ICIC Express Letters
Volume **10**, Number **9**, September **2016**

ICIC International ©2016 ISSN 1881-803X
pp. **2135–2140**

NAÏVE BAYESIAN CLASSIFIERS WITH CONTINUOUS ATTRIBUTES USING MULTI-SMOOTHING PARAMETERS

Jinku Han[1], Kehui Liu[1], Helin Jia[2] and Jianfei Zhang[1,*]

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

[2]Faculty of Information Technology
Macau University of Science and Technology
Macau 999078, P. R. China

Received January 2016; accepted April 2016

Abstract. *In this paper, we propose a method that can be used to construct Naïve Bayesian classifier with multi-smoothing parameters which can improve the classification performance of Naïve Bayesian classifier based on Gaussian kernel function. In the proposed approach, mean integrated square error (MISE) is taken into account, by which we can measure the error between estimated density and actual density. The optimal smoothing parameters can be determined by approximate calculation, so different smoothing parameters of attributes can be used to construct classifiers. The compared experiments results show the proposed method can achieve higher classifying accuracy.*
Keywords: Multi-smoothing parameter, Kernel estimation, Naïve Bayesian classifier

1. **Introduction.** Naïve Bayesian classifier (NBC) is the simple and efficient classifier, but it has the assumption of attributes independence; it limited its classifying accuracy, and a variety of methods have been proposed to improve the performance. So far, a lot of research had been done for discrete attributes NBC, such as tree augmented Naïve Bayesian classifier (TAN) [1], choosing structures by maximizing conditional likelihood [2], and K-dependence Bayesian classifier (KDB) [3]. For the continuous attributes Naïve Bayesian classifier, two methods commonly can be used: both the continuous attributes discretization and attributes density estimation. The main problem of continuous Bayesian classifier is attribute density estimation and dependent relationship between attributes learning. On attribute density estimations, John and Langley [4] proposed Gaussian Naïve Bayesian classifier and flexible Bayesian classifier (fNB) based on Gaussian function and Gaussian kernel. Prezé et al. [5] improved the attribute density estimation based on Gaussian kernel by introducing smooth parameters that used the mean integrated square error (MISE) as statistical standards to optimize the smoothing parameters. Though parameter optimization based on MISE was efficient, the accuracy of classifiers established still needs to be improved further. Huang [6] compared the NBC based on Gauss kernel with support vector machine (SVM) and obtained the result that optimized NBC had higher classifying accuracy than SVM. On the attributes dependency, Guo et al. [7] extended the structure of continuous attributes NBC and improved its classification accuracy based on likelihood score. Wang et al. [8] set up multi-smooth parameter of continuous attributes of fully Bayesian classifier using multi Gaussian kernel function to estimate attributes joint density and proposed a parameter optimization method which combined the classifying accuracy standards and divided different interval entirely searches. It was very important for the Bayesian classifier to estimate the

attribute density by the Gaussian kernel function. So this paper presented a new parameter optimization method. This method makes use of MISE to measure the fitting degree between estimated density and actual density, the optimal parameter can be used by approximate calculation, and finally the multiple smoothing parameter NBC can be established.

This paper presented a new Bayesian classifying method based on multi-smoothing parameters. Firstly, we analyzed Naïve Bayesian classifier based on Gaussian kernel density; secondly, we introduced smoothing parameter optimization and classifier learning; finally, we finished the simulation and experiment.

2. Naïve Bayesian Classifier Based on Gaussian Kernel Density.

NBC assumes that all attributes are conditionally independent with each other. According to Bayesian formula we can obtain (1).

$$p(c|x_1, x_2, \ldots, x_n) = p(c)p(x_1, x_2, \ldots, x_n|c)/p(x_1, x_2, \ldots, x_n) = \alpha p(c) \prod_{i=1}^{n} p(x_i|c) \tag{1}$$

In Formula (1), n is the number of attributes, and α is regularization factor. For continuous attributes NBC, we use Gaussian kernel functions to estimate the attribute marginal density. In order to make the attribute marginal density fit the actual data better, a smoothing parameter is introduced, which can control the marginal density as well as fitting degree in Gaussian kernel. The attribute marginal density can be estimated by Gaussian kernel functions, just shown as Formula (2).

$$p(x_i|c) = \frac{1}{N(c)} \sum_{m=1}^{N(c)} g(x_i; x_{im}, h) = \frac{1}{N(c)} \sum_{m=1}^{N(c)} \frac{1}{\sqrt{2\pi}h} \exp\left[-\frac{(x_i - x_{im})^2}{2h^2}\right] \tag{2}$$

where $N(c)$ is the number of samples which belongs to class c in data set, x_i, x_{im} is the ith attribute and the mth value of attribute i belongs to class c of sample x respectively, and h is smoothing parameter of classifiers. From Formula (1), we can get the $p(x_1, x_2, \ldots, x_d|c)$. Then we can get the posteriori probability, as shown in Formula (3).

$$p(c|x_1, x_2, \ldots, x_d) = \frac{1}{N \cdot (N(c))^{d-1}(2\pi)^{d/2}} \prod_{i=1}^{d} \left\{ \sum_{m=1}^{N(c)} \frac{1}{h_i} \exp\left[-\frac{(x_i - x_{im})^2}{2h_i^2}\right] \right\} \tag{3}$$

where d is the number of attribute, and h_i is the smoothing parameter of each attribute.

3. Smoothing Parameter Optimization and Classifier Learning.

3.1. Smoothing parameter optimization.

Smoothing parameter can control fitting degree between attributes density and actual data. Let smoothing parameter be 0, and kernel function can reflect the distribution of the data well, but the noise in the data will result in over-fitting. On the contrary, the smoothing parameter is bigger, so the fitting degree between estimated conditional density and data will become poor, which may lead to under-fitting. Therefore, it is necessary to optimize smoothing parameter so as to promote the performance of the classifying.

The different smoothing parameters lead to the different errors between the estimated density and the actual density, so we need a standard to measure the error. MISE can comprehensively estimate the relationship between density and actual density, so we choose MISE as a standard to estimate. Gaussian kernel functions are used to estimate the samples distribution, and MISE is a function of smoothing parameters h and the number of samples.

Proof: Suppose $p(x)$ is actual probability that is the actual probability density for the variable, and $\hat{p}(x)$ is estimated density based on Gaussian kernel function. Then the MISE can be expressed as is shown in Formula (4).

$$\begin{aligned} MISE(h) &= E\left[\int \{\hat{p}(x) - p(x)\}^2 dx\right] \\ &= \int \left\{E[\hat{p}^2(x)] - E^2[\hat{p}(x)]\right\} dx + \int \{E[\hat{p}(x)] - p(x)\}^2 dx \end{aligned} \tag{4}$$

For one dimensional Gaussian density variable, the density of x can be expressed as Formula (5).

$$\hat{p}(x) = \sum_{i=1}^{n} g(x; x_i, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} \exp\left[-\frac{(x-x_i)^2}{2h^2}\right] \tag{5}$$

Let $z = \frac{x-x_i}{h}$, then

$$\int g(x; x_i, h)dx = \int g(z)p(x_i)dx_i = \int g(z)p(x - hz)dz \tag{6}$$

According to the Taylor's formula, $p(x-hz)$ can be written as Formula (7). Then we can get $E(\hat{p}(x))$, $E(\hat{p}^2(x))$, then incorporate it into Formula (4), we can obtain Formula (8).

$$p(x - hz) = p(x) - hzp'(x) + \frac{1}{2}h^2z^2p''(x) + o\left(h^2\right) \tag{7}$$

$$\begin{aligned} MISE(h) = \frac{1}{nh}\int \left\{\left[\frac{1}{2\sqrt{\pi}}p(x) + \frac{h^2}{4}p''(x)\right] - \frac{1}{n}\left[p(x) + \frac{1}{2}p''(x)h^2\right]^2\right\} dx \\ + \frac{h^2}{4}\int p^{2\prime\prime}(x)dx \end{aligned} \tag{8}$$

From Formula (8) we can get $\frac{1}{n}\left(p(x) + \frac{1}{2}p''(x)h^2\right)^2 = o\left(n^{-1}\right)$, so when $n \to \infty$, $\frac{1}{n}\left(p(x) + \frac{1}{2}p''(x)h^2\right)^2 = 0$, at this time, MISE can be expressed as Formula (9).

$$\begin{aligned} MISE(h) &= \frac{1}{nh}\int \left[\frac{1}{2\sqrt{\pi}}p(x) + \frac{h^2}{4}p''(x)\right] dx + \frac{h^2}{4}\int p^{2\prime\prime}(x)dx \\ &\approx \frac{1}{2nh\sqrt{\pi}} + \frac{h}{n}\int p''(x)dx + \frac{h^2}{4}\int p^{2\prime\prime}(x)dx \end{aligned} \tag{9}$$

As the number of samples is far larger than h, $\frac{h}{n}\int p''(x)dx \approx 0$ and Formula (9) can be transformed to Formula (10).

$$MISE(h) \approx \frac{1}{2nh\sqrt{\pi}} + \frac{h^2}{4}\int p^{2\prime\prime}(x)dx \tag{10}$$

Using $\hat{p}(x)$ to replace $p(x)$ we can get $p''(x)$, as is shown in Formula (11). If we incorporate Formula (11) into Formula (10), we can get Formula (12).

$$p''(x) = \sum_{i=1}^{n} g''(x; x_i, h)/nh^3 \tag{11}$$

$$\begin{aligned} MISE(h) &\approx \frac{1}{nh}C(x) + \frac{h^4}{4}\left[B(x)\right]^2 \int \left[p''(x)\right]^2 dx \\ &= \frac{1}{2nh\sqrt{\pi}} + \frac{h^4}{4}\cdot\frac{\sqrt{\pi}}{16h^3}\left\{\sum_{i=1}^{n}\sum_{j<i}^{n}\left[(x_i - x_j)^2 - 6h^2\right]^2 - 24h^4\right\} e^{-\frac{(x_i-x_j)^2}{4h^2}} \\ &\quad + \frac{3h^5}{16}\sqrt{\pi} \end{aligned} \tag{12}$$

The proof is completed.

The smaller MISE value is, the smaller deviation between the actual density and the estimated density is, so h is an optimal smoothing parameter when the MISE is the smallest. The derivative of Formula (12) is shown as Formula (13).

$$\frac{d(MISE)}{dh} = \frac{\sqrt{\pi}}{16h^3}\sum_{i=1}^{n}\sum_{j<i}^{n}\left\{\left[(x_i-x_j)^2-6h^2\right]^2-24h^4\right\}\exp\left[-\frac{(x_i-x_j)^2}{4h^2}\right]-4hn \quad (13)$$

Let Formula (13) be equal to 0, and we can get optimal smoothing parameters.

$$h_{best} = \left(\frac{\sqrt{\pi}}{16n}\sum_{i=1}^{N}\sum_{j<i}^{N}\left\{\left[(x_i-x_j)^2-6h^2\right]^2-24h^4\right\}\exp\left[-\frac{(x_i-x_j)^2}{4h^2}\right]\right)^{1/3} \quad (14)$$

3.2. **Multi-smoothing parameters of the Naïve Bayesian classifier.** For given data set, we can get optimal smoothing parameter for each attribute, further we can get the best marginal density estimation of attributes. Finally, we can compute the posteriori probability. The algorithm of multi-smoothing parameter Naïve Bayesian classifier (MSPNBC) is shown as follows.

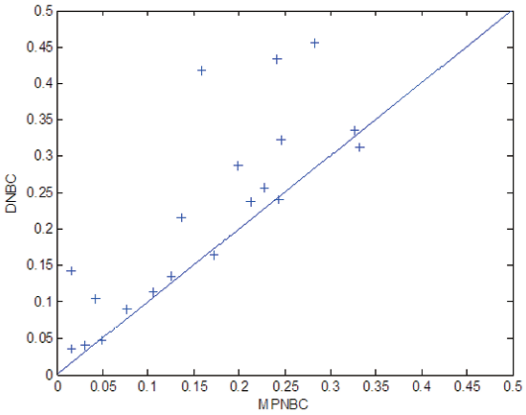
Input: data set $D = (X_1, X_2, \ldots, X_n)$
$//X_i$ is the attribute of the sample, ε is the threshold value
Output: Naïve Bayesian classifier
1: for $i \leftarrow 1 : n$ //Optimized for each property parameter
2: $h_{i0} = 1.06\sigma_i n^{-1/5}$ $//h_{i0}, \sigma_i$ is initial smoothing parameters and variance of X_i
3: $h_{ibest} = p(h_{o0})$
4: while $(|h_{ibest} - h_{i0}| > \varepsilon)$
5: $h_{i0} = h_{ibest}$, $h_{ibest} = p(h_{i0})$, $h_{ibest} = (h_{i0} + h_{ibest})/2$
6: endwhile
7: return h_{ibest}
8: endfor
9: compute $p(c|x_1, x_2, \ldots, x_n)$
10: return $c = \arg\max(p(x_1, x_2, \ldots, x_n|c))$

4. **Experiments.** In our experiment, we used the 20 data sets from UCI and excluded the incomplete data directly [9,10]. Ten-fold cross-validation is used to estimate the error rate of classification. We conduct comparison for the Naïve Bayesian classifer based on the discretization of continuous attributes, the density estimation based on Gaussian function TAN (GTAN), single-smoothing parameter of Naïve Bayesian classifier (fSNBC) [5], smoothing parameter Naïve Bayesian classifier on the basis of fSNBC (fMNBC) and multi-smoothing parameters Naïve Bayesian classifier (MSPNBC) proposed in this paper. The data sets and the results are presented in Table 1. Here S#, A# and C# are the number of samples, attributes and classes respectively.

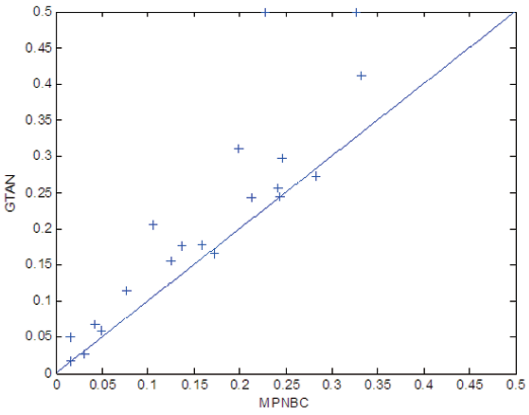
Overall, the error rate of classification in MPNBC reduced 21.8%, 21.8%, 15.6% and 0.55% compared with the other four classifiers respectively, which fully shows that MPNBC has good classification accuracy. The scatter charts that compared the MPNBC and the others are shown in Figure 1 of the data in Table 1. The error rate of classification in MPNBC and fMNBC is lower than fSNBC, because each attribute has its own characteristics and uses multi-smoothing parameters to estimate the marginal density. This method can utilize more and better local information of sample, while single smoothing parameters estimate the density on the whole and ignore the difference between attributes. The error rate of MPNBC is lower than fMNBC, primarily because fMNBC is relatively sensitive for smoothing parameters. When the estimation of h has deviation, it will lead to a large influence on the whole estimation. While MPNBC selected smoothing parameters for every attribute respectively, the smoothing parameters do not affect each other, so the

TABLE 1. Dataset information and compared experiment results

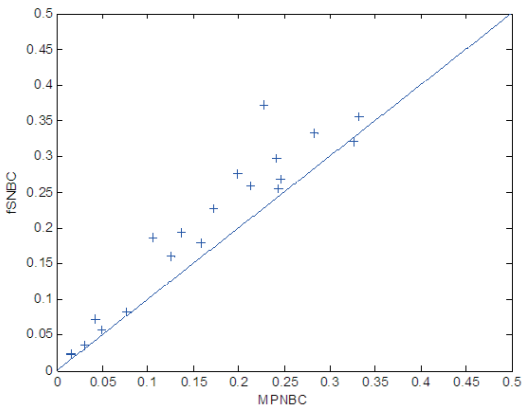
Data Set	*S#;A#;C#*	*DNBC*	*GTAN*	*fSNBC*	*fMNBC*	*MPNBC*
Balance	625;4;3	0.090	0.115	0.083	0.085	0.077
Breast cancer	699;10;2	0.036	0.050	0.024	0.024	0.016
Breast Tissue	106;9;6	0.287	0.310	0.276	0.241	0.199
Ecoli	292;5;4	0.104	0.068	0.072	0.047	0.042
Glass	214;9;6	0.336	0.499	0.321	0.314	0.327
Heart disease	270;13;2	0.135	0.155	0.160	0.149	0.125
Haberman	306;3;2	0.241	0.245	0.255	0.243	0.244
Image	2310;18;7	0.113	0.205	0.187	0.162	0.105
Ionosphere	349;33;2	0.257	0.500	0.372	0.336	0.227
Iris	150;4;3	0.040	0.027	0.036	0.036	0.030
Liver disorder	345;6;2	0.312	0.411	0.356	0.332	0.332
Pima	768;8;2	0.237	0.244	0.260	0.234	0.212
Sonar	208;60;2	0.322	0.298	0.269	0.251	0.246
Spambase	4601;57;2	0.165	0.166	0.228	0.196	0.172
Vehicle	846;19;4	0.433	0.257	0.297	0.265	0.241
Wine	178;12;3	0.142	0.017	0.023	0.019	0.016
Wpbc	198;34;2	0.417	0.177	0.179	0.167	0.158
Waveform	5000;21;3	0.215	0.176	0.193	0.156	0.136
Wdpc	569;31;2	0.047	0.059	0.058	0.055	0.049
Yeast	1484;6;4	0.455	0.273	0.332	0.311	0.283



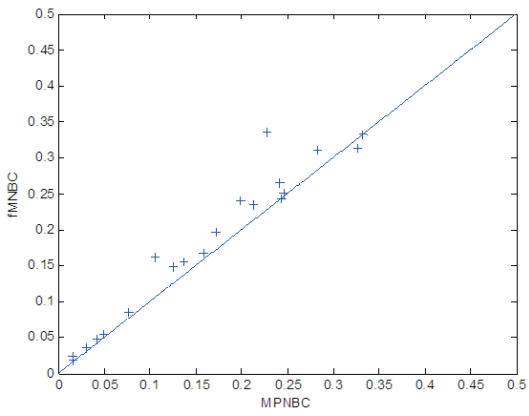
(a) MPNBC and DNBC



(b) MPNBC and GTAN



(c) MPNBC and fSNBC



(d) MPNBC and fMNBC

FIGURE 1. The scatter chart of the classifier error rate

classifier is more stable and has better performance. On most data sets, the classification accuracy of MPNBC is higher than DNBC, but on the data sets haberman, Liver disorder, Spambase, and Wdpc, the classification accuracy of MPNBC is lower than DNBC, so it can be seen that DNBC is more applicable to the data set with more attributes.

5. **Conclusion.** In this paper, we propose an optimal smoothing parameter algorithm, and apply this algorithm to estimating the density of continuous attributes, and then construct the multi-smoothing parameter NBC. The experimental results demonstrate lower error rate of classification in MPNBC. However, the selection of ε in different data sets is various, how to select the reasonable threshold value need further research. The MPNBC can not effectively use the information among the attributes, so the extension of NBC based on Gaussian kernel function needs further research.

Acknowledgments. 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] N. Friedman, S. Geiger and M. Goldszmidt, Bayesian network classifier, *Machine Learning*, vol.29, nos.2-3, pp.131-161, 1997.

[2] D. Griisman and P. Domingos, Learning Bayesian network classifiers by maximizing conditional likelihood, *Proc. of the 21st International Conference on Machine Learning*, Alberta, Canada, pp.361-368, 2004.

[3] G. I. Web, J. R. Boughton, F. Zheng et al., Learning by extrapolation from marginal to full-multivariate probability distributions: Decreasingly Naïve Bayesian classification, *Machine Learning*, vol.86, no.2, pp.233-272, 2012.

[4] G. H. John and P. Langley, Estimating continuous distributions in Bayesian classifiers, *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence*, San Mateo, USA, pp.338-345, 1995.

[5] A. Prezé, P. Larrañaga and I. Inza, Bayesian classifiers based on kernel density estimation: Flexible classifiers, *International Journal of Approximate Reasoning*, vol.50, no.2, pp.341-362, 2009.

[6] S. C. Huang, Using Gaussian process based kernel classifiers for credit rating forecasting, *Expert Systems with Applications*, vol.38, no.7, pp.8607-8611, 2011.

[7] X. S. Li, C. X. Guo and Y. H. Guo, The credit scoring model on extended tree augment Naïve Bayesian network, *Systems Engineering Theory & Practice*, vol.26, no.6, pp.129-136, 2008.

[8] S. C. Wang, J. F. 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.

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

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