Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 72 (2004) 1–8
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Clustering based on kernel density estimation: nearest local maximum searching algorithm

Wei-Jun Wang, Yong-Xi Tan, Jian-Hui Jiang, Jian-Zhong Lu, Guo-Li Shen, Ru-Qin Yu*

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, People's Republic of China

Received 28 February 2003; received in revised form 16 February 2004; accepted 29 February 2004

Available online 26 April 2004

Abstract

Nearest local maximum searching algorithm (NLMSA), an unsupervised clustering algorithm based on kernel density estimation, is proposed. It is designed for detecting inherent group structures with arbitrary shape clusters among multidimensional measurement data without any a priori information. The algorithm is named after its clustering mechanism of converging data points to their corresponding nearest local maxima of the data's density estimate along the ascending gradient direction. Two simulated data sets and two real data sets are employed to validate the performance of the method. A comparison between the clustering results obtained from the proposed algorithm and the K-means cluster analysis shows that the NLMSA possesses quite satisfactory performance.

Keywords: NLMSA; Pattern recognition; Cluster analysis; Kernel density estimation; Local optimization

1. Introduction

For decades, as the hybridization of modern multivariate analysis methods and computer techniques, pattern recognition methods have been the powerful and essential tools to process multidimensional measurement data obtained by utilizing modern analytical instruments [1,2]. It is well known that the area of pattern recognition is made up of two domains: the supervised pattern recognition and the unsupervised pattern recognition [3]. Cluster analysis is considered as a major category of the unsupervised pattern recognition technology [4]. Conceptually, clustering techniques can be divided into several classes, including hierarchical, optimization-partitioning, density-based and others [5].

The optimization-partitioning methods are most popular in chemical applications, e.g., K-means cluster analysis [6], Kohonen neural network [7], etc. To most of these approaches, some criterion functions are formulated and the number of clusters among data is assumed in advance, then the data are divided into partitions subject to the predefined cluster number by reaching the optimal criterion values. Normally, since the nature of the grouping structures among the data is unknown before processing, the assumed cluster number may be unreliable and hence the partitioning methods would not predict the structures of the data very well. Furthermore, a majority of the partitioning methods merely tend to be excellent when they are applied to deal with convex clusters, generally, circle clusters or spherical clusters. They are not effective to discriminate clusters with other types of appearance, for example, ellipsoidal clusters or concave shape clusters, which are deemed to be natural [8] and frequently seen. In a word, the availability of the partitioning methods is restrained to most "real-world" chemical pattern recognition applications.

Unlike the partitioning methods, which are concentrated on measuring and comparing the distances among objects in data, density-based methods discriminate clusters mainly according to the probability distribution in the data. Regions with high densities of objects are recognized as clusters, and areas with sparse distributions of objects are boundaries to keep clusters divided from one another. Density-based methods, particularly, the kernel-density methods are expected to perform clustering procedure without any a priori knowledge concerning the data and are able to identify clusters with any type of shape. Coomans and Massart [9] proposed such a kernel-density based approach, the CLUPOT. Recently, Daszykowski et al. [10] developed another one, the NP.

* Corresponding author. Tel./fax: +86-731-882-2782.
E-mail address: rqyu@hunu.net.cn (R.-Q. Yu).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.02.006

In the present paper, we would like to introduce a new kernel density based clustering algorithm. In a conventional procedure, the centers of clusters are selected from data points (samples) first, and then the cluster belongingness of the data points is investigated one by one. In the newly proposed algorithm, a gradient local maximum search is launched starting from each sample to attain a local maximum of the given data's Gaussian kernel density estimate. This local maximum is the nearest one departing from the sample concerned along the ascending gradient direction and is considered as the representative of the cluster to which the sample is belonging. It is expected that the samples belong to the same cluster would converge to the solution of the same local maximum. As a result, the clusters are identified as all samples converge to their corresponding clusters. We entitle the proposed approach with the name of "nearest local maximum searching algorithm" (NLMSA), for short, the NLMSA. The reliability curve [9] is employed to determine the proper smoothing parameters and to evaluate the clustering results of the proposed method. The NLMSA has been applied to two simulated data sets with ellipsoidal shape and two real data sets with arbitrary shape and compared with the K-means cluster analysis. Comparison between the clustering results show that the proposed method outperforms the K-means cluster analysis, not only in exempting the need of a priori information about the data, but also in better classification precision and stability.

2. Theory

2.1. Kernel density estimation

To implement a density-based clustering method, one ought to estimate the probability density of the data. In most of practical pattern recognition problems, the distribution of data is multimodal and can hardly be classified into any type of classical distribution. Nonparametric density estimation methods are frequently employed because they can deal with the multimodality and do not need to assume the form of the distribution. The kernel density estimation is one kind of the nonparametric methods. Consider N samples (data points) in the d-dimensional space, $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_N$, where the sample vector is $\boldsymbol{x}_i=[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{id}]$, $i=1, 2, \ldots, N$. The probability density of the data is given by the kernel density estimate:

$$P(\boldsymbol{x}) = \frac{1}{Nh^d}\sum_{i=1}^{N}\mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right) \tag{1}$$

In Eq. (1), the kernel function is:

$$\mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right) \tag{2}$$

Where h is the smoothing parameter, also called the bandwidth or the window width. A kernel function represents the contribution of an individual sample $\boldsymbol{x}_i$ to the overall density. In principle, a variety of kernel functions are available, e.g., Gaussian, triangle and rectangle kernel, to name just a few. Here, we choose the Gaussian function as the kernel.

$$\mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right) = \frac{1}{(\sqrt{2\pi})^d}\exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right] \tag{3}$$

And then the density is:

$$P(\boldsymbol{x}) = \frac{1}{Nh^d(\sqrt{2\pi})^d}\sum_{i=1}^{N}\exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2}\right] \tag{4}$$

We prefer use the Gaussian kernel for two reasons. First, the Gaussian kernel is smooth and wherefore the estimated density must be smooth as the overall density is a sum of all individual Gaussian kernels. Second, only one coefficient, the smoothing parameter h (also called the bandwidth or the window width) is required to determine the kernel function.

2.2. Nearest local maximum searching algorithm

2.2.1. Principle of the NLMSA

The principle of the proposed method is illustrated in the Fig. 1 with a one-dimensional example. As one can observe from the figure, all data points could be classified into two clusters based on the density. Each cluster includes the data points within its corresponding hump of the density estimate. If one let a data point which belong to the cluster 1 be the starting point to search the nearest local maximum along the ascending gradient direction, one would find out the Local Max 1, the nearest local maximum departing from the data point. By analogy, the Local Max 2 would surely be found from a data point of the cluster 2 by searching in the same way. Therefore, the Local Max 1 could be regarded as the representative of the cluster 1 because all samples of the cluster 1 would converge to the point. Likewise, the Local Max 2 is the representative of the cluster 2. From the aforementioned reasoning, one can deduce that, to a given data set, the ascending gradient searching from every data points would accomplish the clustering. The number of the clusters would be determined by finding out how many maxima exist, and the cluster belongingness of a data point would be ascertained by examining which local maximum would be attained from the data point.

Noteworthily, deciding on the proper value of the smooth parameter h is vital. The smoothing parameter completely controls the shape and hence the essence of the density. If too small a value of the bandwidth is applied, the kernel would be rather narrow and, as a result, the estimated density would be very spiky and too many clusters would be found. If the width is too large, kernels are overlap and produce a single cluster density that may mask the data structure. In order to avoid both of the two circumstances mentioned above and to determine the significant clustering results, we utilize the reliability curve [9], which is obtained by plotting out the smoothing parameter versus the cluster number that are discovered under different values of the smoothing parameter.

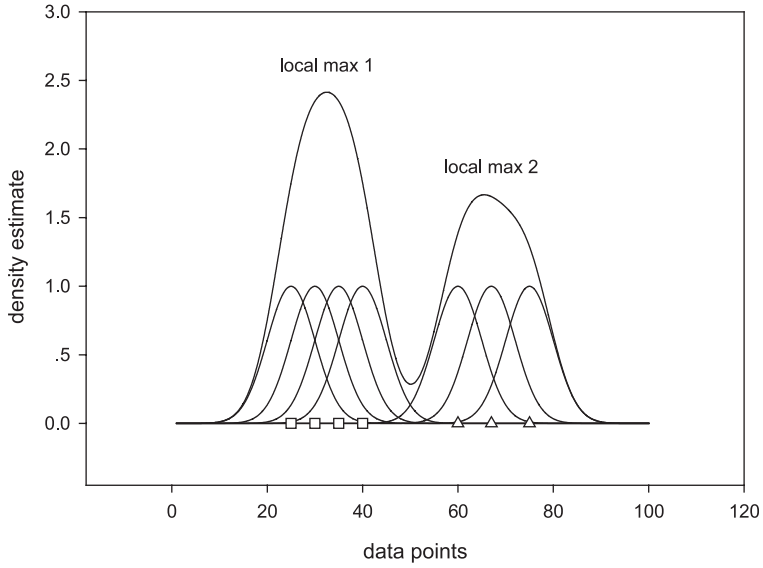

Fig. 1. One-dimensional example of the clustering mechanism of the proposed method. (□) Data points in cluser1, (△) data points in cluster 2.

2.2.2. Procedure of the NLMSA

Gradient search methods are proved to be not likely to attain the global optimum in many optimization applications, because they are easy to sank into local optima. In this paper, this drawback is taken as an advantage and we utilize it for designing the proposed methodology. Consider the data set containing N samples (data points), $\boldsymbol{x}_1$, $\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N$ in the d-dimensional space where the sample vector is $\boldsymbol{x}_i=[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2},\ldots,\boldsymbol{x}_{id}]$, $i=1, 2,\ldots,N$. The NLMSA is carried out orderly in the following steps.

(1) A smoothing parameter h is given. Choose the sample $\boldsymbol{x}_k$ to begin a gradient search, set the iteration number: $k=1$, $k=1, 2,\ldots,N$.
(2) Given the start point $\boldsymbol{x}_k^{(1)}=\boldsymbol{x}_k$ and the convergence error $e>0$, set the iteration number of a single gradient search: $m=1$.
(3) Calculate the search direction $\boldsymbol{d}_k^{(m)}=\nabla P(\boldsymbol{x}_k^{(m)})$, where the gradient $\nabla P(\boldsymbol{x}_k^{(m)})$ is the positive derivative (see Eq. (5)).

$$\nabla P(\boldsymbol{x}_k^{(m)}) = \frac{-1}{Nh^{d+2}(\sqrt{2\pi})^d}\sum_{i=1}^{N}\exp\left[-\frac{\|\boldsymbol{x}_k^{(m)}-\boldsymbol{x}_i\|^2}{2h^2}\right] \times(\boldsymbol{x}_k^{(m)}-\boldsymbol{x}_i) \quad (5)$$

(4) When $m=1$, if $\|\boldsymbol{d}_k^{(1)}\|\le e$, go to step (6); otherwise, set $\lambda^{(1)}=1$ let $\boldsymbol{x}_k^{(2)}=\boldsymbol{x}_k^{(1)}+\lambda^{(1)}\boldsymbol{d}_k^{(1)}$ and go to step (3); When $m\ge 2$, if $\|\boldsymbol{d}_k^{(m)}\|\le e$, go to step (6); otherwise, set the pace

$$\lambda^{(m)}\begin{cases}1.05\lambda^{(m-1)} & P(\boldsymbol{x}_k^{(m)})>P(\boldsymbol{x}_k^{(m-1)})\\ 0.7\lambda^{(m-1)} & P(\boldsymbol{x}_k^{(m)})\le P(\boldsymbol{x}_k^{(m-1)})\end{cases};$$

we set the pace λ by empirical values to avoid too heavy computation burden, which is used to happen when use the linear search approach to determine the λ.
(5) Set $\boldsymbol{x}_k^{(m+1)}=\boldsymbol{x}_k^{(m)}+\lambda^{(m)}\boldsymbol{d}_k^{(m)}$ and $m:=m+1$, then go to step (3).
(6) Obtain the solution vector of the nearest local maximum by searching from the data point $\boldsymbol{x}_k$: $\boldsymbol{y}_k=\boldsymbol{x}_k^{(m)}$, $\boldsymbol{y}_k=\boldsymbol{x}_k^{(m)}$, $\boldsymbol{y}_k=[\boldsymbol{y}_{k1}, \boldsymbol{y}_{k2}, \ldots, \boldsymbol{y}_{kd}]$, the vector is considered as the new representation or the sample $\boldsymbol{x}_k$. Set $k:=k+1$, go to step (1) to repeat the loop.
(7) Finally, N vectors, the new representations of all samples are discovered: $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$. To any two vectors $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ ($\forall a,b=1, 2, \ldots, N$ and $a\neq b$), if $\|\boldsymbol{y}_a-\boldsymbol{y}_b\|\le\varepsilon$, it means that the same local maximum is reached from their corresponding samples, $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$. And the two samples are classified into the same cluster. Based on the criterion mentioned above, the samples are classified into different clusters.

3. Experimental

Two simulated data sets and two real data sets were used to evaluate the NLMSA method. Each of the data sets has been processed by both the proposed method and a typical partitioning method, the K-means cluster analysis, and the clustering performances are compared. In order to validate the clustering results visually, the simulated data sets are generated to be two-dimensional.

3.1. Simulated data set 1

Simulated data set 1 is comprised of three Gaussian clusters, each consisting of 50 samples, which are independent and ellipsoidal distributed. Clusters 1, 2 and 3 have

different expected means of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$, and covariance matrices of Σ_1, Σ_2, Σ_3, respectively.

$$\boldsymbol{u}_1 = (5.5, 0)^t \quad \sum_1 = (0.2, 1)^t \times \mathrm{diag}(1, 1)$$

$$\boldsymbol{u}_2 = (1.1, 0)^t \quad \sum_2 = (0.2, 1)^t \times \mathrm{diag}(1, 1)$$

$$\boldsymbol{u}_3 = (3.3, 2.9)^t \quad \sum_3 = (0.2, 0.9)^t \times \mathrm{diag}(1, 1)$$

Where the cluster 3 is rotated to be reverse relative to the clusters 1 and 2. The diag (1,1) is the diagonal matrix with diagonal elements 1 and 1, and the rest may be deduced by analogy.

3.2. Simulated data set 2

Simulated data set 2 is made up of two clusters. Each follows normal distribution and consists of 100 samples with ellipsoidal covariance. The two clusters have different expected means of $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and covariance matrices of Σ_1, Σ_2, respectively.

$$\boldsymbol{u}_1 = (2, 2)^t \quad \sum_1 = (0.05, 0.3)^t \times \mathrm{diag}(1, 1)$$

$$\boldsymbol{u}_2 (3.5, 2)^t \quad \sum_2 = (0.2, 1.2)^t \times \mathrm{diag}(1, 1)$$

3.3. Chinese tea data set

Liu et al. [11] have studied the pattern recognition of 31 Chinese tea samples by using hierarchical cluster analysis and principal component analysis. All the samples belong to three

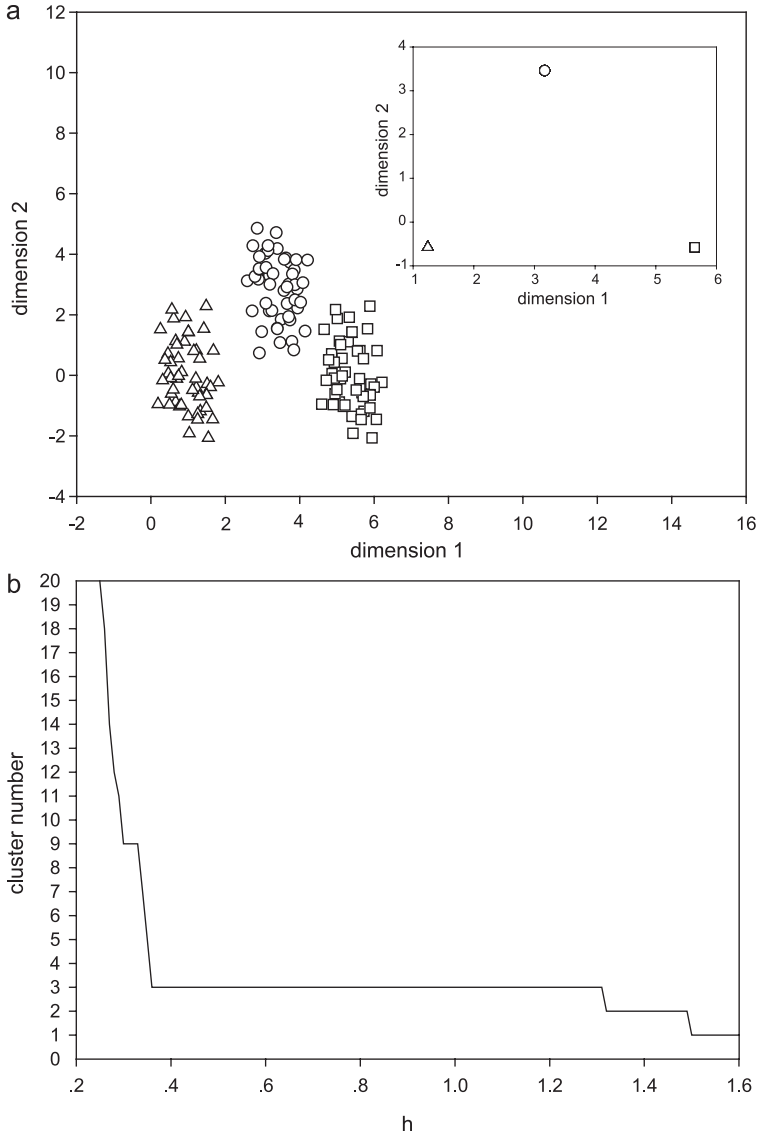

Fig. 2. (a) Simulated data set 1. (□) Data points in cluster 1, (△) data points in cluster 2, (○) data points in cluster 3, (insert) clustering results of the simulated data set 1 ($h = 0.36$). (b) Reliability curve of the simulated data set 1.

different classes: green tea, black tea and oolong tea. Concentrations of six chemical components: cellulose, hemicellulose, lignin, polyphenols, caffeine and amino acids were measured to represent each sample. In our study, the data were reduced to be two-dimensional after applying the principal component analysis to determine whether the reduced data contain apparent clustering structures according to their classes.

3.4. Male–female data set

The male–female data set is a multidimensional data set [12]. To construct the data, five features, the body length, the body weight, the shoe size, the belly outline and the neck size, are measured from total 47 men and women. Hence, the data set is a 47×5 matrix including 47 samples, each represented by a five-dimensional pattern vector. Before clustering, the data are preprocessed by column centering to prevent that the spread of the data in one of the dimensions is much greater that the others.

All calculations were carried out on an IBM/PC compatible computer with the windows 2000 operation system, and all programs were developed and operated in MATLAB 5.3 for windows.

4. Results and discussion

4.1. Simulated data 1

The simulated data set 1 is shown in Fig. 2a. The reliability curve of the simulated data set 1 is presented in

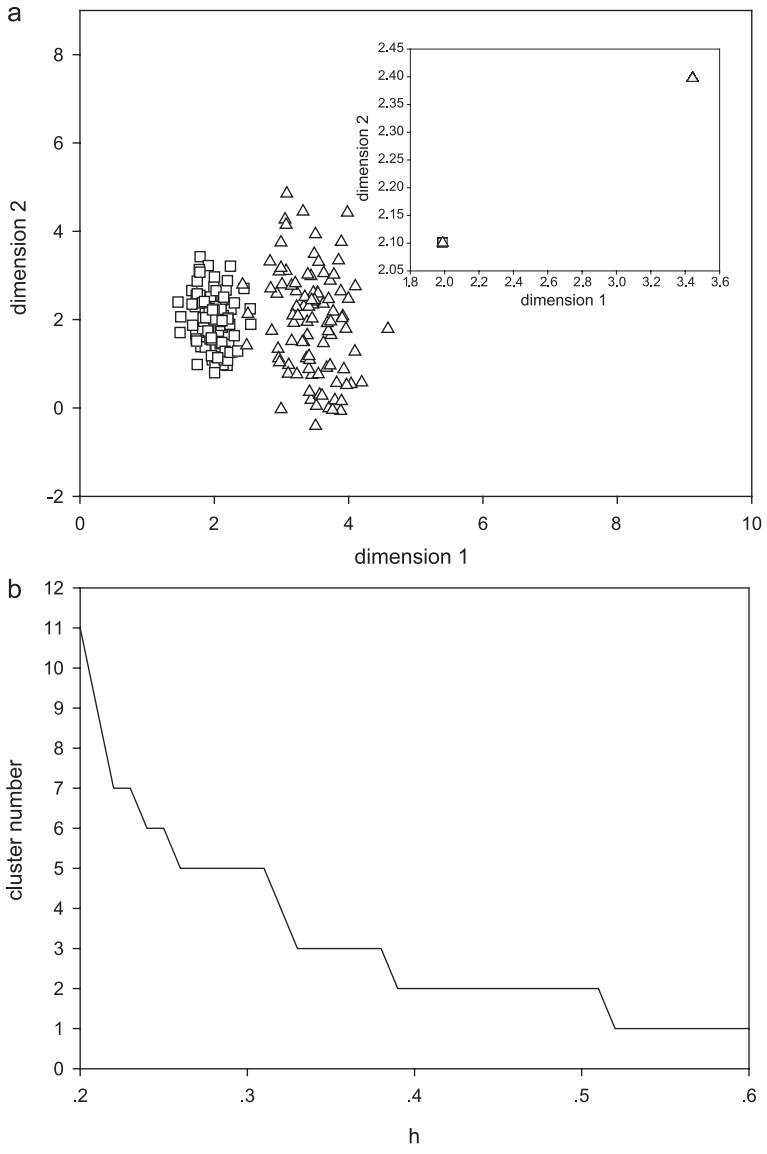

Fig. 3. (a) Simulated data set 2. (□) Data points in cluster 1, (△) data points in cluster 2, (insert) clustering results of the simulated data set 1 ($h = 0.39$). (b) Reliability curve of the simulated data set 2.

Fig. 2b. Two obvious plane sections of the curve suggest their corresponding clustering results are significant. The platform of two clusters are obtained between $h=1.32$ and $h=1.49$. The longest platform, the level of three clusters which is obtained between $h=0.36$ and $h=1.31$ gives the most insightful explanation to the data's structure. Since the greater the bandwidth is, the more amalgamation tendency the data have, we choose the clustering results corresponding to the outsets of the levels to compare with that of the K-means cluster analysis. When the smoothing parameter equals to 0.36, the clustering performance of the NLMSA is shown in Fig. 2a (insert). One can see that all samples converge to three points and hence the three clusters are distinctly perceived. Furthermore, the clustering results show that no sample is mistakenly discriminated. Comparatively, there are five or eight misclassified samples when apply the K-means cluster analysis with predefined cluster number of 3. Since the selection of the original cluster centers is random when the K-means cluster analysis is used, the clustering results of the K-means method are in dependence on the initiation of the cluster centers and therefore are variable.

4.2. Simulated data 2

The simulated data set 2 is shown in Fig. 3a. The reliability curve of the simulated data set 2 is presented in Fig. 3b. The division of two clusters is the most significant results. When the $h=0.39$, there are only four samples that are wrongly classified by the proposed method under this condition (see Fig. 3a, insert). However, the relevant clustering performance of the K-means cluster analysis is unstable. The number of the wrongly classified samples is between 3 and 83. For the convenience of the comparison, we recorded 20 different results obtained by consecutively applying the K-means method and then computed the average number of the mistakenly classified samples, and the average number obtained is above 40 (41.75). The platform of three clusters is between $h=0.33$ and $h=0.39$, and the platform of five clusters is between $h=0.26$ and $h=0.31$. The two levels are about equal length and are the second longest in the reliability curve. They are deemed to be relevant with the significant subgroup structures in the data.

4.3. Chinese tea data

The reduced Chinese tea data are presented in Fig. 4a. According to the reliability curve (see Fig. 4b), we accept that the clustering results at cluster number of 2, 3 and 6 are significant. The platform of two clusters is between $h=0.80$ and $h=1.18$. When choose the width about 0.80, the data is divided into two clusters, one cluster of green tea samples plus red black samples and another cluster of oolong tea samples. The K-means method also can successively identify the two clusters. The platform of three clusters is between $h=0.66$ and $h=0.79$. When choose the width of 0.66, the data form three clusters (see Fig. 4a, insert), each correctly represents a different variety of the

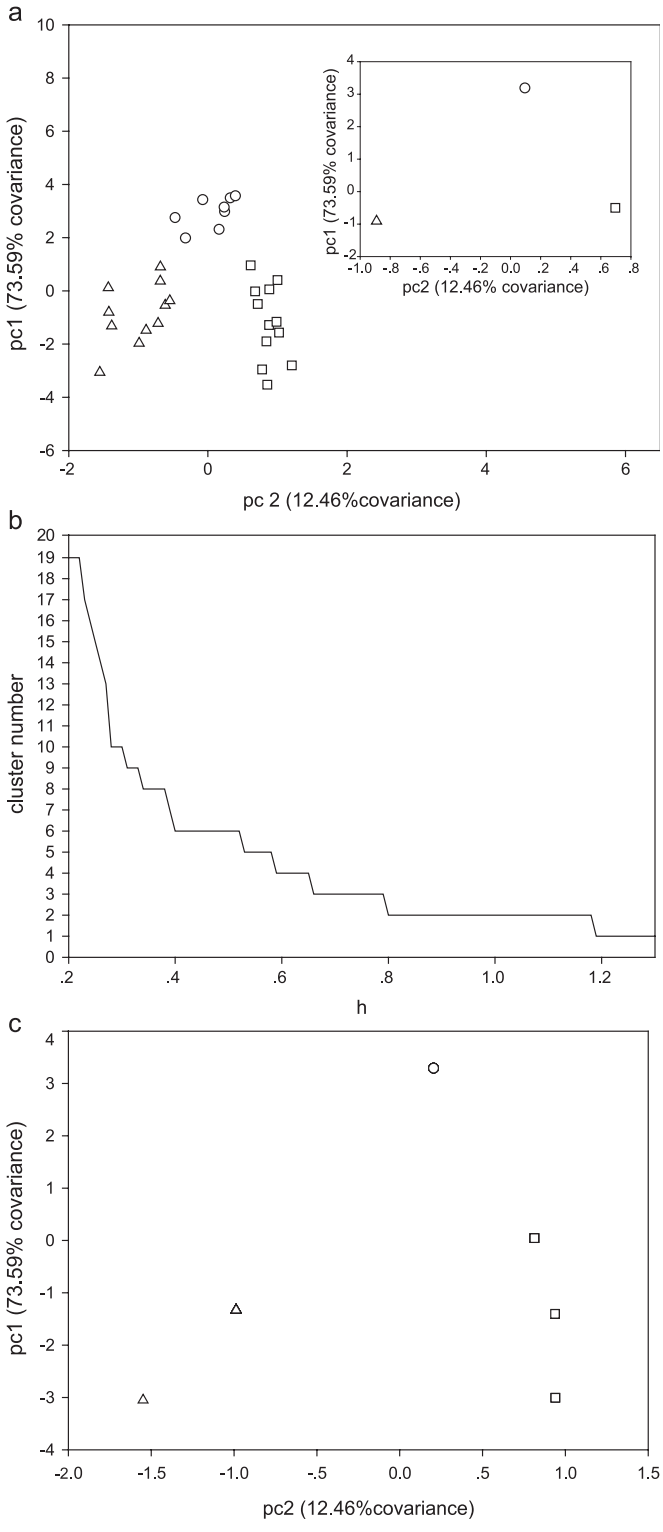

Fig. 4. (a) Chinese tea data set. (□) Green tea sample, (△) black tea sample, (○) oolong tea sample, (insert) clustering results of the Chinese tea data set ($h=0.66$). (b) Reliability curve of the Chinese tea data set. (c) Clustering results of the Chinese tea data set ($h=0.40$).

tea without misclassified samples. According to the results of the K-means approach, the samples C 5, C 6, C 7, H 4 and H 5, which belong to the cluster of green tea, are grouped into the cluster associated with black tea, and the sample F 1, which belongs to the cluster of black tea, is classified into the cluster of green tea. The platform of six clusters is between $h=0.40$ and $h=0.52$, it is expected to represent significant subgroup structures in the data (see Fig. 4c).

4.4. Male–female data

This data's projection on the axes of the two greatest principal components is shown in Fig. 5a. The reliability curve of the data set is given by Fig. 5b. The longest platform of two clusters is obtained between $h=0.81$ and $h=1.14$. When $h=0.81$, the male samples and the female samples are distinguished with only two mistakenly classified samples. Fig. 5c displays the classification results by plotting the first two principal components of the solution matrix. Sample m 11 that belongs to the male cluster is wrongly classified into the female class, and the sample f 24 that should be in the female cluster is incorrectly grouped into the male cluster. The comparative performance given by the K-means approach is that the male samples m 4, m 8 and m 11 are incorrectly grouped into the female cluster, and the sample f 24 is mistakenly classified into the male cluster. The platform of three clusters is between $h=0.69$ and $h=0.80$, the clustering results are shown in Fig. 5d. The divisions of the samples are almost the same as the results of two clusters except that the sample m 19 is isolated and forms an individual cluster. It is thought that the sample m 19 might be an outlier.

5. Conclusion

In this paper, we propose a modified method based on the kernel density estimation, which searches the nearest local maxima of the density estimate to the corresponding data points in a given data space by employing an ascending gradient method. It seeks to discover data's underlying group structure solely depending on the data's own characteristics and hereby requires no a priori information about the data. Moreover, as a density-based method, it is able to deal with arbitrary shape clusters. The presented clustering results demonstrated that the proposed method outperforms the K-means cluster analysis. It is expected to be a feasible and effective technique for chemical pattern recognition applications.

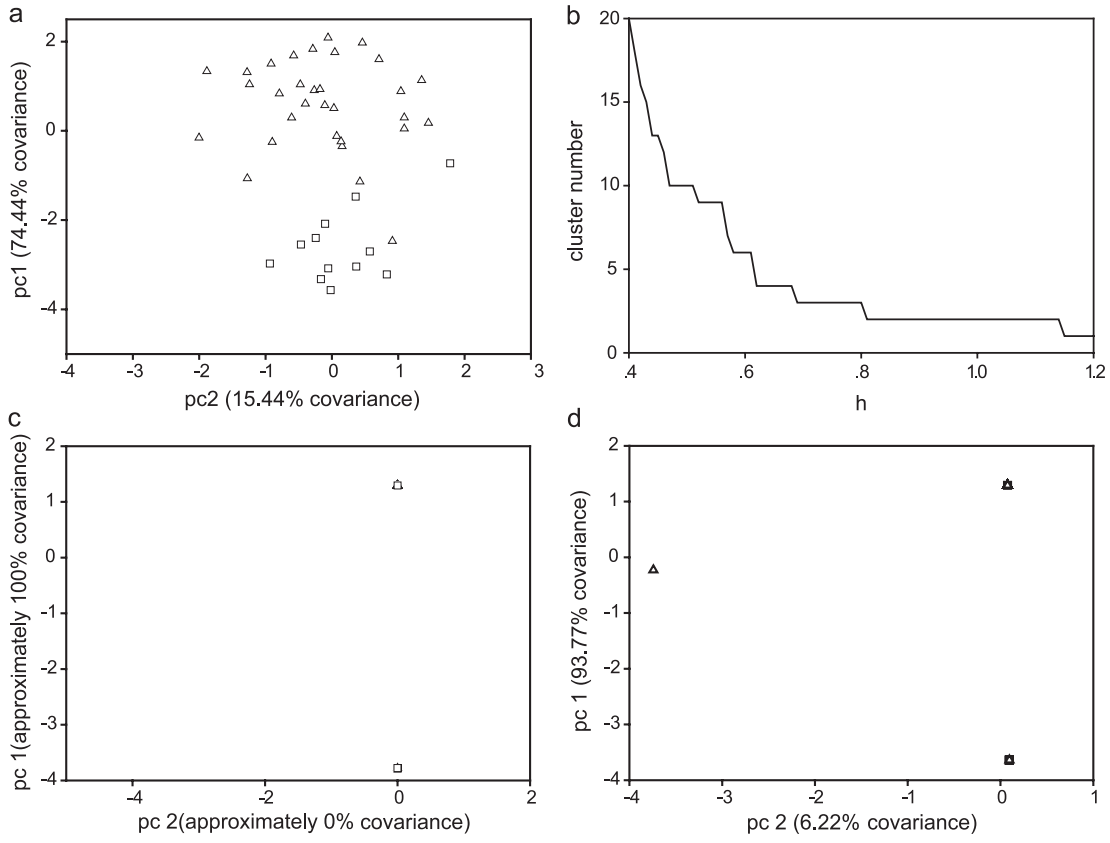

Fig. 5. (a) Male–female data set. (□) Female sample, (△) male sample. (b) Reliability curve of the male–female data set. (c) Clustering results of the male–female data set ($h=0.81$). (d) Clustering results of the male–female data set ($h=0.69$).

Acknowledgements

We are grateful for the financial support from the National Natural Science Foundation of China (Grant numbers: 20375012, 20105007 and 20205005).

References

[1] S.D. Brown, R.S. Bear, T.B. Blank, Anal. Chem. 64 (1992) 22R–49R.
[2] I.E. Frank, J.H. Friedman, J. Chemom. 3 (1989) 463–475.
[3] L.X. Sun, K. Danzer, J. Chemom. 10 (1996) 325–342.
[4] R.Q. Yu, Introduction to Chemometrics, Hunan Education Publishing House, Changsha, 1991.
[5] N. Bratchell, Chemom. Intell. Lab. Syst. 6 (1989) 105–125.
[6] Q. Shen, L. Tang, Introduction Theory of Pattern Recognition, National University of Defense Technology Press, Changsha, 1991.
[7] J. Zupan, J. Gasteiger, Neural Network for Chemists, VCH Publishers, New York, 1993.
[8] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data By the Use of Cluster Analysis, Wiley, New York, 1983.
[9] D. Coomans, D.L. Massart, Anal. Chim. Acta 133 (1981) 224–225.
[10] M. Daszykowski, B. Walezak, D.L. Massart, Chemom. Intell. Lab. Syst. 56 (2001) 83–92.
[11] X. Liu, P.V. Espen, F. Adams, Anal. Chim. Acta 200 (1987) 421–430.
[12] D. Wienke, L. Buydens, Chemom. Intell. Lab. Syst. 32 (1996) 151–164.