Basic Probability Assignment Generation Method Based on Kernel Density Estimation (Postprint)

Authors: Huang Jie, Wei Yongqing, Yi Jing, Liu Mengdi

Date: 2019-05-10T00:00:00+00:00

Abstract

D-S evidence theory is an effective approach for processing uncertain information and has been widely applied across various domains. The D-S combination method operates on basic probability assignments (BPA), and the generation of BPA remains a critical and yet-to-be-resolved primary step in the application of D-S theory. To address BPA generation, this paper proposes a BPA generation method based on Kernel Density Estimation (KDE): training data is employed to construct a data attribute model using kernel density estimation with optimized bandwidth; subsequently, the kernel density model of the training data is utilized to compute the density-distance-distribution values (Tri-D) for test data, and a nested approach is used to allocate Tri-D values to obtain the BPA corresponding to the test data; finally, D-S combination of the BPA yields the final decision, and the effectiveness of the BPA generation method is assessed through classification accuracy. Experiments conducted on UCI datasets, comparing classification accuracy with other methods, validate the effectiveness of the proposed approach.

Full Text

Preamble

Vol. 37 No. 6
Application Research of Computers (ChinaXiv Cooperative Journal)

A New Method for Generating Basic Probability Assignment Based on Kernel Density Estimation

Huang Jie[1],[2], Wei Yongqing[3]†, Yi Jing[1],[4], Liu Mengdi[1],[2]

1. School of Information Science & Engineering, Shandong Normal University, Jinan 250358, China

2. Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250014, China

3. Basic Education Department, Shandong Police College, Jinan 250014, China

4. School of Computer Science & Technology, Shandong Jianzhu University, Jinan 250014, China

Abstract: D-S evidence theory is an effective method for processing uncertain information and has been widely applied across various domains. The D-S combination rule operates on basic probability assignments (BPAs), yet how to generate BPAs remains a critical and unresolved first step in applying D-S theory. This paper proposes a BPA generation method based on kernel density estimation (KDE). Training data is used to construct data attribute models via KDE with optimized bandwidth. The kernel density models of training data are then used to compute density-distance-distribution (Tri-D) values for test data, from which BPAs are obtained through a nested allocation method. Finally, D-S combination of these BPAs yields the final decision, with classification accuracy serving as the metric for evaluating the effectiveness of the BPA generation method. Experiments on UCI datasets, comparing classification accuracy with other methods, validate the proposed approach.

Keywords: basic probability assignment (BPA); kernel density estimation; Tri-D; bandwidth

0 Introduction

D-S evidence theory, proposed by Dempster[1] and developed by Shafer[2], is an effective method for handling uncertain information and combining multiple sources of evidence. Alongside classical Bayesian theory, it represents one of two mainstream frameworks for data fusion. Compared to Bayesian theory, D-S theory offers the advantage of assigning probability to both singleton and compound subsets simultaneously. D-S theory has been successfully applied in numerous fields including data fusion[3], comprehensive evaluation[4,5], classification[6,7], and evolutionary game theory[8–10].

In applying the D-S framework, BPA generation is a critical and fundamental first step that significantly influences final results. However, no universal solution for BPA generation has yet been established, prompting extensive research and diverse proposed methods. Early work by Yager[11] introduced an entire class of fuzzy measures related to D-S belief structures and discussed their entropy. Jiang et al.[12–14] proposed methods for generating BPAs based on generalized fuzzy numbers. Liu et al.[15] combined fuzzy Naive Bayes with nearest-neighbor mean classification to propose the WFDSF algorithm. Reference [16] presented a method for obtaining BPAs using core examples from training data. Reference

[17] proposed an approach based on classifier confusion matrices. Additionally, there are methods for determining BPAs based on clustering[18],[19].

Analysis of these existing methods reveals that BPA generation essentially involves determining the probability of possible events, and kernel density estimation is precisely an effective non-parametric method for probability density estimation based entirely on data itself. Therefore, this paper proposes a BPA generation method based on KDE. First, training data is used to construct KDE models with optimized bandwidth (hereinafter referred to as h). Kernel density values for test data attributes are computed within these models, from which Tri-D values are derived and BPAs are generated through nested allocation[20]. D-S combination of these BPAs yields the final decision, with validation performed on UCI datasets.

1.1 D-S Evidence Theory

We first introduce the fundamental concepts of D-S evidence theory.

a) Frame of discernment.
Let $\Theta = \{c_1, c_2, ..., c_v\}$ denote the frame of discernment, where the hypotheses are exhaustive and mutually exclusive. The power set of Θ contains 2^v subsets.

b) Basic probability assignment (BPA).
Let m be a mapping from the power set 2^Θ to $[0, 1]$ satisfying Eq. (1). Then m is called a basic probability assignment (hereinafter referred to as BPA), also known as a mass function or evidence. Subsets with non-zero m values are called focal elements. The m value represents the degree of support for a subset, with larger values indicating greater support.

c) Combination rule.
For two pieces of evidence, the D-S combination rule is given by Eq. (3), where B_i and C_j are focal elements and k is the conflict coefficient.

1.2 Pignistic Probability[21]

After D-S combination of BPAs, the hypothesis with maximum Pignistic probability is taken as the final result, calculated using Eq. (4), where $|X|$ denotes the cardinality of set X.

Having introduced the fundamentals of D-S evidence theory, we briefly explain its relationship to this work. Our objective is to transform data with different features into BPAs of the form $m(A) = 0.3, m(B) = 0.6, m(C) = 0.1$ that satisfy Eq. (1), where A, B, C are hypotheses in Θ (i.e., possible classes or labels). These BPAs are then combined using Eqs. (3) and (4) to obtain a final BPA. Thus, BPA generation is an indispensable step in D-S applications.

1.3 Kernel Density Estimation (KDE)

Methods for estimating the distribution density of a given sample set include parametric and non-parametric estimation. Kernel density estimation (also known as the Parzen window method) belongs to the latter category. Unlike parametric estimation, which requires the often-inaccurate assumption that data follows specific distributions across possible classes, KDE—proposed by Rosenblatt[22] and Parzen[23]—studies data distribution characteristics from the data itself and has received significant attention in statistics and various application fields.

For independent and identically distributed random variables $x_1, x_2, ..., x_n$, the kernel density estimate of the true probability density function $f(x)$ is expressed as:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is the kernel function and h is a predetermined positive number called the bandwidth or smoothing parameter. For convenience, we denote $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$.

As shown in Eq. (6), the density estimate $\hat{f}$ depends not only on the given sample set but also on the choice of kernel function and bandwidth parameter. Common kernel functions include Gaussian, triangular, uniform, and Epanechnikov kernels, which we will not elaborate on here. This work focuses primarily on the bandwidth parameter h, whose value determines the smoothness of the density curve. [Figure 1: see original paper] shows KDE curves for normally distributed random data under different h values.

As illustrated in [Figure 1: see original paper], when h is too small, the curve becomes overly steep and fluctuating, exhibiting overfitting. As h increases, the curve gradually smooths. When h reaches 2, the curve becomes excessively smooth, masking data characteristics. In summary, selecting an appropriate h is crucial. This paper employs an optimization method to select h, detailed in Section 2.1.

2 BPA Generation Method Based on Kernel Density Estimation

The fundamental requirement for BPA generation is to satisfy subsequent combination processes while extracting and utilizing as much data information as

possible. Different attributes reflect data characteristics from different perspectives; therefore, this work uses attributes as the basic unit for bandwidth optimization and KDE attribute model construction. Subsequent BPA allocation is also based on attribute models (note: the v KDE sub-curves for v classes obtained under a one-dimensional attribute are collectively referred to as an attribute model).

We now introduce our method under the following assumptions: there are v possible hypotheses or final classes, i.e., the frame of discernment is $\Theta = \{c_1, c_2, ..., c_v\}$. The dataset contains n data instances, each with p attributes, and x_t denotes a test instance.

2.1 Attribute KDE Model with Optimized Bandwidth

Existing bandwidth optimization methods are primarily based on minimizing the mean integrated squared error (MISE)[24], defined as:

$$\mathrm{MISE}(h) = E\left[\int (\hat{f}(x) - f(x))^2 dx\right] = \int \left[E\hat{f}(x) - f(x)\right]^2 dx + \int \mathrm{Var}(\hat{f}(x)) dx$$

The first term after decomposition represents the bias between the expected and true values, while the second term represents the variance of the estimate. Solving the above equation yields expressions for bias and variance. Through derivation and minimization, the optimal bandwidth h is obtained as (detailed derivation in reference [24]):

$$h_{opt} = 1.0593\hat{\sigma}n^{-1/5}$$

where $\hat{\sigma}$ and n represent the sample standard deviation and size, respectively. Using Eq. (7), we can compute the optimal bandwidth h_{ij} for attribute j corresponding to class i in the training data, ultimately obtaining an optimized bandwidth matrix H^* of size $p \times v$.

The training data is used to construct attribute KDE models as follows: training data for attribute j is partitioned by different classes. The optimal bandwidths h for different classes under different attributes have been calculated in H^*. Kernel density estimation is performed using Eq. (6), resulting in v sub-KDE models under each one-dimensional attribute.

2.2 Density-Distance-Distribution Value (Tri-D)

With the KDE attribute models for training data constructed, density values for test data attributes can be obtained. While larger density values for a test instance under a particular class indicate higher likelihood of belonging to that class, relying solely on density values for classification can lead to errors when inter-class distances are small or density differences are subtle.

Consider 10 data points randomly generated in interval [4,5] as class A and 4 points in [5.2,5.6] as class B. Their KDE curves are shown in [Figure 2: see original paper]. The figure reveals substantial overlap between classes A and B along the x-axis. As the inter-class distance decreases (moving curve A rightward or curve B leftward), the overlap increases, making density values alone insufficient for correct classification.

In traditional classification and clustering algorithms, KNN uses distance to select k-nearest neighbors, and K-means uses distance between instances and centroids for clustering, demonstrating that distance is a crucial factor for class determination. Moreover, distance distributions among points within the same class should exhibit higher similarity. Therefore, combining density, distance, and distribution characteristics, we propose the Tri-D value concept.

For test instance x_t, the Tri-D value for attribute j regarding class i in the corresponding attribute model is defined as:

$$\text{Tri-D}_{ij}^{t} = f_{ij}(x_t^j) \cdot \frac{1}{|d_{ij}^t - \bar{d}_{ij}|}$$

where $f_{ij}(x_t^j)$ represents the density value of test instance x_t' s attribute j on the class i curve; d_{ij}^t denotes the distance from x_t' s attribute j value to the centroid of class i data for attribute j in the training set; and $\bar{d}_{ij}$ represents the mean distance between any two points in class i data for attribute j of the training data. The centroid matrix is obtained via the K-means++ algorithm[25].

Analysis of this definition reveals: First, a larger $f_{ij}(x_t^j)$ value indicates higher likelihood that x_t belongs to class i under attribute j. Second, the second component comprises two distances: the distance from the test point to the training class centroid, and the average distance between any two points in training class i. Comparing these two distances incorporates distribution characteristics –if they are similar (i.e., $|d_{ij}^t - \bar{d}_{ij}|$ is small), we consider x_t to have a distance distribution similar to that class. Therefore, larger Tri-D values indicate higher likelihood of x_t belonging to class i under attribute j, and vice versa.

2.3 Basic Probability Assignment Generation

Each attribute j of test instance x_t yields v Tri-D values, resulting in a $p \times v$ Tri-D value matrix. Using attribute j of x_t as an example, we explain our nested allocation method[20]: attribute j produces a vector Tri-D_j^t of v Tri-D values, with class vector $C_d = \{c_1, c_2, ..., c_v\}$. Sorting Tri-D_j^t in descending order yields $\text{Tri-D}_j^{t'}$, with corresponding sorted class vector C_d'. The allocation is performed as:

$$
\begin{aligned}
m(\{C'_d[1]\}) &= 1 - \text{Tri-D}_j^{t'}[1] \\
m(\{C'_d[1], C'_d[2]\}) &= \text{Tri-D}_j^{t'}[1] - \text{Tri-D}_j^{t'}[2] \\
&\vdots \\
m(\{C'_d[1], C'_d[2], ..., C'_d[v]\}) &= \text{Tri-D}_j^{t'}[v]
\end{aligned}
$$

This allocation method simultaneously yields both singleton and compound BPA focal elements. To satisfy Eq. (1), normalization is applied to obtain the final BPA. Consequently, each attribute of x_t produces one BPA, resulting in p BPAs from p attributes, which are then combined to obtain the final BPA.

2.4 Algorithm Procedure

The algorithm flow is illustrated in [Figure 3: see original paper]. Details are as follows: Data is split into training data for learning and model construction, and test data for method evaluation. First, the KDE model is built from training data, treating each attribute as an information source. For each attribute j and class i, the optimal bandwidth h_{ij}^* is computed using Eq. (7), yielding the optimal bandwidth matrix H^*. Kernel density estimation is performed for each class' s data under each attribute using the corresponding optimized bandwidth (Gaussian kernel is used in this work, see Section 4.2), constructing the KDE model for attribute j.

For test data: Tri-D values are computed using Eq. (8) based on the trained attribute models, yielding a value matrix for each test instance x_t. Allocation using Eq. (9) produces p BPAs, which are combined via D-S combination to obtain the final BPA. The hypothesis with maximum Pignistic probability is taken as the final result.

---

3 Experiments and Results Analysis

This section presents three groups of experiments on UCI datasets: Section 3.1 illustrates the experimental procedure using the Iris dataset; Section 3.2 determines kernel function selection and demonstrates bandwidth optimization effectiveness; Section 3.3 validates the proposed method through classification accuracy on UCI datasets.

[Figure 3: see original paper] shows the algorithm flow of the proposed method.

3.1 Experimental Example Using Iris Dataset

This section demonstrates the algorithm flow through classification of a test instance from Iris. The Iris dataset is a benchmark classification dataset with

three classes: Setosa (S), Versicolor (C), and Virginica (V). Each instance comprises four attributes: SL, SW, PL, PW, with 50 examples per class (150 total). This example uses 80% training data. A test sample with true label C has attribute values: SL=6.1cm, SW=2.9cm, PL=4.7cm, PW=1.4cm.

First, attribute KDE models are constructed from training data: optimal bandwidths for each attribute and class are computed using Eq. (7), as shown in . [Figure 4: see original paper]-[Figure 7: see original paper] display the KDE models and class-specific curves for each attribute. Second, Tri-D values are computed: K-means++[25] clustering yields class centroids for each attribute (), and the Tri-D matrix is calculated (). Third, BPAs are generated: the nested allocation method produces four BPAs corresponding to the four attributes (). Conflict coefficients between any two BPAs are computed using Eq. (3), yielding $k_{12} = k_{13} = k_{14} = k_{23} = k_{24} = k_{34} = 0$, indicating no conflict. Finally, D-S combination produces the final BPA: $m(S) = m(V) = 0, m(C) = 1$, correctly classifying the test instance as class C.

3.2 Experiments on Kernel Function and Bandwidth Selection

KDE involves two critical factors—bandwidth h and kernel function—both affecting results. This section compares 5-fold cross-validation accuracy on UCI datasets under varying configurations.

First, accuracy comparisons using different kernel functions are shown in [Figure 8: see original paper]. Overall, the Gaussian kernel performs best. Examining individual datasets, performance is comparable across kernels for Iris, Heart, and Australian, with slight differences for Zoo and Sonar.

Second, under Gaussian kernel, [Figure 9: see original paper] demonstrates the impact of different bandwidths. Accuracy using optimized bandwidth significantly exceeds that using default bandwidth, as optimized bandwidths are specifically derived from data characteristics of different attributes and classes, more accurately reflecting data features and yielding higher accuracy.

3.3 Classification Experiments on UCI Datasets

This experiment validates the proposed method by comparing with seven well-known classifiers and three other BPA generation methods. As shown in , the selected UCI datasets vary in instance count, class count, attributes, and missing values, providing comprehensive validation. 5-fold cross-validation results are presented in .

Results demonstrate that the proposed method achieves superior classification accuracy across diverse UCI datasets compared to common classifiers and other BPA generation methods. While all BPA generation methods perform relatively poorly on the Sonar dataset—likely due to many features causing fusion deviation during BPA combination—the proposed method remains effective overall,

confirming its capability to generate BPAs from data with different features for D-S fusion.

4 Conclusion

BPA generation remains an open problem in D-S evidence theory applications. This paper proposes a KDE-based BPA generation method: first constructing optimized KDE attribute models from training data, then computing and allocating Tri-D values to obtain test data BPAs, and finally deriving results through D-S combination. The method offers advantages of low subjectivity and minimal BPA conflict. Experimental results validate its effectiveness, though limitations exist for high-dimensional datasets. Future work will consider improved fusion methods to address this issue.

References

[1] Dempster A P. Upper and lower probabilities induced by a multivalued mapping [J]. *Annals of Mathematical Statistics*, 1967, 38(2): 325-339.
[2] Shafer G. *A Mathematical Theory of Evidence* [M]. New Jersey: Princeton University Press, 1976.
[3] He Jingjing, Cai Desheng, Jie Fei, et al. Synonymous entity recognition based on D-S evidence theory for feature fusion [J]. *Application Research of Computers*, 2018, 35(5): 1429-1433.
[4] Men Zhiyuan, Li Linhong, Zhang Yaohui. State evaluation method of gear technology based on improved evidence theory [J]. *Fire Control & Command Control*, 2018, 43(5): 65-68.
[5] Yu Siqi, Jing Bo, Huang Yifeng. Test-based comprehensive evaluation method based on D-S evidence theory [J]. *Application Research of Computers*, 2014, 31(7): 2071-2073.
[6] Xu Xiaobin, Zheng Jin, Yang Jianbo, et al. Data classification using evidence reasoning rule [J]. *Knowledge-Based Systems*, 2017, 116(1): 144-151.
[7] Li Ying. Weed identification based on multi-classifier DS evidence theory fusion [J]. *Computer Programming Skills & Maintenance*, 2018(2): 145-146, 153.
[8] Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks [J]. *Physics Reports*, 2014, 544(1): 1-122.
[9] Burkov A, Chaib-Draa B. Computing equilibria in discounted dynamic games [J]. *Applied Mathematics & Computation*, 2015, 269(10): 39-49.
[10] Wang Zhen, Wang Lin, Szolnoki A, et al. Evolutionary games on multilayer networks: a colloquium [J]. *European Physical Journal B*, 2015, 88(5): 124.
[11] Yager R R. A class of fuzzy measures generated from a Dempster-Shafer belief structure [J]. *International Journal of Intelligent Systems*, 1999, 14(12): 1239–1247.

[12] Jiang Wen, Chen Yundong, Tang Chao, et al. Basic probability assignment generation method based on sample difference degree [J]. *Control and Decision*, 2015, 30(1): 71-75.
[13] Tu Shijie, Chen Hang, Feng Gang. Ballistic target recognition based on evidence theory and fuzzy function [J]. *Computer Engineering and Applications*, 2017, 53(6): 169-173.
[14] Jiang Wen, Yang Yan, Luo Yu, et al. Determining basic probability assignment based on the improved similarity measures of generalized fuzzy numbers [J]. *International Journal of Computers Communications & Control*, 2015, 10(3): 333-347.
[15] Liu Yuting, Pal N R, Marathe A R, et al. Weighted fuzzy Dempster-Shafer framework for multimodal information integration [J]. *IEEE Trans on Fuzzy Systems*, 2018, 26(1): 338-352.
[16] Zhang Chenwei, Hu Yong, Chan Felix T S, et al. A new method to determine basic probability assignment using core samples [J]. *Knowledge-Based Systems*, 2014, 69(1): 140-149.
[17] Deng Xinyang, Liu Qi, Deng Yong, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem [J]. *Information Sciences*, 2016, 340-341: 18-31.
[18] Gao Zhiyong, Dong Rongguang, Gao Jianmin, et al. Basic probability distribution generation method and application using clustering features [J]. *Journal of Xi' an Jiaotong University*, 2016, 50(10): 8-14.
[19] Chen Yanan, Ren Shengjun. Research on generating probability assignment function of evidence under large sample conditions [J]. *Journal of Information Engineering University*, 2017, 18(4): 118-122.
[20] Mahadevan S, Deng Yong, Xu Peida, et al. A new method to determine basic probability assignment from training data [J]. *Knowledge-Based Systems*, 2013, 46(1): 69-80.
[21] Li Minzheng, Lan Jianping. Improved evidence combination method based on Pignistic probability distance [J]. *Journal of Guilin University of Electronic Technology*, 2017, 37(4): 63-67.
[22] Rosenblatt M. Remarks on some nonparametric estimates of a density function [J]. *Annals of Mathematical Statistics*, 1956, 27(3): 832-837.
[23] Parzen E. On estimation of a probability density function and mode [J]. *Annals of Mathematical Statistics*, 1962, 33(3): 1065-1076.
[24] Bolancé C, Guillen M, Nielsen J P. Kernel density estimation of actuarial loss functions [J]. *Insurance Mathematics & Economics*, 2009, 32(1): 19-36.
[25] Arthur D. K-means++: the advantages of careful seeding [C]// Proc of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007: 1027-1035.

Note: Figure translations are in progress. See original paper for figures.