Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(4): 4853–4873.
DOI: 10.3934/mbe.2024214
Received: 09 December 2023
Revised: 30 January 2024
Accepted: 13 February 2024
Published: 29 February 2024

Research article

Research on dependent evidence combination based on principal component analysis

Xiaoyan Su[1*], Shuwen Shang[1], Leihui Xiong[2], Ziying Hong[1], Jian Zhong[1]

[1] School of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China

[2] State Grid Nanchang Electric Power Supply Company, Nanchang 330069, China

* **Correspondence:** Email: suxiaoyan@shiep.edu.cn.

Abstract: Dempster-Shafer evidence theory, as a generalization of probability theory, is a powerful tool for dealing with a variety of uncertainties, such as incompleteness, ambiguity, and conflict. Because of its advantages in information fusion compared with traditional probability theory, it is widely used in various fields. However, the classic Dempster's combination rule assumes that evidences are independent of each other, which is difficult to satisfy in real life. Ignoring the dependence among the evidences will lead to unreasonable fusion results, and even wrong conclusions. Considering the limitations of D-S evidence theory, this paper proposed a new evidence fusion model based on principal component analysis (PCA) to deal with the dependence among evidences. First, the approximate independent principal components of each information source were obtained based on principal component analysis. Second, the principal component data set was used as a new information source for evidence theory. Third, the basic belief assignments (BBAs) were constructed. As the fundamental construct of evidence theory, a BBA is a probabilistic function corresponding to each hypothesis, quantifying the belief assigned based on the evidence at hand. This function facilitates the synthesis of disparate evidence sources into a mathematically coherent and unified belief structure. After constructing the BBAs, the BBAs were fused and a conclusion was drawn. The case study verified that the proposed method is more robust than several traditional methods and can deal with redundant information effectively to obtain more stable results.

Keywords: information fusion; evidence theory; dependence; basic belief assignment; principal component analysis

1. Introduction

With the popularity of big data, the amount of information that people can obtain increases exponentially. Therefore, it is particularly important to use a large amount of data to analyze and get useful results. As a means of processing big data, information fusion can effectively integrate all kinds of information and guide people's production and life. Evidence theory [1,2], as an influential method for information fusion, can effectively deal with uncertain information [3,4]. This has garnered significant attention and study from scholars [5–8], leading to its application in various domains such as pattern recognition [9–11], risk assessment [12–15], and multi-attribute decision making [16–18].

Evidence theory provides a fusion rule, namely the Dempster's combination rule. It is employed to combine multiple different sources of evidence into a comprehensive body of evidence. The purpose of this rule is to effectively handle and merge various pieces of evidence in situations characterized by uncertainty and incomplete information. However, classic Dempster's combination rule requires that evidences should be independent of each other [19]. This assumption limits the application scope of evidence theory to some extent, because dependence is universal in practice. For example, in the decision-making process, experts will discuss and exchange opinions with each other, so the conclusions they draw will synthesize their opinions and be dependent. If the evidence intersects is not taken into account, the influence of dependence among factors on the results will be calculated repeatedly in the process of evidence combination, leading to incorrect results. In order to solve this problem, many scholars put forward different methods. These methods fall into two main categories: improving the combination rule and modifying the original evidence.

The first type of method is to improve the combination rule. Since Dempster's rule requires evidences to be independent, an evidence fusion method without this constraint can be constructed [20, 21]. Chebbah et al. [21] suggested a method to quantify the degree of independence between evidence sources, which helps to select the most appropriate set of combinatorial rules to fuse evidence information from different sources. However, the method requires a sophisticated process of clustering, matching clusters, and quantifying independence, which might be computationally intensive and complex. Destercke et al. [22] made a more general definition of a rule in possibility theory and proposed the fusion rule, so that it can be used in the fusion of dependent evidence, however, Cattaneo [23] argued that this rule does not satisfy the basic evidence theory. Fu et al. [24] developed a method for fusing dependent interval-valued reliability functions. However, a potential limitation of this approach is its inherent complexity in both implementation and interpretation, particularly in scenarios characterized by significant interdependence among attributes or in dynamic decision-making environments.

The second type of method is to modify the original evidence, the main idea of which is to reduce the dependence among information sources and make the information sources approximately independent by discounting the dependent information, so as to reduce the repeated calculation in the process of fusion [20]. Dempster's combination rule has many advantages, such as satisfying the commutative law, associative law, and other mathematical characteristics, so the second type of method is attracting more and more attention. Su et al. [20] proposed an improved method for combining dependent evidence bodies, which takes the significance of the common information sources into consideration. However, it is constrained to scenarios involving only two sources. Extending this approach to incorporate multiple sources remains an unresolved issue. Su et al. [19] proposed a method to deal with

dependent evidence at the system level, which can grasp internal and external dependence. However, this approach is somewhat subjective. Shi et al. [25] used the rank correlation coefficient to generate the discount coefficient of dependent information. Su et al. [26] proposed an evidence fusion model based on mutual information to discount BBA. However, these methods cannot effectively deal with information redundancy. Kong et al. [27] developed a clinical decision support system (CDSS). Yao et al. [28] proposed a maximum likelihood evidential reasoning (MAKER) framework. These models build evidence bases from clinical historical data. The discount coefficient of evidence is derived from the training model. However, this model requires huge amounts of historical data, and the effect can be affected by the size and quality of the data set.

Principal component analysis (PCA) is a widely used data dimensionality reduction algorithm. The basic idea of PCA is to retain the most important features of the closely dependent high-dimensional data and turn it into the approximately mutually independent low-dimensional data, that is, the principal components, so that a small number of comprehensive indicators can be used to represent the most important information existing in the original data. PCA is widely used in computer science [29–31], engineering [32–34] , and other fields.

Therefore, following the idea of the second type of method, this paper proposes a dependent evidence fusion model based on PCA. Variables in the original data set may be dependent. Through linear combinations of these variables, PCA creates a new set of variables (the principal components) that are orthogonal to each other, meaning their dependence is greatly reduced or eliminated. Although PCA cannot guarantee complete independence (as this requires statistical independence and PCA only eliminates linear correlation), it significantly reduces the depenence between data features. Through coordinate transformation, PCA can effectively deal with the redundancy of information. PCA converts dependent data into approximately independent principal components, which means the evidence processed by PCA can approximately conform to the assumption of Dempster's combination rule.

This paper is organized as follows: Section 2 mainly introduces PCA and related concepts of evidence theory. Section 3 presents the steps of the proposed method. In Section 4, the effectiveness and superiority of this method are demonstrated through several experiments. Section 5 summarizes the method.

2. Preliminaries

2.1. D-S evidence theory [1, 2]

2.1.1. Discernment frame

Define the discernment frame Φ as the set of all possible and independent values of the variable x. Let the number of elements in Φ be μ, then the power set $P(\Phi)$ of Φ has 2^{μ} elements, each of which corresponds to a proposition about the value of x.

$$\begin{cases} \Phi = \{\xi_1, \xi_1, \ldots \xi_N\} \\ P(\Phi) = \{\oslash, \{\xi_1\}, \{\xi_2\}, \ldots, \{\xi_N\}, \{\xi_1, \xi_2, \}, \{\xi_1, \xi_3\}, \ldots \Phi\} \end{cases} \tag{2.1}$$

2.1.2. Basic belief assignment (BBA)

Let any element of $P(\Phi)$ correspond to a number whose value ranges $[0, 1]$. If the mapping is satisfied:

$$m(\oslash) = 0, \sum_{N \in P(\Phi)} m(N) = 1 \tag{2.2}$$

then m is the basic belief assignment (BBA) on Φ. $m(N)$ is a probability function that represents the degree to which the evidence supports proposition N to be true.

The evidence theory supports experts to give the confidence β of their judgment, the BBA m' discounted by confidence β is:

$$\begin{cases} m'(N) = \beta m(N), & \forall N \subset \Phi, N \neq \Phi \\ m'(\Phi) = 1 - \beta + \beta m(\Phi) \end{cases} \tag{2.3}$$

2.1.3. Dempster combination rule

In order to make the BBA of experts work together, it is necessary to establish the corresponding combination rules. Let $m_1, m_2, \ldots, m_h$ be h independent BBAs in Φ, corresponding to the propositions $N_1, N_2, \ldots, N_h$. Then the result of the combination is their orthogonal sum as follows:

$$C = \sum_{N_1 \cap N_2 \cap \cdots N_h = \oslash} \prod_{r=1}^{n} m_r(N_r) \tag{2.4}$$

and the BBA of the fused proposition Y is

$$m(Y) = \frac{1}{1 - C} \sum_{N_1 \cap N_2 \cap \cdots N_h = Y} \prod_{r=1}^{n} m_r(N_r) \tag{2.5}$$

where C is a normalized parameter reflecting the conflict degree between the evidence.

2.1.4. Pignistic probability transformation (PPT)

Let m be a BBA on Θ. Its corresponding pignistic probability function $BetP_m : \Phi \longrightarrow [0, 1]$ is defined as

$$BetP_m(w) = \sum_{N \subseteq \Phi, w \in N} \frac{1}{|N|} \frac{m(N)}{1 - m(\emptyset)}, \quad m(\emptyset) \neq 1, \tag{2.6}$$

where $|N|$ is the cardinality of subset N.

2.1.5. Application of evidence theory

The general usage process of evidence theory in the decision support field is as follows: experts evaluate a discernment frame, and based on the current context, prior knowledge, and experience, they provide judgments on propositions within the discernment frame. Evidence theory supports experts in assigning basic belief assignments (BBA) and confidence levels to any combination of propositions within the frame. Through Dempster's combination rule, the opinions of different experts are integrated to obtain a comprehensive proposition (the fused BBA) that reflects the consensus of all experts.

Take medical diagnosis as an example to illustrate the application process of evidence theory. In medical diagnosis, the discernment frame consists of n possible diseases. Experts provide judgments and confidence levels on which disease it could be based on symptoms and prior knowledge. For instance, Expert 1 might offer an opinion on Disease A with a certain confidence level (BBA 1), while Expert 2 might provide an opinion on either Disease A or B with a different confidence level (BBA 2). Evidence theory integrates the opinions of these experts to arrive at a final judgment opinion (fused BBA). The final opinion reveals the disease and its probability level, incorporating the views of both experts.

2.2. Principal component analysis (PCA) [35]

The specific steps of principal component analysis are as follows:

Step 1. Set the total number of samples in a data set as n and the number of attributes as p, then matrix $X = \left[X_{ij}\right](i \in [1,n], j = [1,p])$ can be obtained from the original data of the samples, where X_{ij} is the observed value of the jth attribute of the ith sample.

Step 2. Standardization of raw data. In general, the contents described by different attributes of data are different, they represent different physical meanings, and have different units of measurement. If this point is ignored and the principal component is calculated directly, it will lead to incorrect results. Therefore, it is necessary to standardize the raw data so that the influence of each attribute is at the same level. The standardized data $Z = \left[Z_{ij}\right](i \in [1,n], j = [1,p])$ is defined as follows.

$$Z_{ij} = \frac{X_{ij} - \frac{1}{n}\sum_{i=1}^{n} X_{ij}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \frac{1}{n}\sum_{i=1}^{n} X_{ij})^2}} \tag{2.7}$$

Step 3. Compute the correlation coefficient matrix R. The correlation coefficient represents the dependence degree between standardized data, and the smaller the value, the smaller the degree. There is little information overlap between variables with low dependence degree, otherwise, the information of variables will overlap, resulting in redundant information. Redundant information affects the objectivity of decision making. In order to identify the dependence of variables, it is necessary to construct matrix $R = \left[R_{ij}\right]_{p\times p}$.

The matrix R is calculated according to the standardized data, where R_{ij} reflects the dependence degree between index Z_i and Z_j, and the expression is as follows:

$$R_{ij} = \frac{\sum_{k=1}^{n}\left(Z_{kj} - \bar{Z}_i\right)\left(Z_{kj} - \bar{Z}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(Z_{kj} - \bar{Z}_i\right)^2\left(Z_{kj} - \bar{Z}_j\right)^2}} \tag{2.8}$$

where $\bar{Z}_i, \bar{Z}_j$ is the sample average of Z_i and Z_j.

Step 4. Compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq 0$ and the eigenvector $V_j = (V_{1j}, V_{2j}, \cdots, V_{pj})^T$ of the correlation coefficient matrix R according to Eq. (2.10).

$$V_1 = \begin{bmatrix} V_{11} \\ V_{21} \\ \vdots \\ V_{p1} \end{bmatrix}, V_2 = \begin{bmatrix} V_{12} \\ V_{22} \\ \vdots \\ V_{p2} \end{bmatrix}, \cdots, V_p = \begin{bmatrix} V_{1p} \\ V_{2p} \\ \vdots \\ V_{pp} \end{bmatrix} \tag{2.9}$$

$$RV = \lambda V \tag{2.10}$$

The eigenvector represents the dependence among the new K-dimensional data and the original N-dimensional data, and the larger the eigenvalue, the more representative the new data is of the corresponding original data. The eigenvalue is the variance of each principal component and represents the contribution degree of each principal component to the final result. V_{ij} is the element in row i and column j of the eigenvector matrix.

Step 5. Let $X_j = (X_{1j}, X_{2j}, \cdots, X_{nj})^T$, then the original data matrix is $X = [X_1, X_2, \cdots, X_P]$. The principal component values are expressed as follows:

$$\begin{cases} F_1 = V_{11} \times X_1 + V_{21} \times X_2 + \cdots + V_{P1} \times X_P \\ F_2 = V_{12} \times X_1 + V_{22} \times X_2 + \cdots + V_{P2} \times X_P \\ \vdots \\ F_P = V_{1P} \times X_1 + V_{2P} \times X_2 + \cdots + V_{PP} \times X_P \end{cases} \tag{2.11}$$

3. The proposed method

In order to fuse the dependence information accurately, this part proposes a method of dependence evidence fusion based on PCA.

Step 1. Acquire raw data

Acquire the original data matrix $X = \left[X_{ij}\right]_{n\times p}$ to serve as the input data for PCA, as folows:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \tag{3.1}$$

Step 2. Obtain principal components through PCA

There are common (redundant) parts among dependent information, which will be calculated repeatedly in the process of information fusion, resulting in unreasonable fusion results. To address this issue, this paper employs principal component analysis in the fusion of evidence. The core principle of PCA involves maximizing the variance of data points through coordinate rotation and transformation. This process enhances the most significant features (principal components) of the data while reducing dimensionality. In this way, PCA is able to identify and prioritize those components that capture the greatest variance in the data, thereby preserving critical information. Moreover, PCA eliminates redundant features from the data by selecting the principal components that best describe the data's variance. This means it selects the features that best explain the variance of the data while disregarding those features that are highly correlated with the selected features. Therefore, this paper use the characteristics of PCA to handle dependence among variables.

Through the basic steps of PCA (see Section 2.2), the original data set is transformed into independent principal components.

Step 3. Build BBAs of principal components

Build the BBA. This paper adopts the method of Xu et al. [36] to generate the BBA based on the data set of independent principal components after PCA processing.

Step 4. Fuse the BBA and draw a conclusion

The results are based on the fusion of the BBA. Eq. (2.5) and Eq. (2.6) are used to obtain the final fusion result.

4. Case study

The iris flower data set was first measured by Edgar Anderson and later used by R. Fisher in his 1936 paper [37] as an example of classification methods in statistics. Since then, this data set has been studied by many scholars, especially in the field of machine learning. The attributes of iris are dependent and can be seen as information sources for building BBAs. The species of iris can be identified through the processing of dependent information. Therefore, the identification of iris is actually research of dependent evidence fusion. Based on the iris species identification, Figure 1 shows the application process of this method.

4.1. Application of iris identification based on the proposed method

Step 1. Acquire raw data

The iris data set contains three species, namely, Setosa, Versicolour, and Virginica. We identify flower species through four different attributes: sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW). The data set provides 50 sample data containing these four attributes for each species of iris, with a total of 150 samples. This paper selects iris data set from a machine learning data set [38] to demonstrate the application process of the method.

Step 2. Obtain independent principal components through PCA

Eq. (2.7) is used for decentralized processing of the data set, and the matrix R of the four attributes is obtained through Eq. (2.8). The eigenvectors of the matrix R are computed as Table 1.

$$R = \begin{bmatrix} 1.0000 & -0.1094 & 0.8718 & 0.8180 \\ -0.1094 & 1.0000 & -0.4205 & -0.3565 \\ 0.8718 & -0.4205 & 1.0000 & 0.9628 \\ 0.8180 & -0.3565 & 0.9628 & 1.0000 \end{bmatrix}$$

Calculate the value of the principal component for each iris sample. By substituting the eigenvectors into Eq. (2.11), the expression of the principal components can be derived as Eq. (4.1).

$$\begin{cases} F_1 = 0.5224 \times X_1 - 0.2634 \times X_2 + 0.5813 \times X_3 + 0.5656 \times X_4 \\ F_2 = 0.3723 \times X_1 + 0.9256 \times X_2 + 0.0211 \times X_3 + 0.0654 \times X_4 \\ F_3 = -0.7210 \times X_1 + 0.2420 \times X_2 + 0.1409 \times X_3 + 0.6338 \times X_4 \\ F_4 = -0.2620 \times X_1 + 0.1241 \times X_2 + 0.8012 \times X_3 - 0.5235 \times X_4 \end{cases} \tag{4.1}$$

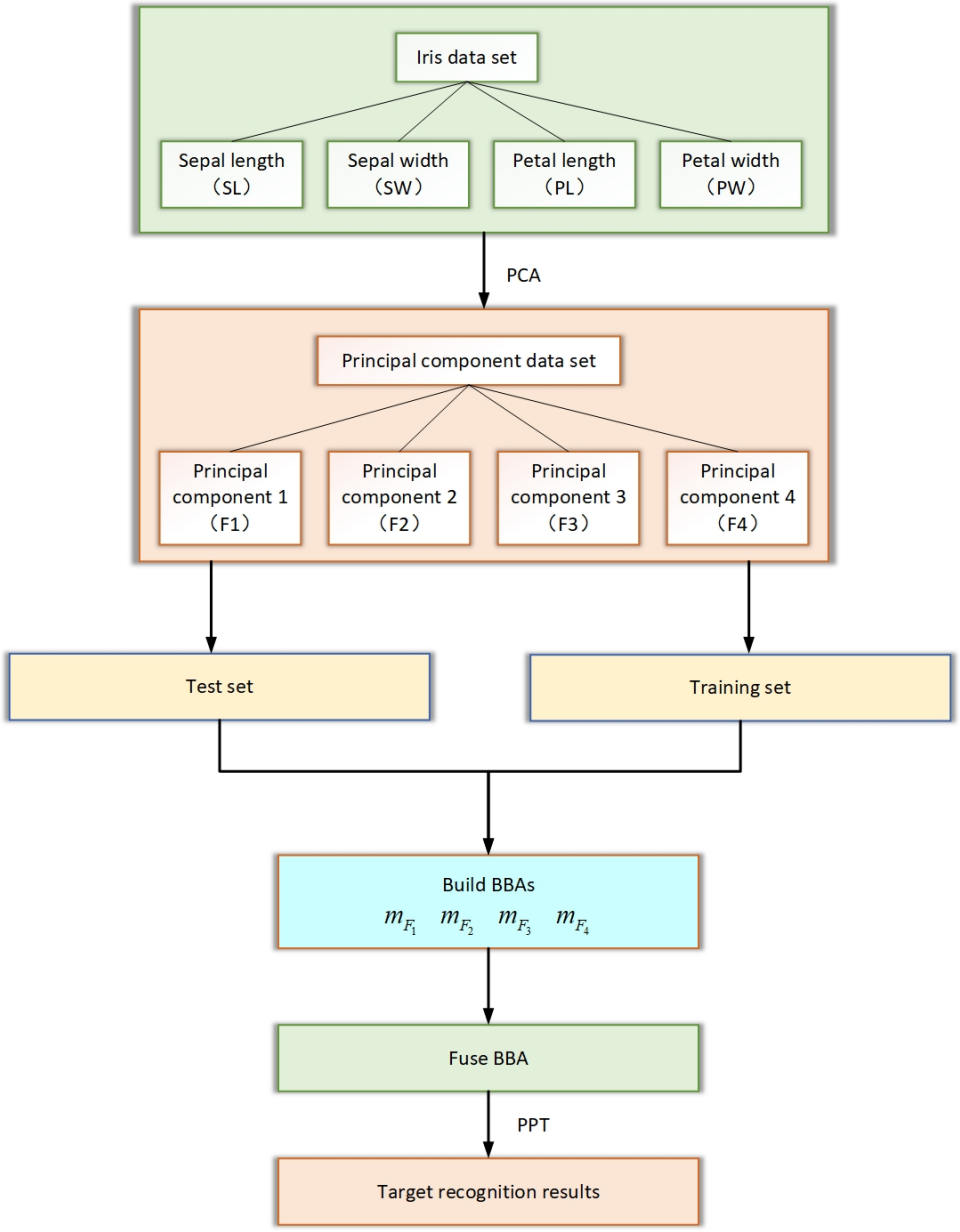

Figure 1. Application of the method in the case study.

Table 1. Eigenvectors of the raw data.

Raw data	Eigenvector V_1	Eigenvector V_2	Eigenvector V_3	Eigenvector V_4
X_1	0.5224	0.3723	-0.7210	-0.2620
X_2	-0.2634	0.9256	0.2420	0.1241
X_3	0.5813	0.0211	0.1409	0.8012
X_4	0.5656	0.0654	0.6338	-0.5235

Step 3. Build BBAs of principal components

The data sets of principal components are divided into training sets and test sets to drive the generation of the BBA of each principal component. Different proportions are assigned to the two sets, such as 20% for the training set and 80% for the test set. We assign the proportion of the training set from 20% to 90% for building BBAs, respectively. Furthermore, considering the inherent randomness in BBA generation due to random sampling, we implement a strategy to minimize this randomness. Each simulation in our study involves random reshuffling of the sample order in the original data set 100 times. This approach is designed to eliminate potential biases introduced by the order of data. To ensure a comprehensive evaluation of the model's performance, each experiment is repeated 100 times, allowing us to calculate the average identification accuracy.

This paper generate BBA through the method in Ref. [36]. The following is the main process: For a raw data set, each sample can be described by the l-dimensional attribute vector $a = [a_b]\,(b = 1, 2, \ldots, l)$. The data of each attribute in the original data set is converted into a normal distribution model. Assume there is a certain sample S_i in the training set or test set, take the intersection point between the line $x = x_b$ and the normal distribution model, and the longitudinal coordinate of the intersection point is the BBA of the corresponding sample's attribute a_b (see Ref. [36] for more details).

Step 4. Fuse the BBA and draw a conclusion

Fuse the BBA according to Eq. (2.5), convert the BBA into a probability value through PPT according to Eq. (2.6), and conduct the iris identification experiment. Table 2 shows the average identification accuracy for different training set proportions.

Table 2. Identification accuracy of training sets with different proportions.

Training set	20%	30%	40%	50%	60%	70%	80%	90%
Identification accuracy	0.9158	0.9171	0.9152	0.9160	0.9172	0.9104	0.9267	0.9156

In order to illustrate the role of PCA, we calculate the iris identification accuracy of evidence theory method (i.e., fuse BBAs applying Dempster's combination rule directly without preprocessing), rank correlation coefficient [25] and mutual information method [26], and compare the proposed method based on PCA with the above methods. Figure 2 represents the average identification accuracy of these four methods for different training set proportions.

From Figure 2, the identification accuracy of evidence theory method is the highest, followed by mutual information method, whereas the proposed method is slightly lower than the previous two, and the accuracy of rank correlation method is the lowest. However, this finding does not indicate that evidence theory method and mutual information method have a better performance in information fusion. The possible reason is that the two methods calculate repeatedly the influence of interrelated parts. This problem is further analyzed and simulated in the next section.

4.2. Additional experiments

To further illustrate the ability of the method to handle redundant information, redundant attribute data is designed and added to the raw data for simulation.

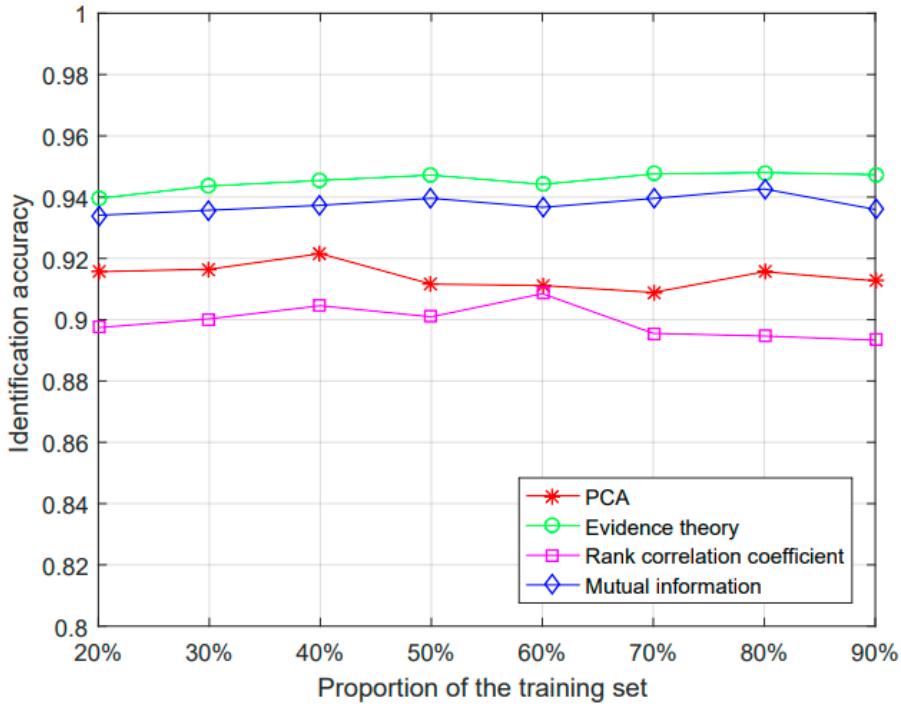

Figure 2. Identification accuracy of the four methods with different proportions of training sets.

The attributes to be repeatedly added are selected with reference to the identification reliability of information sources. The identification reliability represents the degree of consistency between the results obtained by direct identification based on the information of a single information source and the actual results. It is defined as follows:

$$F_{SX_i} = \begin{cases} 1, if\theta_j\{SX_i\} = \theta_j; \\ 0, if\theta_j\{SX_i\} \neq \theta_j \end{cases} \tag{4.2}$$

where F_{SX_i} is the decision factor, representing the consistency between identified category $\theta_j\{SX_i\}$ and real category θ_j . Then we obtain the reliability of information source SX:

$$R_{SX} = \frac{1}{n}\sum_{i=1}^{n} F_{SX_i} \times 100\% \tag{4.3}$$

According to Eq. (4.2) and Eq. (4.3), the identification reliability of each attribute of different training set proportions can be obtained as Table 3.

In this simulation, the proportion of training set is set to 60%. From Table 3, the identification reliability of four attributes is as follows: $R_{SL} = 72.81\%, R_{SW} = 56.26\%, R_{PL} = 95.59\%$, and $R_{PW} = 95.63\%$. It can be seen that the identification reliability of sepal width R_{SW} is the lowest, and that of petal width R_{PW} is the highest.

Case 1. When the attribute SW is redundantly integrated into the original data set, we identify iris species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 100 times following each redundant information integration. Figure 3 shows the changing

Table 3. Identification reliability of each attribute for different proportions of training sets.

Training set proportions	SL	SW	PL	PW
20%	0.7444	0.5933	0.9689	0.9578
30%	0.7407	0.5941	0.9600	0.9659
40%	0.7400	0.5672	0.9589	0.9617
50%	0.7253	0.5747	0.9533	0.9556
60%	0.7281	0.5626	0.9559	0.9563
70%	0.7295	0.5568	0.9533	0.9594
80%	0.7256	0.5531	0.9536	0.9594
90%	0.7309	0.5573	0.9516	0.9615

trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the attribute SW is repeatedly added.

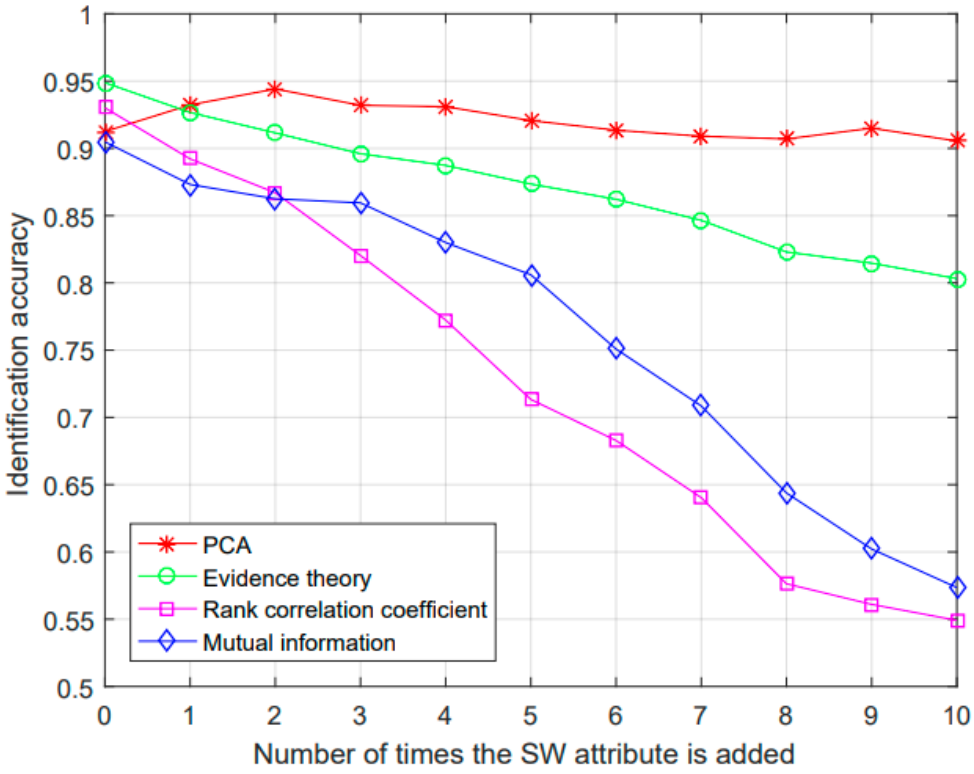

Figure 3. Identification accuracy with repeated addition of SW attributes.

It can be concluded that as the repeated addition of the attribute with the lowest identification reliability, the identification accuracy of evidence theory, rank correlation method, and mutual information method decrease significantly. Among them, evidence theory method dropped to near 0.8, and rank

correlation coefficient method and mutual information method dropped to near 0.55. However, the accuracy of the proposed method remains largely stable in the process of repeated attribute addition.

Case 2. When the attribute *PW* is redundantly integrated into the original data set, we identify iris species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 100 times following each redundant information integration. Figure 4 shows the changing trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the attribute *PW* is repeatedly added.

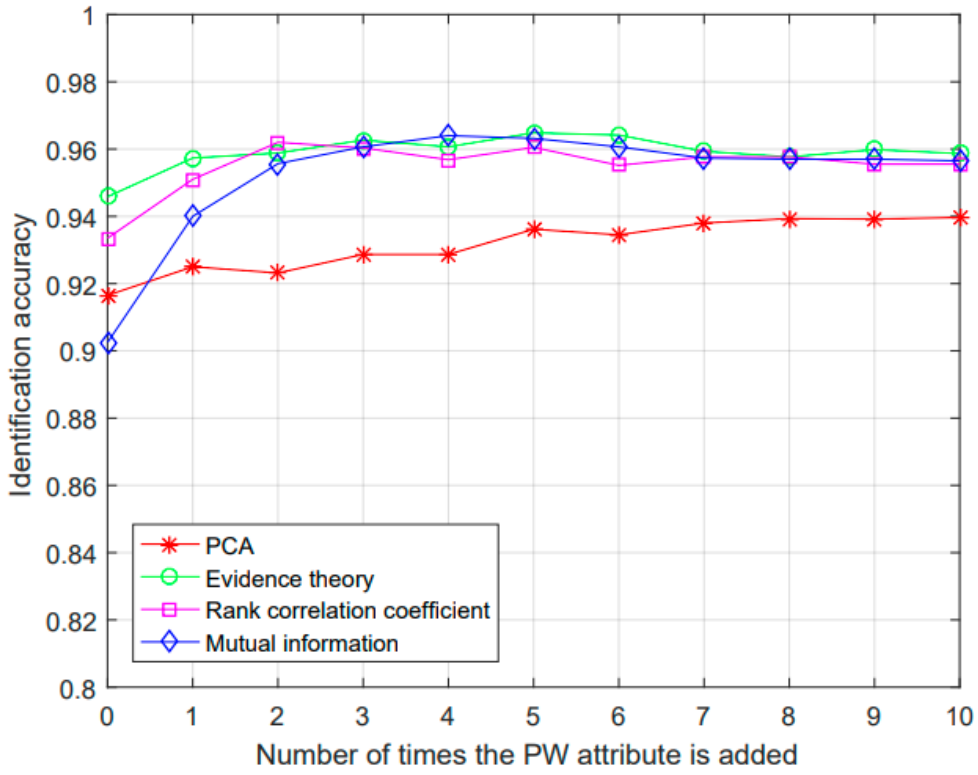

Figure 4. Identification accuracy with repeated addition of *PW* attributes.

Figure 4 shows that as the repeated addition of the attribute with the highest reliability, the identification accuracy of evidence theory method and rank correlation method reached the maximum accuracy value of 0.96 in the second time, and then fluctuated around here. The accuracy of the mutual information method increases rapidly, reaching 0.94 at the first addition and fluctuating also around 0.96 after the second addition. Comparatively, the identification accuracy of the proposed method does not change much in the process of repeated attribute addition.

4.3. Case studies of other data sets

To enhance the persuasiveness of the proposed method's capability in handling redundant information, we additionally conducts identification experiments on the wine and seeds data sets.

The Wine dataset [39] from the University of California, Irvine Machine Learning Repository (UCI Machine Learning Repository) is derived from a chemical analysis of wines grown in Italy, representing three different cultivars. It includes 13 attributes like alcohol content, malic acid, and flavanoids, among others, across 178 instances, categorized into three classes based on the type of cultivar. The

Seeds data set [40], also from the UCI Repository, is focused on the geometric properties of wheat kernels, comprising 7 attributes such as area, perimeter, and compactness. This data set encompasses 210 instances, divided into three classes representing different wheat varieties: Kama, Rosa, and Canadian.

4.3.1. Identification accuracy for the wine and seeds data sets

Following the process of the iris identification experiment in Section 4.1, this section calculates the identification accuracy for the wine and seeds data sets using the four methods (see Section 4.1, Step 4), without adding redundant information. The results are shown in Figure 5 and Figure 6, respectively.

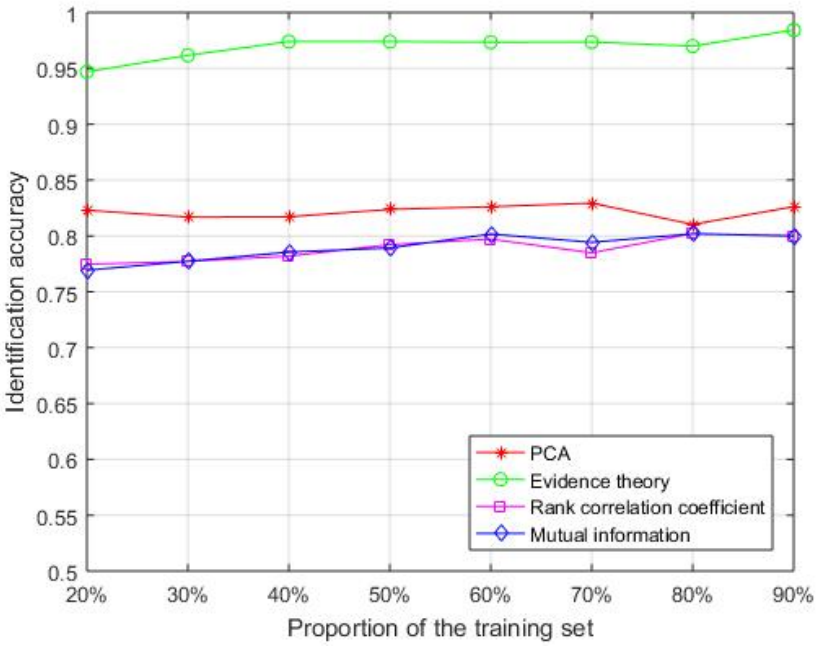

Figure 5. Identification accuracy of the wine data set based on four methods.

Similar to the iris identification experiment, the identification accuracy of the proposed method is not the highest, as this study focuses on the robustness of the method. Experiments involving the addition of redundant information to the wine and seeds data sets are conducted in Section 4.3.2.

4.3.2. Experiments involving the addition of redundant information

(1) The wine data set

The identification reliability of each attribute in the wine data set is calculated according to Eq. (4.2) and Eq. (4.3), as Table 4. Table 4 indicates that in the wine data set, attribute 7 has the highest reliability in identification, with a score of 0.8150, while attribute 3 shows the lowest reliability, scoring 0.4338. Subsequently, experiments are conducted where attribute 7 and attribute 3 are separately added as redundant information to the original dataset for identification purposes.

Case 1. When the 7th attribute (the highest identification reliability) is redundantly integrated into the original dataset, we identify wine species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 20 times following each redundant information integration.

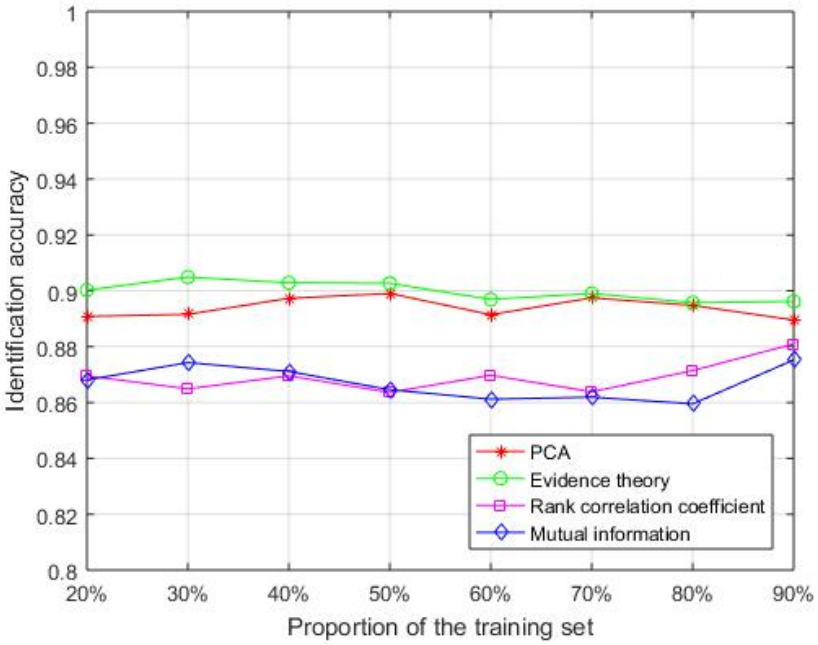

Figure 6. Identification accuracy of the seeds data set based on four methods.

Table 4. Identification reliability of each attribute for the wine data set.

Identification reliability	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
	0.6746	0.6067	0.4338	0.5525	0.4974	0.6531	0.8150
	Attribute 8	Attribute 9	Attribute 10	Attribute 11	Attribute 12	Attribute 13	
	0.5586	0.5496	0.7165	0.6332	0.6860	0.7261	

Figure 7 shows the changing trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the 7th attribute is repeatedly added.

Case 2. When the 3rd attribute (the lowest identification reliability) is redundantly integrated into the original data set, we identify wine species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 20 times following each redundant information integration. Figure 8 shows the changing trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the 3rd attribute is repeatedly added.

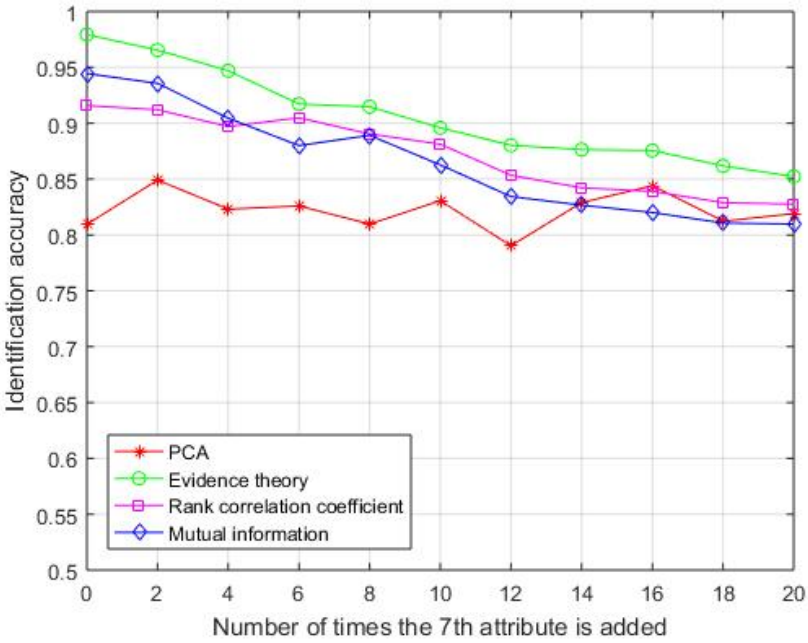

Figure 7. Identification accuracy with repeated addition of the 7th attribute for the wine data set.

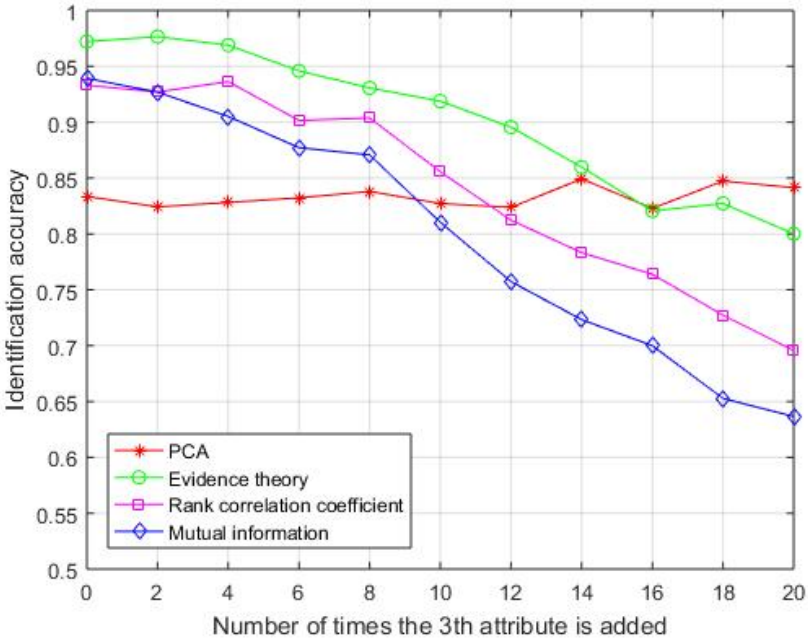

Figure 8. Identification accuracy with repeated addition of the 3rd attribute for the wine data set.

(2) The seeds data set

The identification reliability of each attribute in the seeds data set is calculated according to Eq.

(4.2) and Eq. (4.3), as reported in Table 5. Table 5 indicates that in the seeds dataset, attribute 2 has the highest reliability in identification, with a score of 0.8676, while attribute 6 shows the lowest reliability, scoring 0.5546. Subsequently, experiments are conducted where attribute 2 and attribute 6 are separately added as redundant information to the original data set for identification purposes.

Table 5. Identification reliability of each attribute for the seeds data set.

Identification	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	Attribute 7
reliability	0.8624	0.8676	0.5613	0.7926	0.8245	0.5546	0.7180

Case 1. When the 2nd attribute (the highest identification reliability) is redundantly integrated into the original data set, we identify seed species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 20 times following each redundant information integration. Figure 9 shows the changing trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the 2nd attribute is repeatedly added.

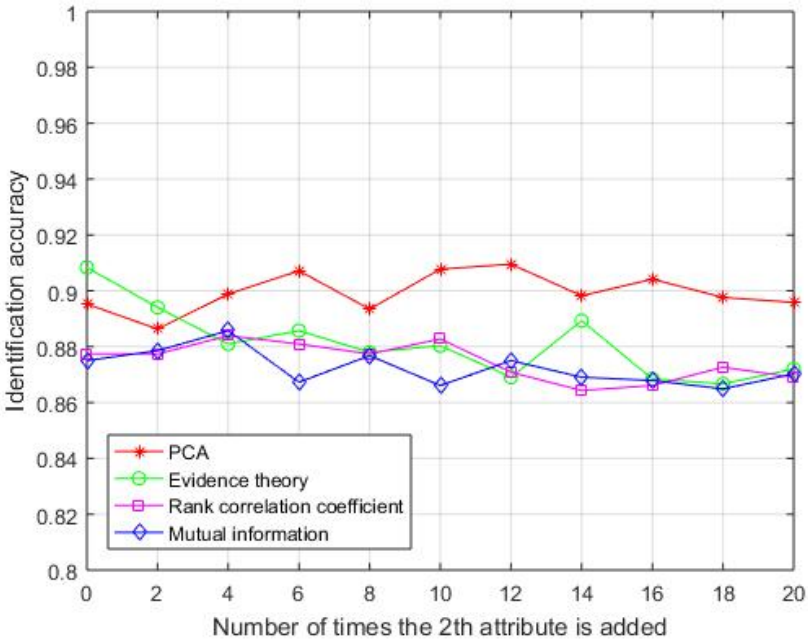

Figure 9. Identification accuracy with repeated addition of the 2nd attribute for the seeds data set.

Case 2. When the 6th attribute (the lowest identification reliability) is redundantly integrated into the original dataset, we identify wine species by the four methods. To ascertain the average identification accuracy, the experiment is repeated 20 times following each redundant information integration. Figure 10 shows the changing trend of the identification accuracy of the four methods when the training set accounts for 60% of the raw data and the 6th attribute is repeatedly added.

From the experiments on adding redundant information to the three data sets, it is observed that with the addition of redundant attributes, the identification accuracies of classical evidence theory, rank correlation coefficient [25], and mutual information method [26] generally change greatly, and

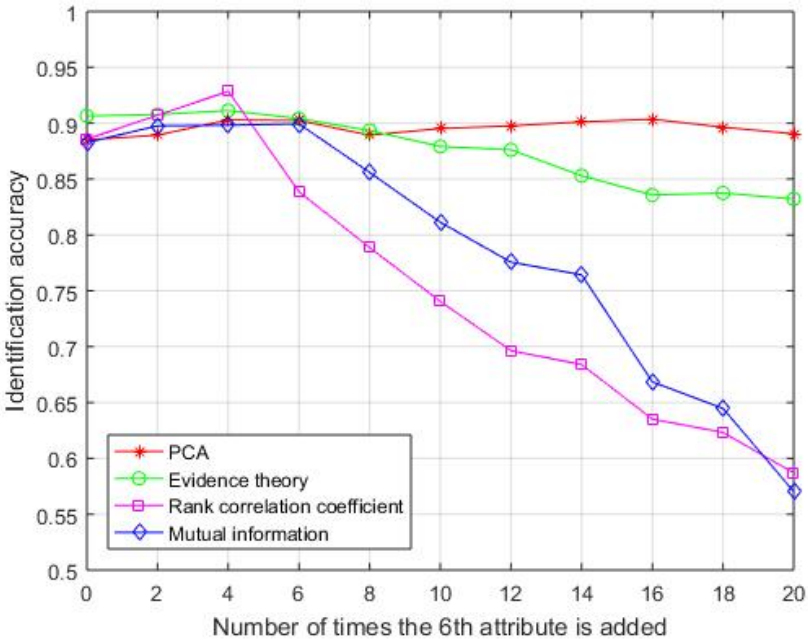

Figure 10. Identification accuracy with repeated addition of the 6th attribute for the seeds data set.

tend to converge toward the inherent reliability of the respective redundant attributes, which is contrary to intuition, while the results of the proposed method are more stable and in accordance with intuition. The other three methods either do not consider the dependence among the information sources (classic evidence theory method) or do not deal with the redundant information among the sources effectively (rank correlation coefficient method and mutual information method), which result in the repeated calculation of the influence of the interrelated parts on the identification result. In contrast, the proposed method has stronger robustness and can effectively deal with redundant information in the process of information fusion.

To further improve the accuracy of the proposed method, the original data can be screened first, and the information sources with particularly low identification reliability can be eliminated, and then the relevant operations of PCA can be carried out.

5. Conclusion

To solve the problem of dependence in evidence theory, this paper proposes a method of dependent evidence fusion based on PCA. This method starts with the original data of the original evidence source, uses PCA to change the previously related variables into new independent principal component variables, and establishes BBA based on the new variables. Finally, the decision is made by integrating evidence based on Dempster's combination rules. The effectiveness and superiority of the method in this paper are illustrated by the experiments of the iris data sets, the wine data sets, and the seeds data sets.

This method has the following advantages: (1) It is a more reliable and efficient way to analyze the

dependence among information sources by using original data, compared with analyzing the dependence by the evidence structure. (2) This paper innovatively proposes a fusion strategy of dependent evidence based on PCA. PCA highlights the main features of data sets and removes redundant information, which can effectively deal with the dependence among information sources and make the fusion results more reasonable and stable.

Our future research endeavors will consider expanding beyond the linear dependencies addressed by PCA, exploring the potential of various nonlinear methods. Alongside this exploration, we intend to assess the compatibility and performance of our method when integrated with established machine learning algorithms, including support vector machine (SVM), k-nearest neighbors (k-NN), extreme gradient boosting (XGB), and random forest. Such integration is anticipated to offer a more comprehensive benchmarking framework, deepening our understanding of our method's adaptability and efficacy across diverse scenarios. In addition, we recognize the importance of a multifaceted evaluation approach. To this end, we aim to refine our evaluation framework by incorporating advanced performance metrics such as area under the receiver operating characteristic curve (AUCROC) and area under the precision-recall curve (AUCPR). These metrics hold significant value, especially in evaluating model performance in situations involving imbalanced data. While it remains to be seen how these modifications will influence our research trajectory, our goal is to ensure a thorough and robust assessment of our method's capabilities, thereby enriching its applicability and contribution to the field.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The work is partially supported by Shanghai Rising-Star Program (Grant No.21QA1403400), Shanghai Natural Science Foundation (Grant No.19ZR1420700), Shanghai Key Laboratory of Power Station Automation Technology (Grant No.13DZ2273800).

Conflict of interest

All authors declare that they have no conflict of interest.

References

1. A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, *Ann. Math. Stat.*, **38** (1967), 325–339. https://doi.org/10.1214/aoms/1177698950
2. G. Shafer, *A Mathematical Theory of Evidence*, Princeton: Princeton University Press, 1976. https://doi.org/10.1515/9780691214696
3. Y. Deng, Uncertainty measure in evidence theory, *Sci. China Inf. Sci.*, **63** (2020), 210201. https://doi.org/10.1007/s11432-020-3006-9

4. F. Xiao, Generalized quantum evidence theory, *Appl. Intell.*, **53** (2023), 14329–14344. https://doi.org/10.1007/s10489-022-04181-0

5. Y. Cui, X. Deng, Plausibility Entropy: A New Total Uncertainty Measure in Evidence Theory Based on Plausibility Function, *IEEE Trans. Syst. Man Cybern. Syst.*, **53** (2023), 3833–3844. https://doi.org/10.1109/TSMC.2022.3233156

6. Y. Deng, Random permutation set, *Int. J. Comput. Commun. Control*, **17** (2022). https://doi.org/10.15837/ijccc.2022.1.4542

7. X. Deng, S. Xue, W. Jiang, A novel quantum model of mass function for uncertain information fusion, *Inf. Fusion*, **89** (2023), 619–631. https://doi.org/10.1016/j.inffus.2022.08.030

8. X. Chen, Y. Deng, A new belief entropy and its application in software risk analysis, *Int. J. Comput. Commun. Control*, **18** (2023). https://doi.org/10.15837/ijccc.2023.2.5299

9. F. Xiao, W. Pedrycz, Negation of the quantum mass function for multisource quantum information fusion with its application to pattern classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 2054–2070. https://doi.org/10.1109/TPAMI.2022.3167045

10. D. Li, Y. Deng, Measure information quality of basic probability assignment: An information volume method, *Appl. Intell.*, **52** (2022), 11638–11651. https://doi.org/10.1007/s10489-021-03066-y

11. C. Zhu, F. Xiao, A belief Hellinger distance for D–S evidence theory and its application in pattern recognition, *Eng. Appl. Artif. Intell.*, **106** (2021), 104452. https://doi.org/10.1016/j.engappai.2021.104452

12. Y. Tao, H. Hu, F. Xu, Z. Zhang, Ergonomic Risk Assessment of Construction Workers and Projects Based on Fuzzy Bayesian Network and DS Evidence Theory, *J. Constr. Eng. Manag.*, **149** (2023), 04023034. https://doi.org/10.1061/JCEMD4.COENG-12821

13. P. Lu, Y. Zhou, Y. Wu, D. Li, Risk assessment of complex footbridge based on Dempster–Shafer evidence theory using Fuzzy matter–element method, *Appl. Soft Comput.*, **131** (2022), 109782. https://doi.org/10.1016/j.asoc.2022.109782

14. S. I. Sezer, G. Elidolu, E. Akyuz, O. Arslan, An integrated risk assessment modelling for cargo manifold process on tanker ships under FMECA extended Dempster–Shafer theory and rule-based Bayesian network approach, *Process Saf. Environ. Prot.*, **174** (2023), 340–352. https://doi.org/10.1016/j.psep.2023.04.024

15. S. I. Sezer, G. Camliyurt, M. Aydin, E. Akyuz, P. Gardoni, A bow-tie extended DS evidence-HEART modelling for risk analysis of cargo tank cracks on oil/chemical tanker, *Reliab. Eng. Syst. Saf.*, **237** (2023), 109346. https://doi.org/10.1016/j.ress.2023.109346

16. L. Fei, Y. Wang, An optimization model for rescuer assignments under an uncertain environment by using Dempster–Shafer theory, *Knowl.-Based Syst.*, **255** (2022), 109680. https://doi.org/10.1016/j.knosys.2022.109680

17. L. Fei, Y. Wang, Demand prediction of emergency materials using case-based reasoning extended by the Dempster-Shafer theory, *Socio-Econ. Plan. Sci.*, **84** (2022), 101386. https://doi.org/10.1016/j.seps.2022.101386

18. R. Zhang, Z. Xu, X. Gou, An integrated method for multi-criteria decision-making based on the best-worst method and Dempster-Shafer evidence theory under double hierarchy hesitant fuzzy

linguistic environment, *Appl. Intell.*, **51** (2021), 713–735. https://doi.org/10.1007/s10489-020-01777-2

19. X. Su, S. Mahadevan, P. Xu, Y. Deng, Handling of dependence in Dempster–Shafer theory, *Int. J. Intell. Syst.*, **30** (2015), 441–467. https://doi.org/10.1002/int.21695
20. X. Su, S. Mahadevan, W. Han, Y. Deng, Combining dependent bodies of evidence, *Appl. Intell.*, **44** (2016), 634–644. https://doi.org/10.1007/s10489-015-0723-5
21. M. Chebbah, A. Martin, B. B. Yaghlane, Combining partially independent belief functions, *Decis. Support Syst.*, **73** (2015), 37–46. https://doi.org/10.1016/j.dss.2015.02.017
22. S. Destercke, D. Dubois, Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory, *Inf. Sci.*, **181** (2011), 3925–3945. https://doi.org/10.1016/j.ins.2011.05.007
23. M. E. G. V. Cattaneo, Belief functions combination without the assumption of independence of the information sources, *Int. J. Approx. Reason.*, **52** (2011), 299–315. https://doi.org/10.1016/j.ijar.2010.10.006
24. C. Fu, S. Yang, The combination of dependence-based interval-valued evidential reasoning approach with balanced scorecard for performance assessment, *Expert Syst. Appl.*, **39** (2012), 3717–3730. https://doi.org/10.1016/j.eswa.2011.09.069
25. F. Shi, X. Su, H. Qian, N. Yang, W. Han, Research on the fusion of dependent evidence based on rank correlation coefficient, *Sensors*, **17** (2017), 2362. https://doi.org/10.3390/s17102362
26. X. Su, L. Li, F. Shi, H. Qian, Research on the fusion of dependent evidence based on mutual information, *IEEE Access*, **6** (2018), 71839–71845. https://doi.org/10.1109/ACCESS.2018.2882545
27. G. Kong, D. Xu, J. Yang, T. Wang, B. Jiang, Evidential reasoning rule-based decision support system for predicting ICU admission and in-hospital death of trauma, *IEEE Trans. Syst. Man Cybern. Syst.*, **51** (2020), 7131–7142. https://doi.org/10.1109/TSMC.2020.2967885
28. S. Yao, J.-B. Yang, D.-L. Xu, P. Dark, Probabilistic modeling approach for interpretable inference and prediction with data for sepsis diagnosis, *Expert Syst. Appl.*, **183** (2021), 115333. https://doi.org/10.1016/j.eswa.2021.115333
29. T. Liu, D. A. Diaz-Pachon, J. S. Rao, J.-E. Dazard, High Dimensional Mode Hunting Using Pettiest Components Analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 4637–4649. https://doi.org/10.1109/TPAMI.2022.3195462
30. J. Zhang, D. Zhou, M. Chen, Self-learning sparse PCA for multimode process monitoring, *IEEE Trans. Ind. Inform.*, **19** (2022), 29–39. https://doi.org/10.1109/TII.2022.3178736
31. A. D. McRae, J. Romberg, M. A. Davenport, Optimal convex lifted sparse phase retrieval and PCA with an atomic matrix norm regularizer, *IEEE Trans. Inf. Theory*, **69** (2022), 1866–1882. https://doi.org/10.48550/arXiv.2111.04652
32. S. Martinović, A. Alil, S. Milićević, D. Živojinović, T. V. Husović, Morphological assessment of cavitation caused damage of cordierite and zircon based materials using principal component analysis, *Eng. Fail. Anal.*, **148** (2023), 107224. https://doi.org/10.1016/j.engfailanal.2023.107224

33. H. B. Bisheh, G. G. Amiri, Structural damage detection based on variational mode decomposition and kernel PCA-based support vector machine, *Eng. Struct.*, **278** (2023), 115565. https://doi.org/10.1016/j.engstruct.2022.115565

34. D. Gedon, A. H. Ribeiro, N. Wahlström, T. B. Schön, Invertible Kernel PCA with Random Fourier Features, *IEEE Signal Process. Lett.*, **30** (2023), 563–567. https://doi.org/10.1109/LSP.2023.3275499

35. L. Shang, S. Wang, Application of improved principal component analysis in comprehensive assessment on thermal power generation units, *Power Syst. Technol.*, **38** (2014), 1928–1933. https://doi.org/10.13335/j.1000-3673.pst.2014.07.032

36. P. Xu, Y. Deng, X. Su, S. Mahadevan, A new method to determine basic probability assignment from training data, *Knowl.-Based Syst.*, **46** (2013), 69–80. https://doi.org/10.1016/j.knosys.2013.03.005

37. R. A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.*, **7** (1936), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

38. *Iris*, R. A. Fisher, 1988. Available from: `https://doi.org/10.24432/C56C76`

39. *Wine*, S. Aeberhard, M. Forina, 1991. Available from: `https://doi.org/10.24432/C5PC7J`

40. *Seeds*, M. Charytanowicz, J. Niewczas, P. Kulczycki, P. Kowalski, S. Lukasik, 2012. Available from: `https://doi.org/10.24432/C5H30K`

AIMS Press