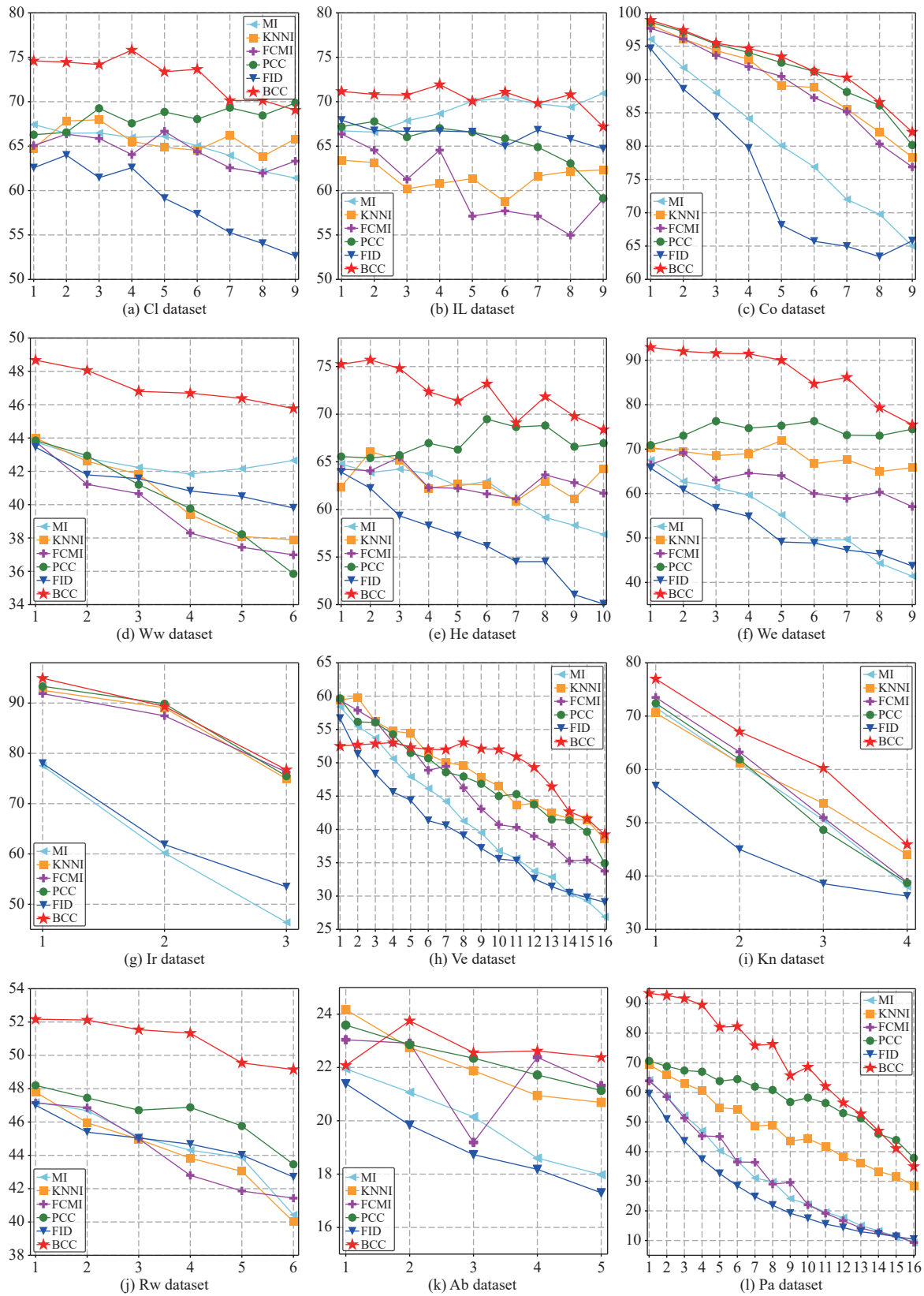


Fig. 3. The effect of parameter ϕ on average accuracy (AC) within different methods over various datasets. The horizontal and vertical axis respectively represents the value of parameter ϕ and AC.

missing data. Consequently, it outperforms conventional classification methods for incomplete data. The core action of BCC is to construct attributes as independent sources, and

each of them is used to train a classifier and thereby predict the class of query patterns. As a result, multiple outputs with different discounting factors for the query pattern are

obtained. The discounting factor includes two parts: the global classifier importance and the local attribute reliability, with which the famous Dempster's rule is employed to fuse the discounted pieces of evidence representing evidential subclasses and then determine the final belief classification for the query patterns. The effectiveness of BCC is demonstrated using various real datasets by comparisons with other conventional methods. The experimental results show that BCC significantly improves the performance in accuracy, precision, recall, and F1 measure. Furthermore, this new method is robust since it does not need to set parameters manually, making it convenient for practical applications.

In recent methods, missing values are usually imputed by value approximation, significantly affected by deep learning approaches. However, the lack of robustness caused by overfitting and underfitting issues has been an obstacle in applying these theoretical methods. The proposed BCC method makes a step forward by taking decisions between specific classes and total ignorance. However, it cannot yet characterize local imprecision [31], [50]. To conclude, the mathematical method given in this paper can somewhat reveal the hidden real world from missing data. In the future, we will employ these classifiers specially designed for high-dimensional data, and we will explore applying a similar methodology to more missing data scenarios beyond the conventional classification. Concerning the robustness problem caused by data missing, a more general framework managing uncertainty and imprecision adaptable to various learning tasks with incomplete patterns is also in the scope of our future work.

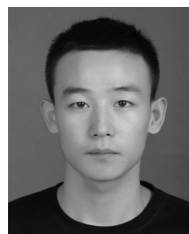
REFERENCES

- [1] K. Y. Chiang, I. S. Dhillon, and C. J. Hsieh, "Using side information to reliably learn low-rank matrices from missing and corrupted observations," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 3005–3039, Jan. 2018.
- [2] N. Städler, D. J. Stekhoven, and Bühlmann, "Pattern alternating maximization algorithm for missing data in high-dimensional problems," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1903–1928, Jan. 2014.
- [3] B. Fekade, T. Maksymyuk, M. Kyryk, and M. Jo, "Probabilistic recovery of incomplete sensed data in IoT," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2282–2292, Aug. 2018.
- [4] J. Pan, C. B. Li, Y. Tang, W. Li, and X. O. Li, "Energy consumption prediction of a CNC machining process with incomplete data," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 5, pp. 987–1000, May 2021.
- [5] L. Chen, L. Q. Wang, Z. Y. Han, J. Zhao, and W. Wang, "Variational inference based kernel dynamic Bayesian networks for construction of prediction intervals for industrial time series with incomplete input," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 5, pp. 1437–1445, Sep. 2020.
- [6] Z. C. Feng, W. He, Z. J. Zhou, X. J. Ban, C. H. Hu, and X. X. Han, "A new safety assessment method based on belief rule base with attribute reliability," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 11, pp. 1774–1785, Nov. 2021.
- [7] R. C. Merton, "A simple model of capital market equilibrium with incomplete information," *J. Finance*, vol. 42, no. 3, pp. 483–510, Jul. 1987.
- [8] J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010.
- [9] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. 3rd ed. New York, USA: Wiley, 2019.
- [10] D. Shen, "Iterative learning control with incomplete information: A survey," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 5, pp. 885–901, Sep. 2018.
- [11] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From predictive methods to missing data imputation: An optimization approach," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 7133–7171, Jan. 2017.
- [12] D. Williams, X. J. Liao, Y. Xue, L. Carin, and B. Krishnapuram, "On classification with incomplete data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 427–436, Mar. 2007.
- [13] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Proc. 6th Int. Conf. Neural Information Processing Systems*, Denver, USA, 1993, pp. 120–127.
- [14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Amsterdam, the Netherlands: Elsevier, 2014.
- [15] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, USA: Springer, 2013.
- [16] G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, Nov. 2000.
- [17] Patidar and A. Tiwari, "Handling missing value in decision tree algorithm," *Int. J. Comput. Appl.*, vol. 70, no. 13, pp. 31–36, May 2013.
- [18] C. Lim, J. H. Leong, and M. M. Kuan, "A hybrid neural network system for pattern classification tasks with missing features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 648–653, Apr. 2005.
- [19] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Netw.*, vol. 18, no. 5-6, pp. 684–692, Jul-Aug. 2005.
- [20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learning Representations*, Scottsdale, USA, 2013.
- [21] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. 3rd ed. Essex, UK: Pearson Education Limited, 2016.
- [22] L. S. Chan and O. J. Dunn, "The treatment of missing values in discriminant analysis-I. The sampling experiment," *J. Am. Stat. Assoc.*, vol. 67, no. 338, pp. 473–477, Jun. 1972.
- [23] L. Brás and J. C. Menezes, "Improving cluster-based missing value estimation of DNA microarray data," *Biomol. Eng.*, vol. 24, no. 2, pp. 273–282, Jun. 2007.
- [24] J. Luengo, J. A. Sáez, and F. Herrera, "Missing data imputation for fuzzy rule-based classification systems," *Soft Comput.*, vol. 16, no. 5, pp. 863–881, May 2012.
- [25] S. G. Liu, J. Zhang, Y. Xiang, and W. L. Zhou, "Fuzzy-based information decomposition for incomplete and imbalanced data learning," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1476–1490, Dec. 2017.
- [26] B. Muzellec, J. Josse, C. Boyer, and M. Cuturi, "Missing data imputation using optimal transport," in *Proc. 37th Int. Conf. Machine Learning*, 2020, pp. 7130–7140.
- [27] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, USA: John Wiley & Sons Inc., 2004.
- [28] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, Dec. 2011.
- [29] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. 35th Int. Conf. Machine Learning*, Stockholm, Sweden, 2018, pp. 5689–5698.

- [30] D. J. Stekhoven and Bühlmann, "MissForest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012.
- [31] Z. G. Liu, Q. Pan, G. Mercier, and J. Dezert, "A new incomplete pattern classification method based on evidential reasoning," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 635–646, Apr. 2014.
- [32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.
- [33] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.*, vol. 38, no. 2, pp. 325–339, Apr. 1967.
- [34] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, USA: Princeton University Press, 1976.
- [35] B. Quost, M. H. Masson, and T. Denoeux, "Classifier fusion in the dempster-Shafer framework using optimized t-norm based combination rules," *Int. J. Approx. Reason.*, vol. 52, no. 3, pp. 353–374, Mar. 2011.
- [36] D. Mercier, G. Cron, T. Denoeux, and M. H. Masson, "Decision fusion for postal address recognition using belief functions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5643–5653, Apr. 2009.
- [37] F. Y. Xiao, "A new divergence measure for belief functions in D-S evidence theory for multisensor data fusion," *Inf. Sci.*, vol. 514, pp. 462–483, Apr. 2020.
- [38] Z. G. Liu, Q. Pan, J. Dezert, J. W. Han, and Y. He, "Classifier fusion with contextual reliability evaluation," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1605–1618, May 2018.
- [39] Smets, "Decision making in the TBM: The necessity of the pignistic transformation," *Int. J. Approx. Reason.*, vol. 38, no. 2, pp. 133–147, Feb. 2005.
- [40] Y. Leung, N. N. Ji, and J. H. Ma, "An integrated information fusion approach based on the theory of evidence and group decision-making," *Inf. Fusion*, vol. 14, no. 4, pp. 410–422, Oct. 2013.
- [41] Z. W. Zhang, H. P. Tian, L. Z. Yan, A. Martin, and K. Zhou, "Learning a credal classifier with optimized and adaptive multiestimation for missing data imputation," *IEEE Trans. Syst., Man, Cybern.: Syst.*, 2021, doi: 10.1109/TSMC.2021.3090210.
- [42] A. Martin and E. Radoi, "Effective ATR algorithms using information fusion models," in *Proc. 7th Int. Conf. Information Fusion*, Stockholm, Sweden, 2004, pp. 161–166.
- [43] T. Denoeux, "Decision-making with belief functions: A review," *Int. J. Approx. Reason.*, vol. 109, pp. 87–110, Jun. 2019.
- [44] M. H. Masson and T. Denoeux, "ECM: An evidential version of the fuzzy c-means algorithm," *Patt. Recognit.*, vol. 41, no. 4, pp. 1384–1397, Apr. 2008.
- [45] Z. G. Su and T. Denoeux, "BPEC: Belief-peaks evidential clustering," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 1, pp. 111–123, Jan. 2019.
- [46] F. J. Li, Y. H. Qian, J. T. Wang, and J. Y. Liang, "Multigranulation information fusion: A Dempster-Shafer evidence theory-based clustering ensemble method," *Inf. Sci.*, vol. 378, pp. 389–409, Feb. 2017.
- [47] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, no. 5, pp. 804–813, May 1995.
- [48] T. Denoeux, "Logistic regression, neural networks and Dempster-Shafer theory: A new perspective," *Knowl.-Based Syst.*, vol. 176, pp. 54–67, Jul. 2019.
- [49] J. Zhao, R. Xue, Z. N. Dong, D. Y. Tang, and W. H. Wei, "Evaluating the reliability of sources of evidence with a two-perspective approach in classification problems based on evidence theory," *Inf. Sci.*, vol. 507, pp. 313–338, Jan. 2020.
- [50] Z. F. Ma, H. Tian, Z. C. Liu, and Z. W. Zhang, "A new incomplete pattern belief classification method with multiple estimations based on KNN," *Appl. Soft Comput.*, vol. 90, Article No. 106175, May 2020.
- [51] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, Sep. 2004.
- [52] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowl.-Based Syst.*, vol. 182, Article No. 104838, Oct. 2019.
- [53] M. J. D. Powell, "A fast algorithm for nonlinearly constrained optimization calculations," in *Numerical Analysis*, G. A. Watson, Ed. Berlin, Heidelberg, Germany: Springer, 1978, pp. 144–157.
- [54] J. Benesty, J. D. Chen, Y. T. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, I. Cohen, Y. T. Huang, J. D. Chen, and J. Benesty, Eds. Heidelberg, Germany: Springer, 2009, pp. 1–4.
- [55] "Spearman rank correlation coefficient," in *The Concise Encyclopedia of Statistics*, New York, USA: Springer, 2008, pp. 502–505.
- [56] M. G. Kendall, *Rank Correlation Methods*. London: Griffin, 1948.
- [57] Smets, "The combination of evidence in the transferable belief model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 447–458, May 1990.
- [58] R. R. Yager, "On the Dempster-Shafer framework and new combination rules," *Inf. Sci.*, vol. 41, no. 2, pp. 93–137, Mar. 1987.
- [59] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," *Comput. Intell.*, vol. 4, no. 3, pp. 244–264, Sep. 1988.
- [60] J. Dezert and A. Tchamova, "On the validity of dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule," *Int. J. Intell. Syst.*, vol. 29, no. 3, pp. 223–252, Mar. 2014.
- [61] Y. M. Yang, "An evaluation of statistical approaches to text categorization," *Inf. Retrieval*, vol. 1, no. 1–2, pp. 69–90, Apr. 1999.
- [62] T. Cover and Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
- [63] J. Twisk, M. de Boer, W. de Vente, and M. Heymans, "Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis," *J. Clin. Epidemiol.*, vol. 66, no. 9, pp. 1022–1028, Sep. 2013.



Zuowei Zhang (Student Member, IEEE) received the master degree from Xi'an University of Architecture and Technology (XAUAT), in 2013, and is currently pursuing the Ph.D. degree at the Key Laboratory of Information Fusion Technology, Ministry of Education, School of Automation, Northwestern Polytechnical University (NPU). He is also pursuing the Ph.D. degree in the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), University of Rennes 1, France. His research interests include belief function theory and its applications in data mining, pattern recognition, and multi-source information fusion.



Songtao Ye received the master degree in automatic control from Xi'an University of Architecture and Technology (XAUAT), in 2021. He is currently working at the Zhejiang Laboratory, Hangzhou. His research interests include pattern recognition and machine vision.



Yiru Zhang is an Associate Professor at CY Cergy Paris University, France. He received the B.E. degree from Huazhong University of Science and Technology (HUST), China, the M.E. degree from Université Pierre et Marie Curie (UPMC), France, and the Ph.D. degree from Université de Rennes 1, France, respectively in 2013, 2015 and 2020. His research interests include imperfect information reasoning with belief function theory, decision theory, data mining, and pattern recognition.



Weiping Ding (Senior Member, IEEE) received the Ph.D. degree in computer science, Nanjing University of Aeronautics and Astronautics (NUAA), in 2013. From 2014 to 2015, he is a Postdoctoral Researcher at the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, China. In 2016, He was a Visiting Scholar at National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney (UTS), Ultimo, NSW, Australia. His research interests include deep neural networks, multimodal machine learning, granular data mining, and medical images analysis. He has published more than 100 journal papers, including *IEEE Transactions on Fuzzy Systems (IEEE T-FS)*, *IEEE Transactions on Neural Networks and Learning Systems (IEEE T-NNLS)*, *IEEE Transactions on Cybernetics (IEEE T-CYB)*, *IEEE Transactions on Systems Man Cybernetics-Systems (IEEE T-SMCS)*, *IEEE Transactions on Biomedical Engineering (IEEE T-BME)*, *IEEE Transactions on Evolutionary Computation (IEEE T-EVC)*, *IEEE Transactions on Industrial Informatics (IEEE T-II)*, *IEEE Transactions on Emerging Topics in Computational Intelligence (IEEE T-ETCI)*, *IEEE Transactions on Cognitive and Developmental Systems (IEEE T-CDS)*, *IEEE Transactions on Intelligent Transportation Systems (IEEE T-ITS)*, *IEEE Transactions on Artificial Intelligence (IEEE T-AI)*. and he has held 20 approved invention patents in total over 35 issued patents. His six

authored/co-authored papers have been selected as ESI Highly Cited Papers. Dr. Ding is vigorously involved in editorial activities. He served/serves on the Editorial Advisory Board of Knowledge-Based Systems and Editorial Board of Information Fusion, Engineering Applications of Artificial Intelligence and Applied Soft Computing. He served/serves as an Associate Editor of *IEEE Transactions on Neural Network and Learning System*, *IEEE Transactions on Fuzzy Systems*, *IEEE/CAA Journal of Automatica Sinica*, *Information Sciences*, *Neurocomputing*, *Swarm and Evolutionary Computation*. He is the Leading Guest Editor of Special Issues in several prestigious journals, including *IEEE Transactions on Evolutionary Computation*, *IEEE Transactions on Fuzzy Systems*.



Hao Wang (Senior Member, IEEE) is an Associate Professor in the Department of Computer Science in Norwegian University of Science and Technology, Norway. He is also a visiting PI in the Research Center for Optical Fibre Sensing of Zhejiang Laboratory, China and a Visiting Professor at the School of Cyber Engineering of Xidian University. He received his Ph.D. degree and a B.Eng. degree, both in computer science and engineering, from South China University of Technology in 2006 and 2000, respectively. His research interests include big data analytics, industrial internet of things, high performance computing, and safety-critical systems. He has published 170+ papers in reputable international journals and conferences. He serves as the editorial board member and Guest Editors for several international journals. He served as a TPC Co-chair for IEEE GPC 2021, CPSCoM 2020, IEEE CIT 2017, ES 2017, IEEE DataCom 2015, a senior TPC member for ACM CIKM 2019, and reviewers for many prestigious journals and conferences. He is a senior member of IEEE and a member of ACM. He has received prestigious awards such as IEEE Outstanding Paper Award and IEEE Outstanding Leadership Award. He is the Chair for Sub TC on Healthcare in IEEE IES Technical Committee on Industrial Informatics.