ICIC Express Letters
Volume **10**, Number **7**, July **2016**

ICIC International ©2016 ISSN 1881-803X
pp. **1643–1648**

BAYESIAN CLASSIFYING ALGORITHM OF CONTINUOUS ATTRIBUTES BASED ON WEIGHTED ERROR RATES

Jianfei Zhang*, Xing Gao and Xiaoxin Du

College of Computer and Control Engineering
Qiqihar University
No. 42, Wenhua Street, Qiqihar 161006, P. R. China
*Corresponding author: Jian_fei_zhang@163.com

Received December 2015; accepted March 2016

Abstract. *The naïve Bayesian classifier is simple and efficient, which suits for discrete and continuous data. As the strong independence assumption limits its classifying performance, this kind of classifying method still needs to be further improved. In this paper, we proposed a new attributes weighting method through analyzing and estimating error rates of each continuous attribute. This method can make the calculation of likelihood probability relatively simple and weaken the adverse effects of the assumption. Simulating experiments show that the method can improve classifying accuracy adaptively.*
Keywords: Continuous attributes classifier, Weighted attribute, Error rates estimation

1. **Introduction.** The classifier is a classifying function which can map the pre-classified data to the given class and it can be constructed by learning from data. Classifying is very important in data mining and pattern recognition. NBC (Naïve Bayesian Classifier) is a simple and efficient classifier, which assumes that attributes are independent under the condition of given classes. Though its assumption can simplify the calculation of the likelihood probability, strict independent assumptions are difficult to apply in more fields. In order to further improve performance of NBC, some methods have been proposed by alleviating the assumptions, such as TAN (Tree-Augmented Naïve Bayesian) [1]. They are derived from NBC, yet they can partly make use of the relationships among the attributes. However, structure learning of classifiers can greatly increase the calculation; in addition, optimal structure learning has been proved to be NP problems [2]. [3] demonstrated that NBC could have good performance though it contains clear dependences between attributes. So optimizing condition probability of likelihood is a valid way on the basic of NBC's simple calculation, because the classifying result is determined by the relatively but not the completely accuracy of likelihood probability estimation. In NBC, each attribute is equally important for classification. In fact, changing the influence of attributes by setting weights respectively can improve classifier performances. [4] first put forward the idea of WNBC (Weighted Naive Bayesian Classifier) classifier called probability adjustment; the other weighted methods included acquire weights by mountain climbing algorithm combined with Monte Carlo method [5], rough set [6], and K-L Measure [4,7]. Although these weighted methods can calculate the weight in different ways, they have the same defect that only emphasizes the effect of attributes to the whole classification, not to the specific class, that only fits for global attributes selection. This paper presented a weighted algorithm based on estimating the error rates; the algorithm can process uncertain information in the training of classifier, so as to improve the performance of classifiers adaptively and get better classification results. Firstly, we analyzed the error rates of each continuous attribute determined by the class distribution in corresponding attributes; secondly, we estimated the error rates in a proper approach as its weight; finally, we finished the experiment to validate the method and show its advantage. This

paper is divided into 5 sections: Section 1 briefly introduces NBC and its improvement approach; Section 2 gives the fundamental theory of WNBC model and error rate analysis; Section 3 constructs the new classifier WNBC-ER based on statistics of the error rate of continuous attributes; Section 4 verifies its performance and analyzes the reason; Section 5 is the conclusions and further works.

2. Fundamental Theory.

2.1. **WNBC model.** Let $D = (X_1, \ldots, X_n, C)$ be a limited dataset, $X_1, \ldots, X_n$ represent property variables, $C = \{c_1, \ldots, c_n\}$ is a class variable, and x_i is the value of X_i. P is probability; p is the value of P. Then the probability of c_j in sample $S_i = \{x_1, \ldots, x_n\}$ can be described as Equation (1), where $\alpha = p(x_1, \ldots, x_n)$ is the regularization factor, $p(c_j)$ is the prior probability of c_j, $p(x_1, \ldots, x_n|c_j)$ is the class-conditional-probability of c_j, $p(c_j|x_1, \ldots, x_n)$ is the posterior probability of c_j, that is, the probability of c_j after information revised. Posteriori probability reflects the influence of sample data on the class c_j. We can assert that each attribute is equal for the classification. However, it is not advisable that classifying does not consider sample distribution and attribute selection to improve the classifiers performances [8,9]. In fact, the class variables distributed in different attributes are diverse, so we can set different classifying contribution ratios on the relationship between attributes and objectives. The assumption of NBC can simplify the calculation of likelihood probability and set each attribute a weight, just shown as Equation (2), where $P^{w_i}(x_i|c)$ is the weighted conditional probability, and w_i is the contribution ratio of attribute i. At present, the major attributes weighted optimization is based on the information in the NBC weighting methods, the method only considers the attributes of the global contribution, but it does not specify each class. With the precondition of the NBC classifying principle, we can start from the error rates of a single attribute, ascertain the relationships between single attributes error rates and the global error rates, and make quantitative analysis of each attribute contribution to the correct classification, and finally we can get a new weighted attributes classifier.

$$p(c_j|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|c_j)p(c_j)}{p(x_1, \ldots, x_n)} = \alpha p(x_1, \ldots, x_n|c_j)p(c_j) \tag{1}$$

$$P(c|X) = \alpha P(c) \prod_{i=1}^{n} P^{w_i}(x_i|c) \tag{2}$$

2.2. **Error rate analysis.** The continuous attribute NBC assumes that the samples obey some distributions, and then estimate the parameter and probability density $p(x)$ as conditional probability (The probability of continuous attribute variable on single point is zero). The continuous attributes NBC error rates can be expressed as Equation (3). R_j is the region where class j has the highest posterior probability.

$$E = 1 - \sum_{j=1}^{m} \int_{R_j} p(X|c_j)\, p(c_j) dx \tag{3}$$

The classifying error rates of classifier refer to the probability that can classify a sample from a certain class to other classes. Since errors are introduced by both the regional errors of the estimation of probability densities and numerical integration, the results are only approximated by finite data [8]. However, the error rates can be reduced by optimizing the two sources of errors. According to the decision rules and the relationships between attributes and classes, we can find a feasible kind of evaluating criteria to measure the properties of the attributes for the correct classification contribution and reflect it to the attributes that can improve the classifying performance. As NBC independence assumption simplifies the likelihood probability calculation, analyzing local error to find

methods to reduce the global error is the most direct way. Here we analyze the classifying error rates of two classes.

Let the number of class $m = 2$, R_1 and R_2 express the conclusive region of class 1 and class 2 respectively, which has higher posterior probability, so two error situations can occur.

(1) $P(X \in R_2, c_1)$, X belong to class 1, the result belongs to class 2;

(2) $P(X \in R_1, c_2)$, X belong to class 2, the result belongs to class 1.

The error rate is the sum of both, as shown in Equation (4).

$$P(e) = \int_{R_2} P_1(e|X)\, p(X)\, dX + \int_{R_1} P_2(e|X)\, p(X)\, dX \tag{4}$$

In Equation (3), the first item on the right refers to the probability that class 1 shows error in the class 2 conclusive region and the second item as well. When $X = x$, the attribute varies into one dimension. Let $p(X|c)\, p(c)$ be the posterior probability, and the distribution can be shown in Figure 1. In it t is the classifying critical point, as shown in Equation (5).

$$p(X|c_1)\, p(c_1) = p(X|c_2)\, p(c_2) \tag{5}$$

In Figure 1, when $p(c_1|X) > p(c_2|X)$, the classifier can put the pre-classifying samples into class 1, and it also can put the samples into class 2. So we can consider that the error rates are equal to $P(c_2|X)$ that the probability of class 2 samples can be assigned to class 1, which can be expressed as Equation (6).

$$P_2(e|X) = P(c_2|X) \tag{6}$$

For the same reasons, error rates can be expressed as Equation (7), when $p(c_2|X) > p(c_1|X)$.

$$P_1(e|X) = P(c_1|X) \tag{7}$$

From Equation (6) and Equation (7), we can get the error rate as Equation (8).

$$\begin{aligned} P(e) &= \int_{R_2} P(c_1|X)\, p(X)\, dX + \int_{R_1} P(c_2|X)\, p(X)\, dX \\ &= \int_{R_2} p(X|c_1)\, p(c_1)\, dX + \int_{R_1} p(X|c_2)\, p(c_2)\, dX \end{aligned} \tag{8}$$

We can further comprehend classifying principle by analyzing the error rates. The classifier error rates can be reduced to a minimum when critical point is located on

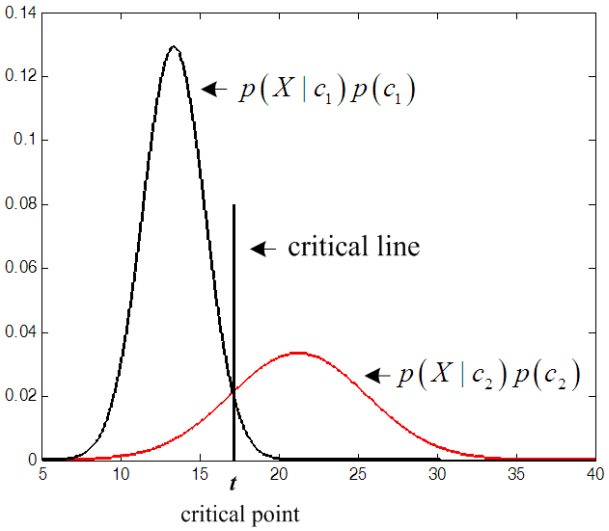

FIGURE 1. The posterior probability distribution of two categories classification

position in Figure 1. We can get the distinct degree of classes by calculating the error rates upper limitation. However, the actual distribution of samples is not in accord with the error rates, and sometime it has some noises. The boundary location uncertainty makes the estimation error rate become more complicated. The basic idea of this paper is to estimate the error rate and set a threshold value, those attributes will be excluded whose error rate is too high, and then calculate the threshold attribute error rate as the evaluation contribution standard of classifier.

3. **WNBC-ER Based on Error Rate Estimation.** Considering the actual classification, in the course of extracting attributes of the pre-classified samples, we always try to obtain attributes as much as possible; however, for information from the information source, some attributes can reduce the uncertainty of information and others may increase the uncertainty. The classifier not only needs to obtain the high reliability information from each attribute, but also reduce unnecessary information, so letting the error rate as the evaluation index of attributes about correct classification is a feasible method. The past conditional probability weighted method is based on discrete attributes samples, we presented a weighted method based on continuous attributes error rates, and this method can divide attributes into different levels by the classification properties that the confused degree between classes is hard to estimate. Calculating the attributes error rates can acquire its weight.

3.1. **Ability of distinguishing classes and undesirable attributes.** Assuming the continuous attribute sample obeys normal distribution $X \sim N(\mu, \sigma^2)$, the probability of random variables in $\mu \pm \sigma$ is $P\{\mu - \sigma < X < \mu + \sigma\} = 68.26\%$. For the case of one attribute classification, it will be not resolved, when the critical point t located in this area. The conditional probability of attribute effect on the likelihood probability is unknown, which is the disadvantage that the NBC inherent conditional independence assumption brings in. In order to reduce the negative impact, this paper adopts weighted method to weaken the influence.

Definition 3.1. *For continuous attributes Bayesian classifier, if the distribution of class i and class j in the attribute satisfies $p(X = \mu_i|c_i)\,p(c_i) < p(X = \mu_i|c_j)\,p(c_j)$, then define the ability of this attribute to distinguish class i and class j is* **bad**.

Definition 3.1 indicates that the maximum probability density of the class i is also less than the class j in this position. In other words, class i is covered by class j around the mean. Assumed $\mu_i < \mu_j$, and then critical point t must locate in $[-\infty, \mu_i]$ or $[\mu_j, +\infty]$. On the contrary, the distinguish ability is higher when t locates in $[\mu_i, \mu_j]$.

Definition 3.2. *In continuous attributes classifier, if the ability to distinguish all classes is bad, define the attribute is* **undesirable attribute**.

Undesirable attribute referring to the distinction between classes is not obvious, and one attribute cannot be used to classify. The division of undesirable attribute not only can weaken the adverse effects of the independence assumption, but also reduce the classification process calculation.

3.2. **Calculation of error rate.** After excluding the undesirable attributes, the calculation of the remaining data set is relatively simple. Let the number of classes be m and the attribute number of the remained data set is n'. Constructing CM (Confusion Matrix) that its attributes number is n' and its dimension is m, which records the attribute error rates between classes. The element of the CM is given as $P_{ij}(e) = \int_{R_j} p(x|c_i)\,p(c_i)dx$, $i \neq j$ and $\bar{P}_{ii}(e) = \int_{R_i} p(x|c_i)\,p(c_i)dx$, where R_i is the region where class i has the highest

posterior probability. The diagonal elements can be used to express the probability of correct classification. Establishing CM can find where the error occurred and the correlation between attributes and correct classification. CM is a tool for calculating weights.

In CM, the row vector expresses that the probability of a variable belonging to corresponding class is assigned to other classes, and the column vector expresses that other class variable is assigned to this column class. The diagonal elements $P_{ii}(e)$ expresses that the probability of corresponding class can be classified correctly. Calculating the weight not only considers the ability about classifying a certain class from classes but also considers that the probability of other class can be assigned to this class. The weight includes the two aspects influence.

3.3. **Method of weight calculation.** This paper presented the weights calculating method, the weight is a description of the attribute that provides a classification of information and the negative impact on the decision-making results, and weight is the arithmetic mean value of them, which can be calculated by $w_i = (w_{i1} + w_{i2})/2$, $i = 1, 2, \cdots, n'$, where w_{i1} and w_{i2} express the average ability of separating a certain class from others and correct classification. In it $w_{i1} = 1 - \sum_{k=1}^{m} \left[p(c_k) \sum_{l=1}^{m} cm^i_{kl,k\neq l} \right] \Big/ m$, $w_{i2} = 1 - \sum_{l=1}^{m} \left[(1 - p(c_l)) \sum_{k=1}^{m} \left(cm^i_{kl,k\neq l} \right) \right] \Big/ m$, where cm^i_{kl} is the ith element of CM.

3.4. **Design of algorithm.** In view of analyzing the problems of continuous attributes of NBC, we put forward a feasible classifier optimization method WNBC-ER (Weighted Naïve Bayesian Classifier by Error Rate) and these steps are as follows:

Step 1. Preprocess the data sets, and delete incomplete data samples and more missing attribute variables (This article selects the data set with missing data items which are very small, so deletion does not affect the classification result);

Step 2. Select training samples and let them obey the normal distribution, and then estimate the mean and variance by the maximum likelihood;

Step 3. According to attribute ability of distinguish classes, estimate position of critical point and select undesirable attributes;

Step 4. Calculate error rate and construct confusion matrix;

Step 5. Calculate attributes weight establishing weighted NBC.

4. **Experimental Design and Analysis.** For verifying the effects of the classifier, we use 15 continuous data sets from UCI [10]. In order to ensure the accuracy and reliability of the experiment, this paper uses 10-fold-cross valid method. Comparison of algorithms is finished by MATLAB. The selected samples information and the experimental results of the error rate are shown in Table 1.

In Table 1, S#, A# and C# are the number of samples, attributes and classes respectively, WNBC-S is a weighted method by integrated information and algebra [6]. It can be seen from Table 1 in most data sets, the WNBC-ER performance is the best, especially classification results are significantly better than other for data sets: banknote, seeds and vehicle. It shows that these data weight values between the different attributes have great difference, and the low error rate attribute and the high one have enormous influence by the weight. WNBC-ER can enhance specially class accuracy by analyzing intermediate variable and CM of classifying process of these outperformances. The weighted method gives good adaptability for this type of data; the classifying effect can be improved.

5. **Conclusions.** This paper presents a weighted method based on attributes error rates. Due to the actual problems error rate in computational complexity, choosing an appropriate estimation method determines the efficiency and performance of classification. However, estimating method of continuous attribute probability density directly affects the

TABLE 1. The data set description and the experimental result

Data Set	*S#*	*A#*	*C#*	*Distribution of the classes*	*NB*	*WNBC-S*	*WNBC-ER*
breast cancer	683	9	2	239; 444	0.0395	0.0381	**0.0366**
wdbc	569	30	2	212; 357	0.0668	0.0668	**0.0650**
ecoli	327	5	5	143; 77; 52; 35; 20	**0.1284**	0.1346	0.1385
banknote	1372	4	2	610; 762	0.1611	0.1588	**0.1403**
wine	178	13	3	59; 71; 48	0.0225	**0.0169**	0.0337
vertebral	310	6	3	60; 150; 100	0.1806	0.1774	**0.1613**
segment	2310	16	7	347; 208; 393; 276; 391; 349; 346	**0.1883**	0.2508	0.2299
seeds	210	7	3	70; 70; 70	0.1000	0.0952	**0.0909**
ionosphere	351	32	2	225; 126	0.1823	**0.1795**	0.1797
yeast	1299	6	4	463; 429; 244; 163	0.4727	0.3923	**0.3692**
iris	150	4	3	50; 50; 50	0.0533	0.0467	**0.0400**
liver	345	6	2	145; 200	0.4377	**0.4290**	0.4464
vehicle	864	18	4	199; 217; 218; 212	0.5496	0.4706	**0.4118**
glass	205	7	4	70; 76; 17; 29	0.5295	**0.4976**	0.5424
haberman	306	3	2	225; 81	0.2516	0.2614	**0.2452**

similarity close degree, and affects the accuracy of error rate estimation. To further improve the performance of the classifier, finding an accurate probability density estimation method that can reflect the actual distribution data is very necessary.

Acknowledgment. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers. This work was supported by Natural Science Fund of Heilongjiang Province of China (F201333), Ministry of Education of Humanities and Social Science Research Youth Fund Projects (14YJC630188).

REFERENCES

[1] D. M. Chickering, Learning Bayesian networks is NP-complete, *Learning from Data*, pp.121-130, 1996.

[2] J. Zhang, X. Han, Q. Zhang and S. Wang, A Bayesian network classifier learning based on dependent analysis, *ICIC Express Letters*, vol.7, no.12, pp.3207-3212, 2013.

[3] S. Wang, J. Zhang and H. Wang, Dynamic Bayesian network method for causal analysis between enterprise operation indexes, *ICIC Express Letters*, vol.7, no.11, pp.3033-3039, 2013.

[4] C. H. Lee, F. Gutierrez and D. Dou, Calculating feature weights in Naïve Bayes with Kullback-Leibler measure, *The 11th IEEE International Conference on Data Mining*, pp.1146-1151, 2011.

[5] W. B. Deng and Y. Wang, Weighted naïve Bayes classification algorithm based on rough set, *Computer Science*, vol.34, no.2, pp.204-206, 2007.

[6] N. A. Daidi, J. Cerquides, M. J. Carman et al., Alleviating naive Bayes attribute independence assumption by attribute weighting, *The Journal of Machine Learning Research*, vol.14, no.1, pp.1947-1988, 2013.

[7] S. Zhang, Learning weighted naive Bayes with accurate ranking, *The 4th IEEE International Conference on Data Mining*, pp.567-570, 2004.

[8] K. Tumer and J. Ghosh, Bayes error rate estimation using classifier ensembles, *International Journal of Smart Engineering System Design*, vol.5, no.2, pp.95-109, 2003.

[9] Y. Cao and Q. G. Miao, Advance and prospects of AdaBoost algorithm, *Acta Automatica Sinica*, vol.39, no.4, pp.745-756, 2013.

[10] L. Murphy and D. W. Aha, *UCI Repository of Machine Learning Databases*, http://archive.ics.uci.Edu/ml/datasets.html, 2015.