An Algorithm for Automatic Recognition of Cluster Centers Based on Local Density Clustering

Ye Xuanzuo, Li Dinghao, He Xiongxiong
Zhejiang University of Technology, 310023 Hangzhou, Zhejiang, P.R. China
E-mail: hxx@zjut.edu.cn

Abstract: Based on the local density clustering (LDC) algorithm, a new automatical local density clustering (ALDC) algorithm is proposed in this paper. Different from the existing LDC algorithm, the ALDC can capture the cluster center automatically. The new algorithm calculates the local density and the distance deviation of every point and expands the difference between the potential cluster center and other points by using these two features. The expansion of difference enables the machine automatically to recognize the cluster centers, and then assigns the remaining points by their closest neighbor of higher density. Experimental results on data sets show that the ALDC algorithm can achieve a accurate clustering and obtain a higher accuracy of clustering result than other two classical cluster algorithms.

Key Words: Cluster, Local Density, Expand Difference, Automatical Recognize

1 INTRODUCTION

Clustering algorithms are used to divide the entities with the similar features into the same category, and the entities in different categories have many unlikeness features. Cluster analysis which can be seen as an unsupervised machine learning process is one of the most important research contents in data mining, pattern recognition and market analysis [1][2].

Up to now, many cluster analysis algorithms have been proposed. K-means [3] is one of the classical partitioning-based methods, which need to determine the cluster center in advance, and the points are divided on the basis of the shortest distance between them and the center. The other typical algorithms based on partitioning includes K-Medoid [4] and CLARANS [5]. DBSCAN [6] is another density-based cluster algorithm, the method to compute the density of a point is counting the number of points which are in the area of a circle whose center is this point, and this circle has a specified radius. Points with density above a specified threshold are considered as a cluster center. The other points which are density-reachable from the cluster center should be divided into the same cluster. The other classical density-based algorithms also include OPTICS [7], DENCLUE [8].

Rodriguez and Laio [9] proposed a new density-based cluster algorithm by calculating the local density and finding the density peaks in 2014. The main idea of this algorithm is that the potential cluster centers have a higher local density than their neighboring points and a relatively greater distance from another higher local density points, and some of them are artificially chosen as cluster centers, then distributing the remaining sample points to their respective cluster. Local density clustering (LDC) has the advantage of simple and efficient, also could recognize any shape and size of the cluster, and the distribution of the points except the cluster centers will be accomplished by one step without any iteration. However, there are still some aspects need to be improved, for instance, the cluster center should be specified artificially and this step will influence the accuracy of the clustering result.

A new clustering algorithm called automatical local density clustering (ALDC) is proposed in this paper. The cluster center can be automatically captured by machine without any human intervention. The experimental results show that this algorithm has good performance of obtaining the clustering results and has satisfying rate of accuracy.

2 BACKGROUND AND RELATED WORK

The local density clustering algorithm considers every point in data set has two features, which can be described as local density and distance deviation. There are two methods to calculate the local density: cut-off kernel and Gaussian kernel. Applying the cut-off kernel, the local density ρ_i of data point i is defined as:

$$\rho_i = \sum_{j \in I_U \setminus \{i\}} \chi(d_{ij} - d_c) \tag{1}$$

Among the formula

$$\chi(a) = \begin{cases} 1 & \text{a<0} \\ 0 & \text{a}\geq 0 \end{cases} \tag{2}$$

where d_c is called the cutoff distance and is set in advance, d_{ij} is the Euclidian distance [10] between two points. Sort the d_{ij} in ascending order like $d_1 <= d_2 <=, ..., d_M$ and make $d_c = d_{f(Mt)}$, where $t \in (0,1)$ is the ratio of average number of neighbors in total points and $f(Mt)$ is expressed for making Mt round up to integer, I_U is the collection of

This work is supported by National Nature Science Foundation under Grant No. 61473262.

978-1-5090-4657-7/17/$31.00 ©2017 IEEE

the whole points in data set. The local density ρ_i of data point i in Gaussian kernel is defined as:

$$\rho_i = \sum_{j \in I_U \setminus \{i\}} \exp(-\frac{d_{ij}}{d_c})^2 \tag{3}$$

where d_c and d_{ij} are the same definition in cut-off kernel and ρ_i represents that the number of points whose distance is smaller than d_c to point i. The distance deviation δ_i is defined as:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & i \text{ is not the highest density point} \\ \max\limits_{j} (d_{ij}) & i \text{ is the highest density point} \end{cases} \tag{4}$$

where δ_i represent the minimum distance between point i to another point which has higher local density than this point i, and the distance of the highest local density point is defined as the maximum distance between the point to any other points to make sure this point will be chosen as a cluster center. The point which has high local density and large distance deviation will be regarded as the cluster center which means the density is higher than its neighbors and the distance between the cluster center and another center is relatively larger. By drawing the two dimensional diagram of ρ_i and δ_i which is called the decision diagram, one can artificially specify the cluster center which is on the position of the upper right of this diagram. At last, each remaining point is distributed to the same cluster as its nearest neighbor which has higher density.

3 AN IMPROVED PROCEDURE

On the basis of theories mentioned above, the measure index γ_i is defined as:

$$\gamma_i = \rho_i * \delta_i \tag{5}$$

where γ_i is the product of ρ_i and δ_i, the point i that have a large value of γ_i will more likely to be regarded as the potential cluster center, but just by observing the value of γ_i to artificially determine the potential cluster center may lead to the problem of getting the wrong number of the cluster centers, and the number is less or more will influence the clustering result seriously. The difference between the value of γ_i is not big enough and may not be easily distinguished whether the point is the cluster center or not. Regarding to the above issues, the new algorithm uses a method of expansion of the difference between the value of γ_i of the point to another to settle them, the E_i is defined as:

$$E_i = \Sigma_{j \in I_U \setminus \{i\}} \sqrt{(\gamma_i - \gamma_j)^2} \tag{6}$$

where $\sqrt{(\gamma_i - \gamma_j)^2}$ is the absolute difference of the point i to any other points in data set, E_i is the sum of the difference values, and the difference between the cluster center and the other points will be expanded by this step, which laid the foundation for next step of machine automatically capturing the cluster center after the expansion. To illustrate this process, we test on a man-made data set and the distribution of E_i is shown in Figure 1, the value of E_i has different order of magnitude and which can be seen in Figure 1. By comparing with each other, the potential cluster center is highlighted on the top, using the monotonically decreasing property of Exp-function with a negative independent variable:

$$Z_i = e^{-E_i} \tag{7}$$

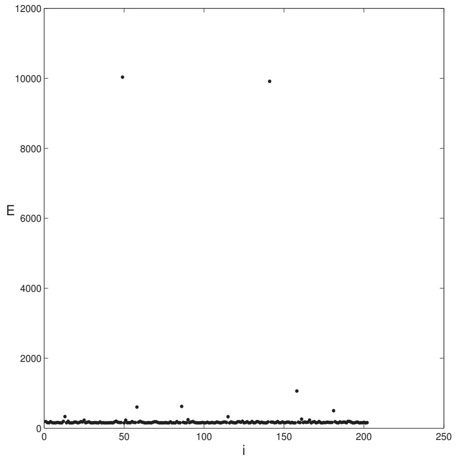


Figure 1: The distribution map of E_i

where Z_i displays the difference between the cluster center and other points, as shown in Figure 2, all the values of Z_i are close to 1 or 0 because of the different order of magnitude of the E_i, and the higher order of magnitude of E_i is, the more closer Z_i to 0 is. It can be given a measure criterion, if the value of Z_i of the point is under the line of 0.1, we can conclude that these points which are lower than this dividing line we set can be considered as the cluster center and machine can automatically capture it without human intervention. In the distribution map of Z_i, different types except the black spots of the point represents the different cluster center and the line in the graph represents the dividing line.

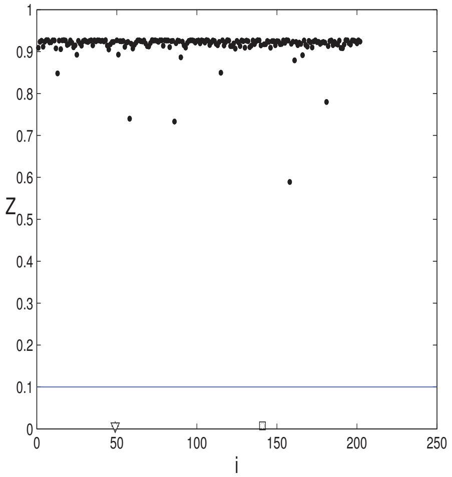


Figure 2: The distribution map of Z_i

This new algorithm steps are as follows:

- *Step I*: Calculate the Euclidean distance between every point in data set
- *Step II*: Calculate the local density and the distance deviation of every points
- *Step III*: Calculate the product of local density and distance deviation according to (5)
- *Step IV*: Expand and show the difference between the potential cluster center and the remaining points according to (6), (7)
- *Step V*: Use the measure criterion to capture the cluster center
- *Step VI*: Assign the remaining points to their nearest neighbor which has higher density

4 EXPERIMENT TEST AND ANALYSIS

In order to verify the performance of the ALDC algorithm, we test the algorithm on the man-made data sets and U-CI data sets respectively and compare with other clustering algorithms.

4.1 Man-made Data Sets

We test on three data sets called Spiral, Aggregation, Flame in our experiments, Table 1 shows the features of these data sets.

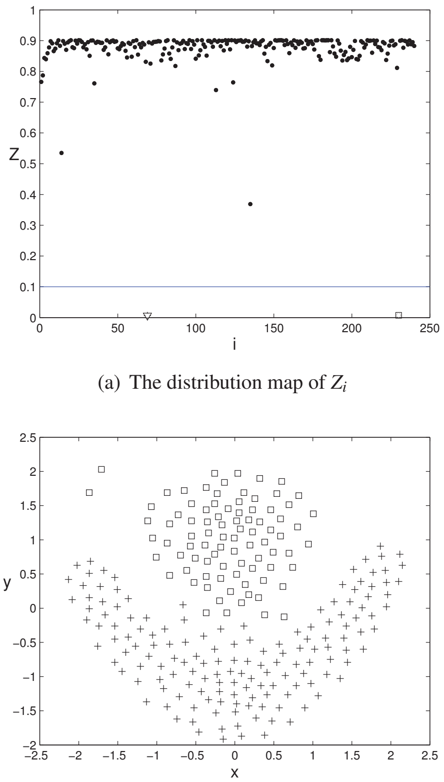
Table 1: The features of Man-made data sets

data set	n	D	k
Spiral	312	2	3
Aggregation	788	2	7
Flame	240	2	2

Where k is the number of clusters, and n is the number of points in data set, and D is the dimension of the data set. We use the accuracy of clustering results [11] to evaluate and compare the performance of the ALDC algorithm, the LDC and other two classical algorithms(K-means, DBSCAN). As defined follows:

$$r = \frac{\sum_{i=1}^{k} * c_i}{n} \quad (8)$$

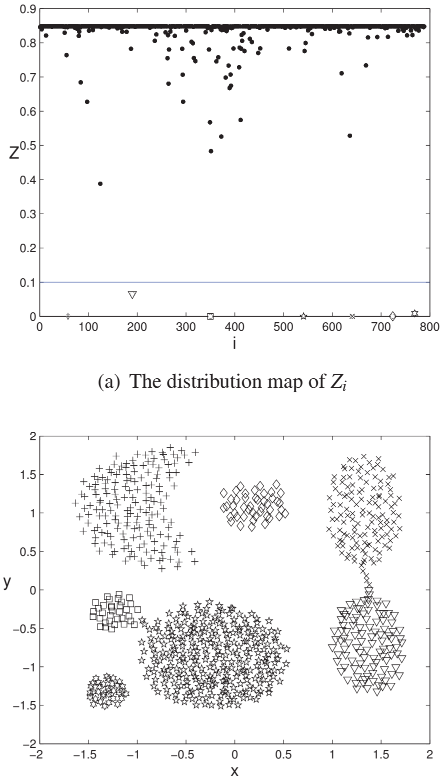
where c_i is the point which is been clustered correctly, k, n are the same definition in Table 1, r is the accuracy of clustering results. Using the three data sets of Table 1 to complete the whole process of the ALDC algorithm, we obtain the distribution maps of the value of Z_i and the graphs of the clustering result, which are shown in Figure 3, Figure 4, Figure 5.



(a) The distribution map of Z_i

(b) The graph of the clustering result

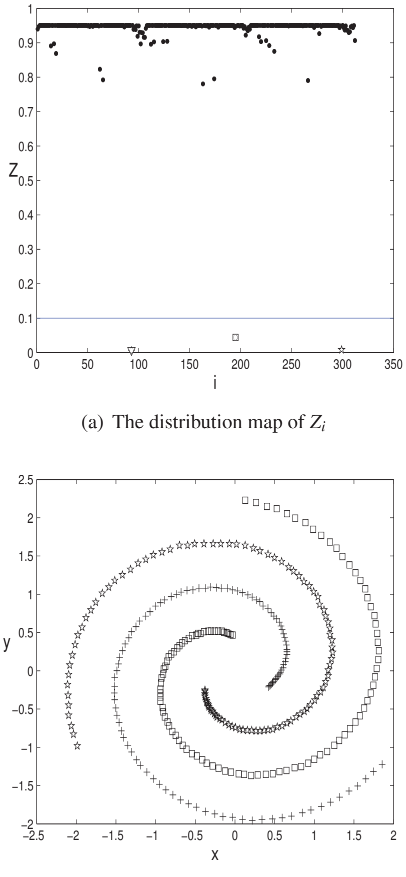
Figure 3: Experimental results of Flame



(a) The distribution map of Z_i

(b) The graph of the clustering result

Figure 4: Experimental results of Aggregation



(a) The distribution map of Z_i

(b) The graph of the clustering result

Figure 5: Experimental results of Spiral

It is obvious that there is a striking difference between the cluster center to other points which is shown in the distribution map of the value of Z_i. The value of Z_i of the cluster center i almost approach 0 and don't need artificially compare the value of γ_i with each others to determine the cluster center. For this characteristic, machine can automatically and easily capture the cluster center, and according to these center points, the remaining points are assigned to the same cluster of their closest neighbor with higher density. At last getting a satisfactory result which are shown in the three graphs of the clustering results. The clustering accuracy of ALDC, LDC, K-means, DBSCAN are shown in Table 2, it can be seen that ALDC and LDC have better performance on these data sets than the other two classical algorithms.

Table 2: The clustering results accuracy of four algorithms about Spiral, Aggregation and Flame

data set	ALDC	LDC	K-means	DBSCAN
Spiral	1.0000	1.0000	0.3395	0.9918
Aggregation	1.0000	1.0000	0.7893	0.7894
Flame	1.0000	1.0000	0.8378	0.9916

4.2 UCI Data Sets

In order to validate the performance of ALDC algorithm and compare with LDC, K-means and DBSCAN, we also use two actual UCI data sets which are called Iris and Seed, Table 3 shows the features of these two data sets.

Table 3: The features of UCI data sets

data set	n	D	k
Iris	150	4	3
Seed	210	7	3

The distribution maps of the value of Z_i are shown in Figure 6, Figure 7 and the comparison table of the accuracy of clustering results is shown in Table 4.

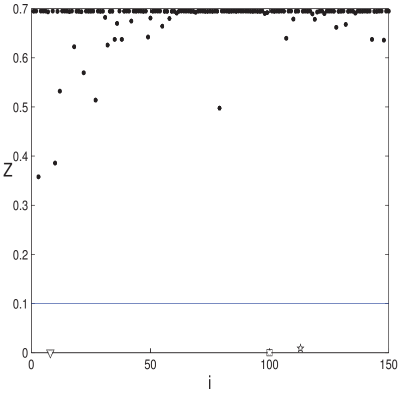


Figure 6: The distribution map of Z_i(Iris)

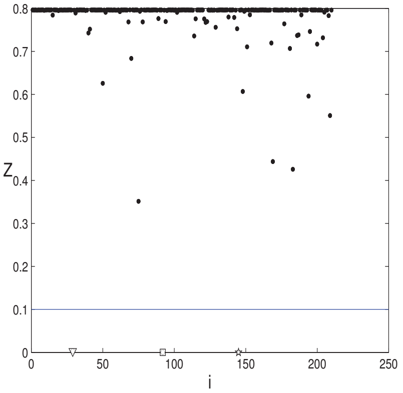


Figure 7: The distribution map of Z_i(Seed)

Table 4: The clustering result accuracy of four algorithms about Iris and Seed

data set	ALDC	LDC	K-means	DBSCAN
Iris	0.9067	0.9067	0.8933	0.6600
Seed	0.8857	0.8857	0.8738	0.3333

On the higher dimension of the actual data sets, ALDC still find the cluster centers accurately and automatically, and the comparison table shows that the accuracy of ALDC is much higher than the other two algorithms on these two data sets like the three data sets we tested on before. ALDC replaces the step of artificial choosing the cluster center in LDC by machine automatical selecting without the human intervention and do not influence the accuracy.

5 CONCLUSIONS

In this paper, a new algorithm, called automatical local density clustering, is proposed in order to improve the local density clustering algorithm by automatical selecting the cluster center replace the way of artificial choosing and eliminating the errors which are caused by human intervention. In addition, ALDC still retain the advantages of LDC: recognize any shapes of the cluster, do not need to specify the number of cluster in advance and assign the remaining points without any iteration. Experimental tests on man-made and actual UCI data sets show that the ALDC algorithm has good performance in getting a satisfactory clustering result in man-made data sets and higher accuracy than other two classical algorithms and which are also validated in actual UCI data sets.

REFERENCES

[1] J G Sun, J Liu, and L Zhao. Clustering algorithms research, Journal of Software, 2008, 19(1): 48-61.

[2] J Wang, S T Wang, and Z H Deng. Survey on challenges in clustering analysis research, Kongzhi Yu Juece/control and Decision, 2012, 27(3): 321-328.

[3] A K Jain. Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters, 2010, 31(8): 651-666.

[4] H S Park and C H Jun. A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 2009, 36(2): 3336-3341.

[5] R T Ng and J Han. CLARANS: A method for clustering objects for spatial data mining, IEEE Transactions on Knowledge and Data Engineering, 2002, 14(5): 1003-1016.

[6] D Birant and A Kut. ST-DBSCAN: An algorithm for clustering spatialCtemporal data. Data Knowledge Engineering, 2007, 60(1): 208-221.

[7] H P Kriegel and M Pfeifle. Hierarchical Density-Based Clustering of Uncertain Data. Proceedings of the 5th IEEE International Conference on Data Mining. Houston, Texas, USA, 2005: 689-692.

[8] A Hinneburg and H H Gabriel. DENCLUE 2.0: fast clustering based on kernel density estimation. Proceedings of the 7th International Symposium on Intelligent Data Analysis. Ljubljana, Slovenia, 2007: 70-80.

[9] Rodriguez A, and Laio A. Clustering by fast search and find of density peaks, Science, 2014, 344(6191): 1492-1496.

[10] E P Xing, A Y Ng, and M I Jordan. Distance metric learning with application to clustering with side-information, Advances in Neural Information Processing Systems, 2003, 15: 505-512.

[11] Z HUANG. Clustering large data sets with mixed numeric and categorical values. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 1997: 21-34.