Improved Clustering Algorithm Based on Local Agglomerative Characteristics

Xi-xian Niu[1] and Yan-ping Cui[2]

[1] Faculty of Information Technology and Propagation, Hebei Youth Administrative Cadres College, Shijiazhuang, China
[2] College of Mechanical & Electronic Engineering , Hebei University of Science and Technology, Shijiazhuang, China
niuxixian@126.com, cuiypkd@163.com

Abstract. Similarity measurement is the bases of clustering analysis. Current clustering algorithms mostly based on density and distance measurement, but these concepts become increasingly difficult to fit more and more complex data set and analysis works. SNN similarity, however, show more flexible ability to deal with different density, shape and multi-dimensions data process problems .In this paper we review mostly popular SNN based clustering method, give the definition of Local Agglomerative Characteristics during the procedure of the clustering, proposed a new clustering algorithm, that is, Improved Clustering algorithm based on Local Agglomerative Characteristics. Apply this clustering algorithm on experimental data set, the result show that it can work well on different type's data objects, can find nature distribute clusters in target data set, can improve the quality of data clustering.

Keywords: Data mining, clustering, SNN density, SNN similarity, local agglomerative characteristics.

1 Introduction

Due to rapid technological development in such areas as computer, network and communication, the researchers face ever-increasing challenges in extracting relevant information from the enormous volumes of available data. The so-called data avalanche is created by the fact that there is no concise set of parameters that can fully describe a state of real-world complex systems studied nowadays by biologists, ecologists, sociologists, economists, etc [1]. Pattern recognition is a primary conceptual activity of the Human being. Even without our awareness, clustering on the information that is conveyed to us is constant [2]. Clustering research has long been a hot research field of Data Mining, is considered the most important unsupervised learning problems. Specifically, clustering techniques are almost indispensable as a tool for data mining [2]. Clustering delineates operation for objects within a dataset having similar qualities into homogeneous groups, it allows for the discovery of similarities and differences among patterns in order to derive useful conclusions about them [3]. Purpose of the Clustering study is to determine the data groups in unlabeled data collection based on its inherent nature.

H. Deng et al. (Eds.): AICI 2011, CCIS 237, pp. 199–206, 2011.

Cluster analysis is a challenging task and there are a number of well-known issues associated with it, e.g., finding clusters in data where there are clusters of different shapes, sizes, and density or where the data has lots of noise and outliers [1]. Determining the structure or patterns within data set is a significant works. Clustering depends critically on density and distance (similarity), but these concepts become increasingly more difficult to define as destination of clustering jobs become more complexity. As a consequence, find new measurement method is a very necessary task for clustering researchers. Hence, SNN(Shared Nearest Neighbors) base similarity measure is proposed, and several more efficient clustering algorithm are designed and implemented, such as, Jarvis-Patrick (SNN similarity based), SNN density based DBSCAN (Density-Based Spatial Clustering of Application with Noise). But these methods still have space to improved, so we can take local cluster feature into account, define new measurement, get more efficient clustering method. Therefore, the purpose of this paper is to analysis existing algorithms' limitation, find its resolution, and design improved algorithm.

2 SNN Similarity Based Jarvis-Patrick Clustering Algorithm

In the Jarvis-Patrick (JP) scheme, a shared nearest neighbor graph is constructed from the proximity matrix as follows. A link is created between a pair of point p and q if and only if p and q have each other in their closest k nearest neighbor lists. This process is called k-nearest neighbor sparsification. The weights of the links between two points in the SNN graph can either be simply the number of near neighbors which shared by the two points, or one can use a weighted version that takes the ordering of the near neighbors into account.

The JP clustering algorithm replaces the proximity between two points with the SNN similarity, which is calculated as described in algorithm 2. A threshold is then used to sparsify this matrix of similarities. At this point, all edges with weights less than a user specified threshold are removed and all the connected components in the resulting graph are our final clusters.

Algorithm 2: JP clustering algorithm

Compute the SNN similarity graph.

Sparsify the SNN similarity graph by applying a similarity threshold.

Find the connected components (clusters) of the sparsified SNN similarity graph.

Because JP clustering is based on the notion of SNN similarity, it is good at dealing with noise and outliers and can handle clusters of different sizes, shapes, and densities. The algorithm works well for high-dimensional data and is particularly good at finding tight clusters of strongly related objects.

A major drawback of the Jarvis – Patrick scheme is that, the threshold needs to be set high enough since two distinct set of points can be merged into same cluster even if there is only one link across them. On the other hand, if the threshold is too high, then a natural cluster may be split into too many small clusters due to natural variations in the similarity within the cluster. Another potential limitation is that not all objects are clustered [4] [6].

3 SNN Density Based Clustering Analysis

The SNN density method can be combined with DBSCAN algorithm to create a new clustering algorithm. This algorithm is similar to the JP clustering algorithm in that it starts with the SNN similarity graph. However, instead of using a threshold to sparsify the SNN similarity graph and then take connected components as clusters. The SNN algorithm, as DBSCAN, is a density-based clustering algorithm. The main difference between this algorithm and DBSCAN is that it defines the similarity between points by looking at the number of nearest neighbors that two points share. Using this similarity measure in the SNN algorithm, the density is defined as the sum of the similarities of the nearest neighbors of a point. Points with high density become core points, while points with low density represent noise points. All remainder points that are strongly similar to a specific core points will represent a new clusters.

The SNN density clustering needs three inputs parameters: K, the neighbors' list size; Eps, the threshold density; MinPts, the threshold that define the core points. After defining the input parameters, the SNN algorithm first finds the K nearest neighbors of each point of the dataset; second apply DBSCAN with user-specified parameters for Eps and MinPts to determine what the data objects' point type should be, that is, core, border, or noise. Two different approaches can be used to implement the SNN algorithms. One that creates the clusters around the core points previously identified. In the other approach the clusters are identified by the points that are connected in a graph. The graph is constructed by linking all the points which similarity is higher than Eps.

The algorithm automatically determined the number of clusters in data space, but not all the points are clustered. SNN density-based clustering finds clusters in which the points strongly related to one another. The strengths and limitations of SNN density-based clustering are similar to those of JP clustering. However, the use of core points and SNN density adds considerable power and flexibility to this approach [6] [7].

4 Improved Clustering Algorithm Based on Local Agglomerative Characteristics

SNN similarity measurement and SNN density based measurement, are both based on the local data space features, and take local configuration characteristics into account, mainly focus on the algorithm's adaptability to the problems, such as the overall context of density, different shapes and sizes etc. Improved LAC (local agglomerative characteristics) clustering algorithm, however, mainly focus on local clusters' configuration and distribution characteristics which displaying in the procedure of the whole clustering analysis, and its application. Analyze the features of the shared neighbors around the data object, such as density, size, shape etc, and redefine data objects' similarity measure, and then improve the algorithm's adaptability and optimize efficiency. The shared nearest neighbors between data objects is a local relatively tight cluster contrast to other data objects that scattered in the local area, since study on its data deploy characteristics is an very meaningful work to determine the data objects' similarity and density.

4.1 Similarity Analysis Based on LAC

Due to JP algorithm take rigid parameter k as the threshold to adjust data objects similarity computation, then it is disadvantage to find local area relative small cluster which has more tightly feature, these situation shown in figure 1(a), data object A and B will be grouped into the same cluster, even though their local view is relatively loosely like, while the parameter k is set equal to 6. In addition, because SNN density based algorithm take strict Eps as input parameter, then, it can't deal with some conditions, it's shown in figure 1(b), here, the objects A and B in the local configuration is more loosely while view directly, but within the limits in the Eps, but data objects C, compared to the object A, it has better connectivity to object B, although this connection is not direct.

In order to better reflect the local distribution of data collection how to affect the data objects' similarity computation; we can assess the data objects' similarity by the following aspects, such as relative density, local distribute shape, and local distance etc. in this notion, we can get the data objects have higher similarity in local area if the shared nearest neighbors have a relatively high density, or if the assessed data objects have a relatively short distance according to the distribute shape of their shared neighbors.

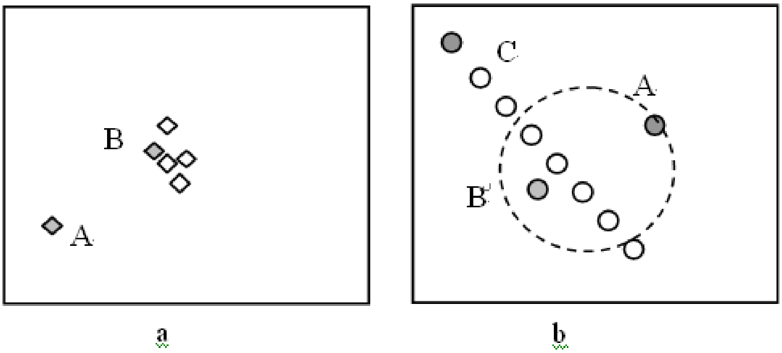

Fig. 1. Data objects' local configuration analysis

4.2 LAC's Definition and Its Measurement

Compared to other objects in target data collection, we can view the whole shared nearest neighbors as a cluster which has relative strong concentration properties in local data space. Therefore, the local clustering properties which get from the shared nearest neighbors can be seen as a measure basis to determine the assessed two objects whether or not have high similarity. In local data area, we can simple consider the two data objects have high similarity, if they have a relatively close distance. Because the distribution of data objects maybe exit different ways, in order to dynamic determine what the relative close distance in local data space is, it is need to analyze the local data's distribution characteristics, such as the shape, size, density of local cluster. Since the cluster size of shared neighbors is pre-determined by user input parameter, k (the number of SNN), which select according to the type of target data collection, then the local characteristics of the data gathering can be simplified to the representation of the local shape and density, where density can be measured by the average distance LAD (Local Average Distance) of all members of the shared neighbors. As the arbitrariness of local data distribution, the distribution shape measure can be reduced to two main aspects, i.e., local maximum distance LMD (Local Maximum Distance) and the local radial distance LRD (Local Radial Distance), the figure 2 give the illustrates of what LMD and LRD are. So we can get the definition of local data features as followings:

$$d_{LMD} = \max\{distance(p-p') \mid p \in C_{SNN}, p' \in C_{SNN}, p \neq p'\} \tag{1}$$

$$\mathrm{d}_{\mathrm{LRD}} = \max\left\{dis\tan ce(p, LineX) \middle| p \in C_{SNN}\right\} \tag{2}$$

$$d_{LAD} = \frac{2}{n(n-1)} \sum_{p \in C_{SNN}, p' \in C_{SNN}, p \neq p'} distance(p, p') \tag{3}$$

Where C_{SNN} is the shared nearest neighbors, n is the number of the data point in C_{SNN}, LineX is the line connected by two objects which have max distance in local cluster.

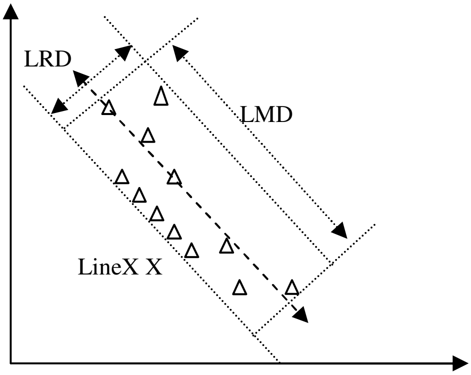

Fig. 2. Local characteristics measure

4.3 Improved LAC Based Clustering Algorithm

Among the previous analysis and definition, the local gathering features can be combined with JP algorithm to create a new clustering algorithm, i.e. improved LAC clustering algorithm. This algorithm is similar to JP and SNN density based clustering algorithms in mainly steps, however, it takes local data distribution into account, using LMD and LAD as a local dynamic threshold to control SNN similarity computation. The steps of improved LAC algorithm are shown in followings:

Identify the k nearest neighbors for each data object (the k data objects most similar to a given object, using a distance function to calculate the similarity).

Utilize LMD and LAD as dynamic control threshold to calculate the SNN similarity between pairs of data object and generate similarity graph.

Find the connected components (clusters) by applying a similarity threshold, and dynamic adjust the cluster membership at the same time.

Applying a similarity threshold to sparsify cluster graph can reduce compute complexity and improve algorithm efficiency in clusters finding. After graph sparsify, there are need a method to find and illustrate what kind cluster in graph. How to find connected components, here we give the pseudo code as followings:

```
While (exist data object not clustered) do
Begin
    Random select a not clustered data point as seed and marked it as a new kind
    cluster member. Put seed in pool.
    While (not empty of seed pool) do
    Begin
        Fetch first seed from pool;
        Search similarity matrix finds all data objects have high similarity with
        fetched seed, marked them the same cluster, put them into seed pool at the
        same time.
    End
End
```

4.4 Experimental Results and Evaluation

The cluster evaluation is a very important part of cluster analysis, because almost every cluster algorithm will find clusters in a data set, even if that data set has no natural cluster structure. Since there are a number of different types of clusters-in some senses, each clustering algorithm defines its own types cluster to fit the destination data set. So design different clustering method to analysis different data objects, there are must have suitable clustering validation. The advantage of distance-based clustering is that distance is easy for computing and understanding. There are several choices for similarity definition between two sets of data points, as follows: [8]

$$Similarity_{\mathrm{rep}}(C_i, C_j) = distance(rep_i, rep_j) \tag{4}$$

$$Similarity_{\mathrm{avg}}(C_i, C_j) = \frac{1}{n_i \times n_j} \sum_{v_i \in C_i, v_j \in C_j} distance(v_i, v_j) \tag{5}$$

.

$$Similarity_{\max}(C_i, C_j) = \max\{distance(v_i, v_j) \mid v_i \in C_i, v_j \in C_j\} \tag{6}$$

.

$$Similarity_{\min}(C_i, C_j) = \min\{distance(v_i, v_j) \mid v_i \in C_i, v_j \in C_j\} \tag{7}$$

To test the effectiveness and accuracy of the proposed clustering method, the splice and synthetic dataset is adapted. In improved LAC algorithm evaluation experiment, total 150 data object being processed. Table 1 has shown the part data object.

Table 1. Part experiment data objects

No.	Standardize X	Standardize Y
1	0.7800872	0.180566049
2	0.61186137	0.43378935
3	0.65880533	0.19837862
4	0.33866129	0.23050593
5	0.89341246	0.92718452
6	0.22261332	0.69293582
7	0.52261332	0.02761889
…	…	…

With the same prerequisite, we run JP algorithm, SNN density method and improved LAC scheme on the experimental data collection, and the experimental result shown in table 2. Through compare with the difference algorithm's running result on statistical data, find the new presented method can get better aggregated clusters to deal with nature distributed data set and objects.

Table 2. The experiment result

Algorithms	Clusters Found	Average Similarity
SNN density	7	0.29
Jarvis-Patrick	6	0.32
LAC-SNN	4	0.35

4.5 Its Strength and Limitation

In some extend the JP method looks like brittle, due to its single link notion; the SNN density based algorithm maybe view small cluster as noise and discarded it, because it depend on ridged input parameters. In our view, both these problems rise from the overlook of clustering's local gathering features. To solve these problems there are need define an improved similarity measure based on the nearest neighbors of the objects. This similarity measure is not only limited to consider the direct nearest neighbors, but also can take into the neighbors' local configuration features. Furthermore, the suggested similarity measure fuzzifies the crisp decision criterion of the Jarvis-Patrick algorithm; and trade off against SNN density based method. Because improved LAC algorithm utilize a SNN similarity-based definition of a cluster with local dynamic threshold control, it is relatively resistant to noise and can handle clusters of arbitrary shapes and sizes, can deal with clusters of different density and natural distribution characteristics. Thus improved LAC not only can find many clusters that could not be found using K-means and DBSCAN, but also can get better analysis result than Jarvis-Patrick and SNN density algorithm, even if the clusters have widely varying or gradually changed densities. Even though widely fit able to different type data set, improved LAC scheme still exist some limitations, some time it can not get best representation of the local agglomerative characteristics in some type data space, so the representative method of local characteristics still need to be improved for some types destination data set.

5 Conclusions

In this paper we review the most widely used and successful SNN based clustering techniques and their related applications, summarize their shortcoming and limitation, , give more detail and reasonable LAC's definition and introduced a new improved clustering algorithms based on LAC. The new algorithm overcome many of the challenges traditionally clustering algorithms, e.g., finding clusters in the presence of noise and outliers and finding clusters in data that has clusters of different shapes, sizes, and density, and finding more natural distribution clusters. The improved algorithm can give strong representative of local agglomerative characteristics of the target data set; can show flexible ability to fit sharply or gradually changed data density environments; can find relative small cluster but group tightly; can realize complete clustering and act outliers detection. De fact, the LAC clustering algorithm provides an effective tool for exploring groups of data.

References

1. Busygin, S., Prokopyev, O., Pardalos, P.M.: Biclustering in data mining. Computers & Operations Research 35, 2964–2987 (2008), http://www.elsevier.com/locate/cor
2. Almeida, J.A.S., Barbosa, L.M.S., Pais, A.A.C.C., Formosinho, S.J.: Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. Chemometrics and Intelligent Laboratory Systems 87, 208–217 (2007)
3. Oyana, T.J.: A New-Fangled FES-k -Means Clustering Algorithm for Disease Discovery and Visual Analytics. EURASIP Journal on Bioinformatics and Systems Biology (2010)
4. Ertoz, L., Steinbach, M., Kumar, V.: A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In: Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining (2002), http://www.bibsonomy.org/bibtex/2ba0e3067111d2e3eea9a4d9fc995e36b/hotho (2011)
5. Hu, T., Xiong, J., Zheng, G.: Similarity-based Combination of Multiple Clusterings. International Journal of Computational Intelligence and Applications 5(3), 351–369 (2005)
6. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education, 427–488 (2006)
7. Moreira, A., Santos, M.Y., Carneiro, S.: Density-based clustering algorithms–DBSCAN and SNN (2011), http://ubicomp.algoritmi.uminho.pt/local/download/SNN&DBSCAN.pdf
8. Qian, W.-N., Zhou, A.-Y.: Analyzing Popular Clustering Algorithms from Different Viewpoints. Journal of Software 13(8), 1382–1394 (2002)