Annals of Data Science (2024) 11(2):709–732
https://doi.org/10.1007/s40745-023-00490-4

Improving Bayesian Classifier Using Vine Copula and Fuzzy Clustering Technique

Ha Che-Ngoc[1] · Thao Nguyen-Trang[2,3] · Hieu Huynh-Van[4,5,6] · Tai Vo-Van[7]

Received: 26 January 2023 / Revised: 13 July 2023 / Accepted: 26 July 2023 /
Published online: 10 August 2023

Abstract

Classification is a fundamental problem in statistics and data science, and it has garnered significant interest from researchers. This research proposes a new classification algorithm that builds upon two key improvements of the Bayesian method. First, we introduce a method to determine the prior probabilities using fuzzy clustering techniques. The prior probability is determined based on the fuzzy level of the classified element within the groups. Second, we develop the probability density function using Vine Copula. By combining these improvements, we obtain an automatic classification algorithm with several advantages. The proposed algorithm is presented with specific steps and illustrated using numerical examples. Furthermore, it is applied to classify image data, demonstrating its significant potential in various real-world

✉ Tai Vo-Van
vvtai@ctu.edu.vn

Ha Che-Ngoc
chengocha@tdtu.edu.vn

Thao Nguyen-Trang
thao.nguyentrang@vlu.edu.vn

Hieu Huynh-Van
hvhieu.sdh221@hcmut.edu.vn

1 Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

2 Laboratory for Applied and Industrial Mathematics, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Vietnam

3 Faculty of Basic Sciences, Van Lang University, Ho Chi Minh City, Vietnam

4 Faculty of Applied Science, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam

5 Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

6 Faculty of Fundamental Science, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

7 College of Natural Science, Can Tho University, Can Tho City, Vietnam

applications. The numerical examples and applications highlight that the proposed algorithm outperforms existing methods, including traditional statistics and machine learning approaches.

Keywords Bayesian method · Bayes error · Prior probability · Vine Copulas

1 Introduction

Data science is a dynamic and rapidly evolving field that plays a crucial role in the modern era of data-driven decision-making. It combines elements of statistics, mathematics, computer science, and domain expertise to extract insights, solve complex problems, and make predictions from large and diverse datasets. One of the key aspects of data science is its interdisciplinary nature. Data scientists employ a diverse range of techniques, including data mining, machine learning, statistical modeling, and data visualization, to extract knowledge and drive informed decision-making. They possess a strong understanding of both the technical aspects of data analysis and the domain-specific context in which they operate [1, 2].

Data science involves the extraction of insights and knowledge from data using various techniques, algorithms, and methodologies. Classification problem is one of the fundamental tasks in data science, where the goal is to categorize or assign a given input or data point to predefined classes or categories [3, 4]. In today's world of scientific and technological advancement, the classification problem is receiving increasing attention from researchers. In traditional statistics, popular classification methods include Fisher, Logistic, and Bayesian methods [5, 6]. These methods have been widely studied and applied in various fields such as finance, medicine, and technology [7–14]. The Fisher method, one of the earliest classification algorithms, can be used for binary and multi-classification problems. However, it requires the same covariance matrices for all groups, which might not always be practical [15]. The Logistic regression method is effective only for binary classification with well-separated data [16].

With the utilization of machine learning and deep learning techniques, a wide array of classification methods has become available, including Quadratic Discriminant Analysis (QDA), XGBoost, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Artificial Neural Network (ANN). QDA is an improvement over the Fisher method as it does not require a common covariance matrix for groups [17]. XGBoost can effectively classify large datasets and mitigate overfitting [18], although its performance may be hindered in cases where there is significant overlap between groups. The k-NN method is complex and offers high classification performance; however, the results heavily depend on the choice of the parameter k and the distance metric used [19, 20]. Furthermore, the computational time of this method can be significant when dealing with large datasets, as it requires calculating distances between the classified element and all elements in the training set. SVM is a robust classification tool in multidimensional space; however, its stability is not guaranteed in all situations [4, 7, 21]. While ANN is considered a powerful method with several improved versions capable of capturing complex non-linear relationships between dependent and independent variables without relying heavily on predefined assumptions, it is prone to

overfitting. Additionally, ANN requires substantial computational resources, and the initial parameter settings are often based on the researcher's expertise, making it more suitable for standard datasets and powerful computer systems [22].

The Bayesian classifier is a versatile method that can be used for binary and multi-classification groups, and it offers several advantages over other methods. It is not restricted by the normal distribution of data or the assumption of equal variance among groups [23]. In recent years, various authors have made improvements to this method for image data, enabling it to compete with deep learning techniques [24, 25]. However, it is important to note that the Bayesian classifier is only suitable for certain cases. To utilize the Bayesian classifier, two key components need to be determined: (i) prior probabilities and (ii) probability density functions (pdfs). Prior probabilities are often chosen based on expert knowledge or previous statistical findings, which can be a limitation in real-world applications [13]. For (ii), pdfs are typically estimated assuming independent variables. However, in real data, strong correlations among variables often exist, which can result in an under-fitting model in many cases [5].

A Copula is a multivariate cumulative distribution function where the marginal pdf of each variable is uniform on the interval [0,1]. Sklar [26] demonstrated that any multivariate pdf can be expressed as a product of the marginal pdfs with a Copula that describes the dependency structure among the variables. Copula has been proposed as a suitable method for describing the dependency properties among variables. However, most of the studies in the literature primarily focus on constructing Copula families for two-dimensional pdfs. As a result, there are only a few Copula families available to express the functions of more than two variables [27–29]. Therefore, the application of Copula to construct the joint pdf in classification problems is becoming challenging.

This paper proposes a new approach for classification problem to address the research mentioned above gap using Bayesian method. Firstly, we utilize Copula to describe the dependency properties among variables to determine the pdfs. However, because Copula has a disadvantage in dealing with more than two variables, we employ the Vine Copula [30, 31] to estimate the Copula of all used variables through the Copulas of pairwise variables. This technique not only overcomes the disadvantage of lacking Copula families for more than two variables but also inherits the advantages of the pair-Copula that have been widely studied. Secondly, for determining the prior probabilities for each observation, we use the fuzzy clustering technique, which can estimate the membership degrees of the observation to the given classes. Previous studies have commonly used a prior probability that is uniformly distributed, the Laplace method, or based on the contribution ratio of each group in the training set [5, 6, 13, 32]. This means that the classified element does not affect the prior probability. Therefore, using the proposed prior probability will have a positive effect on reducing the classification error. Combining these improvements, we offer a new classifier method that can be effectively implemented using the established Matlab code. We have compared this method to others, including traditional statistics and popular deep learning methods, using many datasets with different characteristics. We have found that the proposed method provides outstanding results.

The remainder of this paper is constructed as follows. Section 2 presents the related theoretical issues, including the Bayesian classifier, Bayes error, and Vine Copula. The proposed algorithm is presented in Sect. 3. This section also presents numerical

examples to illustrate the proposed algorithm. Section 4 provides the applications. The conclusion is summarized in Sect. 5.

2 Some Related Issues

2.1 Principle of Classifying and Bayes Error

Let $\{w_1, w_2, \ldots, w_k\}$ be the given k groups, $f_i(x)$ and q_j, $j = \overline{1,k}$ be pdfs and prior probabilities of the jth group, respectively. According to [15, 32], an element x_0 will be assigned to w_j if the following equation holds:

$$g_{\max}(x_0) = q_j f_j(x_0), \tag{1}$$

where

$g_j(x) = q_j f_j(x)$,

$g_{\max}(x) = \max\{g_1(x), g_2(x), ..., g_k(x)\}$.

The theoretical misclassification of Bayesian method, namely Bayes error, is given by (2):

$$Pe^{(q)}_{1,2,\ldots,k} = \sum_{i=1}^{k} \int_{R^n \setminus R_i^n} q_i f_i dx = 1 - \sum_{i=1}^{k} \int_{R_i^n} q_i f_i(x) dx, \tag{2}$$

where

$R_i^n = \{x | q_i f_i(x) > q_j f_j(x), \forall i \neq j, i, j = 1, 2, ..., k\}$, $(q) = (q_1, q_2, ..., q_k)$.

From (2), we have

$$\begin{aligned} Pe^{(q)}_{1,2,\ldots,k} &= \sum_{j=1}^{k} \int_{R^n \setminus R_j^n} q_j f_j(x) dx \\ &= \sum_{j=1}^{k} \left[\int_{R^n} q_j f_j(x) dx - \int_{R_j^n} \max_{1 \le l \le k} \{q_l f_l(x)\} dx \right] \\ &= \int_{R^n} \sum_{j=1}^{k} q_j f_j(x) dx - \sum_{j=1}^{k} \int_{R_j^n} \max_{1 \le l \le k} \{q_l f_l(x)\} dx \\ &= 1 - \int_{R^n} \max_{1 \le l \le k} \{q_l f_l(x)\} dx. \end{aligned} \tag{3}$$

From Formula (3), to calculate Bayes error we must find $g_{\max}(x)$ and its integral. This issue has been interested in many authors [15, 32]. In the case one-dimension, we can

find $g_{\max}(x)$. Therfore, the Bayes error can be found easily. In the multidimensional case, finding the line (surface) of these $g_i(x), i = 1, 2, ..., k$ are very complex (even the case of $k = 2$, and two pdfs with normal distributions). In this article, we integrate function $g_{\max}(x)$ by quasi Monte-Carlo method.

2.2 Some Results About Bayes Error

Theorem 1 *Let $f_i(x), i = 1, 2, ..., k, k \geq 3$ be k pdfs defined on $R^n, n \geq 1, q_i \in (0; 1)$. We have the relationships of Bayes error with other measures as follows:*

(i)

$$m \leq Be \leq M, \tag{4}$$

where $m = \frac{1}{k}[(k-1) - \sum_i \sum_j \left\| g_i, g_j \right\|_1$, $M = 1 - \frac{1}{2} \max_{i<j} \left\| g_i, g_j \right\|_1 - \min_i \{q_i\}$.

(ii)

$$0 \leq Be \leq \max_i \{q_i\}, \tag{5}$$

where $g_i(x) = q_i f_i(x)$, $\left\| g_i, g_j \right\|_1 = \int_{R^n} |g_i(x) - g_j(x)| dx$.

Proof (i) We have $\int_{R^n} \max\{g_1(x), g_2(x), ..., g_k(x)\} dx \geq \max_{i<j} \left\{ \int_{R^n} \max\{g_i(x), g_j(x)\} dx \right\}$.

On the other hand,

$$\begin{aligned}
\max_{i<j} \left\{ \int_{R^n} \max\{g_i(x), g_j(x)\} dx \right\} &= \max_{i<j} \left\{ \frac{1}{2} \left\| g_i, g_j \right\|_1 + \frac{1}{2} \left(q_i + q_j \right) \right\} \\
&\geq \max_{i<j} \left\{ \frac{1}{2} \left\| g_i, g_j \right\|_1 \right\} + \min_{i<j} \left\{ \frac{1}{2} \left(q_i + q_j \right) \right\} \\
&\geq \max_{i<j} \left\{ \frac{1}{2} \left\| g_i, g_j \right\|_1 \right\} + \min_{i<j} \left\{ (q_1, q_2, ..., q_k) \right\}.
\end{aligned}$$

Hence,

$$\int_{R^n} g_{\max}(x) dx \geq \frac{1}{2} \max_{i<j} \left\{ \left\| g_i, g_j \right\|_1 \right\} + \min_{i<j} \{(q_i)\}. \tag{6}$$

We also have

$$\sum_{i<} \sum_j |g_i - g_j| \geq \sum_{j=1}^{k} \left[\max\{g_1, g_2, \cdots g_k\} - g_j \right]$$

$$= k\left[\max\left\{g_1, g_2, \cdots g_k\right\}\right] - \sum_{j=1}^{k} g_j,$$

Therefore,

$$\max\left\{g_1, g_2, ...g_k\right\} \leq \frac{1}{k}\sum_{i<}\sum_{j}\left|g_i - g_j\right| + \frac{1}{k}\sum_{j=1}^{k} g_j.$$

Since $\int\limits_{R^n} g_i(x)dx = q_i$ and $\sum\limits_{i=1}^{k} q_i = 1$, the inequality (5) becomes:

$$\int\limits_{R^n} g_{\max}(x)dx \leq \frac{1}{k}\sum_{i<j}\left\|g_i, g_j\right\|_1 + \frac{1}{k}. \tag{7}$$

Replacing $\int\limits_{R^n} g_{\max}(x) = 1 - Be$ to (6) and (7), we have (4).

(ii) We have $q_i f_i(x) \leq \max\left\{q_1 f_1(x), q_2 f_2(x), ..., q_k f_k(x)\right\} \leq \sum\limits_{i=1}^{k} q_i f_i(x), \forall i = 1, ..., k$. Integrating the above relation, we have:

$$q_i \leq \int\limits_{R^n} g_{\max}(x)dx \leq 1.$$

Above inequality is true $\forall i = 1, ..., k$, so

$$\max\{q_i\} \leq \int\limits_{R^n} g_{\max}(x)dx \leq 1.$$

Replacing $\int\limits_{R^n} g_{\max}(x) = 1 - Be$ in the above relation, we have (5).

□

2.3 Vine Copula

Let $X = (X_1, ..., X_d)$ be a vector of d random variables, with a joint pdf $f(x)$. According to [26], every multivariate distribution F with marginals $F_1, \ldots, F_d$ can be written by (8).

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \tag{8}$$

for some appropriate $d-$dimensional copula C. Using the chain rule, we further have for an absolutely continuous F with strictly increasing continuous marginals

$F_1, \ldots, F_d$ that

$$f(x_1, ..., x_d) = \left[\prod_{k=1}^{d} f_k(x_k)\right] \times c(F_1(x_1), \ldots, F_d(x_d)), \tag{9}$$

where $c(.)$ denotes the copula density.

In practical applications with higher dimensions, the availability of suitable copulas is limited. Multivariate copulas, such as elliptical or exchangeable Archimedean copulas, are often too restrictive and inadequate for modeling dependence. Consequently, there is an increasing demand for more flexible copulas. Bedford, Cooke, Zhang, and others [30, 33, 34] introduced the concept of a regular vine distribution, which was further detailed by Kurowicka [35]. The regular vine distribution involves the specification of a sequence of trees, where each edge represents a bivariate copula known as a pair copula. These pair copulas serve as the fundamental components of the joint regular vine distribution. The regular vine (R-vine) $\mathcal{V}$ on d variables comprises trees $T_1, \ldots, T_{d-1}$ with nodes N_i and edges E_i for $i = 1, \ldots, d-1$, satisfying the following conditions:

(i) T_1 has nodes $N_1 = \{1, \ldots, d\}$ and edges E_1.
(ii) For $i = 2, \ldots, d-1$, the tree T_i has nodes $N_i = E_{i-1}$.
(iii) If two edges in tree T_i are to be joined by an edge in tree T_{i+1} they must share a common node (proximity condition).

Among the studies based on R-vine copula, drawable vines (D-vines) and canonical vines (C-vines), which are two special cases of regular vines, are the most well-known methods. Particularly,

- An R-vine is called a D-vine if each node in T_1 has a degree of at most 2, in which the node's degree is the number of connections between the considered node to other nodes.
- An R-vine is called a C-vine if each tree T_i has a unique node with degree $d - i$, namely the root node.

This gives the following decomposition of a multivariate density

$$f(x) = \prod_{k=1}^{d} f_k(x_k) \times \prod_{i=1}^{d-1} \prod_{j=1}^{d-i} c_{i,i+j|1:(i-1)} \left(F\left(x_i|x_1, \ldots, x_{i-1}\right), F\left(x_{i+j}|x_1, \ldots, x_{i-1}\right)|\theta_{i,i+j|1:(i-1)}\right), \tag{10}$$

where $f_k, k = 1, \ldots, d$ are the marginal densities and $c_{i,i+j|1:(i-1)}$ bivariate copula densities with parameter(s) $\theta_{i,i+j|1:(i-1)}$ (in general $i_k : i_m$ means $i_k, \ldots, i_m$). In the same way, D-vine density which also conveniently decomposes a d-dimensional density (as above the order is w.l.o.g. chosen as $1, \ldots, d$; otherwise nodes can be relabeled) is given by

$$f(x) = \prod_{k=1}^{d} f_k(x_k) \times \prod_{i=1}^{d-1} \prod_{j=1}^{d-i} c_{j,j+i|(j+1):(j+i-1)} \left(F\left(x_j|x_{j+1}, \ldots, x_{j+i-1}\right),\right.$$

$$...., F\left(x_{j+i}|x_{j+1},\ldots,x_{j+i-1}\right)|\theta_{j,j+|(j+1):(j+i-1),}\Big).$$

Based on the structures R-vines and C-D-vines clearly identified problem complex relationships between variables was solved. Since the Vine-copulas structure is based on nodes and edges, some authors used the sequential method to select a C-vine and a R-vine, based on Kendall's tau [36–43]. In this study, the following steps is performed:

Step 1 Input data $(x_{\ell_1},\ldots,x_{\ell_n})$, $\ell=1,\ldots,N$ (realizations of i.i.d. random vectors).
Step 2 Output R-vine (C-vine, D-vine) copula specification.
Step 3 Calculate the empirical Kendall's tau $\hat{\tau}_{j,k}$ for all possible variable pairs $\{j,k\}$, $1\le j<k\le n$.
Step 4 Select the spanning tree that maximizes the sum of absolute empirical Kendall's taus, i.e.,

$$\max \sum_{e=\{j,k\}\text{in spanning tree}} \|\hat{\tau}_{j,k}\|. \tag{11}$$

Step 5 For each edge $\{j,k\}$ in the selected spanning tree, select a copula and estimate the corresponding parameters.
Step 6 Transform to pseudo observations $F_{j|k}(x_{\ell j}|x_{\ell k})$ and $\hat{F}_{k|j}(x_{\ell k}|x_{\ell j})$, $\ell=1,\ldots,N$.

3 The Proposed Algorithm

3.1 The Algorithm

Let $Z=\{z_1,z_2,...,z_N\}$ be a dataset consisting of N elements from k groups $\{w_1,w_2,...,w_k\}$. The partition matrix of Z is denoted by $U=\left[\mu_{ij}\right]_{k\times N}$, where $\mu_{ij}\in[0,1]$ represents the probability of assigning the jth element to the ith group, with $i=1,2,...,k$. The prototype element v_i of the ith group is determined as follows:

$$v_i=\frac{1}{\sum_{j=1}^{n_i}(\mu_{ij})^2}\sum_{j=1}^{n_i}(\mu_{ij})^2 z_j,\quad i=1,2,...,k, \tag{12}$$

where n_i is the number of elements in the ith group.

Let x_0 be a new object. The proposed algorithm to classify x_0 has the following steps:

- *Step 1* Separate the training set to k groups w_i, $i=1,2,...,k$. Find the prototype element for each group by formula (12), and compute the Euclidean distance between each object in Z and v_i, $i=1,2,...,k$.
- *Step 2* Establish the initial partition matrix $U^{(0)}=[\mu_{ij}^{(0)}]_{k\times(N+1)}$, where the first N columns are extracted from known training data with $\mu_{ij}^{(0)}=1$ if the jth object

belongs to w_i, and $\mu_{ij}^{(0)} = 0$ for the opposite. The $(N+1)$th column is the initial prior probability of x_0. We can choose them by uniform distribution in the first time.

- *Step 3* Update the new partition matrix $U^{(1)}$ by the following principle:

$$\mu_{ij}{}^{(1)} = \frac{1}{\sum_{l=1}^{k} \left(d(v_i^{(0)}, z_j)/d(v_l^{(0)}, z_j) \right)^2}, j = 1, 2, ..., N, \tag{13}$$

where $d(.)$ is the Euclidean distance.

- *Step 4* Compute the $S_1 = \|U^{(1)} - U^{(0)}\| = \max_{ij} \left(|\mu_{ij}^{(1)} - \mu_{ij}^{(0)}| \right)$.
  Repeat Step 3 and Step 4 m times until $S_m = \left\| U^{(m)} - U^{(m-1)} \right\| < \varepsilon$,
  where ε is a really small number chosen arbitrarily. This value measures the different of μ_{ij} through two consecutive iterations. In this article, ε is chosen as 0.0001 for all the numerical examples and applications.
  When Step 4 ends, we obtain the matrix that its last column $(\mu_{i(N+1)}, i = 1, 2, ..., k)$ is the prior probability of x_0.
- *Step 5* Estimate tree copula constructions for d variables in data set D.
- *Step 6* Estimate d marginal densities $f_k, k = \overline{1, d}$.
- *Step 7* Estimate the d-dimensional joint densities through tree copula constructions from Step 5 to obtain $f_i(x), i = 1, 2, ..., k$.
- *Step 8* Classify x_0 to group $w_c, c = 1, 2, ..., k$ if

$$\mu_{c(N+1)}.f_c(x_0) = \max\{\mu_{i(N+1)}.f_i(x_0)\}, i = 1, 2, ..., k.$$

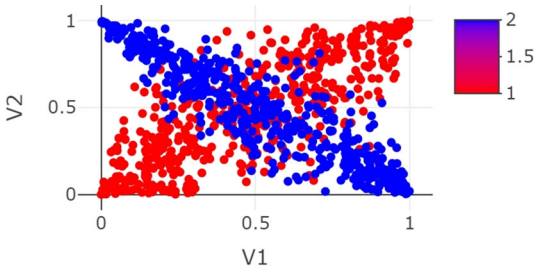
The proposed method consists of four steps for determining the prior probability (Step 1, Step 2, Step 3, and Step 4). Two major approaches have been suggested for determining the prior probabilities in Bayesian classifiers. The first method uses uniform distribution, where each group's prior probabilities are the same. The second method takes into account the information of the training data set and uses the Laplace method $q_i = \frac{n_i+1}{N+n}$ or the ratio of samples method $q_i = \frac{n_i}{N}$, where n_i is the number of elements in w_i, n is the number of dimensions, and N is the number of all objects in the training data [13]. In this study, we propose a new technique for determining the prior probability using fuzzy clustering method [44]. This method considers the new object's and training data's information to give the prior probability. It is expected to contain more information than traditional methods. Steps 5, 6, and 7 of the proposed algorithm estimate pdf with the improved Vine Copula, and the last step is the classification principle.

We have established the Matlab procedure for the proposed algorithm, which can perform automatically and fast for real data.

Table 1 The families and parameters of w_1 and w_2

Population	Family	Parameter
w_1	Gauss	0.8
w_2	Gumbel	4 and rotate of $\pi/2$

Fig. 1 The generated data



3.2 Numerical Example

This section presents two examples to evaluate the performance of the proposed method. In Example 1, we investigate the performance of the proposed process on simulated data, while in Example 2, we test it on four benchmark datasets. To compare the performance of the proposed algorithm with other methods, we use a Bayesian classifier with kernel function (KB) and different prior probabilities (uniform, training set, Laplace), denoted as KB-U, KB-T, and KB-L, respectively. Additionally, we employ several other machines learning algorithms, including linear Supported Vector Machine (LiSVM), Radian basic function Support Vector Machine (RBFSVM), k-Nearest Neighborhoods (k-NN), and Artificial Neural Networks (ANN), as comparative models to provide comprehensive results. Furthermore, we compare the proposed method with Fisher, Tai [5], Thao and Tai [15], Lethikim et al. [25], and Hieu et al. [24] methods. The effectiveness of the methods is based on values including experimental error (EE), F1 Score, and AUC (area under the curve) [45], in which the code for calculating F1 Score and AUC is used for free from https://deepchecks.com/f1-score-accuracy-roc-auc-and-pr-auc-metrics-for-models/.

(a) Example 1

In this example, we apply the proposed algorithm to a simulated dataset consisting of two variables and 1000 observations. The first group (w_1) consists of 500 observations generated from a Gaussian copula with a parameter of 0.8. The second group (w_2) consists of 500 observations generated from a Gumbel copula with $\theta = 4$ and a direction of $\pi/2$. The parameter settings for the two groups and the corresponding generated data are presented in Table 1 and Fig. 1.

The two pdfs estimated by the kernel method (KB) and the proposed method (VC) are illustrated in Figs. 2 and 3.

It can be seen from Figs. 2 and 3 that VC can describe the data in a better way than KB.

Classify by the proposed algorithm and others, and calculate EE, F1 Score and AUC, we obtain Table 2.

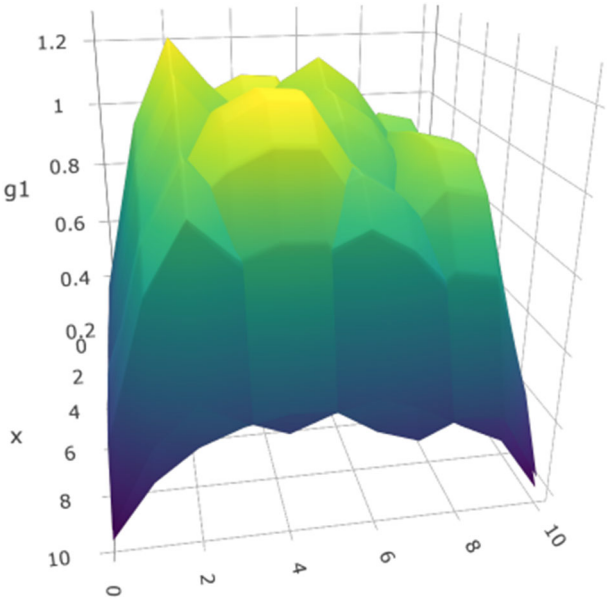


Fig. 2 The pdf estimated by kernel method

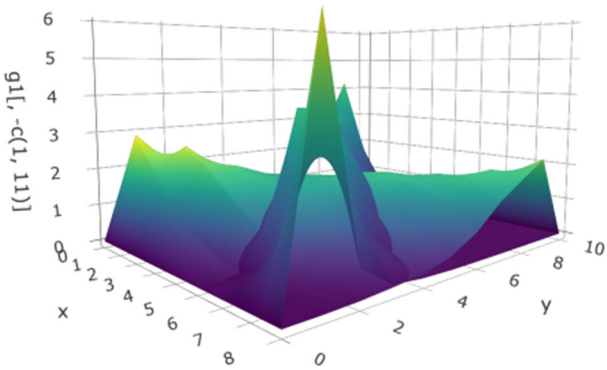


Fig. 3 The pdf estimated by Vine Copula

As shown in Table 2, RBFSVM gives relatively good results with an EE of 0.168, F1 Score of 0.823, and AUC of 0.842. However, the proposed method performs better than RBFSVM with an EE of 0.151, F1 Score of 0.844, and AUC of 0.849. Since the proposed method has the smallest value of EE and the largest values of F1 Score and AUC, it obtains the best result in the application of this data.

(b) Example 2

In this example, the performance of the proposed method is compared to four well-known benchmark datasets: Iris, Pima, Breast Tissue, and User. The Iris flower dataset, also known as Fisher's Iris dataset, is a multivariate dataset introduced by Fisher [8].

Table 2 The EE, F1 Score, AUC of methods in Example 1

Method	EE	F1 score	AUC
KB-U	0.427	0.576	0.573
KB-L	0.409	0.576	0.573
KB-T	0.427	0.576	0.573
LiSVM	0.430	0.478	0.570
RBFSVM	0.168	0.823	0.842
Logistic	0.494	0.511	0.506
Fisher (LDA)	0.494	0.511	0.506
Thao and Tai [15]	0.427	0.576	0.573
Tai [5]	0.401	0.468	0.553
Lethikim et al. [25]	0.210	0.785	0.809
Hieu et al. [24]	0.257	0.767	0.789
k-NN	0.324	0.758	0.760
ANN	0.298	0.745	0.763
Proposed method	0.151	0.844	0.849

It consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and width of the sepals and petals in centimeters. These flowers have significance in the perfume and medicine industries, making it crucial to accurately classify them into their respective groups.

The Pima dataset was originally donated by Vincent Sigillito from the Applied Physics Laboratory at Johns Hopkins University. It was created through a constrained selection process from a larger database maintained by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset comprises female patients of Pima Indian heritage, aged at least 21 years, residing near Phoenix, Arizona, USA. The objective of this dataset is to predict whether a patient would test positive for diabetes based on the World Health Organization criteria, which defines it as having a 2 h post-load plasma glucose level of at least 200 mg/dL. The dataset includes various physiological measurements and medical test results, such as the number of pregnancies, plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure (mm/Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/mL) body mass index (kg/mm), diabetes pedigree function, and age (years). This is a binary classification problem, where Class 1 represents "tested positive for diabetes" and Class 2 represents "tested negative for diabetes." The dataset consists of 500 elements in Class 1 and 268 elements in Class 2.

The Breast Tissue dataset comprises electrical impedance measurements of freshly excised tissue samples from the breast. It consists of nine features, including IO-Impedivity (ohm) at zero frequency, phase angle at 500 KHz, the high-frequency slope of the phase angle, DA-impedance distance between spectral ends, the area under the spectrum, area normalized by DA, maximum of the range, the distance between I0 and the real part of the maximum frequency point, and length of the spectral curve. The

Table 3 General introduction about four bench mark datasets

Data	No of object	No of dimension	No of class
Iris	150	4	3
Pima	768	8	2
Breast	106	9	4
User	403	5	4

Table 4 The EE value of methods in Example 2

Method	Iris	Pima	Breast	User
KB-U	0.033	0.227	0.302	0.116
KB-T	0.033	0.225	0.285	0.120
KB-L	0.033	0.227	0.285	0.120
LiSVM	0.033	0.228	0.216	0.054
RBFSVM	0.027	0.176	0.129	0.054
Fisher (LDA)	0.030	0.219	0.233	0.070
Thao and Tai [15]	0.028	0.225	0.285	0.120
Tai [5]	0.033	0.227	0.285	0.120
Lethikim et al. [25]	0.029	0.217	0.133	0.101
Hieu et al. [24]	0.091	0.199	0.139	0.111
k-NN	0.032	0.265	0.299	0.110
ANN	0.030	0.198	0.138	0.101
Proposed method	0.027	0.125	0.128	0.049

Table 5 The F1 Score value of methods in Example 2

Method	Iris	Pima	Breast	User
KB-U	0.967	0.822	0.724	0.877
KB-T	0.967	0.834	0.736	0.882
KB-L	0.973	0.834	0.736	0.882
LiSVM	0.967	0.834	0.779	0.926
RBFSVM	0.973	0.837	0.862	0.918
Fisher (LDA)	0.970	0.830	0.752	0.905
Thao and Tai [15]	0.970	0.827	0.730	0.882
Tai [5]	0.965	0.822	0.723	0.885
Lethikim et al. [25]	0.969	0.825	0.837	0.875
Hieu et al. [24]	0.879	0.805	0.827	0.860
k-NN	0.969	0.819	0.749	0.863
ANN	0.971	0.826	0.830	0.902
Proposed method	0.973	0.830	0.852	0.920

Table 6 The AUC value of methods in Example 2

Method	Iris	Pima	Breast	User
KB-U	0.983	0.762	0.716	0.965
KB-T	0.983	0.734	0.731	0.959
KB-L	0.983	0.734	0.731	0.959
LiSVM	0.983	0.724	0.790	0.969
RBFSVM	0.987	0.780	0.874	0.964
Fisher (LDA)	0.990	0.734	0.769	0.962
Thao and Tai [15]	0.985	0.734	0.731	0.959
Tai [5]	0.897	0.716	0.731	0.959
Lethikim et al. [25]	0.980	0.732	0.843	0.962
Hieu et al. [24]	0.835	0.752	0.752	0.960
k-NN	0.973	0.728	0.728	0.963
ANN	0.980	0.749	0.754	0.967
Proposed method	0.987	0.772	0.874	0.972

dataset is divided into four classes: car (carcinoma), con (connective), adi (adipose), and the merged class of fad (fibro-adenoma), mas (mastopathy), and gla (glandular).

Finally, the User dataset provides real data on students' knowledge status regarding Electrical DC Machines. All of the datasets mentioned were collected from http://www.is.umk.pl/projects/datasets.html. Table 3 presents a summary of the four datasets.

Comparing the proposed algorithm with others about the EE, F1 Score and AUC, we obtain Tables 4, 5, and 6, respectively.

From Tables 4, 5, and 6, the following results are drawn for the four data sets:

* *Iris:* The values of EE are in the range of [0.027, 0.091], with the proposed method obtaining the smallest value. The F1 Score has values in the [0.967, 0.973] field, where the KB-L and proposed methods give the largest value. As for AUC, its values are in the range of [0.835, 0.987], with the largest value belonging to the RBFSVM and proposed methods.
* *Pima:* The EE has values in the range of [0.125, 0.265], where the proposed method obtains the smallest value. The F1 Score has values in the range of [0.822, 0.837], while the AUC values are in the range of [0.716, 0.780]. Although the RBFSVM method gives the largest value with F1 Score = 0.837 and AUC = 0.780, there is no significant difference with the proposed method, where F1 Score = 0.830 and AUC = 0.772.
* *Breast:* The EE has values in the range of [0.128, 0.302], where the proposed method gives the largest value. The F1 Score has a value range of [0.724, 0.862], where the RBFSVM method obtains the largest value. However, the proposed method has no significant difference (0.862 vs. 0.852). The AUC has values in the range of [0.716, 0.874], and the RBFSVM and proposed methods obtain the largest values.
* *User:* The EE and AUC have values in the range of [0.049, 0.120] and [0.959, 0.972], respectively, where the proposed method obtains the best results in both

Table 7 The used variables and their meaning

Variable	Name	Description
X_1	Ferritin	Ferritin level in blood
X_2	Albumin	Albumin level in blood
X_3	Urea	Urea level in blood
X_4	Time 1	The length of time the patient is on dialysis
X_5	Time 2	The length of time the patient has chronic renal failure
X_6	MCHC	Mean corpuscular hemoglobin concentration
X_7	Creatinine	Creatinine level in blood
X_8	Na	Sodium level in blood
X_9	Ca	Calcium level in blood
X_{10}	Death	1: Will be dead and cannot continue the treatment process
		2: Can continue to live and receive the treatment

Table 8 The EE according to the number of variables (NoV)

Method	NoV 2	NoV 3	NoV 4	NoV 5
KB-U	0.3977	0.3475	0.2664	0.2394
KB-T	0.0849	0.0849	0.0695	0.0656
B-L	0.0849	0.0849	0.0695	0.0656
Proposed method	0.0810	0.0849	0.0734	0.0772
Method	NoV 6	NoV 7	NoV 8	NoV 9
KB-U	0.2278	0.2201	0.2239	0.2010
KB-T	0.0656	0.0656	0.0772	0.0772
KB-L	0.0656	0.0656	0.0772	0.0772
Proposed method	0.0579	0.0502	0.0386	0.0386

values. F1 Score values are in the range of [0.877, 0.926]. Although the proposed method does not give the best effect like the LiSVM method, there is not much difference between the two methods (0.926 vs 0.920).

The four aforementioned datasets are widely recognized datasets with distinct separations between classes. Consequently, the classification results of the methods yield favorable outcomes, and there is minimal variation in the assessed parameters. Nonetheless, the proposed method exhibits advantages over other methods by consistently providing stable values for the evaluated parameters and achieving superior results in the majority of cases. While the RBFSVM method can occasionally rival the proposed method when the data can be linearly separated between classes, its limitations become apparent in practical applications where groups overlap.

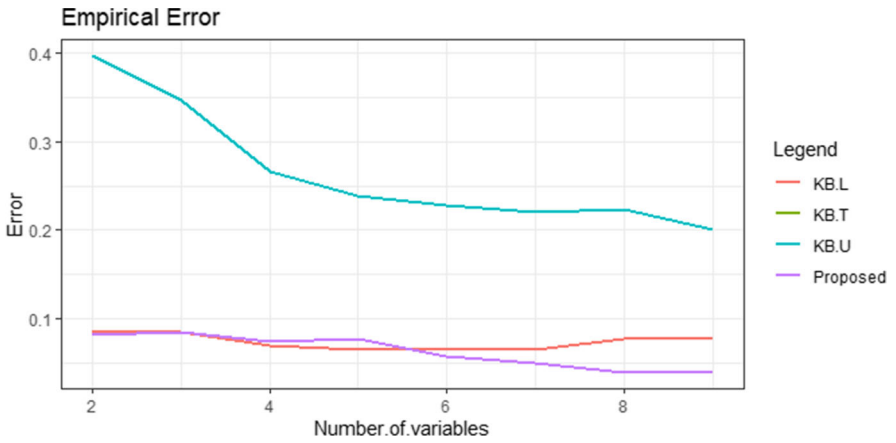


Fig. 4 the values of EE of Bayesian classifiers based on the number of variables

4 Some Applications

4.1 Applying in Medicine from Numerical Data

In this study, we apply the proposed algorithm, along with other methods, to an actual dataset on chronic renal failure. The dataset was collected from a hospital in Can Tho City, Vietnam. It consists of 259 observations (patients) classified into two classes. Class 1 comprises 22 patients who are no longer eligible for treatment, while Class 2 includes 237 patients who can still receive treatment and continue living. The variables chosen based on the expertise of doctors are listed in Table 7.

We initiate the classification process by starting with two variables and gradually incorporating more variables until all nine variables are utilized. The performance metrics, including EE, F1 Score, and AUC values, for each method are presented in Table 8 and Fig. 4.

The analysis of Table 8 and Fig. 4 reveals that KB-U is not effective for classification as it consistently exhibits the highest error rate across all cases. On the other hand, KB-T and KB-L demonstrate relatively lower errors when a small number of variables are employed. However, the performance improvement is not proportional to the number of variables. As depicted in Fig. 4, their error rates increase from the middle of the experiment. These results suggest the presence of the curse of dimensionality, indicating that the interdependencies among variables have not been adequately analyzed using KB-T and KB-L. In contrast, the proposed method initially yields a relatively low error rate that consistently decreases as the number of variables increases. Ultimately, when the entire dataset is utilized, the proposed method achieves the lowest error rate. This outcome highlights the stability and effectiveness of the proposed method when applied to multivariate data with high correlation.

The values of EE, F1 Score and AUC of methods using 9 variables are presented in Table 9.

Table 9 The EE, F1 Score and AUC values of methods in Application 1

Method	EE	F1 Score	AUC
KB-U	0.201	0.880	0.628
KB-T	0.077	0.959	0.628
KB-L	0.077	0.959	0.628
LiSVM	0.084	0.956	0.500
RBFSVM	0.073	0.956	0.568
LDA	0.073	0.959	0.609
Logictic	0.084	0.956	0.500
Thao and Tai [15]	0.204	0.875	0.525
Tai [5]	0.057	0.959	0.617
Lethikim et al. [25]	0.059	0.948	0.597
Hieu et al. [24]	0.054	0.948	0.628
K-NN	0.067	0.927	0.620
ANN	0.057	0.943	0.620
Proposed method	0.039	0.959	0.628

The analysis of Table 9 reveals that the proposed method exhibits the smallest EE and the highest values for F1 Score and AUC. While there may be slight differences in the F1 Score and AUC values among the techniques, some techniques even have the highest F1 Score and AUC values comparable to the proposed method, they lack stability. Taking into account all three evaluation metrics, it is recommended to use the proposed method for classification in this application.

4.2 Applying in Medicine from Image Data

The gray-level co-occurrence matrix (GLCM) is a statistical method used to analyze texture by considering the spatial relationship of pixels. It defines the probability density of pixel pairs with similar or neighboring gray values, thereby reflecting the spatial correlation of texture. The elements of the GLCM are determined using equation (14).

$$p_{d\theta}(i,j) = \#\{\left[(x,y),\left(x+d_x,y+d_y\right)\right] | I((x,y)=i, I(x+d_x,y+d_y)=j\} \quad (14)$$

where

represents taking the number of elements in the following assemble,

$x = 1, 2, ..., N_x$ and $y = 1, 2...N_y$ are pixel element coordinate of a image sized $N_x \times N_y$,

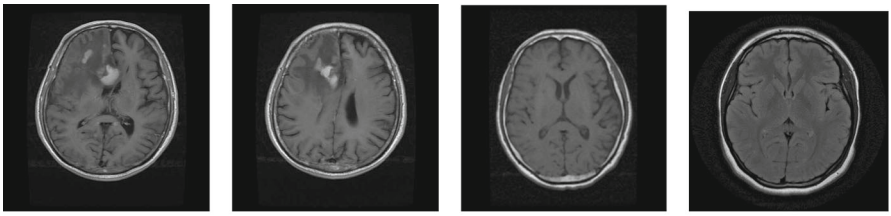
$i, j = 0, 1, ..., L-1$ is the gray value which gray scale is L,

$I(x, y)$ is the value of GLCM at the coordinates (x, y).

d_x and d_y are location displacement calculated by distance d and direction angle θ between matrix pixel pairs. The distance d gets an integer between zero and lower value of N_x and N_y.

Table 10 Four texture features of an image

Features	Variable	Formulas
Energy	X1	$\sum_{i,j} p\,(i,j)^2$
Contrast	X2	$\sum_{i,j} \lvert i-j\rvert^k p^l\,(i,j)$
Uniformity	X3	$\sum_{i,j} \frac{p\,(i,j)}{1+\lvert i-j\rvert}$
Correlation coefficient	X4	$\sum_{i,j} \frac{(i-\mu_i)\left(j-\mu_j\right) p\,(i,j)}{\delta_i \delta_j}$



(a) Benign (b) Benign (c) Malignant (d) Malignant

Fig. 5 The image samples of two groups for Brain Tumors

Table 11 The parameters of the extracted variables for Brain Tumor data

Variable	Mean	Std. Deviation	Variance
X1	0.10	0.13	0.02
X2	0.83	0.17	0.03
X3	0.31	0.14	0.02
X4	0.76	0.16	0.03

Haraclick [46] reported that the GLCM of image textures can have 14 texture features. However, in most subsequent studies, only three or four significant features representing the texture were used. This article employs four image features, namely energy, uniformity, contrast, and correlation coefficient, to characterize each image [29]. The features for the image are presented in Table 10.

where μ_i, μ_j is the mean and standard deviation of the sum of rows and columns in the GLCM matrix, respectively.

This section will apply the proposed method to image data. The dataset consists of 4600 brain tumor images, with 2100 images classified as benign and 2500 as malignant. This dataset is known as the Brain Tumor dataset and is freely available from the website: https://www.kaggle.com. Figure 5 shows sample images from the two groups.

Extracting four image features for all images, and calculating their parameters, we have Table 11.

Table 12 The EE, F1 Score and AUC values of methods in Application 2

Method	EE	F1 Score	AUC
KB-U	0.092	0.942	0.883
KB-T	0.050	0.953	0.924
KB-L	0.050	0.953	0.924
LiSVM	0.321	0.893	0.862
RBFSVM	0.299	0.902	0.863
LDA	0.334	0.789	0.781
Logistic	0.299	0.816	0.802
Thao and Tai [15]	0.074	0.951	0.902
Tai [5]	0.092	0.942	0.883
Lethikim et al. [25]	0.092	0.942	0.883
Hieu et al. [24]	0.065	0.935	0.892
K-NN	0.065	0.942	0.893
ANN	0.085	0.943	0.881
Proposed method	0.012	0.963	0.944

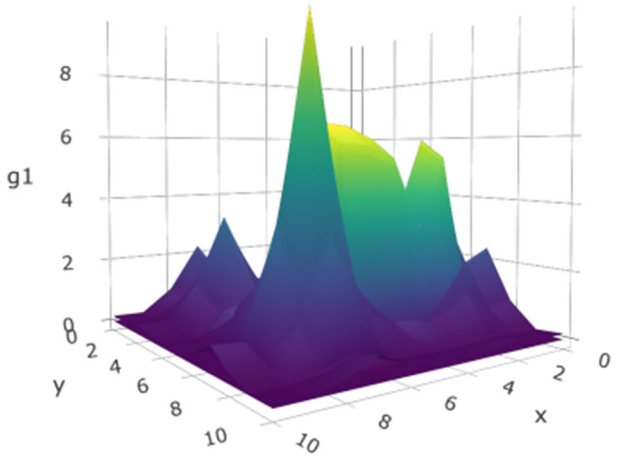


Fig. 6 The estimated pdfs using KB-U

Perform the proposed method and the other methods with four variables, we obtain Table 12:

Table 12 demonstrates that the proposed method outperforms the others with a low value of EE. It also achieves the highest F1 Score and AUC values compared to alternative methods. Classifying image data is a complex problem, and the exceptional results obtained by the proposed approach highlight its significance and potential for practical applications.

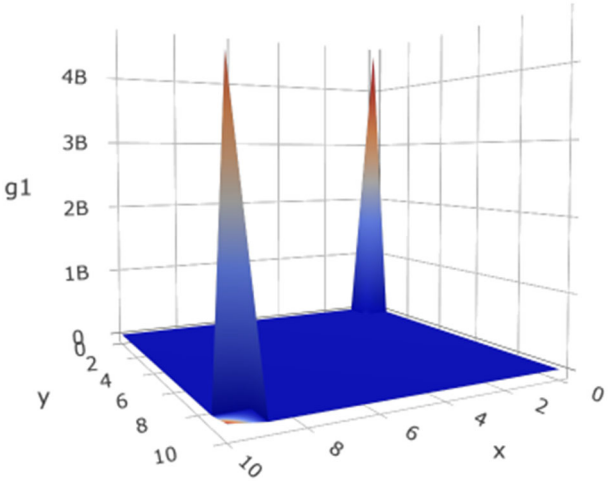


Fig. 7 The estimated pdfs using VC

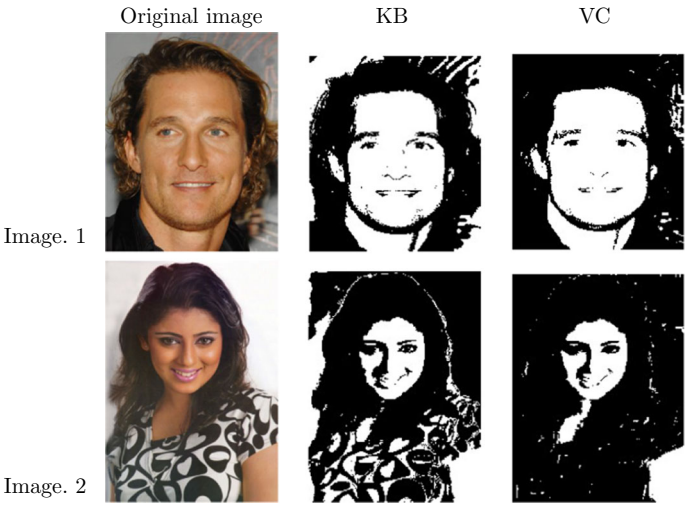


Fig. 8 Application in skin detection

4.3 Applying in Skin Segmentation

Skin segmentation is a technique used to identify human skin regions in an image. It is widely employed in algorithms for face detection, hand gesture analysis, and objectionable image filtering. Consequently, the problem of skin segmentation has gained significant interest in recent years. In this particular study, we explore the suitability of the VC method for skin segmentation. We utilize a benchmark dataset called "Skin" as the training set, where the variables are the Blue (B), Green (G), and Red (R) values at each pixel. The dataset is divided into two classes, labeled as "skin" and "non-skin."

For visualization purposes, we select two variables based on the correlation coefficient matrix derived from the training set. In Figs. 6 and 7, we present the probability density functions (pdfs) estimated by the KB and VC methods. It can be observed that the pdfs estimated by VC demonstrate better separation between the classes compared to those estimated by KB.

We then utilize the VC and KB methods to classify the pixels in two images obtained from the Pratheepan data set [47]. The original test images and the corresponding outcomes are depicted in Fig. 8. It can be observed that the KB method exhibits a high false detection rate. The incorrect detection pixels are evident in the background of Image 1 and the woman's shirt in Image 2. Conversely, the VC method achieves more accurate results at the same pixels and demonstrates relatively good performance compared to the original image. These results affirm the feasibility and applicability of the proposed method in addressing the skin segmentation problem.

The aforementioned applications serve to highlight the advantages of the proposed method. We have also applied the proposed method to other applications and obtained similar results. In our opinion, the proposed method can be considered a suitable solution when working with large and reliable datasets. There are two main reasons to support this claim.

- Other existing methods frequently rely on assumptions that may not be feasible in practical scenarios. For example, Bayesian classifiers and related methods often assume independence among variables. In contrast, the proposed method utilizes a Vine copula structure to account for the dependency property, thereby overcoming the limitations of such assumptions.
- Bayesian classifiers and related methods typically utilize fixed prior probabilities based on either the training dataset or a uniform distribution. However, this approach fails to account for the substantial variation and uncertainty inherent in real-world data. In contrast, the proposed method incorporates a fuzzy clustering approach, enabling the generation of adaptive prior probabilities. These probabilities can vary among observations, accommodating the diverse characteristics of the data and addressing the limitations of fixed prior probabilities.

The proposed method can improve the two above core problem of Bayesian classifier, thereby improving the total performance in several applications.

5 Conclusion

The prior probability and pdf play a significant role in the error rate of Bayesian classification. In order to enhance the performance of this method, the study addresses both of these aspects. Specifically, a method is proposed to determine the prior probability using fuzzy clustering technique. Additionally, the estimation of pdfs incorporates the dependency properties among variables through the Vine Copula structure. By combining these improvements, an effective Bayesian classifier method is developed. Numerical results and real-world applications demonstrate the superiority of the proposed method over alternative approaches. Through the specific steps outlined in the proposed method, especially in comparison with other methods, it has been observed

that the proposed method requires less training data compared to machine learning and deep learning methods. Furthermore, it yields stable results as it is less dependent on parameter tuning during the learning process. Going forward, the plan is to apply the proposed method to various real-world problems in different fields, leveraging its strengths and advantages.

Author Contributions - Ha Che-Ngoc: Build the proposed algorithm.
- Thao Nguyen-Trang: Solve the computation for the proposed algorithm and perform the image examples.
- Hieu Huynh-Van: Perform the numerical examples for the proposed and existing algorithms.
- Tai Vo-Van: Develop the results in theorems and write the manuscript

Funding The authors declare they have no financial interests.

Data Availability The datasets generated and analysed during the current study are not publicly available due the fact that they constitute an excerpt of research in progress but are available from the corresponding author on reasonable request.

Code Availability The codes for computing in this article are established by authors and are available from the corresponding author on reasonable request.

Declarations

Conflict of interest No potential conflict of interest was reported by the author.

Ethical statements: The paper is not currently being considered for publication elsewhere in any form or language.

References

1. Shi Y (2022) Advances in big data analytics: theory, algorithm and practice. Springer, Singapore
2. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4(2):149–178
3. Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York
4. Shi Y, Tian YJ, Kou G, Peng Y, Li JP (2011) Optimization based data mining: theory and applications. Springer, Berlin
5. Tai VV (2017) L^1 - distance and classification problem by Bayesian method. J Appl Stat 44(3):385–401
6. Vovan T, Chengoc H, Ledai N, Nguyentrang T (2022) A new strategy for short-term stock investment using Bayesian approach. Comput Econ 59:887–911
7. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
8. Fisher RA (1938) The statistical utilization of multiple measurements. Ann Eugen 8(4):376–386
9. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W (2018) Applications of support vector machine (svm) learning in cancer genomics. Int J Genomics Proteomics 15(1):41–51
10. Nhu VH, Zandi D, Shahabi H, Chapi K, Shirzadi A, Al-Ansari N, Singh SK, Dou J, Nguyen H (2020) Comparison of support vector machine, Bayesian logistic regression, and alternating decision tree algorithms for shallow landslide susceptibility mapping along a mountainous road in the west of Iran. Appl Sci 10(15):5047
11. Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia M (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Softw 84:240–250
12. Pham BT, Prakash I (2019) Evaluation and comparison of logitboost ensemble, fisher's linear discriminant analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto Int 34(3):316–333

13. Vovan T, Chengoc H, Nguyentrang T (2017) Textural features selection for image classification by Bayesian method. In: 2017 13th international conference on natural computation. fuzzy systems and knowledge discovery (ICNC-FSKD), IEEE, pp 733–139
14. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C (2019) A reliable method for colorectal cancer prediction based on feature selection and support vector machine. Med Biol Eng Comput 57(4):901–912
15. Nguyentrang T, Vovan T (2017) A new approach for determining the prior probabilities in the classification problem by Bayesian method. Adv Data Anal Classif 11(3):629–643
16. Kung JY, Wu CC, Hsu SY, Lee S, Yang CW (2010) Application of logistic regression analysis of home mortgage loan prepayment and default risk. ICIC Express Lett 4(2):325–331
17. Chen Y, Liu C, Chou K, Wang S (2016) Real-time and low-memory multi face detection system design based on naive Bayes classifier using FPGA. In: international automatic control conference (CACS), Berlin pp 7–12
18. Behera DK, Das M, Swetanisha S (2022) Follower link prediction using the XGBoost classification model with multiple graph features. Wirel Pers Commun 127:695–714
19. Gou J, Du L, Zhang Y, Xiong T (2012) A new distance-weighted k-nearest neighbor classifier. J Inf Comput Sci 9(6):1429–1436
20. Imandoust SB, Bolandraftar M (2013) Application of k-nearest neighbor (knn) approach for predicting economic events: theoretical background. Int J Eng Res Appl 3(5):605–610
21. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University, London
22. Neto JG, Ozorio LV, De Abreu TCC, Dos Santos BF, Pradelle F (2021) Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN). Fuel 285:119081
23. Tai VV, Thao NT, Ha CN (2016) The prior probability in classifying two populations by Bayesian method. In: the 1st international conference on applied mathematics in engineering and reliability (Ho Chi Minh City, Vietnam), pp 35–40
24. Hieu HV, Tuan LH, Trong TM, Huy ND, Tai VV (2023) Classifying the lung images for people infected with COVID-19 based on the extracted feature interval. Comput Methods Biomech Biomed Eng Imaging Vis 11(3):856–865
25. Lethikim N, Nguyentrang T, Vovan T (2022) A new image classification method using interval texture feature and improved Bayesian classifier. Multimed Tools Appl 81:36473–36488
26. Sklar M (1959) Fonctions de repartition n dimensions et leurs marges. Univ Paris 8:229–231
27. Qiu H, Hu G, Yang Y, Zhang J, Zhang T (2020) Modeling the risk of extreme value dependence in Chinese regional carbon emission markets. Sustainability 12(19):7911
28. Scheffer M, Weiß GN (2020) Extreme dependence in investor attention and stock returns-consequences for forecasting stock returns and measuring systemic risk. Quant Finance 20(3):425–446
29. Zhang D, Yan M (2018) Financial stress relationships among euro area countries: an R-vine copula approach. Eur J Finance 24:1587–1608
30. Bedford T, Cooke RM (2002) Vines: a new graphical model for dependent random variables. Ann Stat 1:1031–1068
31. Joe H (1996) Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. Lect Notes Ser 28:120–141
32. Pham-Gia T, Turkkan N, Vovan T (2008) Statistical discrimination analysis using the maximum function. Commun Stat Simul Comput 37(2):320–336
33. Bedford T, Cooke RM (2001) Probability density decomposition for conditionally dependent random variables model by vines. Ann Math Artif Intell 32(1):245–268
34. Zhang D, Yan M, Tsopanakis A (2018) Financial stress relationships among euro area countries: an R-vine Copula approach. Eur J Finance 24(17):1587–1608
35. Kurowicka D, Cooke RM (2006) Uncertainty analysis with high dimensional dependence modelling. Wiley, New York
36. Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. Insur Math Econ 44(2):182–198
37. Afifah RH, Noviyanti L, Bachrudin A (2018) Application of selection and estimation regular vine copula on go public company share. J Phys Conf Ser 974:012034
38. Côté MP, Genest C (2015) A copula-based risk aggregation model. Can J Stat 43(1):60–81
39. Dissmann J, Brechmann EC, Czado C, Kurowicka D (2013) Selecting and estimating regular vine copulae and application to financial returns. Comput Stat Data Anal 59:52–69

40. Han D, Tan KS, Weng C (2017) Vine copula models with glm and sparsity. Commun Stat Theory Methods 46(13):6358–6381
41. Mejdoub H, Arab MB (2017) A multivariate analysis for risk capital estimation in insurance industry: vine copulas. Asian Dev Rev 5(2):100–119
42. Mejdoub H, Arab MB (2018) Impact of dependence modeling of non-life insurance risks on capital requirement: D-vine copula approach. Res Int Bus Finance 45:208–218
43. Mensi W, Hammoudeh S, Reboredo JC, Nguyen DK (2015) Are sharia stocks, gold and us treasury hedges and/or safe havens for the oil-based GCC markets. Emerg Mark Rev 24:101–121
44. Phamtoan D, Vovan T (2023) Building fuzzy time series model from unsupervised learning technique and genetic algorithm. Neural Comput Appl 35:7235–7252
45. Powers DMW (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2(1):37–63
46. Haraclick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67:786–804
47. Tan WR, Chan CS, Yogarajah P, Condell J (2011) A fusion approach for efficient human skin detection. IEEE Trans Ind Inform 8(1):138–147