Generating basic probability assignment from the view of distance measures and its application in evidential decision tree

1st Yifan Sun
College of Information Engineering,
Northwest A&F University,
Yangling, China
1fan.sun@nwafu.edu.cn

2nd Mengzhuo Zhang
College of Information Engineering,
Northwest A&F University,
Yangling, China
zhangmz@nwafu.edu.cn

3rd Xiaozhuan Gao
College of Information Engineering,
Northwest A&F University,
Yangling, China
gaoxiaozhuan@nwafu.edu.cn

2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI) | 979-8-3315-2931-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACAI63924.2024.10899481

***Abstract*—In Dempster-Shafer evidence theory, basic probability assignment (BPA) plays a important role in representing uncertain and unknown information. How to generate high-quality BPA is essential, which can promote further application of evidence theory. BPA can be understood as the possibility assigned to each proposition. Distance measures, as an effective tool for quantitatively analyzing inconsistencies between samples and known information, are central to this process. Hence, this paper proposes a novel method to generate BPA by comprehensively considering Gaussian distribution and distance measures. New method is applied to evidential decision tree and its effectiveness can be verified by real world data set.**

***Index Terms*—Basic probability assignment, Distance measure, Evidential decision tree**

I. Introduction

As an extension of traditional probability theory, Dempster-Shafer Theory (DST) [1], [2] provides a broader framework for handling uncertain information. In DST, basic probability assignment (BPA) maps uncertain information to the power set of the identification framework to model imprecise and unknown information. Up to now, DST has attracted more and more attention which can be applied in some fields, such as classification task, target recognition, risk assessment etc.

Generating the hight-quality BPA is essential to apply DST into practical engineering, which can have a direct impact on the experiment results. Up to now, there are some studies about how to generate BPA. Ghafir et al propose a novel method based on the Gaussian and exponential probability density functions, the categorical probability mass function, and the local reachability density [3]. Fu et al use Adaboost to generate BPA which does not consider probability distribution of data [4].Fei et al generate BPA by using K-means method and it is extended by K–nearest neighbor (K-NN) algorithm [5]. Besides, there are some other studies about how to generate BPA, however, which are based on data-driven by building probability distribution models based on the training set. It should be pointed out that those existing methods can not consider distance between sample and known information. Distance measure can effectively quantify the differences between known information and samples. Garg and Rani apply distance into pattern recognition and clustering [6]. Hassanat et al review the specific applications of Hassanat distance metric in supervised and unsupervised learning [7]. It can be seen that distance measure has the better performance when those data is addressed which are contains noise and outliers.

Functionally, the BPA is used to represent uncertainty by assigning mass to subsets, and the effectiveness of this assignment determines the quality of the BPA. Therefore, by adjusting the mass function of each subset based on the distance measures to the test sample, the BPA can more accurately express the degree of membership of the test sample to each category. This, in turn, enhances the BPA's ability to represent uncertain information and improves the quality of generated BPA.

This paper presents a novel approach for generating basic probability assignment. First, the mean and standard deviation of each class with respect to each attribute in the training set are calculated, and Gaussian models are constructed for each attribute. These models are then employed to generate the BPA for the test set by matching the test samples to the corresponding Gaussian distributions. Next, the mean and median values of each class in the training set are computed, and the differences between the test sample and these values are used to define the distances to individual or multiple classes. These distances are subsequently aligned with the power set spatial distribution within the framework of discernment. Finally, the BPA generated from the Gaussian models is combined with the distance values through the computation of their inner product.

To further validate the advantages of the proposed method, it is applied within the context of the evidential decision tree. The performance of the evidential decision tree serves as an indicator of the effectiveness of attribute selection, as well as the quality of the BPA.

The organization of the rest of this paper is shown as fallow. Section 2 is the preliminaries. Section 3 presents the novel method of generating BPA. In section 4, proposed method is applied in the evidential decision tree. Section 5 shows the experimental results by using real world data set. section 6 concludes this paper.

II. Preliminaries

A. D-S evidence theory

(1)Framework of discernment(FoD)

The framework of discernment is a set including exclusive elements. In order to promote the scientific process of decision, the empty set is not involved in framework of discernment because it does not contain any information beneficial for decision making. Then, for $k-element$ framework of discernment $[a_1, a_2, ..., a_{k-1}, a_k]$, its power set spatial distribution contains $n = 2^k - 1$ subsets, namely

$$\begin{aligned}\Omega = \{&\{a_1\}, \{a_2\}, ..., \{a_k\},\\ &\{a_1, a_2\}, ..., \{a_1, a_{k-1}\}, ...,\\ &\{a_1, a_2, ..., a_{k-1}, a_k\}\}\\ &= \{x_1, x_2, ..., x_n\},\end{aligned} \tag{1}$$

here x_n means the nth subset of power set spatial distribution.

(2)Basic probability assignment(BPA)

Basic probability assignments are presented as mass functions m whose function values vary in range [0, 1].

$$m(A) \to [0, 1], A \in \Omega \tag{2}$$

$$\sum_{A \in \Omega} m(A) = 1 \tag{3}$$

$$m(\Phi) = 0 \tag{4}$$

B. Gaussian function

Gaussian function is a probability function which describing the distribution law of random variables. It is defined by two parameters, mean $\bar{X}$ and variance σ. The expression of Gaussian function μ is

$$\mu(x) = \exp\,[-\frac{(x - \bar{X})^2}{2\sigma^2}] \tag{5}$$

III. Generation of BPA

For a m attributes data set with N samples of k classes, n samples are selected from each class as training samples, so as to establish the Gaussian model of each class on each attribute. The remaining samples are used as test samples from which BPA is generated. The procedures of BPA generation are presented as follows.

A. Build Gaussian models on each attribute

The specific process of obtaining Gaussian function $\mu(x)$ is as follows.

(1) For a selected class k and attribute s, respectively calculate the sample mean $\bar{X_{sk}}$ and standard deviation σ_{sk} of all training samples belonging to class k on the attribute s:

$$\bar{X_{sk}} = \frac{1}{m}\sum_{i=1}^{m} x_{sk}^i, \tag{6}$$

$$\sigma_{sk} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_{sk}^i - \bar{X_{sk}})^2}, \tag{7}$$

where, x_{sk}^i represents the value of the ith sample of class k on attribute s.

(2) According to the obtained mean $\bar{X_{sk}}$ and standard deviation σ_{sk}, construct the Gaussian models of class k on the attribute s:

$$\mu_k^s(x) = \exp\,[-\frac{(x - \bar{X_{sk}})^2}{2\sigma_{sk}^2}]. \tag{8}$$

B. Match the test samples to the Gaussian models to get BPA

For example, Gaussian models of a data set with three classes A, B, C on attribute s are built. And there is a piece of test data having a value of v_1 on attribute s. Firstly, calculate the function values of the test value, namely $\mu_A^s(v_1), \mu_B^s(v_1), \mu_C^s(v_1)$. Then sort them in descending order to get the sequence of classes and assign the function value of a class to the subset including itself and classes before it. Finally, the mass functions of these subsets are represented by the function values.

C. Combine the BPA with distance measures

As a means of revealing the intrinsic patterns and structures within data, distance measures provide a quantitative foundation for assessing similarity and dissimilarity between objects. These measures can capture the geometric relationships among different objects, thereby facilitating the identification of potential clustering structures and classification boundaries.

The process of combination is achieved through the calculation of the inner product. For test data $(v_1, v_2, ..., v_m)$ which values on attributes $(A_1, A_2, ..., A_m)$, the final BPA are gained by

$$BPA_{A_m} = Gaussian_{A_m} * Distances, \tag{9}$$

where $Gaussian_{A_m}$ are the BPA directly generated from the Gaussian models and $Distances$ refers to a certain distance sequence consists of distances of subsets in FoD, namely distance measures of $\{x_1, x_2, ..., x_k\}$.

Consider a dataset with three classes A, B, C. The original distances are computed from the test data to the training data on each attribute, as follows:

$$\begin{aligned}dis &= \{d_A^s, d_B^s, d_C^s, d_{A,B}^s, d_{A,C}^s, d_{B,C}^s, d_{A,B,C}^s\}\\ &= \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\},\end{aligned} \tag{10}$$

here d_A^s denotes the distance from the test value to a specific measure of distance for the data in class A of training set, and $d_{A,B}^s$ is the distances of the data in class A and class B.

In this study, the mean and median are chosen as distance measures and are combined with the BPA generated from the Gaussian models. The mean is defined as:

$$mean(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{11}$$

and the median is given by:

$$\begin{aligned}&median(x_1, x_2, \ldots, x_n)\\ &= \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd}\\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{if } n \text{ is even,}\end{cases}\end{aligned} \tag{12}$$

where $x_1, x_2, \ldots, x_n$ refers to a sequence of values.

As an example, when using the mean measure, the elements in dis are defined as:

$$\begin{aligned} d_1 &= d_A^s = v^s - mean(train_A^s), \\ d_4 &= d_{A,B}^s = v^s - mean(train_A^s \cup train_A^s), \end{aligned} \tag{13}$$

here v^s denotes the value of a given test data on attribute s, and $train_A^s$ denotes the part of training data that belongs to class A on attribute s. The remaining elements in dis can be calculated in a similar manner.

To fully leverage the distance information, we apply the negative exponent to the original distance values, thus defining $Distances$ as:

$$\begin{aligned} Distances = e^{(-dis)} = \{&e^{-(d_1)}, e^{-(d_2)}, \\ &e^{-(d_3)}, e^{-(d_4)}, e^{-(d_5)}, e^{-(d_6)}, e^{-(d_7)}\}. \end{aligned} \tag{14}$$

Finally, the BPA with distance measures are gained by combining $Distances$ with the BPA generated from Gaussian models according to Equation (9).

IV. Application in Evidential Decision Tree

Gao et al. [8] introduced a novel method for constructing an evidential decision tree using hierarchical interval estimation, which has been shown to be effective. In this paper, we build upon their work by modifying the attribute selection rule and incorporating the proposed BPA generation method to further enhance the model's performance. The procedures for constructing the modified evidential decision tree are illustrated in Fig. 1.

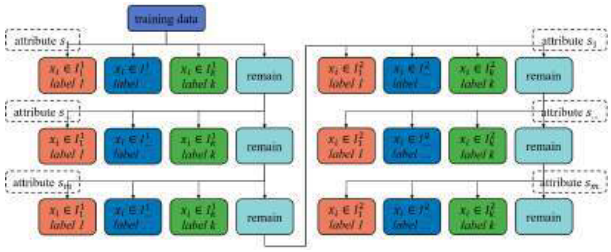


Fig. 1. Process of constructing evidential decision tree

(1)Divide the data set.

For classification problems, the whole data set is usually divided into training set and test set. The data in test set accounts for 20% to 30%.

(2)Select splitting attributes.

In Dempster–Shafer theory, many kinds of entropy methods are proposed to quantify the amount of information contained in uncertain data. In this paper, we choose Nguyen entropy(Equation (15)) and Deng entropy(Equation (16)) to measure the uncertainty.

$$E_{Nguyen} = -\sum_{A \in \Omega} m(A) log_2 m(A) \tag{15}$$

$$E_{Deng} = -\sum_{A \in \Omega} m(A) log_2 \frac{m(A)}{2^{|A|} - 1} \tag{16}$$

Specifically, the entropy of the attribute s is calculated by

$$E(s) = \frac{1}{n} \sum_{i=1}^{n} E(m_i^s), \tag{17}$$

where E refers to Nguyen entropy or Deng entropy, and m_i^s denotes the BPA combined with distance measures. Finally, the attribute $s^* = argmin E(s)$ is selected as the best splitting attribute.

(3)Determine interval estimation criterion.

In this paper, we modify the algorithm proposed by Gao et al. [8]. Both extremum values of attribute data (namely $I^1 = [x_{min}^{sk}, x_{max}^{sk}]$), mid-value μ and width ϵ (namely $I^2 = [\mu_{sk} - \epsilon, \mu_{sk} + \epsilon]$) are used to form splitting intervals. Considering that the value of ϵ can decide the speed at which data are split into different branches, we choose the standard deviation σ as the value of ϵ.

V. Experiment

The classification problem holds a pivotal position in the fields of machine learning and artificial intelligence (AI). It is not only a core driving force behind the advancement of AI technologies but also a crucial factor determining the successful deployment and application of intelligent systems.

To investigate this, we conduct experiments on the classification problem using the Iris dataset. For simplicity, the classes set, ver, vir are denoted as A, B, C, and attributes are denoted as SL, SW, PL, PW.

A. Generate basic probability assignment

Step 1: Determine the framework of discernment.

For the three classes in iris data set, the framework of discernment is $\{A, B, C\}$. Thus its power set spatial distribution is given by:

$$\begin{aligned} \Omega = \{&\{A\}, \{B\}, \{C\}, \{A, B\}, \\ &\{A, C\}, \{B, C\}, \{A, B, C\}\}. \end{aligned} \tag{18}$$

Step 2: Build Gaussian models on each attribute.

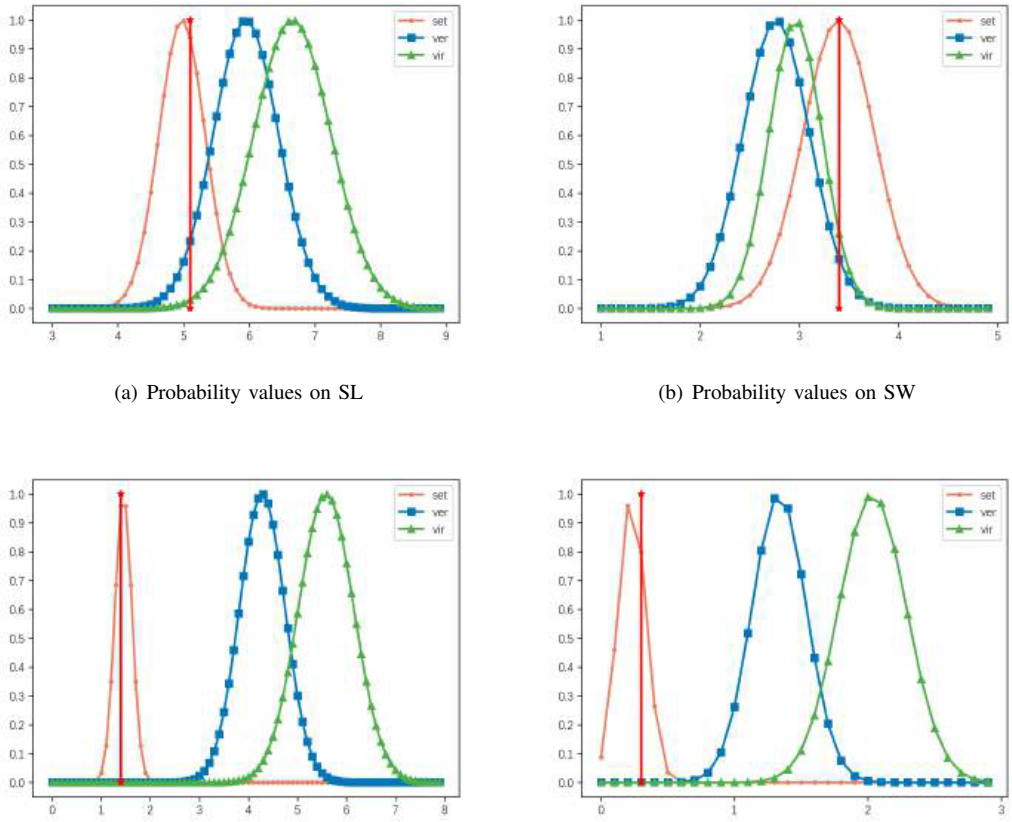
According to Equation(6) and Equation(7), means and standard deviations of each class on each attribute are calculated. With these critical parameters, Gaussian models can be established by Equation(8). For a given data set partitioning situation, the Gaussian models established on each attribute are shown in Fig 2.

Step 3: Match the test samples to the Gaussian model to get BPA.

For instance, consider the test data point (5.1, 3.8, 1.5, 0.3). The corresponding BPA derived from the Gaussian models is shown in Tab I.

Step 4: Combine the BPA with measurement of distance.

For each test data point, compute the distance values for each subset in the power set of the framework of discernment, as specified in Equation (13). Then take the negative exponent of the values and combine them with BPA in Step 3.



(a) Probability values on SL

(b) Probability values on SW

(c) Probability values on PL

(d) Probability values on PW

Fig. 2. Established Gaussian models on each attribute

B. Construct decision tree

Decision trees are constructed on different capacity of training set and their performance is determined by $Accuracy$ which means the ratio of samples correctly classified among all test samples.

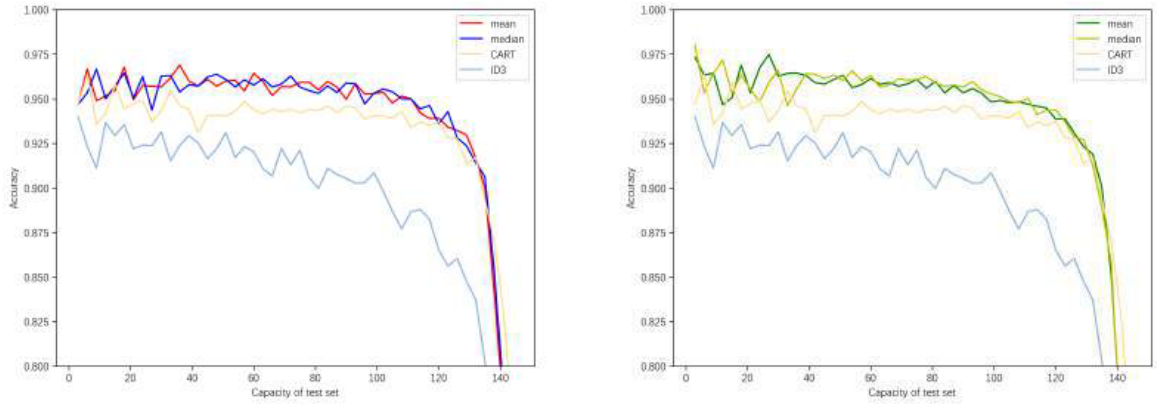
C. Results and discussion

To make comparison with existing methods and verify the performance of proposed method, we also apply it to traditional decision trees such as ID3(with "one vs all" strategy) and CART. For each method, we have done the experience 50 times and calculated the average accuracy of every time in each capacity of test set. To illustrate each methods' dependence on amount of training data and the influence of entropy, the variation of accuracy with capacity is shown in Fig 3.

The experimental results are averaged over the first half of the test set for each capacity value, with these averages serving as the final performance metrics for each method. Specifically, the Nguyen entropy method achieves accuracies of 95.784% for the mean measure and 95.788% for the median measure, while the Deng entropy method yields accuracies of 96.110% and 96.061%, respectively. In contrast, the CART method results in an accuracy of 94.425%, and the ID3 method produces an accuracy of 92.297%.

The proposed method demonstrates enhanced stability as the training set varies, with accuracy oscillating within relatively high ranges. This indicates that the method effectively improves the performance of the decision tree by incorporating uncertain information. In fact, the objective of classification tasks is to determine which category a test sample most closely resembles based on the training data, and the core principle of BPA is to represent uncertainty by assigning mass to subsets. By adjusting the mass function of each category subset according to the distance from the test sample, BPA can more accurately express the degree of membership of the test sample to each category, thereby ultimately improving the performance of the evidential decision tree.

It is also evident that the methods utilizing Nguyen entropy exhibit performance similar to those using Deng entropy. One possible reason for this is that the distributions of BPA generated by both methods are analogous. In this case, the multi-subsets in the generated BPA carry less information, resulting in minimal differences between the two entropy measures. For example, consider the test set data point (5.1, 3.8, 1.5, 0.3); the BPA generated from this point and the



(a) Results by Nguyen entropy

(b) Results by Deng entropy

Fig. 3. Results of each method

corresponding entropy values are recorded in Table I.

TABLE I
BPA GENERATED FROM REAL DATA AND THEIR ENTROPY

	A	AB	AC	ABC	Nguyen	Deng
SL	0.93	0.23	-	0.03	0.7278	1.1823
SW	0.99	-	0.17	0.25	0.9406	1.9391
PL	0.95	nearly 0	-	nearly 0	0.0581	0.0581
PW	0.79	nearly 0	-	nearly 0	0.2595	0.2595

However, let us assume that a group of BPA, where the multi-subsets contain more information, is obtained in some way. Table II presents three forms of BPA. As shown in Table II, the Nguyen entropy values for these BPA are identical, while the Deng entropy values differ. Furthermore, the greater the amount of information contained in the multi-subset, the larger the discrepancy between the Nguyen entropy and Deng entropy values. This can be attributed to the fact that Deng entropy considers the cardinality of the subsets, enabling it to utilize the information from the multi-subset more effectively.

TABLE II
HYPOTHETICAL BPA AND THEIR ENTROPY

	A	B	C	AB	ABC	Nguyen	Deng
α	0.1	0.1	-	0.8	-	0.9219	0.9219
β	0.1	0.1	-	-	0.8	0.9219	3.1678
γ	0.1	-	-	0.1	0.8	0.9219	3.3263

VI. CONCLUSION

In conclusion, this paper proposes a novel method for generating basic probability assignments (BPA). It uses Gaussian models based on the mean and standard deviation of each class to generate BPA for the test set. Distances between test samples and class centroids are calculated, then aligned with the power set distribution in the framework of discernment. These BPA are combined with distance values through their inner product. The method is validated using the evidential decision tree. When applied to the Iris classification problem, the proposed method achieves average accuracies for both entropy types that are 1.011% and 3.319% higher than those of other methods. This result demonstrates its effectiveness in attribute selection for splitting and BPA quality, while also highlighting its advantages in handling uncertain data.

ACKNOWLEDGMENTS

The work is supported by Qin Chuangyuan high-level innovation and entrepreneurship talent program of Shaanxi(Grant No.QCYRCXM-2023-108) and the "Innovation and Entrepreneurship Training Program for college students" project of Northwest A&F University (Project No. X202410712542).

REFERENCES

[1] D. A. P., Upper and lower probabilities induced by a multivalued mapping, Institute of Mathematical Statistics (1967).
[2] G. Shafer, A mathematical theory of evidence, Princeton University Press 42 (1976).
[3] Z. Wang, W. Yang, H. Zhang, Y. Zheng, Spa-based modified local reachability density ratio wsvdd for nonlinear multimode process monitoring, Complexity 2021 (1) (2021) 5517062.
[4] W. Fu, S. Yu, X. Wang, A novel method to determine basic probability assignment based on adaboost and its application in classification, Entropy 23 (7) (2021) 812.
[5] Y. Tang, Y. Zhou, X. Ren, Y. Sun, Y. Huang, D. Zhou, A new basic probability assignment generation and combination method for conflict data fusion in the evidence theory, Scientific Reports 13 (1) (2023) 8443.
[6] H. Garg, D. Rani, Novel distance measures for intuitionistic fuzzy sets based on various triangle centers of isosceles triangular fuzzy numbers and their applications, Expert Systems with Applications 191 (2022) 116228.
[7] A. Hassanat, E. Alkafaween, A. S. Tarawneh, S. Elmougy, Applications review of hassanat distance metric, in: 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), IEEE, 2022, pp. 1–6.
[8] B. Gao, Q. Zhou, Y. Deng, Hie-edt: Hierarchical interval estimation-based evidential decision tree, Pattern Recognition 146 (2024) 110040.