Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Some results of classification problem by Bayesian method and application in credit operation

Tai Vovan

To cite this article: Tai Vovan (2018) Some results of classification problem by Bayesian method and application in credit operation, Statistical Theory and Related Fields, 2:2, 150-157, DOI: 10.1080/24754269.2018.1528420

To link to this article: https://doi.org/10.1080/24754269.2018.1528420

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 150–157
https://doi.org/10.1080/24754269.2018.1528420

Taylor & Francis
Taylor & Francis Group

Some results of classification problem by Bayesian method and application in credit operation

Tai Vovan

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

ABSTRACT

This study proposes some results in classifying by Bayesian method. There are upper and lower bounds of the Bayes error as well as its determination in case of one dimension and multi-dimensions. Based on the proposals for estimating of probability density functions, calculating the Bayes error and determining the prior probability, we establish an algorithm to evaluate ability of customers to pay debts at banks. This algorithm has been performed by the Matlab procedure that can be applied well with real data. The proposed algorithm is tested by the real application at a bank in Viet Nam that obtains the best results in comparing with the existing approaches.

ARTICLE HISTORY
Received 30 November 2017
Revised 16 September 2018
Accepted 22 September 2018

KEYWORDS
Bayesian method; classification; error; credit operation; prior probability

1. Introduction

The classification problem is one of the main subdomains of discriminant analysis and has relation with many fields. Classification is to assign an element to the appropriate population based on the observed variables. It is an important development direction of multivariate statistics and has been applied in many different fields such as medicine and economics. Recently, this problem is interested by many statisticians in both theory and application (Miller, Inkret, & Little, 2001; Nguyen-Trang & Vo-Van, 2017; Pham-Gia, Nhat, & Phong, 2015; Tai, Thao, & Ha, 2016). There are many methods for classifying such as Fisher, logistic regression, Bayes and the machine learning algorithms (Naive–Bayes (NB), Supported Vector Machine (SVM), k-Nearest Neighbour, etc.). For Fisher method, we have to assume the equality of the variance matrices of groups for implementing. This is the drawback of Fisher method in real applications (Fisher, 1936; Marta, 2001). When constructing a logistic regression model, we must constrain on the data conditions that are difficult to satisfy in reality (James, 2001; Jan, Cheng, & Shih, 2010), so it is not suitable for many applications. According to many literatures (Altman, 1991; Hastie & Tibshirani, 1996), the algorithms in machine learning have the following major disadvantages: (i) Error diagnosis and correction: One notable limitation of machine learning is its susceptibility to errors. The actual problem with this inevitable fact is that when they do make errors, diagnosing and correcting them can be difficult because it will require going through the underlying complexities of the algorithms and associated processes, (ii) Time constraints in learning: It is impossible to make immediate accurate predictions with a machine learning system. Remember that it learns through historical data. The bigger the data and the longer it is exposed to these data, the better it will perform, (iii) Problems with verification: Another limitation of machine learning is the lack of variability. Machine learning deals with statistical truths rather than literal truths. In situations that are not included in the historical data, it will be difficult to prove with complete certainty that the predictions made by a machine learning system is suitable in all scenarios, and (iv) Limitations of predictions: Unlike humans, computers are not good story tellers. Machine learning systems know more what they can tell humans. Thus, they cannot always provide rational reasons for a particular prediction or decision. Bayesian method bases the distribution of data, the prior probability and the relation between the classified element with groups to perform. It does not require much historical data as the algorithms in machine learning because it use the prior probabilities in classifying. This method does not also need normal condition for data and can classify for two and more populations. As a result, it has many advantages in classifying (Tai, 2017).

Given k populations w_i with $f_i(x)$ and q_i are the probability density function (pdf) and the prior probability of w_i, respectively. Pham–Gia, Turkkan, and Tai (2008) have used the maximum function of pdfs as a tool to study about Bayesian method

CONTACT Tai Vovan vvtai@ctu.edu.vn College of Natural Science, Can Tho University, Can Tho, Vietnam

and given the important results. Classification principle and Bayes error were established based on the $g_{\max}(x) = \max\{q_1 f_1(x), q_2 f_2(x), \ldots, q_k f_k(x)\}$. The upper and lower bounds of the Bayes error and its relationship with the L^1- distance of the pdfs as well as with the overlap coefficient of the pdfs were also built. The function $g_{\max}(x)$ has a very important role in the classification problem by Bayesian method, so the authors have continued to study on it. Using the Matlab software, Pham-Gia et al. (2015) have given the function $g_{\max}(x)$ of two bivariate normal pdfs. With the given specific parameters of two densities, this method can determine the regions of $g_{\max}(x)$ in R^2 and their boundaries (straight lines, ellipses, parabolas or hyperbolas). However, it can not perform for non-normal distributions. With the similar development, Tai (2017) has proposed the L^1- distance of the $\{q_i f_i(x)\}$ and established its relationship with Bayes error. This distance also was used to calculate Bayes error and to classify a new element. However, we see that the relevant quantities to Bayesian approach have not been surveyed completely yet.

Bayesian method has many advantages, however according to our knowledge, the level for application of this method in practice is less than others. We can find many applications in bank and medicine using Fisher method, logistic method and the algorithms of machine learning model (Altman, 1991; Christopher, 2006; Cristianini & Shawe, 2000; Jan et al., 2010; Marta, 2001). Recently, all statistics software can solve effectively and quickly with big and multivariate data in classifying for above methods, while the Bayesian method does not have this advantage. The cause of this problem is the ambiguity in determining prior probability, estimating pdfs and complex problem in calculating Bayes error. Although all these issues have been discussed by many authors, the optimal methods have not been still found yet (Tai, 2017). In this article, we propose specific approaches to perform all above mentioned problems. From these results, we establish an complete algorithm for evaluating the ability of customers to pay debts at banks from their information. The proposed algorithm is applied for customers of Vietcom bank in Viet Nam. This application gives advantages in comparing to existing approaches. The proposed algorithm can be applied for other domains.

The next section of the article is structured as follows. Section 2 presents the classification principle and the Bayes error. Determining the Bayes error and some its results, finding the function $g_{\max}(x)$ in case of one-dimension and multi-dimensions to calculate Bayes error and to classify a new element are also performed in this section. Section 3 proposes an algorithm to evaluate ability of customers to pay debts at banks and solve calculable problem in applying practice of this algorithm. This section also compares the proposed algorithm with existing ones by many numerical examples. Section 4 applies the proposed algorithm for real data at a bank in Viet Nam. The final section is destined for conclusion of the paper.

2. Classification principle and Bayes error

2.1. Classification principle

Given k populations $w_1, w_2, \ldots, w_k$ with q_i and $f_i(x), i = 1, 2, \ldots, k$ are the prior probability and pdf of w_i, respectively. According to Pham–Gia et al. (2008), an observation x_0 is assigned to population w_i if

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

where $g_i(x) = q_i f_i(x)$, $g_{\max}(x) = \max\{q_1 f_1(x), q_2 f_2(x), \ldots, q_k f_k(x)\}$.

Misclassification probability of this method is called Bayes error. It is given by (2):

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

where

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

From (2), we can prove the following result:

$$Pe^{(q)}_{1,2,\ldots,k} = 1 - \int_{R^n} g_{\max}(x) \, \mathrm{d}x. \tag{3}$$

The correct probability is determined by (4).

$$Ce^{(q)}_{1,2,\ldots,k} = 1 - Pe^{(q)}_{1,2,\ldots,k}. \tag{4}$$

2.2. Determining Bayes error

Theorem 2.1: *Let $f_i(x), i = 1, 2, \ldots, k, k \geq 3$ be k pdfs defined on R^n and let $q_i \in (0; 1)$,*

$$R_1^n = \left\{x \in R^n : q_1 f_1(x) > q_j f_j(x), 2 \leq j \leq k\right\},$$
$$R_k^n = \left\{x \in R^n : q_k f_k(x) > q_j f_j(x), 1 \leq j \leq k\right\},$$
$$R_l^n = \{x \in R^n : q_i f_i(x) > q_l f_l(x), 1 \leq i \leq k,$$
$$2 \leq l \leq k-1, i \neq l\}. \tag{5}$$

The Bayes error is determined by

$$Pe^{(q)}_{1,2,\ldots,k} = 1 - \int_{R_1^n} q_1 f_1(x) \, \mathrm{d}x - \sum_{l=2}^{k-1} \int_{R_l^n} q_l f_l(x) \, \mathrm{d}x$$
$$- \int_{R_k^n} q_k f_k(x) \, \mathrm{d}x. \tag{6}$$

Proof: See Appendix 1. ■

2.3. Bounds of Bayes error

Theorem 2.2: *Let $f_i(x), i = 1, 2, \ldots, k, k \geq 2$ be k pdfs defined on R^n. We have bounds of Bayes error as well as its relationships with other measures as follows:*

(i)

$$Pe^{(q)}_{1,2,\ldots,k} \leq 1 - \frac{1}{k-1} \times \left(1 - \prod_{j=1}^{k} q_j^{\alpha_j} D_T(f_1, f_2, \ldots, f_k)^{\alpha}\right), \quad (7)$$

(ii)

$$Pe^{(q)}_{1,2,\ldots,k} \leq \sum_{i<j} q_i^{\beta} q_j^{1-\beta} D_T(f_i, f_j)^{(\beta,1-\beta)}, \quad (8)$$

(iii)

$$\frac{1}{k}[(k-1) - \sum_{i<j} \|g_i, g_j\|_1] \leq Pe^{(q)}_{1,2,\ldots,k} \leq 1 - \frac{1}{2}\max_{i<j}\{\|g_i, g_j\|_1\} - \min_i\{q_i\}, \quad (9)$$

(iv)

$$0 \leq Pe^{(q)}_{1,2,\ldots,k} \leq \max_i\{q_i\}, \quad (10)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k); \alpha_j, \beta \in (0, 1), \sum_{j=1}^{k} \alpha_j = 1, i, j = 1, 2, \ldots, k$ $D_T(f_1, f_2, \ldots, f_k)^{\alpha} = \int_{R^n} \prod_{j=1}^{k} [f_j(x)]^{\alpha_j} dx$ *is affinity of Toussaint* (1972).

Proof: See Appendix 2. ■

From (7), with $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1/k$, we have the relationship between Bayes error and affinity of Matusita (1967). Especially, when $k = 2$, we have the relationship between $Pe^{(q,1-q)}_{1,2}$ and Hellinger distance.

In addition, we also have the relation between Bayes error and overlap coefficients as well as L^1- distance of $\{g_1(x), g_2(x), \ldots, g_k(x)\}$ (see Tai, 2017). For special case: $q_1 = q_2 = \cdots = q_k = 1/k$, the authors in Pham–Gia et al. (2008) had established expressions about relations between Bayes error and L^1- distance of $f_1(x), f_2(x), \ldots, f_k(x), Pe^{(1/k)}_{1,2,\ldots,k}$ and $Pe^{(1/(k+1))}_{1,2,\ldots,k+1}$.

2.4. Maximum function

To classify a new element by (1) and to determine Bayes error by (3), we must find the $g_{\max}(x)$. Some authors such as Pham-Gia et al. (2015) and Tai (2017) have surveyed relationships between $g_{\max}(x)$ with some related quantities of classification problem. The specific expressions for $g_{\max}(x)$ in some special cases have been found. However, the general expression for all of cases is a complex problem that has not been still found yet.

Given k pdfs $f_i(x)$ and $q_i, i = 1, 2, \ldots, k, q_1 + q_2 + \cdots + q_n = 1$ and let $g_i(x) = q_i f_i(x), g_{\max}(x) = \max\{g_i(x)\}$. The maximum function $(g_{\max}(x))$ is determined in the following two cases:

(i) *For one-dimension*

In this case, we can find $g_{\max}(x)$ by the following algorithm:

Step 1. Solve the equations $g_i(x) - g_j(x) = 0, i = 1, 2, \ldots, k-1, j = i+1, \cdots, k$, to find all roots.

Step 2. With root x_{lm} of equation $g_l(x) - g_m(x) = 0$, compare value $g_l(x_{lm})$ with all the values of $g_j(x_{lm}), j \neq l, m$. If there exists $p \neq l, m$ such that $g_p(x_{lm}) > g_l(x_{lm})$ then we delete x_{lm} and keep x_{lm} for otherwise. Arrange the kept roots in order from small to large, then we have the set $B = \{x_1, x_2, \ldots, x_h\}$.

Step 3. Given $i = 1, 2, \ldots, k; j = 1, 2, \ldots, h$, $g^{(q)}_{\max}(x)$ is determined by the following principles:

If $\max\{g_1(x_1 - \varepsilon_1), g_2(x_1 - \varepsilon_1), \ldots, g_k(x_1 - \varepsilon_1)\} = g_i(x_1 - \varepsilon_1)$ then $g^{(q)}_{\max}(x) = g_i(x)$ for $x \in (-\infty, x_1)$.

If $\max\{g_1(x_j + \varepsilon_2), g_2(x_j + \varepsilon_2), \ldots, g_k(x_j + \varepsilon_2)\} = g_i(x_j + \varepsilon_2) j = 1, 2, \ldots, h-1$ then $g^{(q)}_{\max}(x) = g_i(x)$ for $x \in (x_i, x_{i+1})$.

If $\max\{g_1(x_h - \varepsilon_3), g_2(x_h - \varepsilon_3), \ldots, g_k(x_h - \varepsilon_3)\} = g_i(x_h - \varepsilon_3)$ then $g^{(q)}_{\max}(x) = g_i(x)$ for $x \in (x_h, +\infty)$.

In the above algorithm, ε_1, ε_2, ε_3 are the positive constants such that: $x_1 + \varepsilon_1 < x_2, x_h - \varepsilon_3 > x_{h-1}, x_i - \varepsilon_2 < x_{i-1}, x_i + \varepsilon_2 < x_{i+1}$.

From this algorithm, we have written Matlab procedure to find the $g_{\max}(x)$. When $g_{\max}(x)$ is determined, we will easily calculate Bayes error by (3), as well as classify a new element by (1).

(ii) *For multi-dimensions*

In case of multi-dimensions, it should be very complicated to obtain the closed expression for $g_{\max}(x)$. The difficulty comes from the various forms of the intersection space curves between the surfaces of pdfs. This problem has been interested by the authors in Ghosh (2006), Pham–Gia et al. (2008), Pham-Gia et al. (2015), and Tai (2017). The authors in Pham-Gia et al. (2015) have attempted finding the function $g_{\max}(x)$, however it has been only established for some cases of bivariate normal distribution.

Here, we do not find the expression of $g_{\max}(x)$. We compute Bayes error instead by taking integration of $g_{\max}(x)$ by quasi Monte-Carlo method. An algorithm for doing calculations has been constructed, and a corresponding Matlab procedure is also established.

3. The proposed algorithm in evaluating ability of customers to pay debts

3.1. The proposed algorithm

Based on the Bayesian method, we propose an algorithm to evaluate the ability to repay debt bank of

customers. In bank credit operations, determining the repayment ability of customers is really important. If the lending is too easy, the bank may have bad debt problem. In contrast, the bank will miss a good business. Therefore, this problem is interested of many statisticians and managers.

Given N customers divided to k groups $w_i, i = 1, 2, \ldots, k$. Each customer is considered by n variables. $Z = [z_{ij}]_{n\times N}$ is data set of all customers and x_0 is a new customer that we need to classify. We propose an algorithm to classify x_0 (PAC) as follows:

Step 1: Determine variables that have statistical significance to influence to the ability to repay bank debt of customers.
Setp 2: Find the prototype element v_i for each group by (11):

$$v_i = \left(\sum_j^N {\mu_{ij}}^2 z_j\right) / \left(\sum_j^N {\mu_{ij}}^2\right), \qquad (11)$$

where $i = 1, 2, \ldots, k$, μ_{ij} is the probability of jth element assigned to w_i and z_j is the coordinate of the jth element.
Step 3: Establish the initial partition matrix $U^{(0)} = [\mu_{ij}]_{k\times(N+1)}$, where the first N columns is extracted from known training data with $\mu_{ij} = 1$ if the jth element belongs to the w_i and $\mu_{ij} = 0$ for otherwise. The $(N + 1)$th column is the initial prior probabilities of x_0. We can choose them by uniform distribution.
Step 4: Update the new partition matrix $U^{(1)}$ by the following principle:

$${\mu_{ij}}^{(1)} = \frac{1}{\sum_{l=1}^{k} \left(d_{ij}/d_{lj}\right)^2} \qquad (12)$$

if $d_{ij} > 0$ for $i = 1, 2, \ldots, k$ and for $\mu_{ij}^{(1)} = 0$ for otherwise (d_{ij} is the distance from z_j to v_i).
Step 5: Compute the $\max_{ij}(|\mu_{ij}^{(1)} - \mu_{ij}^{(0)}|)$.
Repeat Step 2, Step 3 and Step 4 until $\max_{ij}(|\mu_{ij}^{(n)} - \mu_{ij}^{(n-1)}|) < \varepsilon$.
Step 6: Estimate pdf f_i for the group w_i and compute $q_i f_i(x_0)$, where $q_i = \mu_{i(N+1)}^{(n)}$, $i = 1, 2, \ldots, k$.
Step 7: If $\max_{1\le i\le k}\{q_i f_i(x_0)\} = q_m f_m(x_0), m = 1, 2, \ldots, k$ then x_0 is assigned to w_m with Bayes error is determined by (3).

The proposed algorithm has two phases to perform. Phase 1 determines the prior probabilities (Step 1 to Step 5) and Phase 2 classifies a new element with specific Bayes error (Step 6 and Step 7). Phase 1 is an important contribution of the proposed algorithm and established based on the fuzzy relation between the classified element and the populations. This phase finishes when the probability of two consecutive iterations is almost the same. Thus, the number of iterations depends on each data set. Computing the Bayes error is complexity of Phase 2. For one dimension, first, we find the $g_{\max}(x)$ by the proposed algorithm in Subsection 2.4 and, and then, compute Bayes error by (2). For multi-dimensions, Bayes error is approximated by quasi Monte-Carlo method.

3.2. Some other related problems of the proposed algorithm

In above algorithm, we need to pay attention to some following problems:

(i) ε is a really small positive number chosen arbitrarily. The smaller it is, the more iterations and the cost time are taken. In this article, we choose $\varepsilon = 0.0001$.
(ii) d_{ij} is the distance from object z_j to the prototype v_i. There are many distances between two elements summarised in Webb (2002). In this paper, we use the L^1- distance (see Pham–Gia et al., 2008) for applications.
(iii) To determine variables having statistical significance of Step 1, we use logistic regression model to perform.
(iv) Normally, in case of non-information, we choose prior probabilities by uniform distribution. Based on the training set, the prior probabilities are often estimated by Laplace method: $q_i = (n_i + n/k)/(N + n)$ and the ratio of sample one: $q_i = n_i/N$, where n_i and N are the number of elements in ith group and training set, respectively, n is the number of dimensions and k is the number of groups. The above mentioned approaches have been studied and applied by many authors, such as, (McLachlan & Basford, 1998; Nguyen-Trang & Vo-Van, 2017; Tai, 2017; Tai et al., 2016). Five steps of the proposed algorithm (Step 1, Step 2, Step 3, Step 4 and Step 5) determine the prior probability for x_0. If the algorithm stops at the fifth step, we will get the matrix of size $k \times (N + 1)$, in which the last column is the prior probability of x_0. Thus, in this algorithm, we have combined the sample data set and classified elements to determine the prior probability. Hence, it contains more information than the ratio of sample and Laplace methods that only depend on training data. Anyway, these prior probabilities which consider the relations between the classified object and all of populations may be more suitable than traditional methods that only base on training set.
(v) There are many parameter and nonparameter methods to estimate pdfs of Step 6. In the examples and applications of this article, we use the kernel function method, a popular one applied in reality nowadays (Inman & Bradley, 1989; Nguyentrang

& Vovan, 2017; Tai, 2017; Tai & Pham-Gia, 2010; Tai et al., 2016).

3.3. Numerical examples for comparison

In this section, three well-known data sets which comprise Pima, Breast Tissue and User are used to test the performance of the proposed method. Pima data was originally donated by Vincent Sigillito, Applied Physics Laboratory, Johns Hopkins University and was constructed by constrained selection from a larger database held by the National Institute of Diabetes and Digestive and Kidney Diseases. All patients represented in this data were females at least 21 years old of Pima Indian heritage living near Phoenix, Arizona, USA. The problem posed here is to predict whether a patient would test positive for diabetes according to World Health Organization criteria (i.e. if the patients 2 hour post load plasma glucose is at least 200 mg/dl.) given a number of physiological measurements and medical test results. The attribute details includes number of times pregnant, 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/m), diabetes pedigree function, age (years). This is a two class problem with class value 1 being interpreted as 'tested positive for diabetes '. There are 500 examples of Class 1 and 268 of Class 2. The Breast Tissue is the data set with electrical impedance measurements of freshly excised tissue samples from the breast. It includes nine features, such as IO-Impedivity (ohm) at zero frequency, phase angle at 500 KHz, high-frequency slope of phase angle, DA-impedance distance between spectral ends, area under spectrum, area normalised by DA, maximum of the spectrum, distance between IO and real part of the maximum frequency point, length of the spectral curve. All observations in this data are divided into four classes: car (carcinoma), con (connective), adi (adipose) and the merged class of fad (fibro-adenoma), mas (mastopathy), gla (glandular). The last data is the real one about the students' knowledge status regarding the subject of Electrical DC Machines. All of considered data sets are collected from www.is.umk.pl/projects/datasets.html. The summary of three data sets is presented in Table 1.

For each data set, we conducted the experiments 10 times and use 30% of objects as the test set at each time, randomly. In addition, the results of the proposed algorithm are compared with Fisher method, logistic method and some machine learning algorithms such as NB, SVM, Radian basic function support vector machine (RBFSVM), linear discriminant analysis (LDA), k-nearest neighbour with $k=1$ (1-NN) and $k=3$ (3-NN). With the considered data sets that have the difference about the features, the number of dimensions and groups, this comparison is very meaningful to evaluate the advantages of the proposed algorithm.

As presented in Table 2, the proposed algorithm provides the best results for all three data sets.

Table 1. Summary of three bench mark data sets.

Data	No of objects	No of dimensions	No of groups
Pima	768	8	2
Breast	106	9	4
User	403	5	4

Table 2. The empirical error of the proposed method and others.

Methods	Pima	Breast	User
SVM	0.2435	0.3688	0.4240
RBFSVM	0.2604	0.2687	0.1926
LDA	0.2335	0.3312	0.3140
1-NN	0.3000	0.2375	0.2545
3-NN	0.2704	0.2625	0.2405
NB	0.2439	0.3625	0.2256
Logistic	0.2319	0.3325	0.2156
Fisher	0.2913	0.4121	0.3210
Proposed algorithm	**0.2226**	**0.2309**	**0.1793**

Note: Bold values highlights the results of the proposed method.

4. Application in credit operation

In this section, we classify customers at Vietcom bank in Viet Nam to illustrate for application of the proposed algorithm. In this article, Bayesian method with prior probabilities calculated by uniform distribution, ratio of samples, Laplace method and proposed algorithm are called BayesU, BayesR, BayesL and BayesC, respectively.

The considered custormers in this application are companies in Can Tho city (CTC), Viet Nam. We collect a data set on 214 enterprises operating in key sectors as agriculture, industry, commerce, including 143 cases of good debt (G) and 71 cases of bad debt (B). Data is provided by responsible organisations of CTC and studied in Tai (2017). Each company is evaluated by 13 independent variables in the expert opinion. The specific variables are given in Table 3.

In this application, the article will use random 70% of the data size (100 elements belong to group G and 50

Table 3. The surveyed independent variables.

Variables	Detail
Financial leverage (X1)	Total debt/total equity
Reinvestment (X2)	Total debt/total equity
Roe (X3)	Net profit/equity
Interest (X4)	(Net income + depreciation)/total assets
Floating capital (X5)	(assets – liabilities)/total assets
Liquidity (X6)	(Cash + Short-term investments)/liabilities
Profits (X7)	Net profit/total assets
Ability (X8)	Net sales/Total assets
Size (X9)	Logarithm of total assets
Experience (X10)	Years in business activity
Agriculture (X11)	Agricultural and forestry sector
Industry (X12)	Industry and construction
Commerce (X13)	Trade and services)

elements belong to group B) as the training set to determine variables which have significance, to estimate pdfs and to find suitable model for classification problem by Bayesian method. 30% of the remaining data will be used as the validation set (43 elements belong to group G and 21 elements belong to group B). The result of Bayesian method is also compared to others.

To assess the effect of the independent variables to the solvency of the companies, we have built the logistic regression model $\log(p/1-p)$ with the independent variables Xi, $i = 1, 2, \ldots, 13$ (p is the probability of repaying bank debt of companies). The analytical results are summarised in Table 4.

Table 4 only shows three variables X1, X4 and X7 have statistical significance at 10% level, so we use three variables to classify. Performing with BayesU, BayesR, BayesL and BayesC, we have the results given in Table 5.

Table 5 shows that the correct probability of BayesC with three variables $X1, X4, X7$ gives the largest value. Comparing this result with that of existing some other methods, we obtain Table 6.

Table 6 shows that BayesC also gives the highest result in comparing with existing method for 1 variable, 2 variables and 3 variables. Using the best model for each case of methods from Table 6 to classify the test set (67 elements), we obtain Table 7.

Once again with test data, BayesC also gives the best result in Table 7.

Table 4. The results of logistic regression model.

Xi	Regression coefficients	Significance level
X1	−2.444	0.003
X2	6.692	0.244
X3	2.566	0.244
X4	2.052	0.034
X5	0.478	0.700
X6	0.340	0.860
X7	4.921	0.093
X8	0.044	0.621
X9	−0.329	0.442
X10	−0.136	0.910
X11	0.009	0.994
X12	−0.007	0.886
X13	2.122	0.360

Table 5. The correct probability (%) in classifying RBD from training set.

Variables	BayesU	BayesR	BayesL	BayesC
X1	86.21	86.14	84.13	87.13
X4	81.12	82.91	86.16	88.19
X7	83.21	84.63	83.14	84.52
X1, X4	87.25	88.72	87.19	89.06
X1, X7	88.16	88.34	83.26	89.56
X4, X7	89.25	89.04	89.02	91.34
X1, X4, X7	91.15	91.53	90.17	93.18

Table 6. The correct probability (%) for optimal models of training set.

Methods	One variable	Two variables	Three variables
Logistic	84.04	88.29	88.69
Fisher	84.73	80.73	79.32
SVM	82.34	82.03	83.07
BayesC	88.19	91.34	93.18

Table 7. Compare the correct probability (%) of test set.

Methods	Correct numbers	False numbers	Correct probability
Logistic	53	11	82.81
Fisher	52	12	81.25
SVM	53	11	82.81
BayesC	57	7	89.06

5. Conclusion

The article has considered completely the classification problem by Bayesian method. Bayes error for one-dimension and multi-dimensions are surveyed in theory and application. The relationships between the Bayes error with affinity of Toussaint are also established. Surveying the function $g_{\max}(x)$ not only adds tool to find Bayes error, classifying the new element but also is the visual illustration for the classification problem. Based on the determinant prior probability, classification principle, computation the Bayes error, we have proposed a new algorithm to evaluate ability of customers to pay debts at banks. This algorithm has been performed by the Matlab procedure that can be applied well with real data. The proposed algorithm is compared with existing algorithms by many benchmark data sets. They show that the proposed algorithm is more advantage than existing approaches. We have also applied the proposed algorithm for customers at Vietcom bank in Viet Nam. This example shows potentiality in real application of the researched problem. In the coming time, we continue to use it to survey the other problems.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Tai Vovan received the Ph.D. degree in theory of probability and statistical mathematics in 2011. He has worked in Can Tho University, Viet Nam, since 1997. His research interests include statistical pattern recognition (classification problem and cluster analysis) and fuzzy time series and their applications in data mining. He has published over 20 papers about these subjects.

References

Altman, D. G. (1991). Statistics in medical journals: Development in 1980s. *Statistical in Medicine*, *10*, 1897–1913.

Christopher, M. B. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Cristianini, N., & Shawe, T. J. (2000). *An introduction to support vector machines and other kernel-based learning method*. London: Cambridge University .

Fisher, R. A. (1936). The statistical utilization of multiple measurements. *Annals of Eugenic*, *7*, 376–386.

Ghosh, A. K. (2006). Classification using kernel density estimates. *Technometrics*, *48*, 120–132.

Hastie, T., & Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(6), 607–616.
Inman, H. F., & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distribution sand point estimation of the overlap of two normal densities. *Communication in Statistics Theory and Methods*, *18*, 3851–3874.
James, J. (2001). *Interaction effects in logistic regression*. London: Sage.
Jan, Y. K., Cheng, C. W, & Shih, Y. H. (2010). Application of logistic regression analysis of home mortgage loan prepayment and default risk. *ICIC Express Letters*, *2*, 325–331.
Marta, E. (2001). Application of Fisher's method to materials that only release water at high temperatures. *Portugaliae Etecfochlmlca Acta*, *15*, 301–311.
Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, *19*(1), 181–192.
McLachlan, G. J., & Basford, K. E. (1998). *Mixture models: Inference and applications to clustering*. New York, NY: Marcel Dekker.
Miller, G., Inkret, W. C., Little, T. T., Martz, H. F., & Schillaci, M. E. (2001). Bayesian prior probability distributions for internal dosimetry. *Radiation Protection Dosimetry*, *94*, 347–352.
Nguyentrang, T., & Vovan, T. (2017). Fuzzy clustering of probability density functions. *Journal of Applied Statistics*, *44*(4), 583–601.
Nguyen-Trang, T., & Vo-Van, T. (2017). A new approach for determining the prior probabilities in the classification problem by Bayesian method. *Advances in Data Analysis and Classification*, *11*, 629–643.
Pham-Gia, T., Nhat, N. D., & Phong, N. V. (2015). Statistical classification using the maximum function. *Open Journal of Applied Statistics*, *5*(7), 665–679.
Pham–Gia, T., Turkkan, N., & Tai, V. V. (2008). Statistical discrimination analysis using the maximum function. *Communications in Statistics Simulation and Computation*, *37*, 320–336.
Tai, V. V. (2017). L^1- distance and classification problem by Bayesian method. *Journal of Applied Statistics*, *4*(3), 385–401.
Tai, V. V., & Pham-Gia, T. (2010). Clustering probability distributions. *Journal of Applied Statistics*, *37*(11), 1891–1910.
Tai, V. V., Thao, N. T., & Ha, C. N. (2016). The prior probability in classifying two populations by Bayesian method. *Applied Mathematics Engineering and Reliability*, *6*, 35–40.
Toussaint, G. T. (1972). Some inequalities between distance measures for feature. *IEEE Transactions on Computers*, *C-21*, 409–410.
Webb, A. (2002). *Statistical pattern recognition*. London: John Wiley & Sons.

Appendices

Appendix 1. Proof of Theorem 2.1

To obtain (6), we need to prove the following two results:

$$R_i^n \cap R_j^n = \phi, \quad (1 \le i \ne j \le k)$$

and

$$\bigcup_{i=1}^{k} R_i^n = R_1^n \cup \left(\bigcup_{i=2}^{k-1} R_i^n\right) \cup R_k^n = R^n, \quad f_{\max}(x) = f_i(x), \quad \forall x \in R_i^n.$$

Let $\overline{A} = R^n \backslash A$, we have

$$\overline{R}_{ij} = \{x \in R^n : q_i f_i(x) \le q_j f_j(x)\},$$
$$R_{ij} = \{x \in R^n : q_i f_i(x) > q_j f_j(x)\}, \quad (1 \le i, j \le k).$$

From (5), we obtain

$$R_1^n = \bigcap_{j=2}^{k} R_{1j}, R_l^n = \bigcap_{i \ne k} \overline{R}_{il}, \quad (2 \le l < k).$$

therefore,

$$R_1^n \cap R_l^n = \left(\bigcap_{j=2}^{k} R_{ij}\right) \cap \left(\bigcap_{i \ne k} \overline{R}_{il}\right) \subset R_{il} \cap \overline{R}_{1l} = \phi$$
$$\Rightarrow R_1^n \cap R_l^n = \phi, \quad (2 \le l < k).$$

On the other hand, from antithesis style of D'Morgan, we have

$$\overline{R_1^n \cup R_l^n} = \left(\bigcup_{j=2}^{n} \overline{R}_{ij}\right) \cup \left(\bigcup_{i \ne k} R_{il}\right) \subset \overline{R}_{il} \cap R_{1l} = \phi$$
$$\Rightarrow R_1^n \cup R_l^n = R^n, \quad (2 \le l < k).$$

Similarly,

$$R_k^n \cap R_l^n = \phi, \quad (2 \le l < k), \quad R_1^n \cap R_k^n = \phi,$$

so

$$\bigcup_{i=1}^{k} R_i^n = R^n, \cup \left(\bigcup_{l=2}^{k-1} R_l^n\right) \cup R_k^n = R_1^n \cup \left(\bigcup_{l=2}^{k-1} R_l^n\right) \cup R_k^n$$
$$= \left(\bigcup_{l=2}^{k-1} R_1^n \cup R_l^n\right) \cup \left(\bigcup_{l=2}^{k-1} R_k^n \cup R_l^n\right)$$
$$= R^n \cup R^n = R^n \Rightarrow \bigcup_{i=1}^{k} R_i^n = R^n.$$

In addition, from (5) we can directly find out

$$g_{\max}(x) = g_i(x), \quad \forall x \in R_i^n, \quad (1 \le i \le k). \tag{A1}$$

Combining (3) and (A1), we obtain (6).

Appendix 2. Proof of Theorem 2.2

(i) For each $j = 1, 2, \ldots, k$, we have

$$\left(\sum_{j=1}^{k} q_j f_j\right)^{\alpha_i} \ge \left(q_i f_i\right)^{\alpha_i}, \quad 1\,2, \ldots, k.$$

Therefore,

$$\left(\sum_{j=1}^{k} q_j f_j\right)^{\alpha_1+\alpha_2+\cdots+\alpha_k} \ge \prod_{j=}^{k} \left(q_j f_j\right)^{\alpha_j} \Leftrightarrow \sum_{j=1}^{k} q_j f_j$$
$$\ge \prod_{j=}^{k} \left(q_j f_j\right)^{\alpha_j}. \tag{A2}$$

On the other hand,

$$\left(\min_{1 \le j \le k} \{q_j f_j\}\right)^{\alpha_1} \le \left(q_1 f_1\right)^{\alpha_1}, \ldots\ldots, \left(\min_{1 \le j \le k} \{q_j f_j\}\right)^{\alpha_k}$$
$$\le \left(q_k f_k\right)^{\alpha_k},$$

So

$$\left(\min_{1 \le j \le k} \{q_j f_j\}\right)^{\alpha_1+\cdots+\alpha_k} \le \prod_{j=1}^{k} \left(q_j f_j\right)^{\alpha_j}.$$

Or

$$\min_{1\le j\le k}\{q_jf_j\} \le \prod_{j=1}^{k}(q_jf_j)^{\alpha_j}. \tag{A3}$$

Combining (A2) and (A3), we obtain

$$0 \le \sum_{j=1}^{k} q_jf_j - \prod_{j=1}^{k}(q_jf_j)^{\alpha_j} \le \sum_{j=1}^{k} q_jf_j - \min_{1\le j\le k}\{q_jf_j\}.$$

Because $\sum_{j=1}^{k} q_jf_j - \min_{1\le j\le k}\{q_jf_j\}$ includes $(k-1)$ terms, we have

$$\sum_{j=1}^{k} q_jf_j - \min_{1\le j\le k}\{q_jf_j\} \le (k-1)\max_{1\le j\le k}\{q_jf_j\}.$$

Thus

$$0 \le \sum_{j=1}^{k} q_jf_j - \prod_{j=1}^{k}(q_jf_j)^{\alpha_j} \le (k-1)\max_{1\le j\le k}\{q_jf_j\}.$$

Integrating the above relation, we obtain:

$$1 - \prod_{j=1}^{k} q_j^{\alpha_j} D_T(f_1,f_2,\ldots,f_k)^{\alpha} \le (k-1)\int_{R^n} g_{\max}(x)\,\mathrm{d}x. \tag{A4}$$

Using $\int_{R^n} g_{\max}(x) = 1 - Pe^{(q)}_{1,2,\ldots,k}$ for (A4), we have (7).

(ii) We have

$$\begin{aligned} Pe^{(q)}_{1,2,\ldots,k} &= \sum_{j=1}^{k}\int_{R^n\setminus R_j^n} q_jf_j(x)\,\mathrm{d}x \\ &= \sum_{j=1}^{k}\sum_{j\ne i}\int_{R_j^n} \min\{q_if_i(x), q_jf_j(x)\}\,\mathrm{d}x \\ &= \sum_{i<j}\int_{R_i^n} \min\{q_if_i(x), q_jf_j(x)\}\,\mathrm{d}x. \end{aligned}$$

Since

$$[\min\{q_if_i(x), q_jf_j(x)\}]^{\beta} \le (q_if_i)^{\beta} \quad \text{and}$$
$$[\min\{q_if_i(x), q_jf_j(x)\}]^{1-\beta} \le (q_if_i)^{1-\beta},$$

then

$$\min\{q_if_i(x), q_jf_j(x)\} \le (q_if_i)^{\beta}(q_jf_j)^{1-\beta}.$$

Integrating the above inequality, we obtain:

$$\begin{aligned} Pe^{(q)}_{1,2,\ldots,k} &\le \sum_{i<j}\int_{R_i^n} [(q_if_i(x))^{\beta}(q_jf_j(x))^{1-\beta}]\,\mathrm{d}x \\ &\le \sum_{i<j} q_i^{\beta}q_j^{1-\beta} D_T(f_i,f_j)^{(\beta,1-\beta)}\,\mathrm{d}x. \end{aligned}$$

(iii) and (iv) The proofs for (9) and (10) can be seen in Pham–Gia et al. (2008).