Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 5060842, 7 pages
https://doi.org/10.1155/2017/5060842

Research Article

Clustering Mixed Data by Fast Search and Find of Density Peaks

Shihua Liu,[1,2] Bingzhong Zhou,[2] Decai Huang,[1] and Liangzhong Shen[3]

[1]*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China*
[2]*Department of Information Technology, Wenzhou Vocational & Technical College, Wenzhou, China*
[3]*School Administration Offices, Wenzhou Business College, Wenzhou, China*

Correspondence should be addressed to Decai Huang; hdc@zjut.edu.cn and Liangzhong Shen; Johnshen0211@163.com

Received 9 February 2017; Accepted 30 April 2017; Published 5 July 2017

Academic Editor: Anna M. Gil-Lafuente

Aiming at the mixed data composed of numerical and categorical attributes, a new unified dissimilarity metric is proposed, and based on that a new clustering algorithm is also proposed. The experiment result shows that this new method of clustering mixed data by fast search and find of density peaks is feasible and effective on the UCI datasets.

1. Introduction

Cluster analysis has been one of the research hotspots in data mining and machine learning. In the big data era, various kinds of data are emerging one after another. Most of them are data with multiple attribute types, such as numerical and categorical attributes. Clustering algorithms such as K-means are mainly used for numerical attribute data. In order to handle the data with mixed attributes, researchers have proposed a lot of solutions, which can be divided into Attribute Conversion methods, Clustering Ensemble methods, prototype-based methods, density-based methods, hierarchical methods, and so forth [1].

The Attribute Conversion method firstly converts other attributes into one attribute and then cluster them, such as SpectralCAT algorithm proposed by David and Averbuch [2], which transforms the numerical attribute into the categorical attribute first and then uses the spectral clustering method to deal with the transformed data.

The idea of Clustering Ensemble methods is to divide a group of objects by various algorithms, and then the results of these algorithms are combined by consensus function to get the final clustering results. It was first proposed by Strehl and Ghosh in 2002 [3] and then becomes one of the main methods for the mixed data clustering. Zhao et al. proposed a mixed data clustering algorithm named CEMC based on Clustering Ensemble method [4]; He et al. proposed a clustering algorithm named CEBMDC [5] based on Clustering Ensemble and Squeezer [6]. The algorithm uses the Squeezer algorithm for clustering the categorical parts and as consensus function.

The K-Prototypes algorithm [7] was proposed by Huang in 1997, and it is mainly based on the idea of k-means algorithm. It combines the clustering centers of the numerical attributes and the mode of the categorical attributes to construct a new mixed center. The data center is a prototype, and a distance metric formula and cost function are constructed on the basis of the prototype. Then the k-means clustering process is used to cluster the mixed data directly. This prototype-based algorithm is simple and efficient. The most important factor for this method is the definition of prototype and the distance metric between the prototype and the data tuples. Cheung and Jia proposed a unified similarity metric method [8], which normalizes the distance metric of the numerical attribute part and makes the value of the similarity measure bound in $[0, 1]$. Then the similarity measure of each categorical attribute is weighted and normalized, respectively, and finally a unified distance measure formula is obtained. Based on this formula, an iterative algorithm OCIL is proposed to cluster the mixed attribute data. At the same time, the OCIL is further improved by introducing the competition and penalty mechanism and proposed a mixed data clustering algorithm PCL-OC which can determine the cluster number automatically. They compared the OCIL algorithm with the K-Prototypes algorithm; the clustering accuracy of OCIL is greatly improved, but the computation complexity

is higher. The prototype-based method mentioned above still needs to determine the number of clusters; it is sensitive to the selection of the original cluster center and the outlier points, and also it cannot find different shape of the cluster.

Li and Biswas proposed SBAC (Similarity Based Agglomerative Clustering) algorithm [9], which is a good agglomerative hierarchical clustering algorithm based on Goodall similarity measure. This method works well but its computational complexity is higher than $O(n^2 \times \log(n))$. Hierarchical clustering algorithms have high time and space complexity and are not reversible.

The RDBC_M algorithm proposed by Huang and Li [10] used dimension-oriented distance formula to calculate the distance of each dimension. It applies Euclidean distance to the numerical attribute and uses expert scoring method for the similarity calculation between the different values of the categorical attributes. The definition of a distance matrix to measure the distance of each categorical dimension requires manual scoring. The MDCDen algorithm [11] and the DC-MDACC algorithm [12] proposed by Chen and He firstly classified the dataset into three categories: numerical dominance, categorical dominance, and balanced. Then, different distance metric functions are defined for each class. They require a priori analysis of the dataset. The similarity measure of the categorical attributes in RDBC_M algorithm needs to be evaluated by the domain expert; the MDCDen algorithm needs to adjust three parameters to obtain the better result.

In 2014, Rodriguez and Laio published a clustering method by fast search and find of density peaks (abbreviated as "DPC algorithm") in *Science* [13]. The algorithm has the advantages of good clustering effect, high efficiency, and few parameters. It can not only find the number of clusters and identify outliers automatically, but also cluster data with different shapes. The input of DPC algorithm is the distance matrix between data points. As long as the problem of distance measurement between data points of mixed data is solved, the algorithm can be applied to cluster analysis directly. However up to now there are still no reports on the clustering of mixed data using DPC algorithm.

In this paper, a new unified distance metric for mixed data points is firstly proposed and used to construct the distance matrix between data points of mixed data. And then based on that, a new method to clustering mixed data by fast search and find of density peaks abbreviated as "DPC_M algorithm" is put forward. Finally, DPC_M algorithm is used to cluster the common UCI mixed datasets. The result shows that the DPC_M algorithm not only has better clustering performance than the traditional K-Prototypes algorithm, but also can automatically find the number of clusters. Moreover, it is not sensitive to the outliers.

This paper is organized as follows: Section 2 introduces principle of the DPC algorithm. Section 3 presents a unified formula for the distance metric of the mixed data points and describes the details of the DPC_M algorithm. Section 4 describes the experimental analysis of the DPC_M algorithm. The final section summarizes our work.

2. Principle of the DPC Algorithm

The DPC algorithm is based on two fundamental assumptions: the cluster center point has a high local density and is surrounded by a point with a lower local density, and the cluster center point is relatively far away from its neighbor points with higher density. Therefore, the DPC algorithm constructs a Decision Graph by computing a local density ρ_i and a relative distance δ_i to discover the cluster center in a dataset. The remaining data points in the dataset are allocated at once to the cluster to which the nearest cluster center belongs.

Suppose $S = \{X_1, X_2, \ldots, X_n\}$ is a dataset for clustering and $d_{ij} = \text{dist}(X_i, X_j)$ is the distance between data points X_i and X_j; the DPC algorithm defines a cutoff distance d_c and a local density ρ_i as formula (1) and distance δ_i as formula (2) for data point i, where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$; otherwise

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right), \tag{1}$$

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \left(d_{ij}\right), & \rho_i < \max\limits_k \left(\rho_k\right), \\ \max\limits_j \left(d_{ij}\right), & \rho_i = \max\limits_k \left(\rho_k\right). \end{cases} \tag{2}$$

Here, the distance δ_i is defined as the distance corresponding to the data point X_i when the local density is not the maximum density but has the minimum value of the distance from the point to the point where all the densities are larger than it, or else it takes the maximum distance to all other points.

When the number of data points in the dataset is small, the effect of calculating local density ρ_i by formula (1) is not ideal. Therefore, in [13], a Gaussian kernel function is given for the dataset with fewer data points as follows:

$$\rho_i = \sum_{j \neq i} \exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right). \tag{3}$$

Based on the local density ρ_i and distance δ_i for each data point, users can explicitly choose the number and the center of the cluster on the Decision Graph. Once the center point is determined, each remaining data point can be classified into the same cluster as its nearest neighbor of higher density.

3. Clustering Mixed Data Based on Density Peaks (DPC_M)

3.1. Definition of Unified Distance Metric. Let $S = \{X_1, X_2, \ldots, X_n\}$ be a mixed dataset with d dimensions and n instances, where the numerical attributes have d_r dimensions and the categorical attributes have $d_c = d - d_r$ dimensions. For two data points X_i and X_j in the dataset, their distance is defined as shown in the following formula:

$$D\left(X_i, X_j\right) = d\left(X_i, X_j\right)_r + d\left(X_i, X_j\right)_c. \tag{4}$$

Formula (5) illustrates the distances computation of the numerical attribute $d(X_i, X_j)_r$ and the categorical attribute $d(X_i, X_j)_c$, respectively:

$$d\left(X_i, X_j\right)_r = 1 - e^{-\operatorname{dist}(X_i^{d_r}, X_j^{d_r})},$$
$$d\left(X_i, X_j\right)_c = \sum_{t=1}^{d_c} \omega_t * \delta\left(x_{it}, x_{jt}\right), \tag{5}$$

where $\operatorname{dist}(X_i^{d_r}, X_j^{d_r})$ denotes the normalized Euclidean distances of the numerical attribute of the data points X_i, X_j. Since the Euclidean distance is nonnegative, it is ensured that the distance value of the numerical attribute is in the interval $[0, 1]$. As for the distance of the categorical attribute, the matching method with the entropy weight is used. The matching distance of the data points X_i, X_j in the tth categorical attribute is calculated by the following formula:

$$\delta\left(x_{it}, x_{jt}\right) = \begin{cases} 0, & \text{if } \left(x_{it} = x_{jt}\right), \\ 1, & \text{if } \left(x_{it} \neq x_{jt}\right). \end{cases} \tag{6a}$$

The importance of a categorical attribute is quantified by its average entropy over each attribute value. The weight of each attribute ω_t is then computed by the following formula:

$$\omega_t = \frac{H_{A_t}}{\sum_{s=1}^{d_c} H_{A_s}}. \tag{6b}$$

Assuming that the total number of categorical values on the tth categorical attribute is m_t, where the probability of occurrence of the sth ($s = 1, 2, \ldots, m_t$) values is $p(a_{ts})$, The entropy weight H_{A_t} can be calculated using the following formula:

$$H_{A_t} = -\frac{1}{m_t} \sum_{s=1}^{m_t} p\left(a_{ts}\right) \log p\left(a_{ts}\right). \tag{6c}$$

3.2. The DPC_M Algorithm. The DPC_M algorithm first calculates the distance of the data points in the dataset by using the unified distance metric (4) and then calculates the local density ρ_i of each data point by formula (3) and the distance δ_i by formula (2). In order to realize the automatic determination of clustering centers, we define $\gamma_i = \rho_i * \delta_i$ and arrange them in a descending order. Then we can get the clustering center by computing the inflection point of γ_i. According to the definition of distance calculation formula (2), we can know that the point with the largest local density is the cluster center point. After the cluster centers have been found, each remaining point is assigned to the same cluster as its nearest neighbor of higher density. The cluster assignment is performed in a single step other than iterative steps.

The input of the algorithm is the mixed dataset S and the neighbor occupation ratio p; the output is the clustered label vector. The specific process is as follows.

Step 1. Formula (4) is used to calculate the distance of each two points in the dataset.

Step 2. The local density of each data point ρ_i is calculated by formula (3) and p, and then the distance δ_i is calculated by formula (2), and at last $\gamma_i = \rho_i * \delta_i$ is calculated.

Step 3. Sort γ_i in a descending order, calculate the inflection point which is used to determine the cluster centers, and set the class labels of centers.

Step 4. The remaining points are assigned to the same class label as their nearest neighbor with higher density one by one, and then the clustering results are achieved.

3.3. Complexity Analysis. For a dataset with n data points, the space complexity of the algorithm is mainly for the storage of distance matrix which requires $3 * n * (n-1)/2$ storage space. The distance matrix has three columns in which column 1 and column 2 are the data point numbers and column 3 is the distance between the two data points. In addition, the algorithm requires three arrays of length n to store the local density ρ, the distance δ, and its product γ, so the space complexity is $O(n^2)$.

The time complexity of DPC_M algorithm is mainly derived from distance calculation in Step 1 and local density computation in Step 2; the time complexity of distance computation and its product calculation is $O(n^2)$; the sort time complexity in Step 3 depends on the sorting algorithm, the minimum $O(n \log(n))$, and the largest $O(n^2)$, so the total complexity is no more than $O(n^2)$; the time complexity of data point allocation in Step 4 is $O(n)$. Therefore, the overall complexity of the algorithm is $O(n^2)$ and is the same as DPC algorithm.

4. Experimental Analysis

In order to verify the effectiveness of the DPC_M algorithm in this paper, a common UCI mixed dataset is used for experimental study, and this algorithm is compared with the K-Prototypes algorithm. We implement both the two algorithms in Matlab 2015a and the experiments were done on a Win 10 computer with Intel Core i5-5200u CPU, 4G DDR3 memory.

4.1. Experiment Datasets. In this study, five datasets of mixed datasets from UCI machine learning repository were investigated, which are Statlog Heart, Cleveland Heart Disease (Cleveland), Statlog Credit Approval (Credit), Acute Inflammations (Acute), and Adult. The brief information about the mixed datasets is shown in Table 1.

The missing data are eliminated before the experiment. In addition, the numerical properties are normalized using the maximum-minimum normalization methods as follows:

$$A_i = \frac{A_i - \min(A)}{\max(A) - \min(A)}, \tag{7}$$

where A denotes a numerical attribute value range and A_i denotes a numerical attribute value corresponding to the ith data point in the dataset.

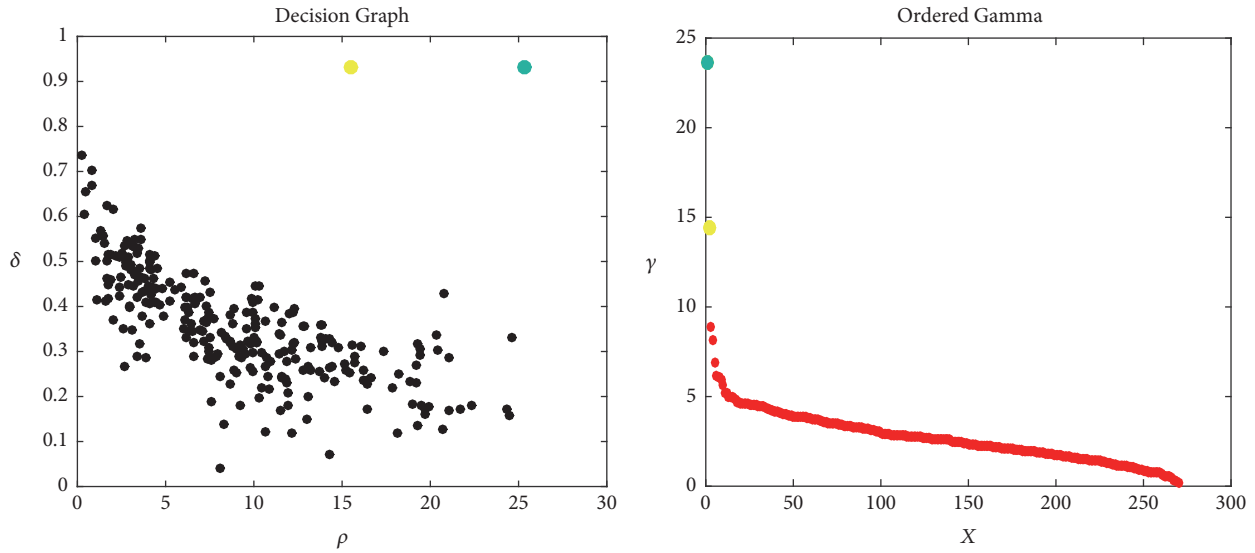

Figure 1: Decision Graph and Ordered Gamma on Statlog Heart dataset.

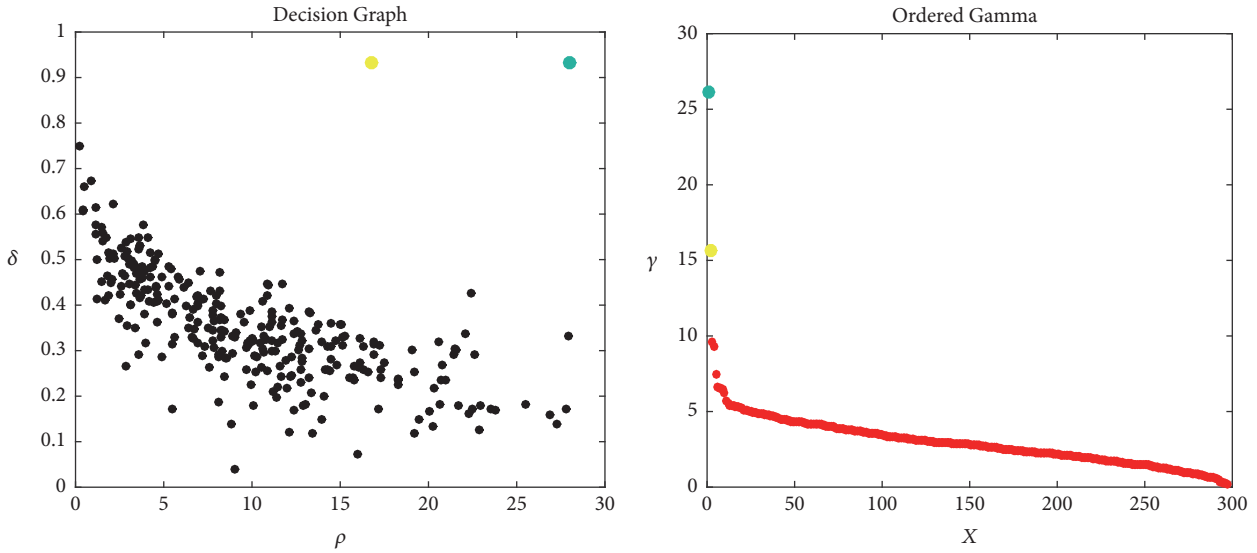

Figure 2: Decision Graph and Ordered Gamma on Cleveland Heart Disease dataset.

4.2. Evaluation Index. Since the UCI datasets used have real class labels, the clustering accuracy rate can be used as the validity index. The clustering accuracy rate ACC is used to calculate the matching degree of the algorithm class label relative to the real class label, which is defined as follows:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{k} a_i}{n}, \tag{8}$$

where a_i denotes the number of samples correctly classified, k denotes the number of clusters, and n denotes the number of instances in the dataset. The higher the clustering accuracy, the better the clustering effect.

4.3. Effectiveness Experiment. The K-Prototypes algorithm and the DPC_M algorithm are used to cluster the dataset described in Section 4.1, respectively, and ACC is calculated by formula (8). According to the literature [7], the important parameter γ of K-Prototypes algorithm takes $1/2\sigma$ (σ represents the mean standard deviation of numerical attributes). The K-Prototypes algorithm runs 100 times and the average clustering accuracy of 100 times is regarded as the comparison index. The parameters of the DPC_M algorithm are given by $p = 1.5\%$ according to the literature [13].

The Decision Graph and γ_i inflexion diagram (Ordered Gamma) of the clustering process for DPC_M algorithm are shown in Figures 1–5.

In the Decision Graph and Ordered Gamma diagram in Figures 1, 2, 3, and 5, the two colorful points are the cluster centers. There are 4 cluster centers in Figure 4.

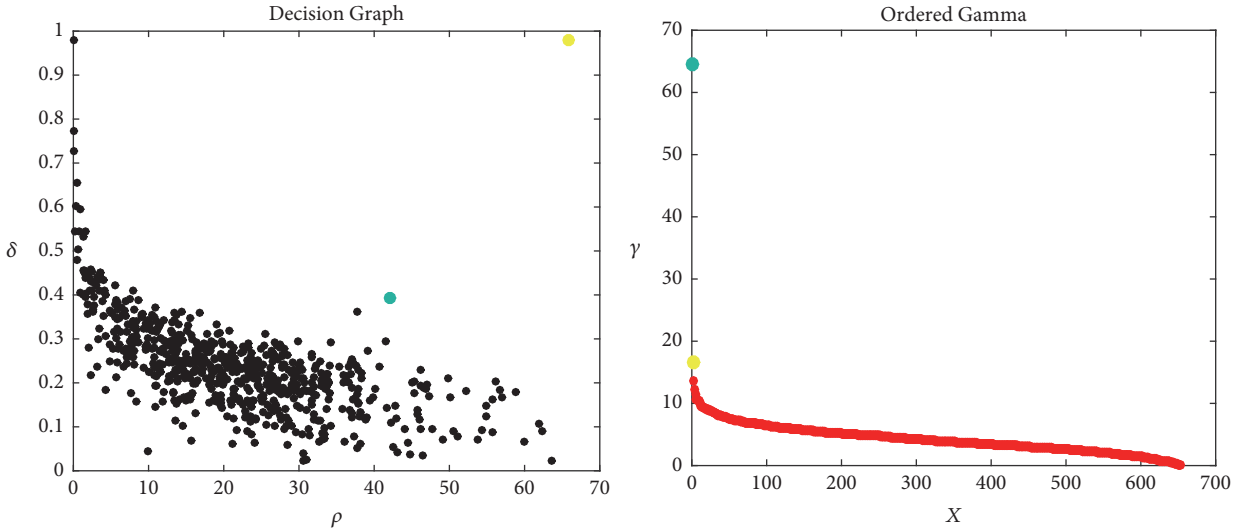

Figure 3: Decision Graph and Ordered Gamma on Statlog Credit Approval dataset.

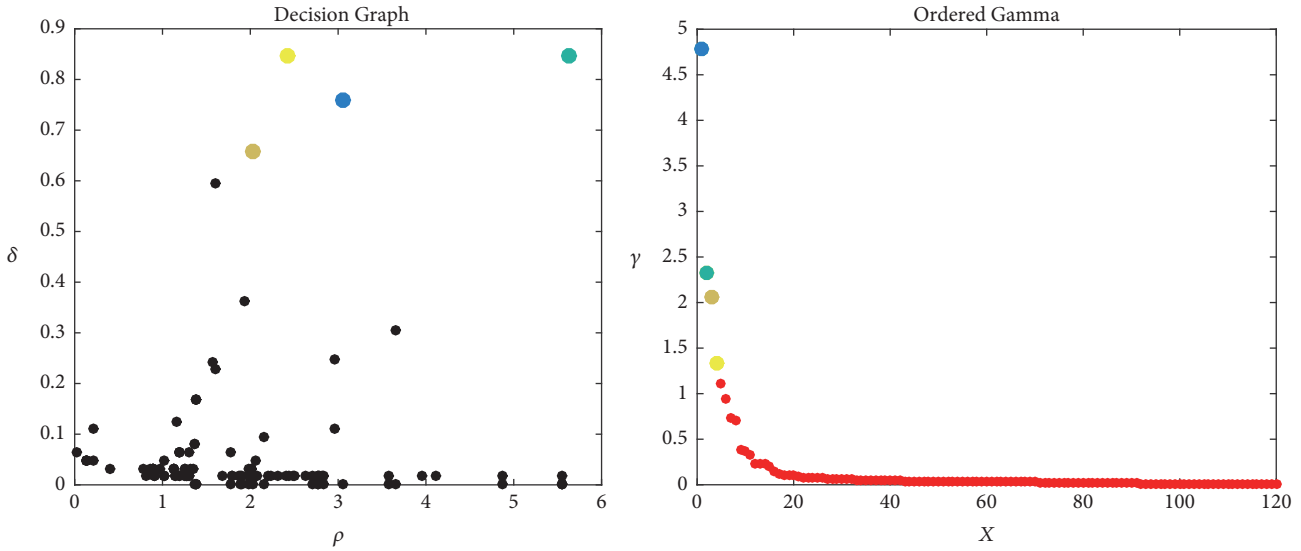

Figure 4: Decision Graph and Ordered Gamma on acute inflammations dataset.

The clustering results are shown in Table 2 which indicates that the accuracy of the DPC_M algorithm is higher than that of the K-Prototypes algorithm in all four datasets. It can be seen that the DPC_M algorithm proposed in this paper can cluster the real mixed attribute datasets and obtain better results, and it can also automatically determine the number of clusters.

4.4. Influence of Parameter p. In this paper, the DPC_M algorithm has a unique parameter neighborhood ratio p. In order to study the influence of parameter p on the clustering effect of the algorithm, we use the credit dataset to carry on the simulation experiment, and let p be 0.5%, 1%, 1.5%, 2%, 3%, 6%, 8%, 10%, 15%, and 20% respectively, and then the DPC_M algorithm is used to cluster the credit dataset. The clustering accuracy ACC is calculated under different conditions.

The clustering results are shown in Figure 6. The x-axis is the value of p, and the y-axis is the cluster accuracy ACC. It can be seen from the figure that when p is less than 6%, the fluctuation of ACC is very small, which means the algorithm has good stability and is less affected by the parameter value. When the value of p is greater than 10%, the clustering accuracy is reduced obviously, which proves the conclusion in [13] that the suitable p is about 1-2%.

5. Conclusion

The key issue for the DPC algorithm proposed in literature [13] is how to define the distance measurement between data points in the mixed dataset. Therefore, the DPC_M algorithm

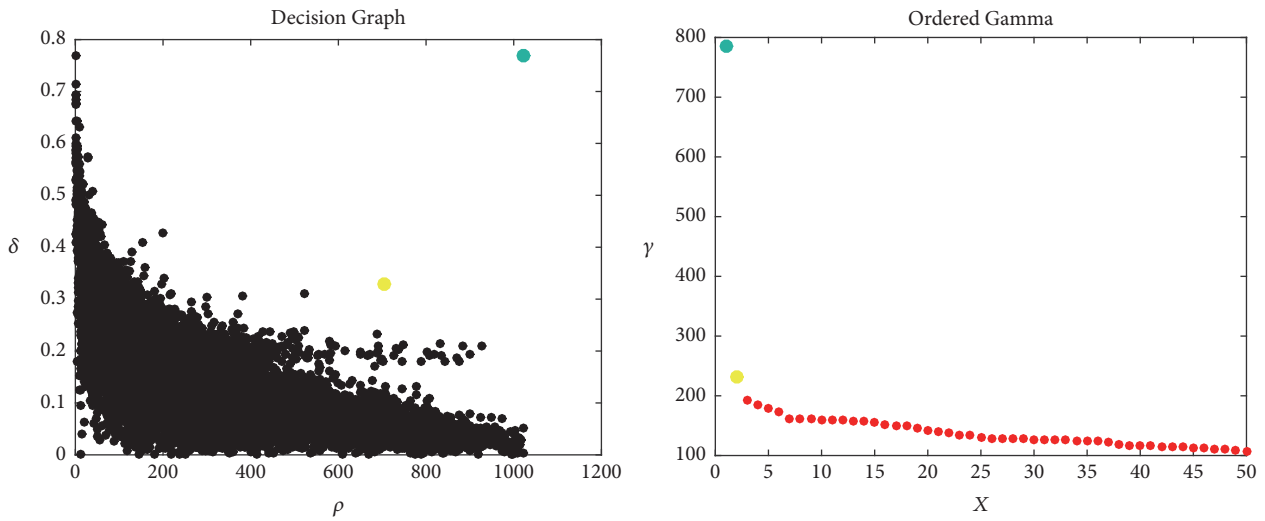

Figure 5: Decision Graph and Ordered Gamma on Adult dataset.

Table 1: Brief information of UCI mixed datasets.

Number	Datasets	Instances	Attributes ($d_c + d_r$)	Class	Demo
(1)	Statlog Heart	270	7 + 6	2	
(2)	Cleveland	303-6	7 + 6	2	6 incomplete instances deleted
(3)	Credit	690-37	9 + 6	2	37 incomplete instances deleted
(4)	Acute	120	5 + 1	4	2 class labels
(5)	Adult	10000	8 + 6	2	Use a subset with 10000 records

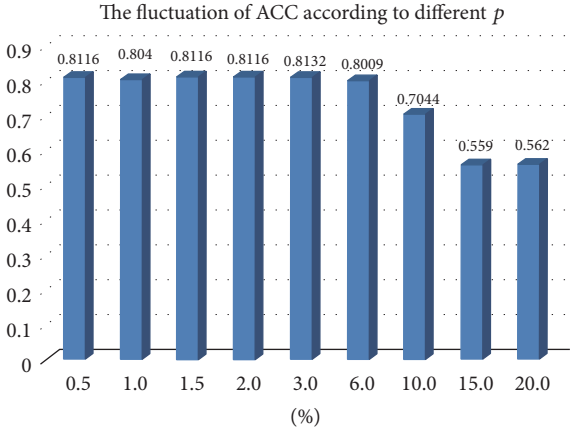

Figure 6: The fluctuation of clustering ACC values according to different p values.

Table 2: The clustering accuracy rate (ACC) on four datasets.

Number	Datasets	k-Prototypes	DPC_M
(1)	Statlog Heart	0.7694	**0.8074**
(2)	Cleveland	0.7720	**0.8013**
(3)	Credit	0.7381	**0.8116**
(4)	Acute	0.7833	**0.9167**
(5)	Adult	0.6099	**0.7354**

designed for the clustering of the mixed data proposed in this paper is constructed by using a new unified dissimilarity metric between the mixed data points. The clustering experiments on the UCI datasets show that the proposed DPC_M algorithm has better clustering performance and higher clustering stability than the traditional K-Prototypes algorithm, and it is not sensitive to the initial selection of original prototypes.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by the Special Fund for Public Welfare Industry Research of the Ministry of Water Resources, China (201401044), the Subproject under National Science and Technology Support Program (2012BAD10B0101), and the General Research Project of the Foundation of Zhejiang Province Educational Committee (no. Y201636767).

References

[1] S. H. Liu, L. Z. Shen, and D. C. Huang, "A three-stage framework for clustering mixed data," *WSEAS Transactions on Systems*, vol. 15, no. 1, pp. 1–10, 2016.

[2] G. David and A. Averbuch, "SpectralCAT: categorical spectral clustering of numerical and nominal data," *Pattern Recognition*, vol. 45, no. 1, pp. 416–433, 2012.

[3] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining partitionings," in *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02), 14th Innovative Applications of Artificial Intelligence Conference (IAAI-02)*, pp. 93–99, 2002.

[4] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Cluster ensemble method for databases with mixed numeric and categorical values," *Journal of Tsinghua Univercity (Science and Technology)*, vol. 46, no. 10, pp. 1673–1676, 2006.

[5] Z. He, X. Xu, and S. Deng, "Squeezer: an efficient algorithm for clustering categorical data," *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611–624, 2002.

[6] Z. He, X. Xu, and S. Deng, "Clustering mixed numeric and categorical data: a cluster ensemble approach," 14 pages, 2005, https://arxiv.org/abs/cs/0509011.

[7] Z. X. Huang, "Clustering Large Data Sets with Mixed and Numeric and Categorical values," in *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '97)*, pp. 21–34, 1997.

[8] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recognition*, vol. 46, no. 8, pp. 2228–2238, 2013.

[9] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 4, pp. 673–690, 2002.

[10] D.-C. Huang and X.-C. Li, "Incremental relative density-based clustering algorithm for mixture datasets," *Control and Decision*, vol. 28, no. 6, pp. 815–822, 2013.

[11] J.-Y. Chen and H.-H. He, "Density-based clustering algorithm for numerical and categorical datawith mixed distance measure methods," *Control Theory and Applications*, vol. 32, no. 8, pp. 993–1002, 2015.

[12] J.-Y. Chen and H.-H. He, "Research on density-based clustering algorithm for mixed data with determine cluster centers automatically," *Acta Automatica Sinica*, vol. 41, no. 10, pp. 1798–1813, 2015.

[13] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.