Optimal Bandwidth Selection for Density-Based Clustering

Hong Jin[1], Shuliang Wang[1,2,*], Qian Zhou[2], and Ying Li[3]

[1] State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430079, China
slwang2005@whu.edu.cn
[2] International School of Software, Wuhan University, Wuhan 430079, China
[3] School of Mathematics and Statistics, Wuhan University, Wuhan 430079, China

Abstract. Cluster analysis has long played an important role in a wide variety of data applications. When the clusters are irregular or intertwined, density-based clustering is proved to be much more efficient. The quality of clustering result depends on an adequate choice of the parameters. However, without enough domain knowledge the parameter setting is somewhat limited in its operability. In this paper, a new method is proposed to automatically find out the optimal parameter value of the bandwidth. It is to infer the most suitable parameter value by the constructed model on parameter estimation. Based on the Bayesian Theorem, from which the most probability value for the bandwidth can be acquired in accordance with the inherent distribution characteristics of the original data set. Clusters can then be identified by the determined parameter values. The results of the experiment show that the proposed method has complementary advantages in the density-based clustering algorithm.

Keywords: Density-based clustering, Bayesian posterior probability estimation, Optimal bandwidth selection.

1 Introduction

The rapid advance in spatial data acquisition, transmission and storage results in the growth of vast computerized datasets at unprecedented rates. For numerous data–based applications, efficient methods of data analysis can make use of the information implicitly contained in the data [1]. As a primary means of data analysis, cluster analysis helps to understand the natural grouping and structure in a dataset [2]. The clustering algorithms can be regarded as an approach to get insight into the distribution of a data set. According to the different criteria of similarity measurement and clustering evaluation, the commonly used clustering algorithms may be based on partition, hierarchy, density and grid [3]. The density-based algorithms are to discover the clusters of arbitrary shape and it is easy to be extended. Each cluster corresponds to a relatively dense area of data distribution, by looking for low-density regions separated by the connectivity of high-density area [4].

* Corresponding author.

J. Xu et al. (Eds.): DASFAA Workshops 2011, LNCS 6637, pp. 156–167, 2011.

Meanwhile, it is not sensitive to the existence noise. However, the quality of its clustering result mainly depends on the input parameters. DENCLUE is such a representative.

DENCLUE (DENsity based CLUstEring) is a generic clustering algorithm based on kernel density estimation. By means of adjusting the bandwidth of the kernel function, the density-based clustering algorithm is able to efficiently get insight into the distribution of a data set. Since the effectiveness of kernel density estimation depends on the selection of bandwidth, the algorithm is supposed to optimize the selection of the bandwidth in order to improve the accuracy. In this paper a new approach is proposed to optimize the bandwidth selection by using Bayesian inference. So the appropriate clustering results can be acquired more quickly in accordance with the inherent distributed characteristics of the original data set. Theoretical analysis and experimental results show that the approach has good clustering quality and computing performance, and the parameter selection is more objective with good robustness.

The rest of the paper is organized as follows. In section 2, the related principles are introduced such as kernel density estimation and Bayesian inference. And it illustrates how the parameter estimation model can be constructed with respect to the above mentioned principles. In section 3, it is the process of the proposed algorithm that includes its theoretical foundations such as Bayesian posterior density estimation and MCMC (Markov Chain Monte Carlo) method as well as the rationality of the parameter setting method. In section 4, an experimental evaluation is provided. For the experiments, analog data is used as the related paper commonly used. The results are concluded in section 5 along with some issues for future work.

2 Related Principles

Density-based clustering is to model the distribution density of dataset as the sum of the influences of individual data objects by using the functions under kernel density estimation [5]. For kernel density estimation, the contribution of each point to the overall density function is expressed by an influence or kernel function [6]. The overall density function is simply the sum of the influence functions associated with each data point.

2.1 Basic Idea of Density Based Clustering Algorithm

DENCLUE is a clustering algorithm on a group of density distribution functions [7]. Given a space Ω containing dataset $D=\{x_1, x_2, \ldots\ldots, x_n\}$ in d -dimensional space, the basic idea of the algorithm is followed.

(1) The kernel density estimator of the overall density function

Assume that the probability distribution associated with each observed data point uniformly distributes in different dimensions. $\forall x \in \Omega$, the probability density can be estimated as equation (1).

$$\hat{f}^{D}(x) = \frac{1}{nh^{d}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

Where, K(x) is the kernel function in terms of product kernel that will be explained in the subsequent part. It generally chooses a symmetric density function that has single peak at the origin such as square wave function and Gaussian function. Constant h is called the bandwidth of the kernel function [8]. In accordance with the above hypothesis, the bandwidth value in different dimensions can be viewed as the same.

(2) Center-Defined Cluster
Given a density-attractor x^*, if there exists C⊆D satisfying the condition that $\forall x \in C$, x is density attracted by $x*$ and $f^D(x^*) \geq \xi$ where ξ is the preset parameter noise threshold, and C is called the cluster centered with x^*.

(3) Arbitrary-Shape Cluster.
An arbitrary-shape cluster for the set of density-attractors X is a subset $C \subseteq D$ where

① $\forall x \in C, \exists\, x^* \in X$: $f^D(x^*) \geq \xi$, x is density-attracted to x^*, and
② $\forall\, x_i^*, x_j^* \in X(i \neq j)$: $\exists$ a path $P \subset Q$ from x_i^* to x_j^* with $\forall y \in P$: $f^D(y) \geq \xi$.

Obviously, there are two important preset parameters in the algorithm such as the bandwidth and the noise threshold. The bandwidth affects the efficiency of the overall density function estimator as well as the number of the density-attractors or the clusters. Let h_{max} represents the maximum of the bandwidth under the condition that the density function $f^D(x)$ has only one density-attractor. While h_{min} represents the minimum of the bandwidth under the condition that the density function $f^D(x)$ has n density-attractors. Each value in the interval $[h_{min}, h_{max}]$ corresponds to an appropriate clustering result about the dataset [8]. Consequently, the value of the bandwidth can be selected from the interval $[h_{min}, h_{max}]$ in order to naturally acquire hierarchy clustering result. Considering the optimal bandwidth, it is acknowledged that the bandwidth value in the maximal interval $I \subset [h_{min}, h_{max}]$ which keep the number of density-attractors remain constant corresponds to an appropriate clustering result. When the bandwidth h is ready, the clustering result can be determined by the noise threshold ξ.

2.2 Parameter Estimation Model

It is important to set the parameter for the density-based clustering algorithm. Here is the bandwidth to be estimated under Bayesian Theorem. First, the kernel density estimation is used to equationte the likelihood function. Then, the parameter estimation is modeled by choosing an empirical prior density function, along with MCMC (Markov Chain Monte Carlo) method to sample the parameter space.

(1) Bayesian Inference
Bayesians views unknown parameter values as random quantities using probability distributions to represent its uncertainty [9].

Let D represent the observed data and θ represent the model parameters. The joint probability distribution P(D, θ) over all random quantities is equation (2), in which θ is able to be multi-dimensional [10].

$$P(D,\theta) = P(\theta)P(D|\theta) \tag{2}$$

Where we call P(θ) the prior density and P(D|θ) the likelihood function. Once given the observed data D, the posterior distribution of the parameter θ can be acquired as equation (3) according to Bayes Theorem.

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d_\theta} \tag{3}$$

It represents the distribution of θ condition on the observed data D. Since the denominator of equation (3) is not relevant to θ, it can be simplified as being proportional to the prior times the likelihood and formalized as equation (4).

$$P(\theta|D) \propto P(\theta)P(D|\theta) = P(\theta)L(\theta; D) \tag{4}$$

Seen from equation (4), the posterior is a conditional distribution for the model parameters given the observed data.

(2) Parameter Space Sampling

By obtaining samples x_t(t=0,1,…,n) from the distribution $P(x)$, various features of the distribution $P(x)$ can be calculated. For a Bayesian, x is comprised of model parameters and $P(x)$ is called a posterior distribution [9]. From equation (3), with MCMC it has to know the distribution of x up to the constant of the normalization [11]. The notation t expresses an ordering or sequence to the random variables in MCMC. When x_t are independent, the approximation can be made as accurate as needed by increasing n. Under the condition that x_t are not independent, it doesn't limit its usefulness as long as they are sampled from the entire domain of $P(x)$ in correct proportions [12]. By means of constructing a Markov Chain taken $P(x)$ as its stationary distribution, this can be resolved.

In the MCMC methods, the key is how to construct chains that the stationary distribution is the interested one. In this paper, the random-walk metropolis- hastings sampler are chosen to construct the Markov Chain when generating a sequence samples of the target distribution referred to as the posterior distribution on the model parameter.

3 Density-Based Clustering Algorithm Using the Optimal Bandwidth Selection

The effectiveness of the density-based algorithm depends on the subjective preset of the two parameters bandwidth and noise threshold. The choice of the bandwidth has significant impact on the estimation result of the overall density function causing difference with respect to the number and the pattern of the clusters. If the choice of the bandwidth is closer to the original distribution of the data set, the natural clustering results and the number of the categories can be acquired. Suitable value of the noise threshold makes the algorithm focusing on the calculation of high-density area in order to decrease the computing time.

3.1 The Structure of the Algorithm

Regarding the bandwidth as the parameter to be estimated, the parameter estimation is modeled by using Bayesian method and MCMC sampling. Such the estimated

bandwidth may make the overall density function better fit the inherent distribution of the original data set. When the data space is multi-dimensional, it is easy to be further extended [13]. According to the estimated bandwidth, the noise threshold can be subjectively preset before starting the clustering algorithm. With the estimated bandwidth and the existing clustering results, the noise threshold can be further adjusted to acquire a more accurate clustering pattern. Besides these, during the process of searching density-attractors, it uses conjugate gradient hill-climbing method instead of the gradient hill-climbing method to accelerate the convergence speed. The structure of the proposed approach is as Fig.1 shows.

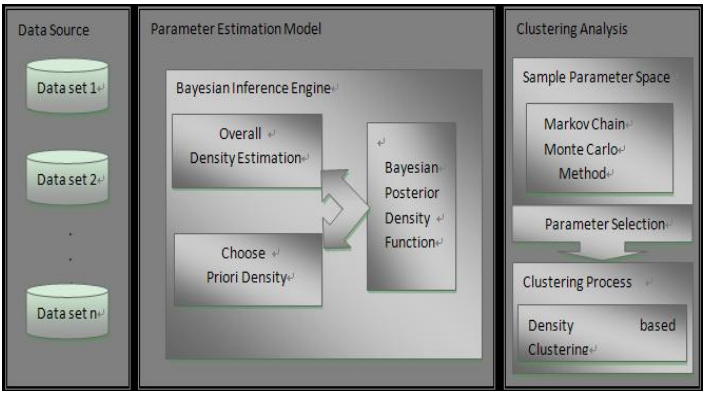

Fig. 1. Density-based Clustering Algorithm Using the Optimal Bandwidth Selection

3.2 Optimal Bandwidth Selection Model

The key step of the approach is how to select the optimal bandwidth value. In this section, the modeling process of the parameter estimation is illustrated by Bayesian Theorem and MCMC method. It first discusses the typically calculated form logarithm of the likelihood function for the parameter to be estimated. Then, with the assumed parameter prior density function, the Bayesian posterior density function of the parameter can be constructed. Using the MCMC simulations to sample the parameter space, the expected value of the parameter can be obtained.

Note that the bandwidth matrix can be restricted to a class of positive definite diagonal matrix with the corresponding kernel function known as a product kernel [14]. When choosing a full bandwidth matrix, it is identical to pre-rotating the original data with an optimal amount and then still using a diagonal bandwidth matrix. Consequently, the general form of kernel density estimator can be transformed to be as equation (5) shows.

$$\hat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{h_j}K\left(\frac{x - X_{ij}}{h_j}\right) \tag{5}$$

In particular, $K(\cdot)$is univariate kernel density function associated with product kernel, and h_j represents the different bandwidth value in each dimension.

According to the above kernel density estimator of *f(x)*, the log pseudo-likelihood function for the bandwidth matrix *H* can be got as equation (6).

$$L\left(x_1, x_2, ..., x_n \middle| H\right) = \sum_{i=1}^{n} \log \hat{f}_{H,i}(x_i) \tag{6}$$

Where the leave-one-out estimator is as equation (7):

$$\hat{f}_{H,i}(x_i) = \frac{1}{(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} \prod_{m=1}^{d} \frac{1}{h_m} K\left(\frac{x_i - X_{jm}}{h_m}\right) \tag{7}$$

Regarding the non-zero elements of the bandwidth matrix as parameters, the posterior density of the parameters based on the log pseudo-likelihood function can be obtained according to equation (4).

Assume that the prior density of each non-zero component of *H* is as probability distribution function (8) shows:

$$P(h_j) \propto \frac{1}{1 + h_j^{\,2}} \tag{8}$$

It is proved to be effective that the above priors can put low probability on the region of the parameter space where the likelihood function is flat [15]. We can get the joint prior of all elements of H in the product form of these marginal priors. Then, using Bayes Theorem, the logarithmic posterior of H is as equation (9) shows.

$$P(H|D) \propto \sum_{j=1}^{d} \log P(h_j) + \sum_{i=1}^{n} \log \hat{f}_{H,i}(x_i) \tag{9}$$

In case of a diagonal bandwidth matrix, all elements of H can be sampled through the Metropolis-Hastings algorithm with the acceptance probability computed through (9). Meanwhile, the corresponding kernel function known as a product kernel.

4 Case Study

To demonstrate the effectiveness and efficiency of the proposed method, an experiment is performed using synthetic data. In this section, it starts the algorithm described in section 3 via several bivariate data sets. Given a dataset generated from simulation, we sample the diagonal bandwidth matrix from its corresponding posterior density defined in equation (9) using the random-walk metropolis-hastings algorithm.

After the sample paths of *H* for each dataset are obtained, the posterior mean acts as an estimation of optimal bandwidth are calculated. With the estimated bandwidth, the density based clustering algorithm is initialized. And for another parameter the noise threshold is subjectively set to a certain value which can be adjusted by the existed clustering result.

4.1 The Procedure of Optimal Bandwidth Selection

Taken two-dimensional dataset as an example, the process of optimal bandwidth selection can be instantiated as follows. According to the above information, the

parameter estimation model by Bayesian method and MCMC sampling can be constructed as equation (9). By means of simulation, it provides three data sets that are commonly used in the relative papers. In the experiment, the three synthetic sample databases depicted in Fig.2 are used.

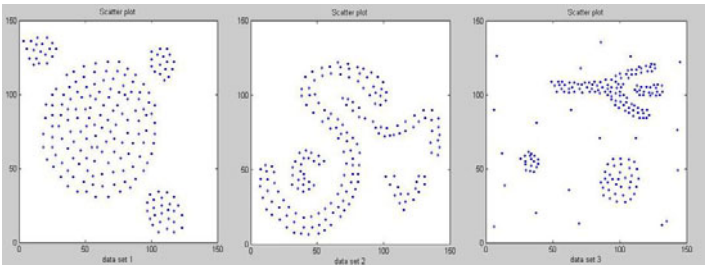

Fig. 2. The Original Data Sets

Therefore, the accuracy of the proposed algorithm is evaluated by visual inspection. Judging from the morphological, for sample dataset 1 there are four ball-shaped clusters with significantly different sizes. For sample dataset 2 it contains four clusters of non-convex shape. While in sample dataset 3 it has four clusters of different shape and size with additional random noise. In order to clearly distinguish the different clusters in the clustering results, it visualizes each cluster found by different color.

As for each dataset the optimal bandwidth are calculated by the parameter estimation model. With respect to the corresponding data set, the optimal bandwidth can be acquired by the expected value of the sample points in the generated Markov Chains.

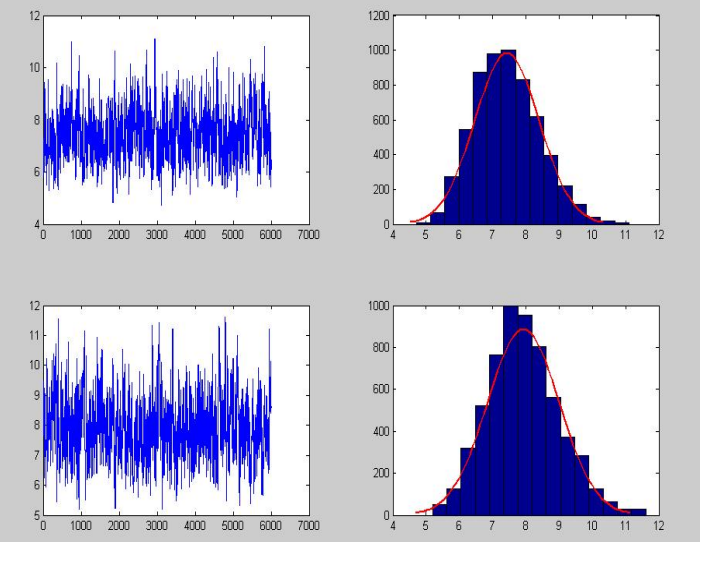

Fig. 3. The Markov chain and Statistic histogram of dataset 1

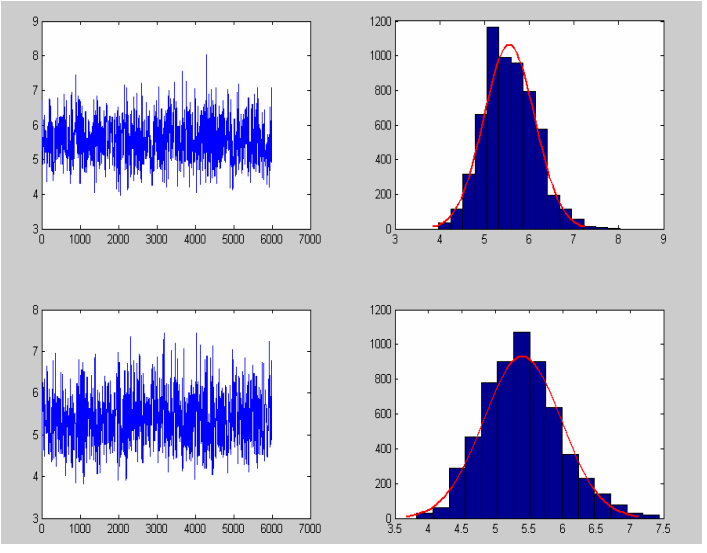

Fig. 4. The Markov chain and statistic histogram of dataset 2

For dataset 1, the left panel in Fig.3 shows the Markov Chains in two dimensions, while the right panel represents the Statistic Histogram to the relevant dimension.

On the basis of the calculation result, the optimal bandwidth for data set1 is equal to 7.9. As shown in Fig.3, it indicates that the bandwidth values approximately the same in different dimensions. The sample values for the bandwidth in both dimensions centralized in the interval [6, 10, 16]. The corresponding Statistic Histograms reflect the most likely values for the bandwidth.

Taking the arbitrary shape clustering into consideration, an experiment on dataset 2 is also given. The same as stated above, in Fig.4 the simulated result is shown. It can be seen that the bandwidth values in different dimensions are fairly close to each other.

In accordance with the calculated result, the optimal bandwidth for data set2 is equal to 5.4. Seen from the corresponding statistic histogram, the values around 5.5 is the most frequently values sampled in the parameter space. And the possible values for the bandwidth in both dimensions distribute in interval [4, 6].

Especially, in the third case a specific realization of data set3 with random noise is provided. The simulated result is as Fig.5. Obviously, the bandwidth values in different dimensions exactly not distribute in the approximate interval which is differ from the above two cases. It just reflects that the overall density function of this dataset is affected by the random noise.

According to the simulation results, the bandwidth values in different dimension are respectively equal to 13.2 and 5.8. It demonstrates the inherent distribution of the original data. Here the related statistic histograms represent the most probability values of the bandwidth in different dimensions. During the process of clustering analysis, a discussion on both dimensions will be given.

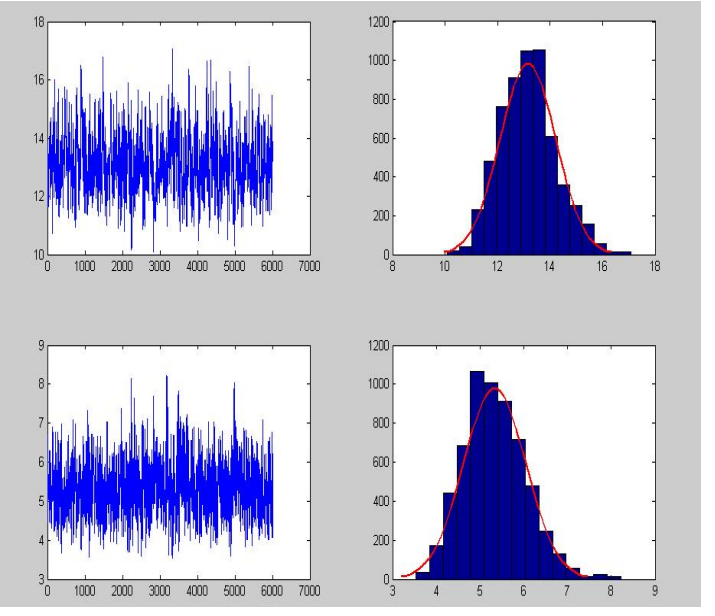

Fig. 5. The Markov chain and statistic histogram of dataset 3

4.2 Clustering Analysis

Density-based clustering algorithm needs two parameters such as the noise threshold and the bandwidth. Together with the optimal bandwidth acquired by the above simulation, the preset noise threshold value may be used to initialize the density-based clustering algorithm. Comparing to the selection of bandwidth, the selection of noise threshold is less important when determining the clustering results.

The calculated bandwidth value for each dataset can be used to start the clustering process. For data set1, the noise threshold ξ=2 and the optimal bandwidth σ=7.9, the clustering result of which is in Fig.6. Obviously, the clusters found by this approach

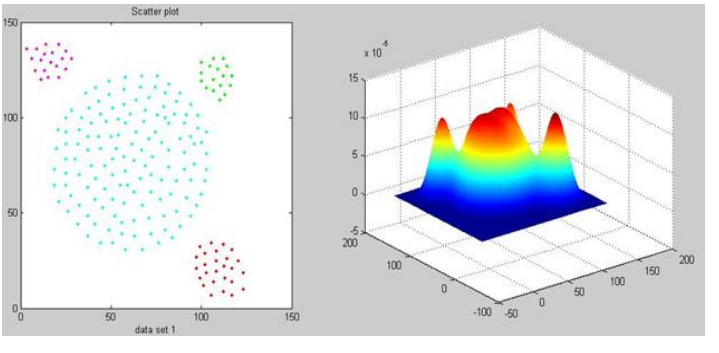

Fig. 6. The clustering result of dataset 1

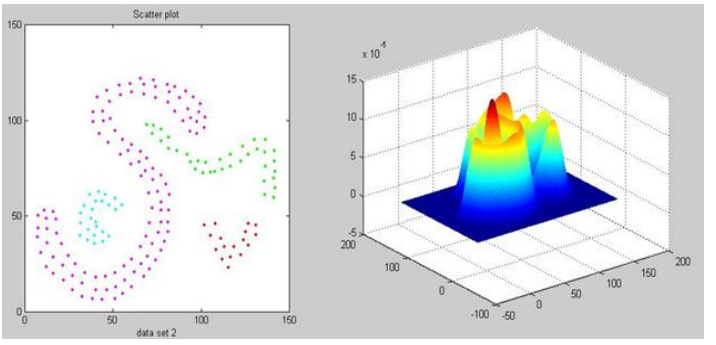

Fig. 7. The clustering result of dataset 2

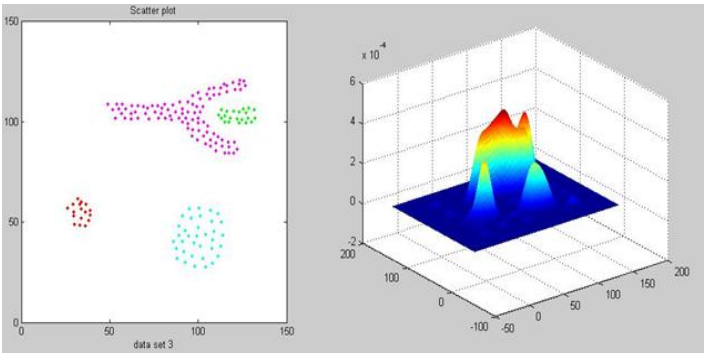

Fig. 8. The clustering result of dataset 3

well reflect the distribution of the original data set. Meanwhile, the density function on the basis of kernel estimator for dataset 1 is also given.

The experiment process on dataset 2 mainly reflects the applicability of arbitrary shape clusters. According to the above bandwidth value calculated for dataset 2, the clustering result can be acquired as Fig.7 shows where the preset parameter ξ=2 and σ=5.6. For visualization, the density function based on kernel estimator is given simultaneously.

When there is noise, the experiment on data set3 is given. Based on the above analysis, the bandwidth values in different dimensions are not the same. One is 5.3 and the other is 13.2, which indicates that the distributed characteristic of the original dataset in two dimensions are heterogeneous. In consistent with the basic idea of density-based clustering, the bandwidth in two dimensions is regarded as the same. Therefore, two cases are respectively considered in simulation. Under the parameter values ξ=2 and σ=13.2, it can be seen that the original dataset is divided into four clusters as Fig.8 shows. While under the conditions that ξ=2 and σ=5.3, the original dataset is divided into several clusters affected by the noise which are not given here. To maintain consistency, the density function of dataset 3 is given too.

Generally speaking, the effect and efficiency of density-based clustering algorithm depends on the carefully selection of the parameter value. The choice of the related parameters has reliance on domain knowledge or subjectivity. Nevertheless, for the proposed approach in this paper, one of the important parameters such as the bandwidth can be automatically adjusted in accordance with the original dataset.

5 Conclusions

In this paper, a cluster analysis method was proposed on the density-based clustering algorithm. By means of treating the elements of the bandwidth matrix as parameters to be estimated, it constructed a parameter estimation model by Bayes Theorem. It provides MCMC algorithms to sample the parameter space. Numerical simulations show that the resulting bandwidths are superior and it has no increased difficulty as the dimension of data increases. Additionally, its main advantage is that the bandwidth selected by this parameter estimation model can more accurately reflect the distribution of the original dataset. Though Denclue is not fundamentally a grid-based technique, it does employ a grid-based approach to improve efficiency. The length for the grid-base is amount to the value of the bandwidth. In the further research, the length of the grid-base can be different in the corresponding dimension on the basis of the bandwidth matrix. Obviously, under this condition it will be more efficient and accurate in the clustering result in consistent with the original dataset.

Acknowledgements. This paper is supported by National 973 (2007CB310804), National Natural Science Fund of China (6074300), Best National Thesis Fund (2005047), and Natural Science Fund of Hubei Province (CDB132).

References

1. Ankerst, M., Breuing, M.M., Kriegel, H.P.: OPTICS: ordering points to identify the clustering structure. In: Proc. of the 1999 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM Press, New York (1999)
2. Hinneburg, A., Keim, D.A.: An efficient approach to clustering in large multimedia databases with noise. In: Proc of the 4th International Conference on Knowledge Discovery and Data mining, pp. 58–65. AAAI Press, Menlo Park (1998)
3. George, K., Han, E.H., Kumar, V.: CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. IEEE Computer 27(3), 329–341 (1999)
4. Ester, M., Kriegel, H.P., Sander, J.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc.of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press, Menlo Park (1996)
5. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education, London (2006)
6. Gentle, J.E.: Computational Statistics. Springer, New York (2001)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
8. Gan, W.Y., Li, D.Y.: Hierarchical Clustering based on Kernel Density Estimation. Journal of System Simulation 16(2), 302–309 (2004)

9. Dellaportas, P., Forster, J.J., Ntzourfras, I.: On Bayesian model and variable selection using MCMC. Statistic and Computing 12(2), 27–36 (2002)
10. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. Chapman&Hall, London (2004)
11. Chen, M.H., Shao, Q.M., Ibrahim, J.G.: Monte Carlo Methods in Bayesian Computation. Springer, New York (2000)
12. Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Introducing Markov chain Monte Carlo. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.T. (eds.) Markov Chain Monte Carlo in Practice, pp. 1–19. Chapman and Hall, London (1996a)
13. Terrell, G.R., Scott, D.W.: Variable kernel density estimation. Annals of Statistics (20), 1236–1265 (1992)
14. Duong, T., Hazelton, M.L.: Plug-in Bandwidth Selectors for Bivariate Kernel Density Estimation. Journal of Nonparametric Statistics (15), 17–30 (2003)
15. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, Visualization. Wiley, New York (1992)
16. Fang, M., Wang, S.L., Jin, H.: Spatial Neighborhood Clustering Based on Data Field. In: Cao, L., Feng, Y., Zhong, J. (eds.) ADMA 2010, Part I. LNCS, vol. 6440, pp. 262–269. Springer, Heidelberg (2010)