Research Article

Nonuniform Sparse Data Clustering Cascade Algorithm Based on Dynamic Cumulative Entropy

Ning Li, Yunxia Gu, and Zhongliang Deng

Beijing University of Posts and Telecommunications, Electronic Engineering School, Beijing, China

Correspondence should be addressed to Yunxia Gu; guyunxia_123@bupt.edu.cn

Received 15 June 2016; Revised 23 September 2016; Accepted 9 October 2016

Academic Editor: Rita Gamberini

Copyright © 2016 Ning Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A small amount of prior knowledge and randomly chosen initial cluster centers have a direct impact on the accuracy of the performance of iterative clustering algorithm. In this paper we propose a new algorithm to compute initial cluster centers for k -means clustering and the best number of the clusters with little prior knowledge and optimize clustering result. It constructs the Euclidean distance control factor based on aggregation density sparse degree to select the initial cluster center of nonuniform sparse data and obtains initial data clusters by multidimensional diffusion density distribution. Multiobjective clustering approach based on dynamic cumulative entropy is adopted to optimize the initial data clusters and the best number of the clusters. The experimental results show that the newly proposed algorithm has good performance to obtain the initial cluster centers for the k -means algorithm and it effectively improves the clustering accuracy of nonuniform sparse data by about 5%.

1. Introduction

Clustering is an important discovery technique of exploratory data mining and a common technique for statistical data analysis. Iterative clustering algorithm is one kind of the clustering algorithms. And k -means is the most popular and the fast method in iterative clustering algorithms. Because of the simplicity of k -means algorithm, it is used in many fields, including machine learning, medicine, image analysis, pattern recognition, information retrieval, bioinformatics, and computer. For example, in the medical field, cancer genomics [1], cell signaling [2], and viral genomes [3] use k -means as a data analysis tool; in the bioinformatics field, bioanalytical chemistry [4], the vibrational spectra of biomolecules [5], and the nervous system [6] use k -means to mine potential information; in the image analysis field, imaging techniques [7] use k -means to partition a given set of points into homogeneous groups; in the pattern recognition field, automatic system for imbalance diagnosis in wind turbines [8] uses k -means to suggest the optimum number of groups. Reference [9] uses k -means to analyze network data. Reference [10] generates profiles by k -means to group together days with a similar pattern of request arrivals.

Although k -means algorithm has been developed to solve a wide range of different problems, it has three major drawbacks:

- (1) It needs to predetermine the cluster number k by user. In practice, due to little prior knowledge, k value is generally difficult to determine.
- (2) It is sensitive to selection of the initial cluster centers. That is, k -means selects different initial cluster centers with different results. Because of randomly chosen initial clusters centers, populations are generally composed of low quality individuals exclusively.
- (3) The effect of k -means algorithm for nonuniform sparse data processing is not good.

To overcome these drawbacks, many evolutionary algorithms such as GA, TS, and SA have been introduced. Kao et al. have proposed a hybrid technique based on combining the k -means algorithm [11]. Bahmani Firouzi et al. have introduced a hybrid evolutionary algorithm based on combining PSO, SA, and k -means to find optimal solution [12]. Niknam and Amiri have proposed a hybrid algorithm based on a fuzzy adaptive PSO, ACO, and k -means for cluster analysis [13]. Niknam et al. have purposed a novel algorithm

that is based on combining two algorithms of clustering: k -means and Modified Imperialist Competitive Algorithm [14]. Evolutionary algorithms require large amounts of data to study; however, many real-world problems are like black boxes; hence no sufficient data about their internals is available.

In addition, to solve the problem of selection of the initial cluster centers, Bianchi et al. have proposed two density-based k -means initialization algorithms for nonmetric data clustering [15]. Tunali et al. have proposed an improved clustering algorithm for text mining: multicluster spherical k -means [16]. Tvrdík and Křivý have proposed a new algorithm combining differential evolution and k -means [17]. Rodriguez and Laio proposed selection of the initial cluster centers by density peak. Khan and Ahmad [18] have proposed the cluster center initialization algorithm for k -means clustering. But the computing of the above methods is laborious.

In order to solve the situation that a small amount of prior knowledge and randomly chosen initial cluster centers have a direct impact on the accuracy of the performance of iterative clustering algorithm. In this paper, we propose a new algorithm for nonuniform sparse data clustering based on cascade entropy increase and decrease. It designs Euclidean distance sparse degree of aggregation density control factor, determines the initial cluster center of nonuniform sparse data, and groups initial data clusters by multidimensional diffusion data distribution density. Multiobjective clustering approach is adopted to compensate the clustering error of initial data clusters. The experimental results show that the new data clustering algorithm can effectively improve the clustering accuracy of nonuniform sparse data clusters.

2. Nonuniform Sparse Data Clustering Cascade Algorithm Based on Dynamic Cumulative Entropy

In order to obtain the optimal clustering results of nonuniform sparse data, in this paper, we use multidimensional diffusion density distribution to obtain the initial data clusters, while the Euclidean distance control factor based on aggregation density sparse degree is put forward to solve the problem that multidimensional data is easy to misjudge. The initial data clusters are more than the real cluster. So we need to select optimal initial cluster centers by decision graph and then execute k -means on complete data set based on multiobjective clustering approach.

2.1. Initial Data Clustering Using Multidimensional Diffusion Density Distribution. In iterative clustering algorithms, choosing initial cluster centers is extremely important as it has a direct impact on the formation of final clusters. It is dangerous to select some samples as initial centers, which are away from normal samples. In this paper, first we define the multidimensional diffusion density distribution of samples and the comprehensive distance, according to which we get the initial data clusters.

2.1.1. Multidimensional Diffusion Data Normalization. Different attributes of multidimensional data have different units of measurement and value ranges, which has a serious impact on clustering formation. So in order to avoid the above situation at first we need to normalize clustering data to make all attributes have the same weights. Normalization is done using the following formula.

Calculate the average of the absolute deviation: D is the set of n data elements described with attributes $A_1, A_2, \dots, A_m$ where m is the number of attributes and all attributes are numeric. X'_{1A_1} is the measured value of the first attribute belonging to the first data:

$$S_{A_1} = \frac{1}{n} \left(\sqrt{(X'_{1A_1} - M_{A_1})^2} + \sqrt{(X'_{2A_1} - M_{A_1})^2} + \dots + \sqrt{(X'_{nA_1} - M_{A_1})^2} \right) \quad X_1, X_2, \dots, X_n \in D, \quad (1)$$

where

$$M_{A_1} = \frac{1}{n} (X'_{1A_1} + X'_{2A_1} + \dots + X'_{nA_1}), \quad (2)$$

$$X_{1A_1} = \frac{X'_{1A_1} - M_{A_1}}{S_{A_1}}.$$

2.1.2. Multidimensional Diffusion Density Distribution. In iterative clustering algorithms the function adopted for density is the cut-off kernel [19]:

$$\rho_i = \sum_{j \in D/\{i\}} \chi(d_{ij} - d_c), \quad (3)$$

where

$$\chi(X) = \begin{cases} 1, & X < 0, \\ 0, & X \geq 0, \end{cases} \quad (4)$$

$$d_{ij} = \sqrt{(X_{iA_1} - X_{jA_1})^2 + (X_{iA_2} - X_{jA_2})^2 + \dots + (X_{iA_m} - X_{jA_m})^2},$$

where d_c is the cut-off distance determined by user and $d_c > 0$.

X_{iA_1} is the i th data with attributes A_1 and X_{jA_1} is the j th data with attributes A_1 . d_{ij} measures the Euclidean distance between X_{iA_1} and X_{jA_1} . So ρ_i is the number of data whose distance to the data i is less than d_c .

The clustering results of multidimensional data have shortcoming using the above functions adopted for density. Some multidimensional data are misjudged for the functions without considering the difference among attributes. If few attributes of some data change a lot but other attributes are close to other data, these mutating data can be misjudged into the cluster that they are not similar to for the changing attributes are ignored. In order to solve the above problems,

we put forward the Euclidean distance control factor based on aggregation density sparse degree E :

$$E = \log \left(\frac{\max \sqrt{(X_{iAq} - X_{jAq})^2}}{\min \sqrt{(X_{iAp} - X_{jAp})^2}} \right) d_{ij}, \quad (5)$$

where q is the q th attribute of data about which X_{iA} and X_{jA} have the maximum distance and p is the attribute of data about which X_{iA} and X_{jA} have the minimum distance.

According to the above views, we propose the optimized density formula:

$$\rho_i = \sum_{j \in D/\{i\}} \chi((1 + E) d_{ij} - d_c). \quad (6)$$

Formula (6) makes distinction among different attributes using the Euclidean distance control factor based on aggregation density sparse degree E . The attributes with large differences are given more weights' computing density, which reduces the risk of misjudged data. The optimized density is multidimensional diffusion density:

$$Z_\rho = \bar{\rho} + \delta_\rho, \quad (7)$$

where $\bar{\rho}$ is the average of multidimensional diffusion density ρ_i and δ_ρ is the standard deviation of multidimensional diffusion density. Z_ρ is standard value of multidimensional diffusion density. Let the data whose multidimensional diffusion density is bigger than Z_ρ in a collection named B :

$$\sigma = \begin{cases} \sigma_i = \min(d_{ij}), & \rho_i < \rho_j, \\ \sigma_i = \max(d_{ij}), & \rho_i \max \end{cases} \quad X_i, X_j \in B. \quad (8)$$

σ is the distance between X_i and X_j , and both of them are from collection B :

$$Z_\sigma = \bar{\sigma} + \delta_\sigma, \quad (9)$$

where $\bar{\sigma}$ is the average of distance between the data from collection B . And δ_σ is the standard deviation of distance between the data from collection B . Z_σ is standard value of distance. Let the data whose distance is bigger than Z_σ be in a collection named C . Reference [19] shows that the initial cluster centers have the large density and the far distance. So the data in collection C are likely to be initial cluster centers. We choose them as the initial cluster centers of initial data clusters.

2.1.3. Obtaining Initial Data Clusters by Multidimensional Diffusion Density Distribution. In this subsection we present execution steps of our proposed initial data clustering using multidimensional diffusion density distribution for k -means clustering. This algorithm consists of two parts. The first part deals with the the initial cluster centers of the initial data. Then we execute the second part of the algorithm to group the data.

$D = \{x_1, x_2, \dots, x_n\}$, which is the set of n data elements described with attributes $A1, A2, \dots, Am$ where m is number of attributes and all attributes are numeric. Compute ρ_i of each data, mean $\bar{\rho}$, standard deviation δ_ρ , and Z_ρ . Choose the data whose ρ_i is bigger than Z_ρ , and put them into collection B . Compute σ of each piece of data in B collection, $\bar{\sigma}$, standard deviation δ_σ , and Z_σ . Choose the data whose σ is bigger than Z_σ and put them into C collection, $C = \{c_1, c_2, \dots, c_{k'}\}$ is the set of the initial cluster centers of the initial data, and k' is the number of the initial cluster centers:

$$d(X_i, c_j) = \sqrt{\sum_{a=1}^m (X_{iAa} - c_{jAa})^2}, \quad (10)$$

$$i = 1, \dots, n; \quad j = 1, \dots, k'.$$

$d(X_i, c_j)$ measures the Euclidean distance between a pattern X_i and its cluster center c_j that is in the C collection.

k -means algorithm minimizes the function which is defined as object function [18]:

$$J_{\text{MSE}} = \sqrt{\sum_{i=1}^{k'} \sum_{x_j \in C_i} (x_j - c_j)^2}. \quad (11)$$

The k -means algorithm groups the data iteratively as follows. Choose the data in C collection as the initial cluster centers and the number of them is k' , the number of initial data clusters. Decide membership of the patterns in one of the K -clusters according to the minimum distance from cluster center criteria. Then calculate new c_j centers as:

$$c_j = \frac{1}{N} \sum_{X_i \in c_j} X_i. \quad (12)$$

N is the number of data from which the j th cluster center of. Repeat the previous steps till there is no change in cluster centers. The clusters are the initial data clusters. Let $W = \{w_1, w_2, \dots, w_{k'}\}$, $W_i = \{x_1, x_2, \dots, x_N\}$ be a clusters set of the initial data clusters.

2.2. Multiobjective Clustering Approach Based on Dynamic Cumulative Entropy. Obtaining the initial data clusters by multidimensional diffusion density distribution is the primary data processing. To get the accurate clustering results we need to cluster again on the basis of it, and the initial data clusters are the basic elements.

2.2.1. Multiobjective Clustering Function Based on Dynamic Cumulative Entropy. In the k -means algorithm the object function is JMSE. But it is not suitable for optimizing the primary clusters of data and determining k , the number of clusters because JMSE decreases monotonously for the increase of the number of clusters. When JMSE reduced to the global minimum, the initial cluster centers are putted away from their data, which makes single data a cluster. So proposed objective function refers to the information of each initial data cluster in the process of the deepening clustering.

The paper tries to solve the above problems by studying the clusters' structure by informational entropy theory and the principle of Maximum Informational Entropy. We propose a multiobjective clustering function J based on dynamic cumulative entropy to determine the final clustering results. Firstly, we define the information entropy formula of cluster:

$$H_i = -\sum_{i=1}^k \frac{P_i}{n} \ln \frac{P_i}{n}, \quad (13)$$

where P_i is the number of data in the i th initial data clusters, n is the number of data to be clustered, and k is the number of clusters.

According to the principle of Maximum Informational Entropy, if clusters have no elements, the information entropy is 0, $H = 0$. When clusters tend to be stable and meet the condition that the numbers of elements belonging to different clusters P_i are very similar, the information entropy is maximum; $H_{\max} = \ln k$. Based on the information entropy formula (14) of cluster, we define the cluster structure equilibrium degree S :

$$S_i = \frac{H_i}{H_{\max}} = -\sum_{i=1}^k \frac{((P_i/n) \ln (P_i/n))}{\ln k'}. \quad (14)$$

S is time-varying entropy-to-maximum entropy ratio, which shows the balance degree of clusters and $0 \leq S \leq 1$. When $S_i = 0$, cluster is in the most uneven condition. When $S_i = 1$ clusters are in ideal equilibrium state:

$$Y = 1 - S. \quad (15)$$

Y is the degree of equilibrium gain, which reflects the cluster occupancy of data.

Finally, we get the multiobjective clustering function based on dynamic cumulative entropy J :

$$J = J_{\text{MSE}} + \sum_{i=1}^k Y_i \mathfrak{R}, \quad (16)$$

where

$$\mathfrak{R} = \frac{2\beta}{k(k-1)} \left(\sum_{i=1}^k \sum_{j=1}^P (d(c_i, c_j)) \right). \quad (17)$$

$\mathfrak{R}$ is the average distance among the clusters. The multiobjective clustering function considers not only the distance among the cluster centers and its data influence on the clustering results, but also the amount of information. JMES decreases monotonically with the decreasing of distance among the cluster centers and its data influence on the clustering results. On the contrary, $\sum_{i=1}^k Y_i \mathfrak{R}$ increases monotonically with the increasing of the number of the global clusters. So $\sum_{i=1}^k Y_i \mathfrak{R}$ as restricting factor prevents the number of clusters k more than the real k for JMES decreasing. It improves the clustering accuracy and reliability when the number of clusters k is not determined.

FIGURE 1: Initial data cluster.

2.2.2. Multiobjective Clustering Approach Based on Dynamic Cumulative Entropy. The initial data clusters $W = \{w_1, w_2, \dots, w_{k'}\}$ by multidimensional diffusion density distribution need to be clustered again. The initial cluster centers $C = \{c_1, c_2, \dots, c_{k'}\}$ are determined again by the distance-information decision graph.

The less information of a cluster has, the smaller its uncertainty is and the more likely it is to be the final cluster. At the same time, the global classification is more stable. Formula for information of the initial data clusters is

$$I_{c_i} = -\log \left(\frac{P_i}{n} \right),$$

$$\sum_{i=1}^{k'} P_i = n, \quad (18)$$

$$i = 1, 2, \dots, k'.$$

Calculate the distance among the initial data clusters:

$$\beta = \begin{cases} \beta_i = d \min(M_i, M_j), & I_i < I_j, \\ \beta_i = d \max(M_i, M_j), & I_i \max, \end{cases} \quad ij \in C, \quad (19)$$

where $M_i = (1/N) \sum_{f=1}^N X_f$, X_f is the data from which the i th cluster center of.

By the nature of the clustering, we can know that the cluster's uncertainty should be small and the distance among clusters should be large. According to this, we can draw distance-information decision graph to determine k and the initial cluster centers. Take two-dimensional data for example, such as shown in Figure 1. Multidimensional diffusion density distribution obtains initial data clusters A, B, C, D, E, F, G, H, I, J, and K. Figure 2 is distance-information decision graph with the inverse of information of the initial data cluster as the horizontal axis ($1/I$) and the distance among initial data clusters as the longitudinal axis (β).

By the distance-information decision graph we can see that initial data clusters A, H, D, and K have compared low

FIGURE 2: Distance-information decision graph.

information and large distance among clusters because the larger k value can expand the searching scope of solution space, which avoids ignoring the sparse clusters. The selection of the initial cluster centers can be expanded according to the distance-information decision graph. Select the respective average values m of initial data clusters A, H, D, K, G, and B as the initial cluster centers and k_1 as 6; minimize the multiobjective clustering function with result J_1 . Select the respective average values m of initial data clusters A, H, D, K, and G as the initial cluster centers and k_2 as 5; minimize the multiobjective clustering function with result J_2 . Select the respective average values m of initial data clusters A, H, D, and K as the initial cluster centers and k_3 as 4; minimize the multiobjective clustering function with result J_3 . Then compare J_1 , J_2 , and J_3 ; the minimum is the best choice.

3. Experimental Results and Analysis

3.1. Experimental Objects and Related Settings. All experiments were performed on Intel® Core™ i5 with 3.30 GHz CPU and 4.00 GB of random access memory (RAM). All programs were coded by standard MATLAB language and operating system was Windows 7. To show the accuracy of the proposed algorithm, it has been applied to the two types of data sets. One is the AR artificial data set from experimental data and the other is UCI data. To define the quality of proposed algorithm we use 4 indicators. They are accuracy, Adjust Rand Index, MSE, and BIC. The “Adjusted Rand index” is the best clustering validity evaluation criterion [20]. And BIC is often used as an accurate evaluation [21]. Because the value of MSE is large, the results of it are scaled down. The MSE is different from other indicators; it is the smaller the better.

In this paper, an experimental system is used to evaluate the effectiveness of the proposed approach. A wireless information collection system of field soil temperature and humidity is used in this work to obtain the real data sets. The measurement system is shown in Figure 3. The field soil temperature is measured by TM-100 and humidity is measured by SP40A. The TM-100 measuring range is from 0% to 100%, and the SP40A measuring range is from -20°C to 60°C . The JN5148 is used for receiving and transmitting the soil temperature and humidity data. Then the data is controlled and recorded digitally by a pc. The measurements were

TABLE 1: The artificial data AR.

Cluster	Cluster center	Data range	Number of samples
Cluster 1	(10, 25)	[2, 15]	110
Cluster 2	(24, 25)	[15, 29]	80
Cluster 3	(18, 9)	[10, 25]	120
Cluster 4	(35, 10)	[27, 40]	80

FIGURE 3: Field soil temperature and humidity collection system.

FIGURE 4: Four fields for measurement.

performed in four fields. So the record data are from four real data sets. The collecting places are shown in Figure 4. Table 1 shows the measurements data named artificial data AR. The artificial data AR that sets two-dimensional distribution is shown in Figure 5.

The AR artificial data set clusters by the proposed algorithm in this paper. Multidimensional diffusion density distribution obtains initial data clusters. The proposed Z_p of AR is 37, and the Z_σ of AR is 2.97. The data samples with ρ_i bigger than Z_p and σ bigger than Z_σ are putted into C. And $k' = 11$, $C = \{(5.6, 28.0), (5.1, 24.6), (10.7, 24.4), (19.2, 25), (24.8, 27.4), (26.3, 22.5), (18.9, 12.8), (11.3, 8.2), (19.1, 6.6), (31.7, 14.3), (35.4, 6.5)\}$. The elements in C are the

FIGURE 5: The artificial data AR.

FIGURE 6: Initial data clusters.

initial cluster centers of the initial data. We obtain initial data clusters A, B, C, D, E, F, J, H, I, G, and K by using multidimensional diffusion density distribution. The results are shown in Figure 6.

The optimal clustering result is obtained by using the multiobjective clustering approach based on dynamic cumulative entropy. The information and the distance among the initial data clusters are obtained by the proposed formula.

The distance-information decision graph is shown in Figure 7. From the graph we can clearly see that the uncertainty of initial data clusters G, C, E, and J is low and the distance among them is large. So we select the respective average values m of initial data clusters G, C, E, J, and H as the initial cluster centers and k_1 value as 5 and calculate the minimum of J . Then we select the respective average values m of initial data cluster G, C, E, J as the initial cluster centers and k_2 value as 4 and calculate the minimum of J . From the

FIGURE 7: Distance-information decision graph.

FIGURE 8: The optimal clustering result.

proposed formula $J_1 = 247.49$, $J_2 = 218.36$, so the optimal number of clusters is 4. And the optimal clustering result is shown in Figure 8.

We get the cluster centers of AR. They are (10.2, 26.2), (23.7, 26.4), (17.8, 9.1), and (36, 10.5). And the real cluster centers are (10, 25), (24, 25), (18, 9), and (35, 10). The similarity between our cluster centers and the real cluster centers is 95.582%, 97.628%, 99.337%, and 96.968%. The average value of similarity of cluster centers is 94.72%, the clustering accuracy is 92.14%, Adjust Rand Index is 0.7840, and MSE (Mean Squared Error) is 3.82.

The AR artificial data is clustered by [11, 15]. The comparisons of cluster center's similarity and clustering result's accuracy and Adjust Rand Index and MSE, using our algorithm and literature [11, 15], for AR data sets, are shown in Figures 9, 10, 11, 12, and 13.

The comparison of similarity of initial cluster centers computed using our algorithm and [11, 15], for the AR data set, is shown in Figure 9. The algorithm with high value will have a good selection of cluster centers. The result of our algorithm

TABLE 2: Comparison of running time on UCI data set.

DS	Reference [11]	Reference [15]	Our algorithm
Soybean • small	1.83	2.056	2.169
Iris	2.89	2.962	2.984
Segmentation	2.883	3.055	3.087
WDBC	3.142	3.436	3.541
Pima Indians Diabetes	4.780	4.985	5.001
Wine	7.106	7.938	7.942
Ionosphere	9.488	9.647	9.884
Yeast	17.219	17.863	17.894

FIGURE 9: The comparisons of cluster center's similarity.

has the highest value of similarity of initial cluster centers. What is more, for the other numerous measures used for clustering validation, the accuracy, Adjust Rand Index MSE, and BIC, our algorithm has better performance than other algorithms.

Obviously, our algorithm can find the optimal clustering number k and the optimal cluster center to get the best clustering result, which provides the accuracy of clustering.

3.2. Experimental Results and Comparative Analysis of UCI Data Sets. In order to verify the effectiveness of the our algorithm, this paper selects 8 data sets of UCI data to carry out the experiment. They are Iris ($150 \times 4 \times 3$), Soybean • small ($47 \times 35 \times 4$), Segmentation ($210 \times 19 \times 7$), WDBC ($569 \times 30 \times 2$), Pima Indians Diabetes ($768 \times 8 \times 2$), Wine ($178 \times 13 \times 3$), Ionosphere ($351 \times 34 \times 2$), and Yeast ($1484 \times 8 \times 10$). Clustering results use the clustering time, clustering accuracy, the “Rand index Adjusted,” and MSE for evaluation.

FIGURE 10: The comparisons of clustering results accuracy.

Table 2 shows the comparison of average of the running time of clustering algorithms; the unit is S. Figures 14, 15, 16, and 17 are the comparisons about the clustering accuracy, the “Rand index Adjusted,” MSE, and the BIC of our algorithm and [11, 15].

According to clustering results of UCI data we can see that for all the tested data sets, the proposed algorithm gets improved and consistent clusters for all data sets in comparison to [11, 15]. Its accuracy for complicated data sets also was high. However, the algorithm of [11] is quicker than the proposed algorithm.

4. Conclusion

We have presented an algorithm for iterative clustering algorithm with a small amount of prior knowledge for nonuniform sparse data. This procedure is based on the dynamic cumulative entropy. However, the outliers are more susceptible to a change in cluster membership. We propose

FIGURE 11: The comparisons of Adjust Rand Index.

FIGURE 13: The comparisons of BIC.

FIGURE 12: The comparisons of MSE.

FIGURE 14: Comparison of the accuracy for UCI data.

multidimensional diffusion density distribution in computing initial cluster centers, which generate the initial data clusters that may be more than the number of desired clusters. So we need to cluster again on the basis of them for getting the accurate clustering results by multiobjective clustering approach based on dynamic cumulative entropy. Experimental results show improved and consistent cluster

structures. And the experimental results show that compared with full search, our proposed method may get production up by 5 percent for nonuniform sparse data set from field soil temperature and humidity. The superiority of our method over other algorithms is more remarkable, when a data set with higher data dimension and larger number of clusters is used. And our method handles clustering with little prior knowledge, so our method is at high complexity.

Following our paper, several interesting problems deserve to be explored, for example, how to reduce computation and enhance the algorithm speed with a similar approximation ratio. To generalize our method to the setting of arbitrary partition is also challenging but useful in some scenarios.

FIGURE 15: Comparison of the Adjust Rand Index for UCI data.

FIGURE 16: Comparison of the MSE for UCI data.

FIGURE 17: Comparison of the BIC for UCI data.

Competing Interests

The authors declare that they have no competing interests.

References

- [1] Y. Şenbabaoğlu, G. Michailidis, and J. Z. Li, "Critical limitations of consensus clustering in class discovery," *Scientific Reports*, vol. 4, article 6207, 2014.
- [2] A. Oravec, A. Apostolov, K. Polak et al., "Ikaros mediates gene silencing in T cells through Polycomb repressive complex 2," *Nature Communications*, vol. 6, article 8823, 2015.
- [3] S. Weingarten-Gabbay, S. Elias-Kirma, R. Nir et al., "Systematic discovery of cap-independent translation sequences in human and viral genomes," *Science*, vol. 351, no. 6270, pp. 240–253, 2016.
- [4] B. Laforce, C. Carlier, B. Vekemans et al., "Assessment of ovarian cancer tumors treated with intraperitoneal cisplatin therapy by nanoscopic X-ray fluorescence imaging," *Scientific Reports*, vol. 6, Article ID 29999, 2016.
- [5] J.-X. Cheng and X. S. Xie, "Vibrational spectroscopic imaging of living systems: an emerging platform for biology and medicine," *Science*, vol. 350, no. 6264, Article ID aaa8870, pp. 1054–1064, 2015.
- [6] S. Thakurela, A. Garding, R. B. Jung et al., "The transcriptome of mouse central nervous system myelin," *Scientific Reports*, vol. 6, article 25828, 2016.
- [7] S. Jesse, M. Chi, A. Belianinov et al., "Big data analytics for scanning transmission electron microscopy ptychography," *Scientific Reports*, vol. 6, Article ID 26348, 2016.
- [8] E. P. de Moura, F. D. Melo, F. F. R. Damasceno et al., "Classification of imbalance levels in a scaled wind turbine through detrended fluctuation analysis of vibration signals," *Renewable Energy*, vol. 96, pp. 993–1002, 2016.
- [9] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [10] F. Ferrucci and S. Bock, "Pro-active real-time routing in applications with multiple request patterns," *European Journal of Operational Research*, vol. 253, no. 2, pp. 356–371, 2016.
- [11] Y.-T. Kao, E. Zahara, and I.-W. Kao, "A hybridized approach to data clustering," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1754–1762, 2008.
- [12] B. Bahmani Firouzi, M. Sha Sadeghi, and T. Niknam, "A new hybrid algorithm based on PSO, SA, and k-means for cluster analysis," *Innovative Computing Information and Control*, vol. 6, no. 4, pp. 1–10, 2010.
- [13] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis," *Journal of Applied Soft Computing*, vol. 10, no. 1, pp. 183–197, 2010.
- [14] T. Niknam, E. T. Fard, S. Ehrampoosh, and A. Rousta, "A new hybrid imperialist competitive algorithm on data clustering," *Sadhana—Academy Proceedings in Engineering Sciences*, vol. 36, no. 3, pp. 293–315, 2011.
- [15] F. M. Bianchi, L. Livi, and A. Rizzi, "Two density-based k-means initialization algorithms for non-metric data clustering," *PAA. Pattern Analysis and Applications*, vol. 19, no. 3, pp. 745–763, 2016.
- [16] V. Tunali, T. Bilgin, and A. Camurcu, "An improved clustering algorithm for text mining: multi-cluster spherical k-means," *International Arab Journal of Information Technology*, vol. 13, no. 1, pp. 12–19, 2016.

- [17] J. Tvrđík and I. Křivý, “Hybrid differential evolution algorithm for optimal clustering,” *Applied Soft Computing*, vol. 35, pp. 502–512, 2015.
- [18] S. S. Khan and A. Ahmad, “Cluster center initialization algorithm for K-means clustering,” *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293–1302, 2004.
- [19] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
- [20] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [21] D. Pelleg and A. Moore, “X-means: extending K-means with efficient estimation of the number of clusters,” in *Proceedings of the 17th International Conference on Machine Learning (ICML '00)*, pp. 727–734, 2000.

Hindawi

Submit your manuscripts at
<http://www.hindawi.com>