

An Improved Clustering Algorithm Based on Density Distribution Function

Jianhao Tan (Corresponding author)

Electrical and Information Engineering College, Hunan University

Changsha, Hunan 410082, China

E-mail: tanjianhao96@sina.com.cn

Jing Zhang

Electrical and Information Engineering College, Hunan University

Changsha, Hunan 410082, China

Weixiong Li

Electrical and Information Engineering College, Hunan University

Changsha, Hunan 410082, China

The research is financed by National Natural Science Foundation (Approval No. 60634020) and Hunan Province Natural Science Foundation (Approval No. 08JJ3132).

Abstract

Some characteristics and weak points of traditional density-based clustering algorithms are deeply analysed, then an improved way based on density distribution function is put forward. K Nearest Neighbor (KNN) is used to measure the density of each point, then a local maximum density point is defined as the center point. By means of local scale, classification is extended from the center point. For each point there is a procedure to find whether it is a core point by a radius scale factor. Then the classification is extended once again from the core point until the density descends to the given ratio of the density of the center point. The tests show that the improved algorithm greatly improves the sensitivity of density-based clustering algorithms to parameters and enhances the clustering effect of the high-dimensional data sets with uneven density distribution.

Keywords: Clustering algorithms, KNN, Density distribution function, OPTICS, DENCLUE, Local scale, Radius scale factor

1. Introduction

Clustering analysis is a very active research topic in data mining, and has extensive applications in such areas as mode identification, picture processing, market marketing, etc. Some algorithms are only able to find globular clusters but difficult in detecting arbitrary-shape ones. Density-based clustering ways provide a way of solving the clustering of arbitrary-shape data sets. There are such typical algorithms as DBSCAN, OPTICS and DENCLUE among density-based clustering methods. DBSCAN divides districts with enough high density into clusters and can detect arbitrary-shape clusters in space databases with noises. But the algorithm asks users determine input parameters according to their experiences, which is not very available toward real high-dimensional data sets. In addition, the algorithm is much sensitive to parameter values, whose tiny changes can produce clustering results with great differences (Chen Yan, Geng Guohua, Zheng Jianguo, 2005, pp12-16; Hong Shaorong, Xiao Wenjun, 2008, pp24-27; Ma Shuai, 2003, pp34-37). OPTICS calculates a clustering-ordering for auto and alternative clustering analysis, and the ordering represents the density-based clustering structure of data, which includes these information equal to density-based clustering required from a comprehensive parameter set scope. Because OPTICS is equivalent with DBSCAN in structure, they have the same time complexity. DENCLUE, whose running speed is quicker than DBSCAN and OPTICS, is a clustering analysis way based on a group of density distribution function, so can more formally define center-defined clusters and arbitrary-shape ones. However the algorithm requires selecting carefully density parameters and noise threshold, like DBSCAN, being sensitive to parameters (Rong qiusheng, Yan Junbiao, Guo Guoqiang, 2004, pp12-16). By means of combining the characteristics of OPTICS and DENCLUE, an improved way based

on density distribution function is put forward in the paper. KNN is used to measure the density of each point, then a local maximum density point is defined as the center point. By means of local scale, current clustering is extended the edge of density which is decided by a scale factor. For each point there is a procedure to find whether it is a core point by a radius scale factor. The tests show that the algorithm is an auto and alternative quick clustering way.

2. Traditional Density-based Clustering Algorithms

2.1 OPTICS

OPTICS(Ordering Points to Identify the Clustering Structure) doesn't directly produce a data set cluster, and it calculates a clustering-ordering for auto and alternative clustering analysis, and the ordering represents the density-based clustering structure of data, which includes these information equal to density-based clustering required from a comprehensive parameter set scope.

To set up the sets or ordering of density-based clustering, a group of distance parameter values are processed at the same time by extending DBSCAN. For sake of building different clustering simultaneously, objects are processed by a specific order, by which density-reachable objects with minimum ϵ are selected in order that high density clustering can be firstly completed. According to this viewpoint, the two values of each object, core-distance and reachability-distance, are saved.

(a) the core distance of object p is minimum ϵ to make the p become a core object. If the p isn't a core object, the core distance of the p isn't defined.

(b) the reachability-distance of object q to object p is the greater one between the core distance of the p and the Euclid distance of the p to the q . If the p isn't a core object, the reachability-distance between the p and the q isn't defined.

OPTICS establishes an ordering of objects among databases, and saves the core-distance and suitable reachability-distance of each object. By the ordering information, OPTICS abstracts clusters (Chen Yan, Geng Guohua, Zheng jianguo, 2005, pp12-16; Hong Shaorong, Xiao Wenjun, 2008, pp24-27; Ma Shuai, 2003, pp34-37; Rong qiusheng, Yan Junbiao, Guo Guoqiang, 2004, pp12-16).

2.2 DENCLUE

DENCLUE(DENsity-based CLUstEring) is a clustering algorithm based on a group of density distribution function. The algorithm is mainly based on the following ideas: (1) the influence of each data point can be formally imitated by a mathematics function, which describes the influence of a data point to its neighbor, called influence function; (2) the whole density of data space can be modeled into the total of influence function of all data points; (3) clustering can be gotten by determining a density attractor, which is the local biggest point of the global density function.

By density function, the gradient and density attractor of the function can be defined. A point is density-attracted by a density attractor if there are a group of points $x_0, x_1, \dots, x_k, x_0 = x, x_k = x^*$, for $0 < i < k$, the gradient of x_{i+1} is along x_i . For a continuous differential influence function, density attractors of a group of data points can be calculated with the help of the hill climbing algorithm using the gradient.

Based on these concepts, the center-defined cluster and arbitrary-shape one can be formally defined. The center-defined cluster of a density attractor x^* is a sub-set C density-abstracted by x^* , whose density function value is no less than the threshold ξ ; Otherwise, namely, if its density function value is less than the threshold ξ , it is called an isolated point. An arbitrary-shape cluster is a set of the sub-set C . From one area to another, there exists a path, along which the density function value of each point is no less than ξ (Chen Yan, Geng Guohua, Zheng jianguo, 2005, pp12-16; Hong Shaorong, Xiao Wenjun, 2008, pp24-27; Ma Shuai, 2003, pp34-37; Rong qiusheng, Yan Junbiao, Guo Guoqiang, 2004, pp12-16).

2.3 Weak Points of OPTICS and DENCLUE

Because OPTICS is an extension of DBSCAN and equivalent with DBSCAN in structure, they have the same time complexity, i.e., being $O(n \log n)$ while adopting space index, otherwise being $O(n^2)$, although OPTICS can realize auto and alternative clustering and is not sensitive to parameters. So slow running speed is one of its drawbacks.

DENCLUE applies grid units and only saves the information of grid units combining data points. It administrates

these units by means of a tree-shaped access structure. So it is quicker than DBSCAN and OPTICS. But it requires a careful choice of density parameter σ and noise threshold, it is evident that it is parameter-sensitive.

3. Improved Clustering Algorithm Based on Density Distribution Function

Aimed at the problems put forward above, in the improved clustering algorithm based on density distribution function, the idea of local scale is conducted (Ergun Bicici, Deniz Yuret, 2007; Zhou Shui-geng, Zhou Aoying, Jin Wen, Fan Ye, Qian Weining, 2000, pp735-744), namely, clustering is realized with the help of local statistics. So-called local scale means the ratio of the average distance of KNN points of some point in the data set to the average distance of KNN points of the center point. By local scale, the scale factors among clusters with different density can be found, and then the simulation relation (matrix) of the data set, which is reflexive and symmetrical, can be created. On the basis, according to the concept, transitive closure, the simulation relation can be transformed into an equivalence relation. Then the *ratio* level cut set of the equivalence relation can be gotten by selecting suitable scale threshold *ratio*. At last, the classification of the points can be determined according to the level cut set. The algorithm is related with the following concepts (Li Fei, Xue Bin, Huang Yalou, 2002, pp94-96; Zhou Yonggeng, Zhou Aoying, Cao Jing, 2000, pp1153-1159; Chen Ning, Chen An, Zhou Longxiang, 2002, pp1-7).

1) Density Distribution Function of Points

Suppose p and i be objects in d -dimensional characteristic space. The influence function of p and i is an one $Density_B^i : F^d \rightarrow R_0^+$, and can be defined according to a basic influence function.

$$Density_B^i(p) = Density_B(p, i) \quad (1)$$

In principle, the influence function can be an arbitrary function, and can be determined by the distance between two points in some neighbor. The distance function $d(p, i)$ should be reflexive and symmetrical like the Euclid distance, applied to calculate a square wave influence function

$$Density_{Square}(p, i) = \begin{cases} 0 & \text{if } d(p, i) = 0 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

or a Gauss influence function

$$Density_{Gause}(p, i) = e^{-\frac{d(p, i)^2}{2\sigma^2}} \quad (3)$$

The density function of an object $p(p \in F^d)$ is defined as the total of the influence of all data points. Given k data objects $D = (x_1, x_2, \dots, x_k) \subset F^d$, the density function of p is defined as follows

$$Density_B^D(p) = \sum_{i=1}^k Density_B^i(p) \quad (4)$$

and the density function of p obtained by the Gauss influence function is

$$Density(p) = Density_B^D(p) = \sum_{i=1}^k e^{-\frac{d(p, i)^2}{2\sigma^2}} \quad (5)$$

where $d(p, i)$ is the Euclid distance between p and its No.1 nearest neighbor point. The smaller the distance is, the more tightly packed the neighbor of the point is and the bigger the density of the point is.

2) Core Point

Such a point is defined as a core point when some given ratio *ratio* of the distance between p and its No.K nearest neighbor point is as the radius *radius* and the number of the points among the nearest neighbor of the circle, which uses the point as the center and the *radius* as the radius, is bigger than some given threshold *MinPts*. Classification is extended from the core point.

3) Center Point

The center point is the one at present with the biggest density, which is defined as p_core .

[Idea of the algorithm] the p_core is as a center point, extended to its KNN, then its KNN is added to the candidate list. If the density value of some q of the candidate list is bigger than the product of the density of the center point and density threshold, namely,

$$density(p) \geq \alpha \times density(p_core) \quad (6)$$

The point just is the core point, and then the classification is extended from the KNN of the point, which is added to the candidate list. Otherwise, the point is given a classification label and deleted from the seed list. Each classification extension ends when the density descends to the given ratio α of the center point. The process is circulated until clustering finishes.

[Formal Description of the Algorithm]

**** SearchCorePoint()----The Core-point p_core with the maximum density and without a classification label is returned

**** IsCorePoint()----Judge whether some point is a core point

**** AddtoSseeds()----Add to a seed list

**** DelfromSeeds()----Delete from a seed list

**** input: Dataset, $MinPts$, k , $ratio$, α

**** output: clustering Dataset

While $theCorePoint \neq 0$

$theCorePoint = SearchCorePoint()$;

if $theCorePoint \neq 0$

$theCorePoint.class = cluster_id$

$seeds = theCorePoint.neighbors(k)$

while $seeds.length > 0$

$cur_point = seeds(1)$

if $cur_point.density \geq \alpha * theCorePoint.density$

if $cur_point.class = 0$

$cur_point.class = cluster_id$

if $IsCorePoint(cur_point)$

$AddtoSeeds(cur_point.neighbors(k))$

end

end

end

end

end

$DelfromSeeds(cur_point)$

end

$cluster_id = cluster_id + 1$

end

where $Dataset$ is the data set to be clustered, $MinPts$ is the density threshold, k is the number of the neighbor, $ratio$ is the ratio threshold, and α is the density boundary threshold.

4. Examples of the Algorithm

The clustering results of DENCLUE are very sensitive to parameters, whose tiny changes may lead to a very different result (Zhang Li, Zhou Weida, Jiao Licheng, 2002, pp587-590; Zhang Wei, Liao Xiaofeng, Wu Zhongfu, 2002, pp114-116). In the paper, clustering is carried out by making three group of data sets given different parameters in order to investigate the sensitivity of the improved clustering algorithm based on density

distribution function. The test results are showed as Figure.1.

Seen from the test results, the algorithm of the paper effectively modifies the sensitivity of density-based clustering ways to parameters.

To compare the clustering effect of the algorithm in the paper and that of DENCLUE, the same data set is respectively clustered using the two algorithms. The clustering results are showed as Figure.2.

Seen from Figure.2, DENCLUE has better effect in finding the classification of these districts with even density, otherwise, the effect is not very ideal. But the algorithm in the paper evidently modifies the week point.

5. Conclusion

In the paper, an improved clustering algorithm based on density distribution function is set up using the ideas of local scale and boundary threshold. The main characteristics has: it has a solid mathematics foundation and generalizes many other clustering ways, such as partitioning method, hierarchical one, density-based one, grid-based one, model-based one; For the data set with a lot of “noise”, it has an excellent clustering performance; For the clustering of arbitrary-shape high-dimensional data sets, it gives a concise mathematics description; The point with maximum density is as the center-point, from which, the classification is extended to density boundary threshold. The purpose of the paper is aimed to improve the sensitivity of DENCLUE to parameters and enhances the clustering effect of the high-dimensional data sets with uneven density distribution. The tests show that these problems above are greatly modified.

References

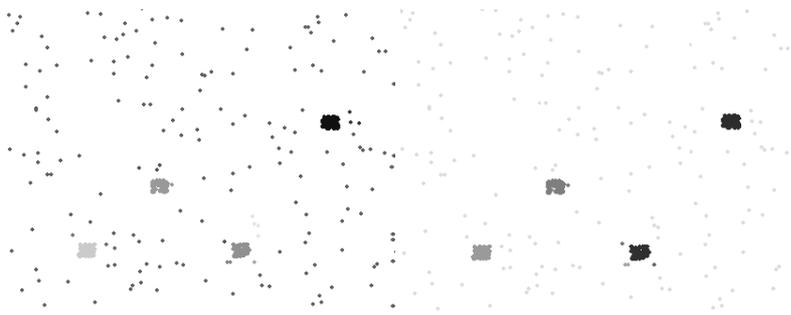
- Chen Ning, Chen An, Zhou Longxiang. (2002). An Incremental grid Density Based Clustering Algorithm. *Journal of Software*, 2002,13(1):1-7.
- Chen Yan, Geng Guohua, Zheng jianguo. (2005). An Improved Density-based Clustering Algorithm. *Micro-computer Development*, 2005, 3(15):12-16.
- Ergun Bicici, Deniz Yuret. (2007). Locally Scaled Density Based Clustering. *Proceedings of the 8th international conference on Adaptive and Natural Computing Algorithms, Part I. Apr*, 2007.
- Hong Shaorong, Xiao Wenjun. (2008). A New Way Enhancing the Performance of DBSCAN Clustering Algorithm. *Journal of Xidian University (Natural Science)*, 2008, 3(35):24-27.
- Li Fei,Xue Bin, Huang Yalou. (2002). K-Means Clustering Algorithms of Optimization Based on the Initial Center. *Computer Science*, 2002, 29(7):94-96.
- Ma Shuai. (2003). A quick Clustering Algorithm Based on Reference Points and Density. *Journal of Software*, 2003, 11(6):34-37.
- Rong qiusheng, Yan Junbiao, Guo Guoqiang. (2004). Research and Application of DNSCAN-based Clustering Algorithms. *Journal of Computer Applications*, 2004, 4(24):12-16.
- Zhang Li, Zhou Weida, Jiao Licheng. (2002). Nucleus Clustering Algorithm. *Computer Science*, 2002, 25(6):587-590.
- Zhang Wei, Liao Xiaofeng, Wu Zhongfu. (2002). A New Clustering Algorithm Based on Genetic Algorithms. *Computer Science*, 2002, 29(6):114-116.
- Zhou Shui-geng, Zhou Aoying, Jin Wen, Fan Ye, Qian Weining. (2000). FDBSCAN: A Fast DBSCAN Algorithm. *Journal of Software*, 2000, 11(6):735-744.
- Zhou Yonggeng, Zhou Aoying, Cao Jing. (2000). DBSCAN Algorithm Based Data Partition. *Journal of Computer Research and Development*, 2000, 37(10):1153-1159.



(a)

$MinPts = 4, k = 15, a = 0.9, ratio = 0.1$

$MinPts = 7, k = 20, a = 0.9, ratio = 0.1$



$MinPts = 4, k = 15, a = 0.9, ratio = 0.1$

$MinPts = 7, k = 20, a = 0.9, ratio = 0.1$

(b)

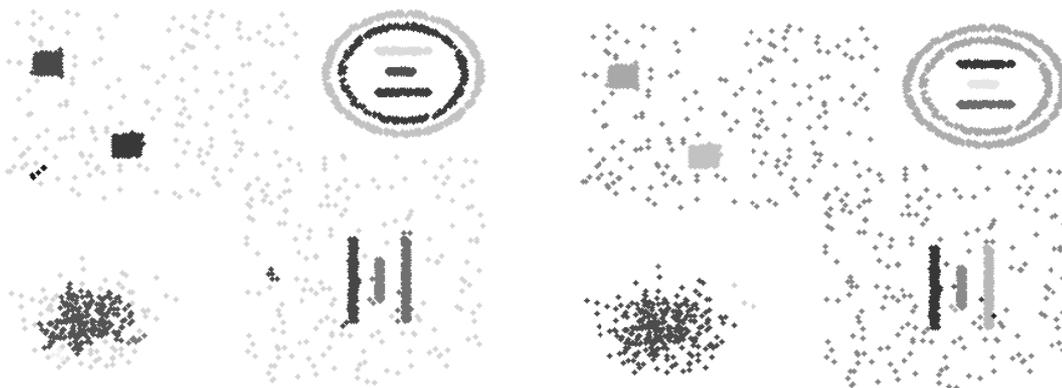


$MinPts = 4, k = 15, a = 0.9, ratio = 0.1$

$MinPts = 7, k = 20, a = 0.8, ratio = 0.2$

(c)

Figure 1. Clustering Results of Improved Clustering Algorithm



$MinPts = 4, \epsilon = 0.015$

$MinPts = 4, k = 12, a = 0.9, ratio = 0.1$

(a) Clustering Result of DENCLUE (b) Clustering Result of the Improved Clustering Algorithm

Figure 2. Comparison of Clustering Results of DENCLUE and the Improved Clustering Algorithm