ICIC Express Letters
Volume **11**, Number **1**, January **2017**

 ISSN 1881-803X
pp. **247–252**

NAÏVE BAYESIAN CLUSTERING BASED ON THE OPTIMIZATION OF SMOOTHING PARAMETERS

Ming Liu[1], Jianfei Zhang[1,*], Kehui Liu[1] and Shuangcheng Wang[2]

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

[2]College of Mathematics and Information
Shanghai Lixin University of Commerce
No. 2800, Wenxiang Road, Shanghai 201620, P. R. China

Received June 2016; accepted September 2016

Abstract. *The Naïve Bayesian clustering with Expectation Maximization (EM) algorithm is sensitive to the initial value. It has proved to be easy to fall into local optimization. Gibbs sampling can reduce the error of clustering. Both the methods take advantage of Gaussian density to estimate conditional density. Data structure information and sample fitting degree will decline when sample density and Gaussian density have a larger difference. It leads to the decrease of clustering effect. In this paper, on the basis of making use of Gaussian kernel function to deal with clustering, a Naïve Bayesian clustering with Smoothing Parameter Optimization Algorithm (SPOA) is presented. Experiment and analysis are done by using data set in the UCI machine learning repository. The results show that the Naïve Bayesian clustering based on the optimization of smoothing parameters has higher prediction accuracy.*

Keywords: Naïve Bayesian clustering, Gaussian kernel function, Smoothing parameter optimization algorithm

1. **Introduction.** Clustering is a type of special classification. In a clustering problem, we need to divide the data into mutually exclusive categories or clusters, and we have to avoid every aspect of the work, both the homogeneity between clusters and the heterogeneity within the clusters [1]. It is an unsupervised learning method because the number of categories is uncertain. Its main purpose is to divide the similar data into corresponding classes according to a rule. Generally, clustering analysis is a process of repeatedly maximizing the resemblance between intracluster components and the dissimilarity between intercluster components [2]. These similarity-based clustering methods can calculate similarity using a specific distance function for components with continuous attributes and calculate similarity measures for components with qualitative analysis. Two main approaches are well known among the similarity-based methods. These are the hierarchical approach (e.g., Ward's method, single linkage method) and the partition approach (e.g., K-means) [3,4]. Naïve Bayesian network clustering is a classical method. It is based on the Naïve Bayesian network and EM algorithm [5] to determine the value of the class variable during the clustering process. Small data clusters are merged into other clusters. Finally, the optimal number of clusters is obtained. However, the EM algorithm is sensitive to the initial value, so it is easy to fall into local extreme point. The iteration of the parameters may converge to the boundary of the parameter space, and then the convergence of the deception is produced. Wang et al. proposed using Gibbs sampling method to reduce the error of clustering [6,7]. Both the methods used Gaussian density to estimate conditional density. Data structure information and sample fitting degree will decline when sample density and Gaussian density have a larger difference. [8]

proposed a clustering method using Gaussian kernel density, which can obviously improve the accuracy of clustering. However, the method is only related to the number of the current samples, and cannot be used to effectively utilize the information of the samples. In this paper, we propose a Naïve Bayesian clustering with SPOA on the basis of Gaussian kernel.

This paper presented a Naïve Bayesian clustering method based on the optimization of smoothing parameters. Firstly, we analyzed the algorithm SPOA; secondly, we introduced the clustering process on the basis of SPOA; finally, we finished the simulation and experiment to validate the method and show its advantage.

2. Problem Statement and Preliminaries.

2.1. SPOA.

Smoothing parameter can adjust and control the fitting degree between attributes of density and actual data. If the value of smoothing parameter is zero, kernel function can reflect the distribution of the data better, but overfitting can appear when there is noise in data. If the value of smoothing parameter is big, under fitting can appear because the fitting degree of the probability density and the actual density will be worse. The different selection of smoothing parameter lead to that the error is different between probability density and the actual density. We need to choose a suitable standard to determine the error. So we take account of both Gaussian kernel functions and Mean Integrated Square Error (MISE) [9,10]; the former is used to estimate the samples distribution, and the latter is used to estimate the relationship between probability density and the actual density. MISE is a function of smoothing parameters h and the number of samples.

The SPOA algorithm is shown as follows. For a given data set $D = (X_1, X_2, \ldots, X_n)$, we can optimize the smoothing parameter for each attribute by (1), so we can get the best marginal density estimation of attributes. We can compute the posteriori probability finally.

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

where n is the number of the samples attributes, N is the number of samples, and h is the smoothing parameter.

2.2. Clustering process on the basis of SPOA.

Generally, Naïve Bayesian network clustering algorithm makes use of Gibbs sampling method to estimate the joint probability density of the samples, then uses Naïve Bayesian network to determine the category. The conditional probability density is represented by a Gaussian kernel density. The clustering process is described as shown in Figure 1.

The algorithm was originally described as follows.

We use standard Gibbs sampling method to sampling to the full conditional distribution, the complexity of the sampling increases with the increase of the number of samples attributes at exponential rates, under the Naïve Bayes star-shaped structure, it can effectively reduce the complexity of the sampling through decomposition and calculation for the joint probability, and the decomposition formula is described as shown in (2):

$$\begin{aligned} f(c|x_1, \ldots, x_n, S) &= \frac{f(c, x_1, \ldots, x_n|S)}{f(x_1, \ldots, x_n|S)} \\ &= \alpha f(c|S) f(x_1, \ldots, x_n|c, S) \\ &= \alpha f(c|S) \prod_{i=1}^{n} f(x_i|c, S) \end{aligned} \tag{2}$$

where α has nothing to do with the class variables, $f(c|S)$ is class marginal density, and $f(x_i|S)$ is attribute conditional density.

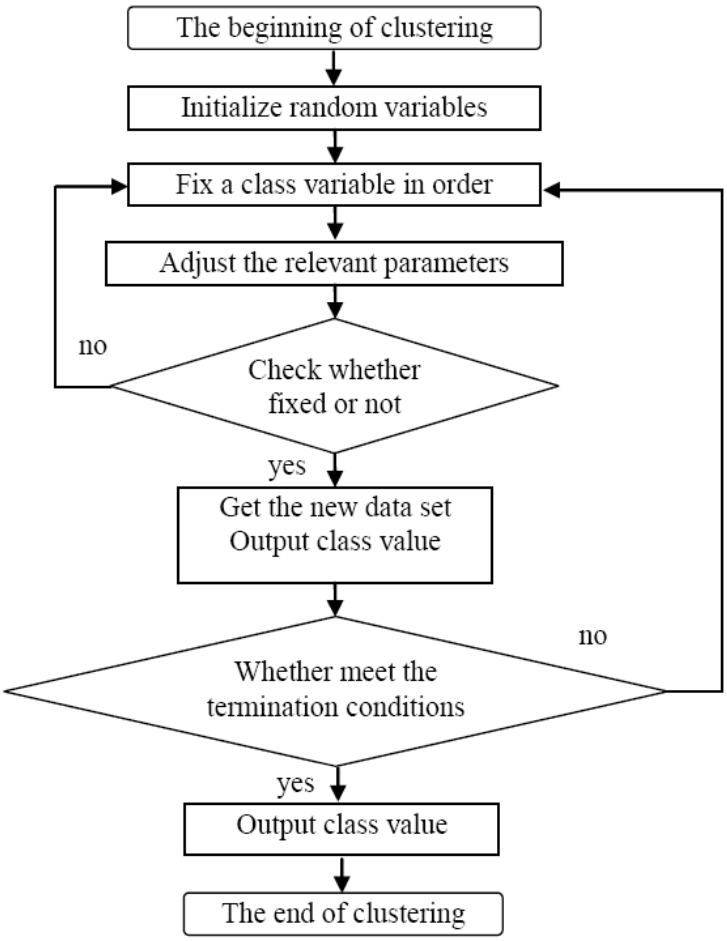

FIGURE 1. The clustering process

Clustering using Gibbs sampling is a process which updates the category of each sample in order, which proceeds according to the following steps.

1) Initialize the value of all categories C in dataset randomly.

2) Update the data of the first unknown category. Suppose c_1 is the value of category, x_i is the value of the attributes, c_1' is the value after updating, possible values for the c is c^1, $c^2, \ldots, c^r$, D is the dataset before iteration, D' is the dataset after updating. The conditional density can be estimated by Gaussian kernel function, which can be described as Equation (3).

$$p(x_{i1}|c_1, D) = \frac{1}{N(c_1)} \sum_{v \neq 1} sign(c_v) g\left(x_{i1}, \mu_i(c_v), \sigma_i(c_v)|D\right) \tag{3}$$

where $N(c_1)$ is the sample size when $c = c_1$ in the dataset, $sign(c_v) = \begin{cases} 1, & c_1 = c_v \\ 0, & c_1 \neq c_v \end{cases}$, $\mu_i(c_v) = x_{iv}$, $\sigma_i(c_v) = 1 \Big/ \sqrt{N(c_1)}$, $g\left(x_{i1}, \mu_i(c_v), \sigma_i(c_v)|D_{m-1}^k, S\right) = \frac{1}{\sqrt{2\pi}\sigma_i(c_v)} e^{-\frac{(x_{im}-\mu_i(c_v))^2}{2\sigma_i^2(c_v)}}$.

By calculating each attribute, we can get (4) by (2).

$$f(c_1|x_{11}, \ldots, x_{1n}, S) = p(c_1|D, S) \prod_{i=1}^{n} p(x_{1i}|c_1, D, S) \tag{4}$$

Calculate the posterior probability when c_1 has different values $f\left(c_1^j|x_{11},\ldots,x_{1n},S\right)$, $j=1,\ldots,r_c$, and then carry out normalization processing for (2), which can be described as Equation (5).

$$\begin{aligned} w(h) &= \frac{f\left(c_1^j|x_{11},\ldots,x_{1n},D,S\right)}{\sum_{j=1}^{r_c} f\left(c_1^j|x_{11},\ldots,x_{1n},D,S\right)} \\ &= \frac{f\left(c_1^j|D,S\right)\prod_{i=1}^{n} f\left(x_{1i}|c_1^j,D,S\right)}{\sum_{j=1}^{r_c} f\left(c_1^j,|D,S\right)\prod_{j=1}^{n} f^k\left(x_{1m}|c_1^j,D,S\right)} \end{aligned} \tag{5}$$

where $h=1,2,\ldots,r_c$.

For random number c_1, correct values for the first sample category can be obtained by (6).

$$c_1' = \begin{cases} c^1, & 0 < c_1 \le w(1) \\ c^k, & \sum_{t=1}^{k-1} w(t) < c_1 \le \sum_{t=1}^{k} w(t) \\ \cdots & \cdots \\ c^{r_c}, & c_1 > \sum_{t=1}^{r_c-1} w(t) \end{cases} \tag{6}$$

Conformance testing method is as follows.

Suppose that the class variable between adjacent two iterations are c_1^k, c_2^k, ..., c_N^k and c_1^{k+1}, c_2^{k+1}, ..., c_N^{k+1} respectively, for a given threshold $\delta_0 > 0$, if Formula (7) is workable then the result of two kinds of clustering methods is consistent, else the result of two kinds of clustering methods is inconsistent.

$$\frac{1}{N}\sum_{i=1}^{N} sign\left(c_i^k, c_i^{k+1}\right) \le \delta_0 \tag{7}$$

where $sign\left(c_i^k, c_i^{k+1}\right) = \begin{cases} 0, & c_i^k = c_i^{k+1} \\ 1, & c_i^k \ne c_i^{k+1} \end{cases}$.

For a Naïve Bayesian clustering based on Gaussian kernel density, the size of the smoothing parameters $\sigma_i(c_v)$ determines the estimation and computation of joint probability, so it has a great influence on the clustering results. In general Naïve Bayesian clustering, the same smoothing parameter ($\sigma_i(c_v) = 1/\sqrt{N(c_m)}$) is selected, so some extended approaches can achieve better clustering results. The smoothing parameter has a relation to the number of the current samples in the approach. Yet it cannot reflect the samples information comprehensively.

In order to make use of the sample information better, we proposed the approach which selects the smoothing parameter of kernel function for each attribute based on SPOA, and we can improve the effects of clustering on the basis of the estimation of the density of the sample.

3. **Experiment and Analysis.** The 10 datasets for classification can be found in the UCI repository of machine learning databases [11], we remain five percent of the class labels, and cluster the rest of the category samples with generally Naïve Bayesian and optimize the parameters of Naïve Bayesian clustering. The forecast accuracy can be got by comparing the prediction results with the actual category of samples. Table 1 shows the accuracy of the magnitude of the situation, and the latter is better than the former.

From Table 1, we can see that Naïve Bayesian clustering with optimization parameters has higher correct rate compared with the general Naïve Bayesian clustering. Because the

Table 1. Experiment results

Data set	The size of data set	The number of instances	Accuracy on NB clustering (%)	Accuracy on NB clustering based on theoptimization of smoothing parameters (%)
heart-disease	270	14	78.38	81.62
hepatitis	155	8	85.73	86.38
Soybean	307	16	97.30	98.16
tic_tac_toe	958	48	57.72	60.22
Voting-records	435	22	87.96	89.10
New_thyroid	215	11	94.56	95.60
Iris	150	8	92.93	93.22
Wdbc	569	28	91.39	92.36
Wine	178	9	94.07	96.44
Glass	214	10	51.35	52.65

general Naïve Bayesian network clustering method uses the number of the current sample as a smoothing parameter, which can only reflect partial information of the samples. The Naïve Bayesian clustering method is used to select the smoothing parameters for each attribute with the SPOA algorithm, which can describe each attribute better, so it has better clustering effect.

4. **Conclusions.** In this paper, we have investigated a Naïve Bayesian clustering method based on the optimization of smoothing parameters. On the basis of using Gaussian Kernel function, a Naïve Bayesian clustering method with SPOA has been proposed. The optimization of smoothing parameter can promote the performance of the clustering. The optimization problem is based on the idea of looking for the smallest MISE value. The comparison with NB clustering has been carried out and the clustering effect has been investigated. It shows that the proposed approach has a strong clustering accuracy. It is useful for a more detailed exploration of the clustering method in terms of the Gaussian kernel. This is a topic for our future research.

Acknowledgments. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers. This work was supported by Ministry of Education of Humanities and Social Science Research Youth Fund Projects (14YJC630188).

REFERENCES

[1] N. Dogru and A. Subasi, Comparison of clustering techniques for traffic accident detection, *Turkish Journal of Electrical Engineering & Computer Sciences*, vol.23, pp.2124-2137, 2015.

[2] C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, vol.97, no.458, pp.611-631, 2002.

[3] B. Depaire, G. Wets and K. Vanhoof, Traffic accident segmentation by means of latent class clustering, *Accident Analysis & Prevention*, vol.40, no.4, pp.1257-1266, 2008.

[4] R. Xu and D. Wunsch II, Survey of clustering algorithms, *IEEE Trans. Neural Networks*, vol.16, no.3, pp.645-678, 2005.

[5] G. H. John and P. Langley, Estimating continuous distributions in Bayesian classifier, *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, pp.338-345, 1995.

[6] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.6, no.6, pp.721-742, 1984.

[7] S. Wang et al., Restricted Gaussian classification network, *Acta Automatica Sinica*, vol.41, no.12, pp.2164-2176, 2015.

[8] S. Wang et al., Dependency extension of Naïve Bayesian classifiers based on Gaussian kernel function, *Control and Decision*, vol.30, no.12, pp.2280-2284, 2015.
[9] D. W. Scott and G. R. Terrell, Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, vol.82, no.400, pp.1131-1146, 1987.
[10] A. Perez, P. Larranagaa 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.
[11] L. Murphy and D. W. Aha, *UCI Repository of Machine Learning Databases*, http://archive.ics.uci.Edu/ml/datasets.html, 2015.