Evidence integration credal classification algorithm versus missing data distributions

Zuo-Wei Zhang, Zhe Liu, Zong-Fa Ma, Ji-Huan He, Xing-Yu Zhu

Evidence Integration Credal Classification Algorithm Versus Missing Data Distributions

Zuo-wei Zhang[a,b,*], Zhe Liu[c], Zong-fa Ma[c], Ji-huan He[c], Xing-yu Zhu[d]

[a] *School of Automation, Northwestern Polytechnical University, Xi'an, China.*
[b] *Institut de Recherche en Informatique et Systèmes Aléatoires, University of Rennes 1, France.*
[c] *School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China.*
[d] *Department of Network and telecommunication, IUT of Lannion, University of Rennes 1, France.*

Abstract

In complex incomplete pattern classification, the classification results produced by a single classifier and used for decision-making may be quite unreliable and uncertain due to the random distribution of missing data. This paper proposes a new evidence integration credal classification algorithm (EICA) for multiple classifiers working on different attributes, aiming to reduce the negative impact on incomplete pattern classification by modeling the missing values locally. In EICA, the dataset is first grouped into several subsets, and missing values in each subset are estimated by similar subpatterns with different weights. The similarity is measured by discounting the overall similarity of subpatterns and the local similarity of attributes on the basis of fully exploiting the distribution characteristics of the attributes. The greater the variation in distribution across classes, the greater the weight. The classification results of the edited subpatterns with different discounting factors obtained by the optimization function can often provide (more or less) useful information for the classification of the query pattern. Thus, these discounted pieces of evidence (outputs) represented by basic belief assignments (BBAs) are globally fused to classify the query pattern on the basis of evidence theory. The validity has been demonstrated with various real datasets.

Keywords: Missing data, Evidence theory, Classifier fusion, Credal classification, Incomplete pattern.

1. Introduction

Missing data classification is an important branch in statistical multivariate analysis and supervised machine learning, with broad applications in various fields such as financial fraud, medical diagnosis, image processing, information retrieval, and bioinformatics, etc. To make traditional classifiers applicable to missing data, preprocessing is considered the dominant technique, and a number of methods [1, 2] have been developed based on the three mechanisms [1]. The simplest method is the deletion of incomplete patterns [2], and it is applicable only in cases where the number of incomplete patterns is small (less than 5% of the overall data).

The estimation strategy is a common method for classifying incomplete patterns in many cases [3]-[12]. Missing values in these methods are estimated by the observed values, and the chosen classifier can then classify these patterns with estimations as if they were complete data. For example, mean imputation (MI) [4, 5] can replace

*Corresponding author.
Email address: zhangzuowei0720@gmail.com. (Zuo-wei Zhang)

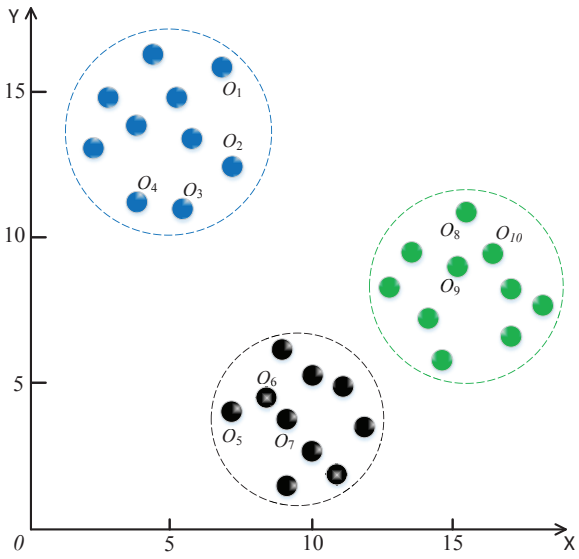


Figure 1: Illustration of similarity analysis for missing data.

missing values with the average of the corresponding attributes in the entire dataset. The use of similar patterns to estimate missing values for incomplete patterns, also known as K-nearest neighbor imputation (KNNI)[6, 7], is considered to be one of the most popular methods for classifying missing data. In [8, 9], an interesting estimation method based on a machine learning process, named fuzzy c-mean imputation (FCMI), is introduced to fill in the missing values with the clustering centers produced by FCM. In particular, a new method, named fuzzy-based information decomposition (FID) [11], has been developed to solve both class imbalance and missing value problems, which consists of two steps: weighting and recovery. In the weighting step, the weights generated by the fuzzy membership function are used to quantify the contribution of the observations to the missing estimates. In the recovery step, the different contributions of the observations will be considered to estimate the missing values. Interestingly, in [12], a local linear approximation (LLA) method for processing incomplete pattern is proposed, which adopts the optimal weight of KNNs gained from local linear reconstruction to fill the missing values.

In complex incomplete pattern classifications, however, the classification results produced by a single classifier and used for final decision-making can be quite unreliable and uncertain due to the random nature of the missing data distribution. For instance, in Fig. 1, if the x-dimension of the pattern O_5 is missing, one can obtain the neighbors, i.e., O_6 and O_7, by calculating the similarity when KNNI is adopted here. By contrast, if the y-dimension is missing, the neighbors of O_5 will become the patterns O_1 and O_2, which is far away from the truth. A similar erroneous conclusion can be drawn from the O_8 pattern without the x-dimension attribute, since then the neighbors are O_3 and O_4. In this case, it would be an excellent option to model the uncertainty caused by missing values and to minimize the negative impact of missing values.

Based on the above analysis, we propose a new weighted evidence integration credal classification algorithm (EICA) for multiple classifiers working on different attributes (features) of patterns, aiming to reduce the negative impact of missing data distributions on incomplete pattern classification by modeling the missing values locally. In

EICA, the dataset is grouped into several subsets based on prior knowledge (i.e., the correlation between different attributes) or random combinations. Missing values in each subset are imputed from similar subpatterns (neighbors), where the choice of neighbors is a compromise between the overall similarity of the subpatterns and the local similarity of the attributes. The (edited) subpatterns in each subset are then classified, and the results of several subclassifications with different discounting factors, represented by basic belief assignments (BBAs), are fused to make the credal classification for the query pattern.

The rest of this paper is organized as follows. The basics of evidence theory will be introduced in Section 2, and the EICA method is introduced in the Section 3. The performance of the proposed method is then tested and compared with several other widely used methods in Section 4, followed by conclusions.

2. Related works

2.1. Some other classical methods

Although many estimation strategies have been reviewed, there are still a number of methods are designed to deal with incomplete data without estimation [13, 14], such as model-based algorithms [15, 16]. For example, a (supervised) logistic regression algorithm is proposed in [15] to deal with missing values, where missing values are estimated by performing analytic integration with an estimated conditional density function. The conditional density functions are estimated with a Gaussian mixture model (GMM), the parameters of which are obtained by Expectation-Maximization (EM) and variational Bayesian EM (VB-EM). In [13], the method based on Support Vector Machine (SVM) [17] adopts a non-parametric perspective by defining a modified risk taking into account the uncertainty of the predicted outputs when missing values are involved. The method is extended to the multivariate case of fitting additive models using componentwise kernel machines. All of these methods have performed well to some extent. The models however are difficult to obtain sometimes, thus estimation strategies are still the main pre-processing methods in classifying missing data.

2.2. The basics of evidence theory

Evidence theory, also called Dempster-Shafer theory (DST) or belief function theory [18]-[21], has been widely used in fusion decision-making [22]-[28], and also good at clustering [29]-[32] and classification [33]-[35] since it can well characterize the uncertain and imprecise information. In evidence theory, one starts with a frame of discernment $\Omega = \{\omega_1, ..., \omega_c\}$, which consists of a finite discrete set of mutually exclusive and exhaustive hypotheses (classes), and the power-set of Ω denoted 2^Ω is the set of all the subsets of Ω. For example, if $\Omega = \{\omega_1, \omega_2, \omega_3\}$, then $2^\Omega = \{, \omega_1, \omega_2, \omega_3, \omega_1 \cup \omega_2, \omega_1 \cup \omega_3, \omega_2 \cup \omega_3, \Omega\}$. The singleton element (e.g., ω_i) represents a specific class. The disjunctions (union) of several singleton elements represents the partial uncertainty in 2^Ω (e.g., $\omega_i \cup \omega_j$ or $\omega_i \cup \omega_j \cup \omega_k$, etc), and they are called meta-classes.

In evidence theory, the pattern can be well associated with different singleton elements and sets of elements according to a basic belief assignment (BBA), which is a function $m(.)$ from 2^Ω to $[0, 1]$ satisfying $m(\emptyset) = 0$ and the normalization condition satisfying $\sum_{A \in 2^\Omega} m(A) = 1$. The belief function $Bel(.)$ and the plausibility function

$Pl(.)$ represent the lower and upper bounds of the imprecision probability associated with BBA's. $[Bel(.), Pl(.)]$ is interpreted as the imprecision interval of the unknown probability $P(A)$ of any element A of 2^{Θ}. These bounds are defined for all $A \in 2^{\Omega}$ as:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{1}$$

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{2}$$

where $Bel(.)$ and $Pl(.)$ can be used for decision-making support when adopting pessimistic or optimistic attitudes if necessary.

In a multi-classifier system, the output of each classifier can be considered as an evidence represented by a BBA. The well-known DS rule is still widely applied for combining multiple BBA's mainly because of its commutative and associative properties, which makes it relatively easy to implement. In ET framework, Shafer proposed that the different pieces of evidence represented by BBA's should be combined using Dempster's rule, which is commonly denoted by DS rule and expressed by $\oplus$ symbol. DS rule offers a compromise between the specificity and complexity for the combination of BBA's. The DS combination of two distinct sources of evidence characterized by the BBA's $m_1(.)$ and $m_2(.)$ over 2^{Ω} is denoted $\mathbf{m} = \mathbf{m}_1 \oplus \mathbf{m}_2$, and it is mathematically defined by $m_{DS}(\emptyset) = 0$ and for $A \neq \emptyset, B, C \in 2^{\Theta}$ by:

$$m_{DS}(A) = [m_1 \oplus m_2](A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \tag{3}$$

DS rule is exchangeable and correlative, making a compromise between the particularity and complexity of BBA composition. All the conflicting beliefs $\sum_{B \cup C = \emptyset} m_1(B) m_2(C)$ are proportionally redistributed back to the focal elements by using this rule through a classical normalization step. However, some paradoxes under the high conflicting cases and under some special low conflicting cases maybe produced because of this redistribution. Many different rules of combination have emerged to overcome its limitations. Among the possible alternatives of DS rule, some rules such as the well-known Smets' conjunctive rule (used in his transferable belief model (TBM) [19, 20]), Dubois-Prade (DP) rule [36], and more recently the more complex Proportional Conflict Redistributions (PCR) rules [37, 38] are found. Inspired by the above methods, the EICA method treats the classification result of each subpattern as one evidence, and then the final class of the query pattern is determined by the integration of these evidences.

2.3. Some methods based on evidence theory

Due to the advantages in modeling and combining the uncertain information, evidence theory [18]-[21] has been widely used in evidence integration [22]-[26]. In this theory, different pieces of evidence (i.e., classifiers) can provide often provide (more or less) useful supplementary information to reduce the error rate and enhance the robustness of classification. Given the non-independence between classifiers, for example, the literature [23] investigates an approach that combines other operators with the Dempster's rule, aiming to mediate their behavior

between Dempster's rule and the cautious rule, where two strategies (i.e., a single combination rule and a two-step fusion method) are investigated for learning an optimal combination scheme, based on a parameterized family of t-norms. These integration methods, based on evidence theory, achieved good performance, however, they were designed for complete data. Only a few methods for incomplete patterns based on evidence theory [27, 28] have been proposed. In [27], a prototype-based credal classification (PCC) method is proposed. The incomplete pattern is first edited into c versions, for a c-class problem, with the class prototypes obtained by training patterns. Then the c classification results are globally fused with different discounting factors to finally determine the class of the incomplete pattern. In [28], the missing values in wireless sensor networks are estimated with the regularized extreme learning machine. Then the estimations are converts into multiple classification results on the basis of the distance between interval numbers. These results are fused to classify the query pattern by an evidential reasoning rule. Although these methods use evidence integration locally, they do not fully consider the negative impact of incomplete data distribution on the accuracy of estimations. For example, in [27], some of the missing values can be filled in with class centers, but others may be unreliable.

3. Evidence Integration Credal Classification

Let us consider a query set $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ being classified on the frame of discernment $\Omega = \{\omega_1, ..., \omega_c\}$ according to a set of labeled patterns, i.e., $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_H\}$. Each pattern $\mathbf{x}_h$ has $\mathcal{S}$ different attributes (features) and the class label is represented by $L(\mathbf{x}_h)$. We assume that both the training set $\mathcal{X}$ and the test set $\mathcal{Y}$ contain missing data. To avoid the possible negative impacts of missing value distributions, the EICA method first groups the dataset into non-overlapping Φ subsets based on the prior knowledge (i.e., correlations between different attributes) or random combinations, e.g., $\mathcal{X} = \mathcal{X}^1, ..., \mathcal{X}^\Phi$. Of course, different subversions (subpatterns), i.e., $\mathbf{x}_h^1, ..., \mathbf{x}_h^\Phi$, share the same class label $L(\mathbf{x}_h)$. For the incomplete pattern $\mathbf{y}$, the missing values of the subversion $\mathbf{y}^\varphi$ are estimated by the neighbors in the subset $\mathcal{Y}^\varphi$. Each neighbor has a corresponding weight depending on the Euclidean distance from it to $\mathbf{y}^\varphi$. In addition, the importance of different attributes is also considered when measuring similarity (seeking neighbors). By doing this, one thereby can obtain Φ soft outputs for $\mathbf{y}$. These outputs can be the probabilistic membership (for classifiers under the probabilistic framework, e.g., Naive Bayes [39]) or belief degree (for evidential classifiers working with evidence theory, e.g., EK-NN [34]). After that, the Φ outputs are discounted with different discounting factors obtained by a global optimization function. The global fusion of the Φ discounted outputs, represented by the basic belief assignments (BBAs), is used to make the final dicision-making for $\mathbf{y}$.

3.1. Estimate missing values on Φ subsets

As we have previously analyzed, the selection of true similar patterns in the same class is a key factor for effective imputation. Thus, we should prioritize mining some a priori information about the distribution of data attributes to aid in the selection of true neighbors. In general, the greater the difference in the distribution of an attribute across classes, the greater its weight, as it provides us with more information about the class. By contrast, the more similar the distribution of an attribute in different classes, the smaller its weight, because the attribute does

not provide us with much useful information to find the neighbors in the class. Many methods exist to calculate the weighting factor α_s for the s-th attribute. Here we argue that the greater the mean-variance of an attribute across classes, the greater the weight of that attribute. For $s = 1, ..., \mathcal{S}$, the mean-variance $\mathcal{V}^s$ of the s-th attribute is defined by:

$$\mathcal{V}^s = \frac{1}{c}\sum_{j=1}^{c}(\frac{1}{\mathcal{C}_j}\sum_{i=1}^{\mathcal{C}_j}x_i^s - \frac{1}{H}\sum_{h=1}^{H}x_h^s)^2 \tag{4}$$

where $\frac{1}{H}\sum_{h=1}^{H}x_h^s$ is the mean value of the s-th attribute in the training set $\mathcal{X}$, and H is the number of training patterns. $\frac{1}{\mathcal{C}_j}\sum_{i=1}^{\mathcal{C}_j}x_i^s$ is the mean value of the s-th attribute in the j-th class. $\mathcal{C}_j$ is the number of the training patterns in the j-th class, $j = 1, ..., c$.

Since the dataset (e.g., $\mathcal{X}$) is grouped into Φ subsets, and the missing values for each pattern (e.g., $\mathbf{x}^\varphi$) in each subset (e.g., $\mathcal{X}^\varphi$) may be different, it is necessary to calculate the attribute weighting factor α_s of each incomplete pattern $\mathbf{x}^\varphi$ separately. From these mean-variance $\mathcal{V}^s$ for $s = 1, ..., \mathcal{S}$, for example, we can define the weighting factor α_s of the s-th attribute of $\mathbf{x}^\varphi$ in the $\mathcal{X}^\varphi$ partition by:

$$\alpha_s = \frac{\mathcal{V}^s}{\sum\limits_{\exists s, s=1}^{\mathcal{S}_\varphi - \mathcal{S}_m}\mathcal{V}^s} \tag{5}$$

subject to

$$\begin{cases} \alpha_s > 0, \\ \sum\limits_{\exists s, s=1}^{\mathcal{S}_\varphi - \mathcal{S}_m}\alpha_s = 1. \end{cases} \tag{6}$$

where $\mathbf{x}^\varphi$ is an incomplete pattern in $\mathcal{X}^\varphi$, i.e., the subversion of $\mathbf{x}$ in the subset $\mathcal{X}^\varphi$. α_s represents the weighting factor of the s-th known attribute of $\mathbf{x}^\varphi$. $\mathcal{S}_\varphi$ is the number of attributes included in $\mathcal{X}^\varphi$, and $\mathcal{S}_m$ is denoted as the number of the missing values of $\mathbf{x}^\varphi$. $\mathcal{V}^s$ is the mean-variance of the s-th attribute obtained from E.q. (4). Then we can select the neighbors of $\mathbf{x}^\varphi$ in the same class using the Euclidean distance $d(\mathbf{x}^\varphi, \mathbf{x}_k^\varphi)$, defined by:

$$d(\mathbf{x}^\varphi, \mathbf{x}_k^\varphi) = \sqrt{\sum_{\exists s, s=1}^{\mathcal{S}_\varphi - \mathcal{S}_m}\alpha_s(x_s^\varphi - x_{k,s}^\varphi)^2} \tag{7}$$

where x_s^φ and $x_{k,s}^\varphi$ represent the s-th attribute values of the patterns, respectively.

Since the contribution of different neighbors to the missing values is different, before estimating the missing values of $\mathbf{x}^\varphi$, we need to calculate the reliability β_k of each neighbor $\mathbf{x}_k^\varphi$. In fact, the smaller distance $d(\mathbf{x}^\varphi, \mathbf{x}_k^\varphi)$ generally leads to the bigger reliability β_k. Therefore, a rational way that has been widely applied in many works is adopted here to estimate the reliability β_k. For $k = 1, ..., \mathcal{K}$, the reliability β_k is defined by:

$$\beta_k = \frac{e^{-d(\mathbf{x}^\varphi, \mathbf{x}_k^\varphi)}}{\sum\limits_{k=1}^{\mathcal{K}}e^{-d(\mathbf{x}^\varphi, \mathbf{x}_k^\varphi)}} \tag{8}$$

submit to

$$\begin{cases} \beta_k > 0, \\ \sum\limits_{k=1}^{\mathcal{K}} \beta_k = 1. \end{cases} \tag{9}$$

where $\mathcal{K}$ is the number of the selected neighbors for estimating the missing values of $\mathbf{x}^{\varphi}$. Then, the o-th missing value of $\mathbf{x}^{\varphi}$ is given by:

$$x_o^{\varphi} = \sum_{k=1}^{\mathcal{K}} \beta_k \cdot x_{k,o}^{\varphi} \tag{10}$$

By doing this, it is possible to obtain all the incomplete patterns with estimations in each subset of the training sets, and then use these edited patterns to train Φ basic classifiers. Since we are not intended to improve the performance of the chosen classifier itself, the classifiers that satisfy the probabilistic framework or similar probabilistic frameworks can be used here as basic classifiers, e.g., K-NN [40], NB [39], EK-NN [34], etc.

After obtaining the mean-variance $\mathcal{V}^s$, the missing values for different partitions (e.g., $\mathcal{Y}^{\varphi}$) in the test pattern (e.g., $\mathbf{y}^{\varphi}$) can also be imputed by the similar method to the training pattern $\mathbf{x}^{\varphi}$. The difference, however, is that we will find the neighbors for $\mathbf{y}^{\varphi}$ globally, aiming to select the similar patterns as much as possible in the same class. In other words, the candidate database is extended to all patterns in the training subset $\mathcal{X}^{\varphi}$ and the test subset $\mathcal{Y}^{\varphi}$.

3.2. Optimize reliability of different classifiers

Since the classifiers to be integrated learn based on different attribute knowledge, they may have different classification capabilities, so the reliability of the classification results provided by Φ subversions (e.g., $\mathbf{y}^{\varphi}$) is different. Therefore, each classifier should be given appropriate reliability (discounting factor) in the fusion to achieve the best classification performance. In general, the closer the trained basic classifier Θ^{φ} is to the truth, the more valuable the information provided by the classifier Θ^{φ}, i.e., the higher reliability γ^{φ} should be given to the classifier Θ^{φ}. Here the complete patterns in the training set $\mathcal{X}$, i.e., $\mathcal{X}_{com} = \{\mathbf{x}_1, ..., \mathbf{x}_{\mathcal{H}}\}$, are used to optimize the discounting factor $\boldsymbol{\gamma} = [\gamma^1, ..., \gamma^{\Phi}]$ of Φ basic classifiers to eliminate the possible negative effects of missing values. The optimal discounting vector $\boldsymbol{\gamma}$ should make the combination results as close as possible to the truth of the complete training patterns. The classification result of the subpattern $\mathbf{x}_h^{\varphi}$ in the subset $\mathcal{X}^{\varphi}$ is given by:

$$\boldsymbol{\mathcal{P}}_h^{\varphi} = \Theta^{\varphi}(\mathbf{x}_h^{\varphi} | \mathcal{X}^{\varphi}) \tag{11}$$

where Θ represents the chosen classifier. $\boldsymbol{\mathcal{P}}_h^{\varphi}$ is the classification result of $\mathbf{x}_h^{\varphi}$ using the classifier Θ^{φ} trained by the set $\mathcal{X}^{\varphi}$. $\boldsymbol{\mathcal{P}}_h^{\varphi}$ can be a Bayesian BBA if the chosen classifier works under probability framework (e.g., K-NN [40], NB [39]), and it can also be a regular BBA with having some mass of beliefs committed to the ignorant class Ω if the classifier works under evidence theory (e.g., EK-NN [34]). Therefore, a BBA (probability) matrix $\boldsymbol{\mathcal{P}}_h$ of $\Phi \times c$

for $\mathbf{x}_h$ can be obtained with Eq. (11), and defined as:

$$\boldsymbol{\mathcal{P}}_h = \begin{bmatrix} p_{11}^h & \cdots & p_{1c}^h \\ \vdots & \ddots & \vdots \\ p_{\Phi 1}^h & \cdots & p_{\Phi c}^h \end{bmatrix} \tag{12}$$

subject to

$$\begin{cases} p_{\varphi j}^h \in [0,1], \\ \sum\limits_{j=1}^{c} p_{\varphi j}^h = 1. \end{cases} \tag{13}$$

where $p_{\varphi j}^h$ is defined as the support degree (probability) that the subpattern $\mathbf{x}_h^{\varphi}$ is classified into class ω_j by the basic classifier Θ^{φ}, $j = 1, ..., c$, and $h = 1, ..., \mathcal{H}$. In order to optimize the reliability matrix $\boldsymbol{\gamma} = [\gamma^1, ..., \gamma^{\Phi}]$, we assume that the Φ subpattern $\mathbf{x}_h^{\varphi}$ have the same label $L(\mathbf{x}_h)$ characterized by the binary vector $\boldsymbol{T}_h = [T_{h1}, ..., T_{hc}]$. It can be found that for the pattern $\mathbf{x}_h$ belonging to the class $\mathbf{x}_h$, all components of $\boldsymbol{T}_h$ are equal to zero, but $T_{hj} = 1$. For a 3-class problem, for example, the vector $\boldsymbol{T}_h$ should be represented as $\boldsymbol{T}_h = [0, 0, 1]$ if the training pattern $\mathbf{x}_h$ belongs to the class ω_3. For all patterns in $\mathcal{X}_{com}$, we can construct a set of equations as follows:

$$\begin{cases} \|\boldsymbol{\gamma}\mathcal{P}_1 - \boldsymbol{T}_1\| = \theta_1 \\ \|\boldsymbol{\gamma}\mathcal{P}_2 - \boldsymbol{T}_2\| = \theta_2 \\ \qquad\vdots \\ \|\boldsymbol{\gamma}\mathcal{P}_{\mathcal{H}} - \boldsymbol{T}_{\mathcal{H}}\| = \theta_{\mathcal{H}} \end{cases} \tag{14}$$

where $\|.\|$ stands for Euclidean distance and θ represents the deviation value. $\mathcal{H}$ is the number of the complete pattern in the set $\mathcal{X}_{com}$. The sum of all deviations is defined as a global objective function, and the global objective function $\mathcal{O}$ is expressed as follows:

$$\mathcal{O} = \sum_{h=1}^{\mathcal{H}} \|\boldsymbol{\gamma}\mathcal{P}_h - \boldsymbol{T}_h\| \tag{15}$$

The reliability matrix $\boldsymbol{\gamma} = [\gamma^1, ..., \gamma^{\Phi}]$ can be obtained by minimizing the global objective function $\mathcal{O}$, i.e., the sum of the expected deviation values is as close to zero as possible. Therefore, we obtain $\boldsymbol{\gamma}$ by the following formula:

$$\boldsymbol{\gamma} = \arg\min_{\boldsymbol{\gamma}} (\sum_{h=1}^{\mathcal{H}} \|\boldsymbol{\gamma}\mathcal{P}_h - \boldsymbol{T}_h\|) \tag{16}$$

In the optimization process, the Sequential Quadratic Programming (SQP) method is used and the estimation of the Lagrange-Hessen equation is updated by using Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula (quasi-Newton method), where the objective function $\mathcal{O}$ must satisfy the constraint Eq. (13). In Matlab^{TM} software, the function *fmincon* can be called directly to solve this type of constraint optimization problem. Setting the initial value of $\boldsymbol{\gamma}$ to the unit matrix has no effect on the result.

3.3. Global fuse Φ discounted evidence

For the query (test) pattern $\mathbf{y}$, the Φ classification results obtained from the basic classifier Θ may be in conflict due to the missing values. In other words, the Φ classification results may support that the query pattern $\mathbf{y}$ belongs to different classes. Thus, some discounting techniques are needed to weight the impact of these sources of evidence differently in the global fusion process. In general, the greater the reliability of the classifier γ^{φ}, the more reliable the φ-th output for $\mathbf{y}$. From these discounting factors γ^{φ} for $\varphi = 1, ..., \Phi$, we can then define the relative reliability (discounting factors) by:

$$\hat{\gamma}^{\varphi} = \frac{\gamma^{\varphi}}{\max\limits_{j} \gamma^{j}} \tag{17}$$

where $\max\limits_{j} \gamma^{j} = \max\{\gamma^{1}, ..., \gamma^{\Phi}\}$. The discounting method proposed by Shafer in [18] is applied here to discount the BBA of each evidence (output) depending on the discounting factors $\hat{\gamma}^{\varphi}$, and the discounted masses of belief (i.e., ${}^{\hat{\gamma}^{\varphi}}\mathbf{m}$) for φ-th evidence can be defined as follows:

$$\begin{cases} {}^{\hat{\gamma}^{\varphi}}\mathbf{m}(A) = \hat{\gamma}^{\varphi}\mathbf{m}(A), \forall A \subset \Omega \\ {}^{\hat{\gamma}^{\varphi}}\mathbf{m}(\Omega) = \hat{\gamma}^{\varphi}\mathbf{m}(\Omega) + 1 - \hat{\gamma}^{\varphi} \end{cases} \tag{18}$$

In (18), the conflict information is transferred into the total ignorance denoted by the framework of discernment, i.e., Ω. The advantage of Ω is not only to reduce the possible negative effects of high conflict but also to maintain neutrality in the fusion process. For the query pattern $\mathbf{y}$, the global fusion results of the Φ BBAs are given by the DS rule, for a focal element $A \in 2^{\Omega}$, defined by:

$$\mathbf{m}_{\mathbf{y}}(A) = [{}^{\hat{\gamma}^{1}}\mathbf{m} \oplus ... \oplus {}^{\hat{\gamma}^{\Phi}}\mathbf{m}]_{\mathbf{y}}(A) \tag{19}$$

where $\mathbf{m}_{\mathbf{y}}$ is the final BBAs of credal classification for the query pattern $\mathbf{y}$. $\oplus$ represents the DS combination defined in (3). Since the DS rule is associative, these BBAs can be combined sequentially using (3) and the sequential order does not matter.

3.4. Some discussions in applications

We discuss some issues that may be encountered in applications of the proposed EICA method in this subsection.

1) Guideline for attribute combinations: For a particular problem, if some a priori information is available, then reconstructing a subset of Φ with correlations between attributes is the best solution, as it provides a good estimate of the missing values of incomplete patterns in subsets $\mathcal{X}^{\varphi}$ or $\mathcal{Y}^{\varphi}$. Random combinations can also be used as an alternative method if prior knowledge of attribute correlation is insufficient. In order to verify the validity of the proposed EICA method in this paper, several simulation experiments are conducted on both combinations.

2) The effectiveness of local imputation: In this paper we present the idea of local modeling of missing values, aiming to form the dataset into Φ subset and estimate missing values in the subset using similar patterns (neighbors). This approach can effectively avoid the negative effects of irrelevant attributes, especially for high dimensional data. It is well known that the selection of neighbors is highly dependent on methods that measure

similarity, such as Euclidean distance. Since the distribution of attributes can be very different, this sometimes leads to observed "neighbors" that are not actual neighbors and thus to estimations that are far from the truth. In such cases, if a part of attributes can be isolated into a subset based on correlation or relatedness, the negative influence of irrelevant attributes can be avoided and the attributes in each subset are (possibly) linearly correlated, which greatly guarantees the reliability of the estimations. Of course, the importance of different attributes in one subset is also different. Similar to principal component analysis, we give different weights to different attributes, which can further observe the neighbors that are relevant to the classification and thus provide more reasonable estimations.

Table 1: Basic information of the used datasets.

Data	#Class	#Attr.	#Inst.
Wine (Wi)	3	13	178
Heart (He)	2	13	270
Knowledge (Kn)	4	5	403
Wdbc (Wd)	2	30	569
Hayes-Roth (Hay)	3	5	160
Pima (Pi)	2	8	768
Red wine quality (Rwq)	6	11	1599
White wine quality (Wwq)	7	11	4898
Iris (Ir)	3	4	150
Ionosphere (Io)	2	34	351
Tae (Ta)	3	5	151
Movement-libras (Ml)	15	90	360
Vehicle (Ve)	4	18	946
Parkinson (Pa)	21	26	2938
German (Ge)	2	24	1000
Segment (Seg)	7	19	2310
Seeds (Se)	3	7	210
sonar (So)	2	60	208

4. Experiment applications

Three experiments have been carried out to test and evaluate the performance of the proposed method with that of some classical methods, i.e., mean imputation (MI) method [4, 5], K-Nearest neighbors imputation (KNNI) method [6], fuzzy c-means imputation (FCMI) method [8, 9], fuzzy-based information decomposition (FID) method [11], prototype-based credal classification (PCC) method [27] and locally linear approximation method [12].

In our experiments, we assume the presence of partially incomplete patterns in both the training set and the test set. In MI, the missing values are replaced by the mean values of the same attributes of the complete patterns in the training and test sets. In KNNI, the missing values in the different sets are estimated using the neighbors. In FCMI, the missing values of incomplete patterns in each set are filled based on the centers generated by FCM and the distances between the pattern and the cluster centers. In FID, the missing values of incomplete patterns are estimated by the observed data (complete attributes) in different sets and the different contributions (weights) of the observed data obtained through the membership function. In PCC, the missing values of incomplete patterns

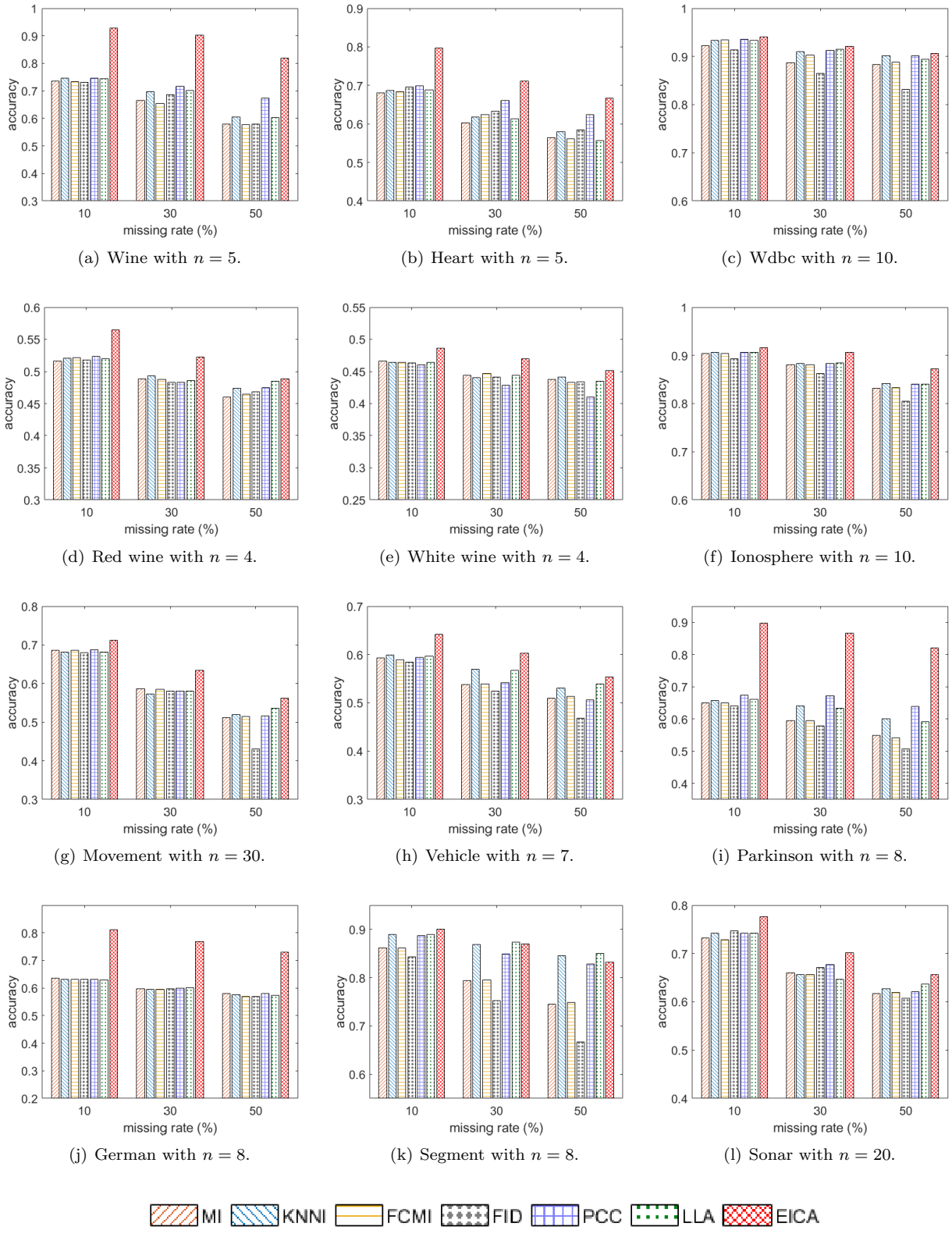
in the training set are replaced by the mean value of the same class of complete patterns, while the others are estimated in multiple versions separately for each prototype in the training set. In LLA, the missing values are estimated by the optimal weights of the KNNs obtained from local linear reconstruction of the different sets. In addition, the default parameters of these methods are given as follows. Specifically, we take $K = 11$ in KNNI, $\beta = 2$, $\epsilon = 10^{-3}$ in FCMI, $\epsilon = 0$ in PCC, $K = 11$, iter=10 in LLA, and $K = 11$ in EICA.

In this paper, K-Nearest neighbor (K-NN) [40], Naive Bayesian (NB) [39] and Evidence K-Nearest neighbor (EK-NN) [34] classifier are employed as basic classifiers to classify test patterns. $K = 11$ is default in KNNs, K-NN and EK-NN, and the parameters of EK-NN are automatically optimized by the method introduced in [34].

In our simulations, the classification accuracy of the test patterns is used as the main indicator to assess the effectiveness of these methods. In addition, the precision (P), recall (R), F1 score (F1), and random index (RI) are also employed as indicators in Experiment 3 to show the robustness of the proposed method. 18 well-known real datasets obtained from the UCI repository are used here as the benchmarks. Half of each dataset is randomly selected as training patterns and the rest consists of test patterns, some of which randomly lose the n attributes. Here ten training and test sets are randomly generated for the same dataset, and the average of the evaluation index is reported. In PCC, we adjust the meta-class threshold $\epsilon = 0$ so that it can produce certain classification results. The basics of the used dataset, including the number of classes (#Class.), attributes (#Attr.), and instances (#Inst.), are shown in Table 1, all details of which can be found at http://archive.ics.uci.edu/ml/. In the follow-up, each experiment used part of these datasets, as determined by how the attributes are combined. For example, the Iris dataset has only four attributes, which does not lend itself to random combinations. Therefore, this dataset is not used in Experiments 1 and 2. By contrast, the datasets with prior knowledge of attribute correlations are tested in Experiment 3.

Table 2: Classification accuracy of different methods with different Φ (In %).

Data	n	MI	KNNI	FCMI	FID	PCC	LLA	EICA ($\Phi = 2$)	EICA ($\Phi = 3$)	EICA ($\Phi = 4$)	EICA ($\Phi = 5$)
Wi	5	66.52±3.36	69.66±3.76	65.39±2.70	68.55±2.13	71.69±1.31	70.11±1.83	**89.89**±1.01	**90.56**±2.20	**90.56**±1.53	**90.11**±1.10
Wi	7	60.90±3.05	65.17±2.36	63.60±2.72	66.79±2.16	68.54±1.23	65.62±2.03	**86.29**±1.65	**84.94**±1.68	**86.52**±1.74	**84.94**±2.08
He	5	60.30±2.13	61.78±2.23	62.37±3.16	63.26±2.55	66.07±2.79	61.33±1.71	**68.74**±1.84	**71.70**±1.43	**70.96**±0.86	**71.70**±1.71
He	7	58.81±2.83	58.07±2.32	59.11±2.75	62.77±1.24	64.15±1.91	60.44±1.37	**67.85**±0.89	**70.96**±2.36	**70.81**±1.29	**69.48**±1.84
Wd	10	88.63±1.23	91.02±0.82	90.32±1.01	86.53±0.96	91.30±0.93	91.44±0.85	**92.14**±1.51	**92.70**±0.34	**91.79**±0.61	**91.79**±0.98
Wd	20	87.16±1.51	90.60±0.81	89.54±1.71	81.26±1.31	**90.95**±0.81	**91.02**±1.34	90.88±0.74	**92.14**±0.82	90.95±0.60	90.88±1.28
Rwq	4	48.85±1.51	49.29±1.47	48.75±1.52	48.27±1.06	48.32±1.46	48.55±1.41	**52.49**±1.69	**51.86**±1.31	**53.03**±1.09	**54.07**±1.34
Rwq	6	49.34±0.66	48.93±1.42	48.17±0.60	47.88±1.59	47.81±1.19	49.52±0.78	**51.88**±1.01	**52.26**±1.14	**52.14**±0.62	**51.35**±1.53
Wwq	4	44.44±0.85	44.00±0.27	44.71±0.94	44.09±0.45	42.80±0.65	44.39±0.77	**47.24**±1.43	**47.93**±1.66	**46.93**±0.73	**45.67**±1.86
Wwq	6	44.21±1.00	44.97±0.38	44.87±0.92	43.90±0.94	39.75±0.77	44.20±0.81	**46.48**±1.48	**47.39**±0.85	**46.92**±0.82	**46.80**±1.02
Io	10	88.07±1.30	88.30±1.37	88.07±1.08	86.25±1.88	88.30±1.33	88.41±1.51	**90.34**±1.65	**90.68**±1.22	**90.45**±1.36	**91.02**±1.16
Io	20	80.68±2.03	85.34±0.43	80.80±2.11	80.57±1.66	83.64±2.01	84.43±0.45	**86.36**±1.25	**87.27**±1.32	**87.16**±1.37	**87.27**±1.17
Ml	30	58.67±1.43	57.33±1.38	58.56±1.43	58.11±1.43	58.00±2.01	58.11±1.34	**62.67**±1.02	**63.11**±0.57	**63.44**±1.51	**63.55**±1.30
Ml	50	53.00±1.20	52.89±2.09	52.89±1.38	46.89±1.30	52.00±2.45	54.22±2.15	**56.78**±1.33	**58.11**±1.71	**57.56**±1.59	**56.56**±1.42
Ve	7	53.81±0.51	56.93±1.67	53.90±0.42	52.44±1.65	54.18±0.61	56.69±0.75	**60.75**±1.76	**61.23**±1.53	**60.28**±1.73	**59.57**±1.79
Ve	10	53.43±1.46	57.49±0.77	55.13±1.06	51.77±1.69	53.43±1.39	56.03±1.18	**58.49**±0.66	**59.81**±1.41	**58.96**±1.53	**58.20**±0.83
Pa	8	59.52±0.47	64.06±0.97	59.44±0.91	57.88±1.29	67.23±1.36	63.35±1.29	**88.14**±0.63	**86.59**±1.99	**86.83**±0.89	**85.13**±1.18
Pa	12	57.24±0.78	61.50±0.69	56.23±0.85	52.86±1.78	65.85±1.03	61.49±0.75	**84.97**±1.17	**83.18**±1.08	**84.66**±0.64	**82.69**±1.63
Ge	3	61.56±0.81	61.76±1.66	61.32±0.80	61.64±1.50	61.96±1.02	62.03±1.23	**75.28**±1.92	**79.76**±1.41	**80.80**±2.08	**81.44**±0.73
Ge	8	59.64±1.18	59.40±1.41	59.48±1.54	59.72±0.97	59.84±1.46	60.20±0.97	**73.96**±1.99	**77.12**±1.03	**77.36**±1.93	**79.04**±0.60
Seg	4	85.35±0.59	**89.04**±0.71	85.44±0.67	81.19±0.89	**88.21**±0.85	**89.07**±0.91	**90.94**±1.01	**89.32**±0.50	**89.31**±0.83	88.19±0.60
Seg	8	79.41±0.79	**86.89**±0.89	79.57±0.85	75.19±1.48	84.90±1.17	**87.38**±1.08	**87.86**±0.78	**87.80**±1.60	**87.43**±0.64	85.31±1.05
So	20	65.96±1.56	65.58±1.65	65.58±1.96	67.11±2.94	67.69±1.68	64.62±2.61	**69.61**±1.56	**69.81**±1.98	**70.19**±1.61	**70.77**±1.31
So	40	61.54±1.36	62.50±0.86	61.54±1.36	61.73±2.23	63.27±2.23	63.65±2.05	**66.73**±1.44	**66.54**±1.54	**68.46**±1.86	**68.84**±1.98
Ave		63.63±1.40	65.52±1.35	63.95±1.44	62.78±1.55	65.83±1.40	64.30±1.30	**72.78**±1.31	**73.45**±1.36	**73.47**±1.21	**73.11**±1.31
WT		0	2	0	0	2	3	23	24	23	21



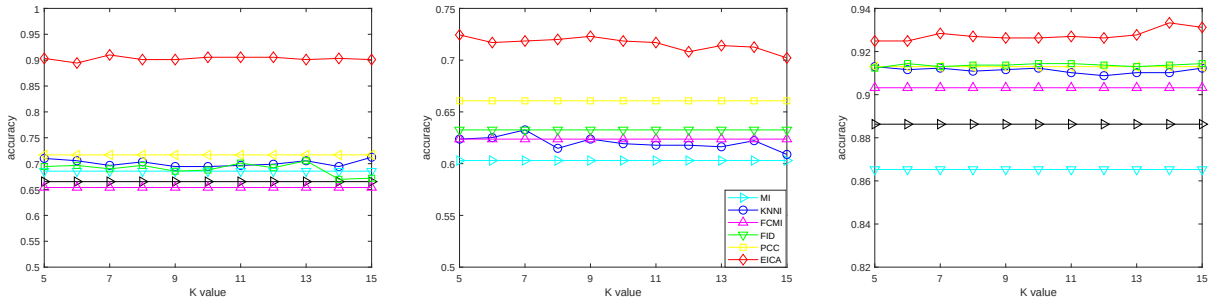
(a) Wine with $n = 5$. (b) Heart with $n = 5$. (c) Wdbc with $n = 10$.

(d) Red wine with $n = 4$. (e) White wine with $n = 4$. (f) Ionosphere with $n = 10$.

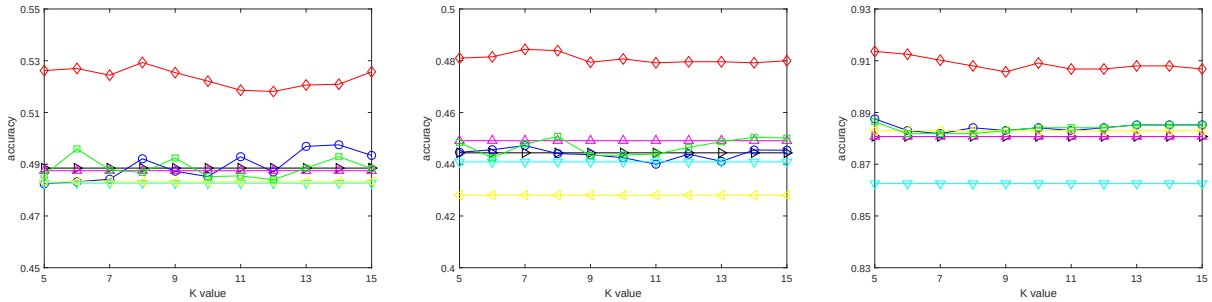
(g) Movement with $n = 30$. (h) Vehicle with $n = 7$. (i) Parkinson with $n = 8$.

(j) German with $n = 8$. (k) Segment with $n = 8$. (l) Sonar with $n = 20$.

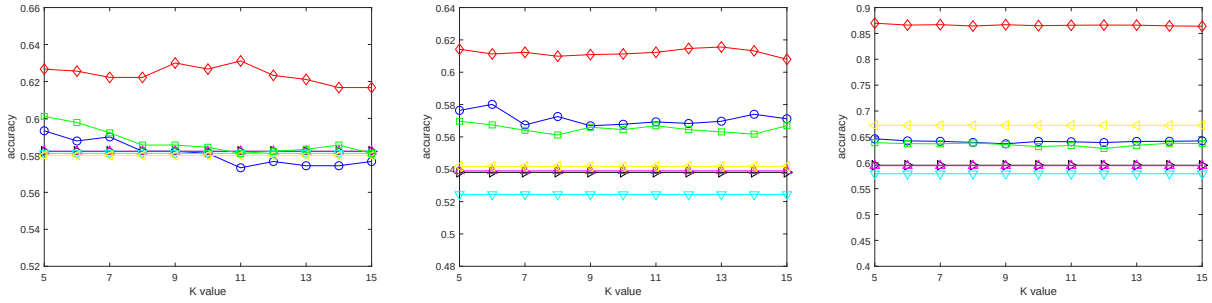
Figure 2: Classification accuracy of different methods with various missing rates.



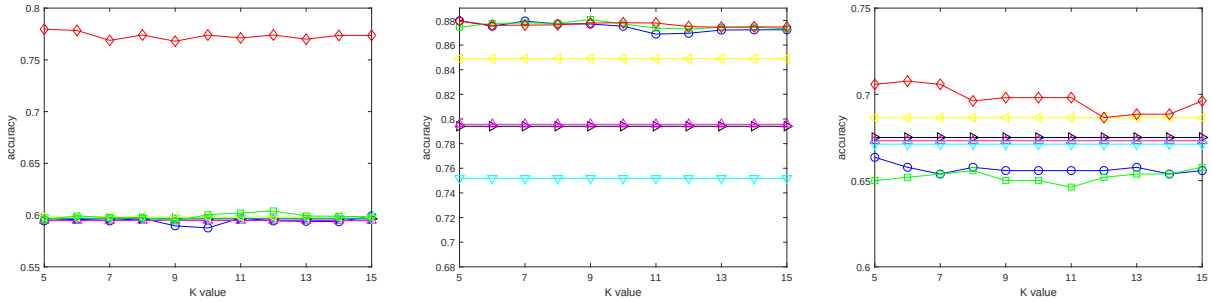
(a) Wine with $n = 5$ (b) Heart with $n = 5$ (c) Wdbc with $n = 10$



(d) Red wine with $n = 4$ (e) White wine with $n = 4$ (f) Ionosphere with $n = 10$



(g) Movement with $n = 30$ (h) Vehicle with $n = 7$ (i) Parkinson with $n = 8$



(j) German with $n = 8$ (k) Segment with $n = 8$ (l) Sonar with $n = 20$

Figure 3: Classification results of different methods with various K.

4.1. Experiment 1

In this experiment, we use 12 real datasets to test the performance of EICA with respect to MI, KNNI, FCMI, FID, PCC, and LLA. In each set, randomly selected 30% patterns are used for missing values in different dimensions. Here EK-NN is chosen as the basic classifier. The average accuracy (including standard deviation) of the different methods is shown in Table 2. In EICA, it can be seen that with different values of Φ, the level of accuracy varies (i.e., Φ = 2, 3, 4, 5). Here we assume that the correlation between attributes is not known in advance, i.e., the dataset is randomly divided into Φ subsets, as evenly as possible.

From Table 2, it can be seen that the EICA method generally has higher accuracy than that of other methods. That is, evidence integration classification yields higher robustness than single classification, suggesting that different classifiers can provide useful and complementary information to improve classification performance. It is worth noting that an increase in the number of missing values (i.e., n) in the training and test sets generally leads to a decrease in accuracy. The higher the number of missing values, the higher the likelihood that the estimate is wrong, due to the inevitable deviation between the estimations and the true. In the process, EICA uses local modeling, which can effectively reduce the possible negative impact of the distribution of missing data.

Table 3: Classification accuracy of different methods with K-NN classifier (In %).

Data	n	MI	KNNI	FCMI	FID	PCC	LLA	EICA
Wi	5	65.17±2.56	69.21±3.38	65.84±2.62	67.79±0.95	70.11±1.52	68.31±3.58	**90.78**±1.31
Wi	7	61.35±2.08	65.39±1.93	63.37±3.52	66.92±1.82	69.66±1.00	64.94±2.18	**84.95**±2.20
He	5	60.59±2.36	61.93±1.66	62.07±3.05	63.26±2.42	63.85±1.51	61.33±1.84	**71.85**±3.86
He	7	59.85±1.44	59.85±1.90	60.74±1.33	62.52±1.21	63.70±1.81	60.59±2.01	**70.67**±2.12
Hay	1	44.54±1.54	44.54±1.54	44.54±1.54	44.54±1.54	44.54±1.54	44.55±1.55	**54.24**±1.77
Hay	2	42.42±1.92	41.51±2.81	42.42±2.71	41.82±2.64	42.73±2.61	41.52±2.61	**43.94**±3.59
Pi	1	70.94±2.05	71.82±1.81	71.20±1.85	70.47±2.50	71.15±2.19	72.45±2.49	**74.95**±1.44
Pi	3	69.64±1.02	69.53±0.89	69.01±1.27	69.53±1.66	69.11±1.10	68.59±1.99	**72.24**±1.14
Rwq	4	49.59±1.86	49.47±1.37	49.75±1.89	48.68±0.68	48.80±2.12	49.08±1.30	**52.88**±1.40
Rwq	6	49.87±0.90	49.64±1.67	49.01±0.41	48.80±1.37	49.19±1.14	50.25±1.25	**52.72**±1.41
Wwq	4	44.87±0.83	44.28±0.19	45.00±0.75	44.33±0.67	44.22±0.93	44.54±0.77	**49.92**±0.88
Wwq	6	44.60±0.94	44.93±0.28	44.89±1.03	44.07±0.74	41.01±1.19	44.32±0.81	**48.75**±0.54
Io	10	82.39±1.24	83.52±1.19	82.27±1.04	82.96±1.24	84.20±0.91	84.09±1.48	**84.77**±0.98
Io	20	72.84±2.05	76.93±3.57	72.50±2.29	72.73±2.96	77.50±1.89	75.68±3.44	**80.12**±1.19
Ve	7	55.13±1.01	57.45±0.72	54.80±1.09	52.63±1.08	54.70±1.17	56.93±1.59	**61.09**±1.36
Ve	10	53.57±0.93	57.26±1.21	55.08±1.17	52.10±1.28	54.33±1.17	56.64±0.80	**60.38**±1.22
Pa	8	60.81±0.28	65.41±1.11	60.75±0.96	59.04±1.31	67.43±1.36	64.36±1.58	**86.90**±1.38
Pa	12	58.74±0.69	62.21±0.68	57.40±0.69	54.28±1.54	66.52±0.92	62.60±0.77	**83.37**±1.24
Ge	3	61.36±1.23	61.72±1.84	61.16±0.91	61.28±1.00	61.92±1.59	61.83±1.45	**78.64**±1.01
Ge	8	60.28±0.56	59.24±1.64	59.84±0.97	59.76±0.92	60.88±1.37	60.36±0.67	**76.12**±0.64
Seg	4	85.18±0.49	88.88±0.67	85.25±0.55	80.83±0.42	87.81±0.54	88.80±0.45	**90.60**±0.97
Seg	8	78.87±0.78	86.74±0.86	78.93±0.76	74.72±1.59	85.45±1.13	87.15±1.03	**88.69**±1.31
So	20	63.27±1.28	62.12±2.62	62.50±1.05	65.19±2.61	64.42±1.72	61.35±2.68	**67.12**±1.41
So	40	58.46±1.86	58.46±2.46	58.65±1.61	59.04±2.48	59.23±1.15	59.81±2.23	**65.58**±0.95
Ave		60.60±1.33	62.17±1.58	60.71±1.46	60.30±1.53	62.60±1.40	62.09±1.73	**70.47**±1.47
WT		0	0	0	0	0	0	24

Fig. 2 shows the accuracy of these methods with different missing rates. The x-axis indicates the missing rate (%) and the y-axis indicates the level of accuracy, where EICA reports the average accuracy of $\Phi \in \{2, 3, 4, 5\}$. We can intuitively see that EICA has better classification performance than that of other methods. It can also be seen that there is a tendency for the accuracy of the different methods to decrease as the proportion of missing values increases. This is consistent with the increase in n values in Table 2. Here the average accuracy (denoted by Ave)

Table 4: Classification accuracy of different methods with NB classifier (In %).

Data	n	MI	KNNI	FCMI	FID	PCC	LLA	EICA
Wi	5	92.13±1.23	92.81±1.15	92.58±1.68	92.08±1.55	93.71±1.52	93.03±1.31	**94.16**±1.93
Wi	7	90.11±1.65	89.89±2.56	90.11±2.18	90.26±3.07	**91.91**±1.65	89.44±1.96	90.78±1.31
He	5	80.74±0.66	80.59±1.09	81.04±0.76	80.96±1.18	81.19±1.29	80.00±1.93	**82.66**±0.75
He	7	78.52±1.05	79.41±1.78	78.22±0.59	78.26±2.74	78.96±1.79	79.41±1.84	**81.78**±1.66
Hay	1	59.74±2.61	60.61±1.92	60.00±1.55	59.70±2.06	60.91±2.01	59.70±2.06	**62.73**±1.21
Hay	2	55.46±1.82	57.27±2.61	56.36±1.77	54.85±3.51	55.45±2.81	56.06±2.71	**59.40**±2.94
Pi	1	72.45±2.11	72.81±1.99	72.71±1.84	72.40±2.95	72.71±2.02	73.02±1.85	**75.00**±1.24
Pi	3	70.99±1.02	71.82±1.36	71.04±1.35	70.36±1.28	71.72±1.24	71.25±1.58	**72.29**±1.26
Rwq	4	53.46±1.93	53.82±1.83	53.51±1.71	51.63±1.67	52.93±1.71	53.74±1.87	**55.62**±0.80
Rwq	6	51.40±0.87	51.55±1.03	50.87±1.18	51.20±2.01	51.70±1.00	51.30±0.75	**53.94**±1.53
Wwq	4	44.36±0.67	44.00±0.78	44.01±0.81	44.17±0.89	44.39±0.72	44.15±0.73	**49.04**±0.68
Wwq	6	43.41±0.54	42.90±0.84	43.12±0.70	44.42±1.16	43.03±0.46	43.00±0.82	**48.76**±1.41
Ve	7	45.53±1.53	42.79±2.03	43.88±2.00	41.84±2.85	44.02±1.87	43.07±1.77	**49.64**±2.79
Ve	10	43.36±2.04	43.07±1.97	43.26±2.15	40.57±2.09	42.98±2.77	43.07±1.91	**45.63**±1.84
So	20	65.96±1.78	66.73±0.47	66.15±1.65	67.12±0.72	66.92±2.16	66.35±1.05	**70.96**±1.66
So	40	63.85±2.76	64.42±2.02	63.65±3.41	61.35±3.82	65.38±2.51	64.04±2.32	**69.23**±2.58
Ave		63.22±1.52	63.41±1.59	63.16±1.58	62.57±2.10	63.62±1.72	63.16±1.65	**66.35**±1.60
WT		0	0	0	0	1	0	15

is given in the penultimate row to represent the general performance of the corresponding method and winning times[1] (denoted by WT) of different methods on these datasets are also reported in the last row to demonstrate the generalization capability of the proposed method. Although EICA experienced the same trend, the accuracy level is less affected by the missing values, as seen in Fig. 2, which proves the robustness of EICA.

Fig. 3 shows more clearly and intuitively the effect of different K values in KNN on the classification results of the KNNI, LLA, and EICA methods. The x-axis corresponds to the K values, ranging from 5 to 15, and the y-axis corresponds to the average accuracy in the classification method, denoted by $[0, 1]$. Among them, the 12 datasets shown in Fig. 3 are randomly missing in different dimensions. It can be observed that the accuracy of EICA is much higher than that of the other methods. The classification results associated with different K values do not change much in EICA, which further indicates that the classification performance is not sensitive to the setting of K values. This also indicates that the EICA method is robust to the choice of K values so that it can be taken from 5 to 15 in practice.

4.2. Experiment 2

In this experiment, we test and evaluate the performance of different methods using the 12 real datasets in Table 1. K-NN, NB[2], and EK-NN are chosen as the basic classifiers, and we choose $\Phi = 3$ in this experiment. The average accuracy (including standard deviation) of the different basic classifiers is reported in Tables 3-5.

As can be seen from the Tables 3-5, the accuracy of the EICA method using K-NN, EK-NN, and NB is higher than that of the other methods in most cases. In addition, as the missing values (i.e., n) increase, this may lead to a decrease in accuracy, but EICA still provides better performance than the other methods. Since conflicting information from different sources is effectively discounted to the full ignorance during evidence integration, it can

[1] If one method produces the maximum accuracy compared with the other methods, it wins one time.

[2] Naive Bayes does not apply to the Io, Pa, Ge, and Seg datasets because the within-class variance of several attributes is not positive.

Table 5: Classification accuracy of different methods with EK-NN classifier (In %).

Data	n	MI	KNNI	FCMI	FID	PCC	LLA	EICA
Wi	5	66.52±3.36	69.66±3.76	65.39±2.70	68.55±2.13	71.69±1.31	70.11±1.83	**90.56**±2.20
Wi	7	60.90±3.05	65.17±2.36	63.60±2.72	66.79±2.16	68.54±1.23	65.62±2.03	**84.94**±1.68
He	5	60.30±2.13	61.78±2.23	62.37±3.16	63.26±2.55	66.07±2.79	61.33±1.71	**71.70**±1.43
He	7	58.81±2.83	58.07±2.32	59.11±2.75	62.77±1.24	64.15±1.91	60.44±1.37	**70.96**±2.36
Hay	1	45.45±1.66	45.45±1.66	45.45±1.66	45.45±1.66	45.45±1.66	45.45±1.66	**58.49**±2.64
Hay	2	42.12±1.13	42.12±1.13	42.12±1.13	42.12±1.13	42.12±1.13	42.12±1.13	**50.91**±2.06
Pi	1	71.56±1.64	72.34±1.74	71.88±1.21	70.68±2.59	71.72±1.64	72.66±2.04	**74.27**±1.41
Pi	3	70.31±1.66	70.36±1.25	69.79±2.06	69.90±2.22	70.36±1.25	69.74±2.10	**70.62**±1.05
Rwq	4	48.85±1.51	49.29±1.47	48.75±1.52	48.27±1.06	48.32±1.46	48.55±1.41	**51.86**±1.31
Rwq	6	49.34±0.66	48.93±1.42	48.17±0.60	47.88±1.59	47.81±1.19	49.52±0.78	**52.26**±1.14
Wwq	4	44.44±0.85	44.00±0.27	44.71±0.94	44.09±0.45	42.80±0.65	44.39±0.77	**47.92**±1.66
Wwq	6	44.21±1.00	44.97±0.38	44.87±0.92	43.90±0.94	39.75±0.77	44.20±0.81	**47.39**±0.85
Io	10	88.07±1.30	88.30±1.37	88.07±1.08	86.25±1.88	88.30±1.33	88.41±1.51	**90.68**±1.22
Io	20	80.68±2.03	85.34±0.43	80.80±2.11	80.57±1.66	83.64±2.01	84.43±0.45	**87.27**±1.32
Ve	7	53.81±0.51	56.93±1.67	53.90±0.42	52.44±1.65	54.18±0.61	56.69±0.75	**61.23**±1.53
Ve	10	53.43±1.46	57.49±0.77	55.13±1.06	51.77±1.69	53.43±1.39	56.08±1.18	**59.81**±1.41
Pa	8	59.52±0.47	64.06±0.97	59.44±0.91	57.88±1.29	67.23±1.36	63.35±1.29	**86.59**±1.99
Pa	12	57.24±0.78	61.50±0.69	56.23±0.85	52.86±1.78	65.85±1.03	61.49±0.75	**83.18**±1.08
Ge	3	61.56±0.81	61.76±1.66	61.32±0.80	61.64±1.50	61.96±1.02	62.03±1.23	**79.76**±1.41
Ge	8	59.64±1.18	59.40±1.41	59.48±1.54	59.72±0.97	59.84±1.46	60.20±0.97	**77.12**±1.03
Seg	4	85.35±0.59	89.04±0.71	85.44±0.67	81.18±0.89	88.21±0.85	89.07±0.40	**89.32**±0.50
Seg	8	79.41±0.79	86.89±0.89	79.57±0.85	75.19±1.48	84.90±1.17	87.38±1.08	**87.80**±1.60
So	20	65.96±1.56	65.58±1.65	65.58±1.96	67.11±2.94	67.69±1.68	64.62±2.61	**69.81**±1.98
So	40	61.54±1.36	62.50±0.86	61.54±1.36	61.73±2.23	63.27±2.23	63.65±2.05	**66.54**±1.54
Ave		62.21±1.43	62.96±1.38	61.36±1.46	60.92±1.65	63.22±1.38	63.02±1.54	**71.29**±1.52
WT		0	0	0	0	0	0	24

effectively characterize the uncertainty caused by missing values while remaining neutral during the fusion process. Thus, the proposed method is effective in reducing the classification error. The average accuracy of different methods on these datasets of the same classifier is given in the penultimate row of the Tables 3-5. In addition, the winning times of each method on these datasets is reported in the last row of Tables 3-5. It can be seen that the EICA method is well adapted to the three basic classifiers: K-NN, NB, and EK-NN. In other words, the EICA method is robust and can be applied to a variety of basic classifiers. However, in the case of large amounts of data, we find that the NB classifier is less time consuming than K-NN and EK-NN. This is because the K-NN and EK-NN methods impose a heavy computational burden in such a case.

4.3. Experiment 3

In this experiment, we combine 11 real datasets into Φ subsets according to the correlation between attributes, as shown in Table 6. Here K-NN is chosen as the basic classifier. The average classification accuracy (including standard deviation) of the different methods is given in Table 7.

As shown in Table 7, our proposed EICA method outperforms other methods in terms of accuracy because EICA fully considers the correlation between attributes and thus divides the dataset into several reasonable subsets. By doing this, it can effectively avoid the negative effects of attribute distribution diversity and provide useful information for the final decision. EICA reduces error rates and improves classification performance by combining useful information of different classifiers under the framework of evidence theory. PCC, while also working within the framework of evidence theory, may not be accurate enough based on class prototype estimations, and thus

Table 6: Specific attribute combinations of different datasets.

Data	Attritube
Wi	{Magnesium, Nonflavanoid phenols, Proanthocyanins, OD280/OD315 of diluted wines}
	{Alcohol, Flavanoids, Proline}
	{Total phenols, Color intensity, Hue}
	{Malic acid, Ash, Alcalinity of ash}
He	{Resting blood pressure, Serum cholestoral, Fasting blood sugar, Oldpeak, The slope of the ST segment}
	{Age, Sex, Number of major vessels, Thal}
	{Chest pain type, Resting electrocardiographic, Maximum heart rate, Exercise induced angina }
Kn	{STR, LPR, PEG}
	{STG, SCG}
Hay	{Hobby, Educational level}
	{Age, Marital status}
	{Name}
Pi	{Pregnancies, Insulin, BMI, Age}
	{Glucose, Blood Pressure, Skin Thickness, Diabetes Pedigree Function}
Rwq	{Fixed acidity, Volatile acidity, Citric acid, Residual sugar}
	{Chlorides, pH, Sulphates, Alcohol}
	{Free sulfur dioxide, Total sulfur dioxide, Density}
Wwq	{Fixed acidity, Volatile acidity, Citric acid, Residual sugar}
	{Chlorides, pH, Sulphates, Alcohol}
	{Free sulfur dioxide, Total sulfur dioxide, Density}
Ir	{Sepal length in cm, Sepal width in cm }
	{Petal length in cm, Petal width in cm}
Ta	{Native English speaker, Course instructor, Class size}
	{Course, Summer or regular semester}
Ve	{Scaled variance along major axis, Scaled variance along minor axis, Skewness about major axis, Skewness about minor axis, Kurtosis about major axis, Kurtosis about minor axis}
	{Verage perim, Distance from border, Axis aspect ratio, Length aspect ratio, Scatter ratio}
	{Elongatedness, Axis rectangularity, Length rectangularity, Hollows ratio}
	{Average radius, Radius ratio, Scaled radius of gyration}
Se	{Area, Perimeter, Compactness, Asymmetry coefficient}
	{ Width of kernel, Length of kernel, Length of kernel groove}

PCC has a lower level of accuracy than EICA. Our tests and analyses illustrate the interest of EICA in classifying incomplete patterns.

Since EICA and PCC both works under the framework of evidence theory, we will further analyze the performance of them separately. In addition to comparing the accuracy on 11 real datasets, the precision (P), recall (R), F1 score ($F1$), and random index (RI) are also employed here to show the robustness of the proposed method. Fig. 4 shows the mean values of P, R, $F1$, and RI obtained by EICA relative to PCC. As can be seen in Fig. 4(a), EICA outperforms PCC on the 11 datasets in terms of the mean P value. In Fig. 4(b), the mean R of EICA is higher than that of PCC in most cases except for the Pima dataset. Similarly, Fig. 4(c) and Fig. 4(d) also verify the superiority of EICA in terms of $F1$ and RI. In Fig. 4(d), the mean RI value of EICA is higher than that of PCC on the 10 datasets except for the white wine dataset.

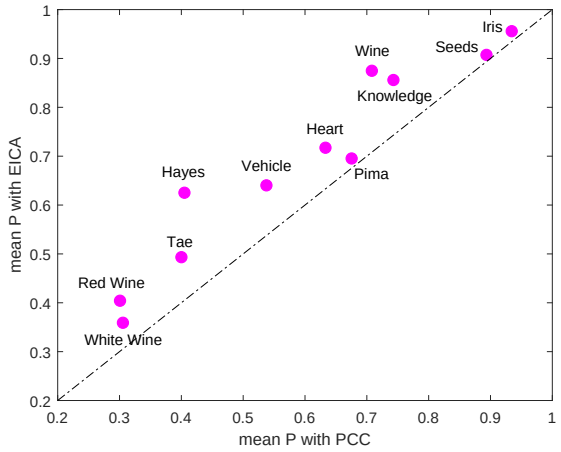
5. Conclusion

In this paper, a new evidence integration credal classification algorithm (EICA) is proposed to classify incomplete patterns thanks to the evidence theory. The proposed EICA method is dedicated to solve the classification problem

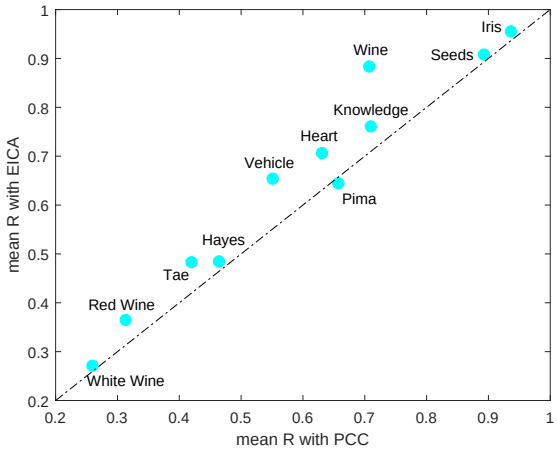
Table 7: Classification accuracy of different datasets with specific combinations (In %).

Data	n	MI	KNNI	FCMI	FID	PCC	LLA	EICA
Wi	5	65.17±2.56	69.21±3.38	65.84±2.62	67.79±0.95	70.11±1.52	68.31±3.58	**89.66**±0.84
Wi	7	61.35±2.08	65.39±1.93	63.37±3.52	66.92±1.82	69.66±1.00	65.62±1.83	**84.72**±2.62
He	5	60.59±2.36	61.93±1.66	62.07±3.05	63.26±2.42	63.85±1.51	61.63±1.53	**71.41**±2.55
He	7	59.85±1.44	59.85±1.90	60.74±1.33	62.52±1.21	63.70±1.81	62.52±3.06	**69.33**±2.22
Kn	1	77.52±1.79	77.82±1.15	77.72±1.96	77.13±1.30	77.03±1.20	77.72±1.50	**83.27**±1.75
Kn	2	68.91±2.06	69.41±2.33	69.41±2.45	67.43±2.65	70.59±0.80	70.10±2.11	**77.53**±1.52
Hay	1	44.54±1.54	44.54±1.54	44.54±1.54	44.54±1.54	44.54±1.54	44.55±1.55	**50.91**±1.55
Hay	2	42.42±1.92	41.82±2.64	42.42±2.71	41.82±2.64	42.42±2.71	41.52±2.81	**49.70**±2.94
Pi	1	70.94±2.05	71.82±1.81	71.20±1.85	70.47±2.50	71.15±2.19	72.45±2.49	**72.66**±0.96
Pi	3	69.64±1.02	69.53±0.89	69.01±1.27	69.53±1.66	69.11±1.10	68.59±1.99	**70.99**±0.39
Rwq	4	49.59±1.86	49.47±1.37	49.75±1.89	48.68±0.68	48.80±2.12	49.08±1.30	**56.24**±1.56
Rwq	6	49.87±0.90	49.64±1.67	49.01±0.41	48.80±1.37	49.19±1.14	50.25±1.25	**55.78**±1.66
Wwq	4	44.87±0.83	44.28±0.19	45.00±0.75	44.33±0.67	44.22±0.93	44.54±0.77	**48.81**±0.73
Wwq	6	44.60±0.94	44.93±0.28	44.89±1.03	44.07±0.74	41.01±1.19	44.32±0.81	**49.73**±0.84
Ir	1	90.40±3.20	94.67±1.46	94.67±1.46	91.20±3.11	95.20±1.36	94.40±2.44	**96.80**±1.36
Ir	2	86.40±1.96	89.07±1.55	90.67±1.69	87.73±2.59	91.47±0.65	90.93±3.09	**93.86**±1.81
Ta	1	42.63±2.71	40.00±1.78	42.89±2.44	42.37±2.93	43.42±1.66	43.42±1.86	**50.00**±3.22
Ta	3	38.42±1.29	38.95±1.97	36.32±1.78	38.95±2.58	38.42±1.93	38.42±0.53	**45.26**±3.39
Ve	7	55.13±1.01	57.45±0.72	54.80±1.09	52.63±1.08	54.70±1.17	56.93±1.59	**66.38**±1.48
Ve	10	53.57±0.93	57.26±1.21	55.08±1.17	52.10±1.28	54.33±1.17	56.64±0.80	**63.07**±1.23
Se	1	90.29±1.40	90.86±0.47	90.67±0.71	89.71±1.11	90.67±0.71	91.43±1.04	**91.81**±0.47
Se	3	88.00±1.55	87.43±1.11	87.43±1.11	87.62±1.60	87.62±0.85	88.00±1.55	**89.52**±0.85
Ave		61.58±1.70	62.52±1.50	62.16±1.72	61.80±1.75	62.78±1.38	62.79±1.79	**69.43**±1.63
WT		0	0	0	0	0	0	22

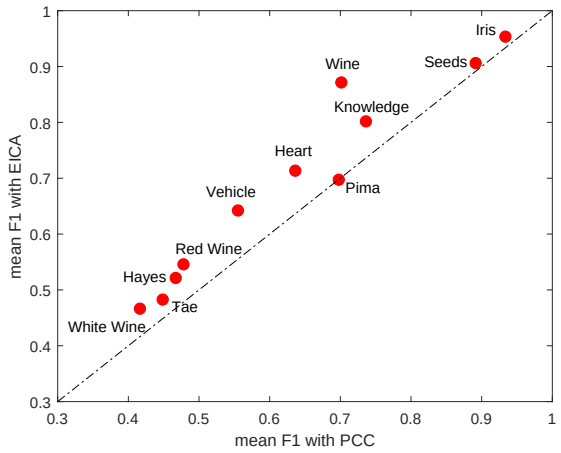
with missing values in both the training and test sets. In EICA, we first group the query set into multiple subsets depending on the correlation between attributes or random combinations without prior knowledge. This kind of local modeling for incomplete patterns can reduce the negative impact of attributes distribution diversity and provide more or less complementary information for decision-making. Then, we model the missing values in each subset using similar subpatterns, in which we also fully consider the importance of different attributes to minimize the negative impact caused by missing values. Finally, the discounted classification results of multiple subversions, represented by basic belief assignment (BBA), are fused globally to determine the final class of incomplete patterns. Three experiments with real datasets have been done to evaluate the performances of EICA with respect to other classical methods. The results show that local modeling of missing values is effective for the classification of incomplete patterns. In this paper, we give the combinations based on prior knowledge or randomly. This however may not suitable for some specific applications. In the future, we will consider a more effective credal classification method for incomplete patterns, especially in some specific cases, such as imbalanced data [41, 42], mining the original distribution information of data attributes in multiple subversions from adaptive selection and imputation methods. In addition, we find large differences in the distribution characteristics across attributes, suggesting that local modeling, while reducing the negative impact of missing values, has the potential for further improvement. It is thereby considered to model the missing values at the attribute-level to fully consider the distribution characteristics of the attributes and to improve the effective identification of incomplete patterns.



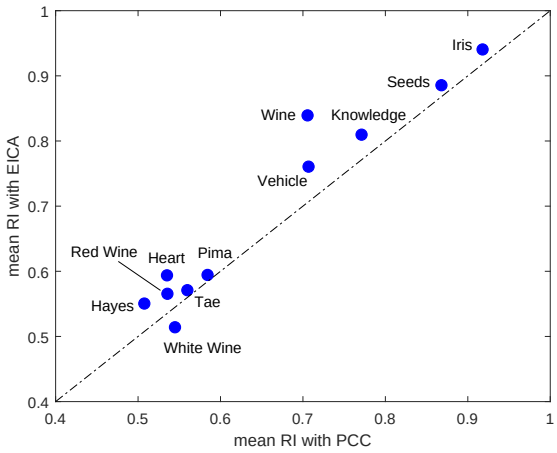
(a) P with EICA vs. P with PCC.



(b) R with EICA vs. R with PCC.



(c) $F1$ with EICA vs. $F1$ with PCC.



(d) RI with EICA vs. RI with PCC.

Figure 4: The mean $P, R, F1$ and RI values via different methods.

6. Acknowledgement

This work has been partially supported by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University, China (No. CX201953).

REFERENCES

[1] J. G. Pedro, S. José-Luis, R. F. Aníbal, Pattern classification with missing data: a review. Neural Computing & Applications, 19 (2) (2010) 263-282.

[2] R. J. A. Little, D. B. Rubin, Statistical analysis with missing data, Wiley, (1986).

[3] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognition, 41 (12) (2008) 3692-3705.

[4] D. J. Mundfrom, A. Whitcomb, Imputing missing values: The effect on the accuracy of classification, MLRV, 25 (1) (1998) 13-19.

[5] A. R. T. Donders, J. M. G. van der Heijden Geert, T. Stijnen, K. G. M. Moons, Review: a gentle introduction to imputation of missing values., Journal of Clinical Epidemiology, 59 (10) (2006) 1087-1091.

[6] O. Troyanskaya, M. Cantor, G. Sherlock, Brown, P., Hastie, T., Tibshirani, R., Missing value estimation methods for dna microarrays, Bioinformatics, 17 (6) (2001) 520-525.

[7] C. H. Cheng, C. P. Chan, Y. J. Sheu, A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction, Engineering Applications of Artificial Intelligence, 81 (2019) 283-299.

[8] L. Julián, F. H. José A Sáez, Missing data imputation for fuzzy rule-based classification systems, Soft Computing, 16 (5) (2012) 863-881.

[9] D. Li, J. Deogun, W. Spaulding, B. Shuart, Towards missing data imputation: a study of fuzzy k-means clustering method, Rough Sets & Current Trends in Computing, International Conference, Rsctc, Uppsala, Sweden, June, (2004) 573-579.

[10] I. B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences, 233 (2013) 25-35.

[11] S. Liu, J. Zhang, Y. Xiang, W. Zhou, Fuzzy-based information decomposition for incomplete and imbalanced data learning, IEEE Transactions on Fuzzy Systems, 25 (6) (2017) 1476-1490.

[12] J. H. Dai, H. Hu, Q. H. Hu, W. Huang, N. G. Zheng and L. Liu, Locally linear approximation approach for incomplete data, IEEE Transactions on Cybernetics, 48 (6) (2018) 1720-1732.

[13] K. Pelckmans, J. D. Brabanter, J. A. K. Suykens, B. D. Moor, Handling missing values in support vector machine classifiers, Neural Networks, 18 (5-6) (2005) 684-692.

[14] P. Chan, O. J. Dunn, The treatment of missing values in discriminant analysis, Journal of the American Statistical Association, 67 (338) (1972) 473-477.

[15] D. Williams, X. Liao, Y. Xue, L. Carin, B. Krishnapuram, On classification with incomplete data, IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (3) (2007) 427-436.

[16] M. Ramoni, P. Sebastiani, Robust learning with missing data, Machine Learning, 45 (2001) 147-170.

[17] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning, 20 (3) (1995) 273-297.

[18] G. Shafer, Mathematical Theory of Evidence, Princeton, NJ, USA: Princeton Univ. Press, (1976).

[19] P. Smets, The combination of evidence in the transferable belief model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (5) (1990) 447-458.

[20] P. Smets, Analyzing the combination of conflicting belief functions, Information Fusion, 8 (4) (2007) 387-412.

[21] L. G. De O. Silva, A. T. De Almeida-Filho, A multicriteria approach for analysis of conflicts in evidence theory, Information Sciences 346-347 (2016) 275-285.

[22] T. Denœux, Decision-making with belief functions: a review, International Journal of Approximate Reasoning, 109 (2019) 87-110.

[23] B. Quost, M. H. Masson, T. Denœux, Classifier fusion in the dempster-shafer framework using optimized t-norm based combination rules, International Journal of Approximate Reasoning, 52 (3) (2011) 353-374.

[24] Y. Leung, N. N. Ji, J. H. Ma, An integrated information fusion approach based on the theory of evidence and group decision-making, Information Fusion, 14 (4) (2013) 410-422.

[25] J. Zhao, R. Xue, Z. N. Dong, D. Y. Tang, W. H. Wei, Evaluating the reliability of sources of evidence with a two-perspective approach in classification problems based on evidence theory, Information Sciences, 507 (2020) 313-338.

[26] Z. He, W. Jiang, An evidential Markov decision making model, Information Sciences, 467 (2018) 357-372.

[27] Z. G. Liu, Q. Pan, G. Mercier, J. Dezert, A new incomplete pattern classification method based on evidential reasoning, IEEE Transactions on Cybernetics, 45 (4) (2015) 635-646.

[28] Y. Zhang, Y. Liu, H. Chao, Z. J. Zhang, Z. Y. Zhang, Classification of incomplete data based on evidence theory and an extreme learning machine in wireless sensor networks. Sensors, 18(4), (2018) 1046-1061.

[29] M. H. Masson, T. Denœux, ECM: an evidential version of the fuzzy c-means algorithm, Pattern Recognition, 41 (4) (2008) 1384-1397.

[30] Z. G. Su, T. Denœux, BPEC: belief-peaks evidential clustering, IEEE Transactions on Fuzzy Systems , 27 (1) (2019) 111-123.

[31] Z. W. Zhang, Z. Liu, Z. F. Ma, Y. R. Zhang, H. Wang, A new belief-based incomplete pattern unsupervised classification method, (2021), DOI 10.1109/TKDE.2021.3049511.

[32] Z. W. Zhang, Z. Liu, A. Martin, Z. G. Liu, K. Zhou, Dynamic evidential clustering algorithm, 213 (2021), DOI 10.1016/j.knosys.2020.106643.

[33] X. B. Xu, D. Q. Zhang, Y. Bai, L. L. Chang, J. N. Li, Evidence reasoning rule-based classifier with uncertainty quantification, Information Sciences, 514 (2020) 462-483.

[34] T. Denœux, A k-nearest neighbor classification rule based on dempster-shafer theory, IEEE Transactions on Systems, Man and Cybernetics, 25 (5) (1995) 804-813.

[35] T. Denœux, P. Smets, Classification using belief functions: relationship between case-based and model-based approaches, IEEE Transactions on Cybernetics, 36 (6) (2007) 1395-1406.

[36] D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures, Computational Intelligence, 4 (3) (1988) 244-264.

[37] F. Smarandache, J. Dezert, Information fusion based on new proportional conflict redistribution rules, in Proceedings of International Conference on Information Fusion, Philadelphia, PA, USA, July, (2005).

[38] F. Smarandache, J. Dezert, On the consistency of PCR6 with the averaging rule and its application to probability estimation, in Proceedings of International Conference on Information Fusion, Istanbul, Turkey, July, (2013).

[39] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, second ed., Prentice Hall, (2003).

[40] B. W. S. C. Jones, E. Fix and J. Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation: commentary on fix and hodges (1951), International Statistical Review / Revue Internationale de Statistique, 57 (3) (1989) 233-238.

[41] J. Liang, L. Bai, C. Dang, F. Cao, The k-means-type algorithms versus imbalanced data distributions, IEEE Transactions on Fuzzy System, 20 (4) (2012) 728-745.

[42] H. He, E. A. Garcia, Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9) (2009), 1263-1284.