An improved compatible clustering algorithm

Renxia WAN[1, 2,a *] , Duoqian MIAO[1] and Caixia LI[3]

[1]School of Electronics and Information Engineering, Tongji University,Shanghai,201804,China

[2] Beifang University of Nationalities, Yinchuan, Ningxia,750021,Chinan

[3] Informatization Office, Donghua University, Shanghai 201620, China

[a]wrx1022@mail.dhu.edu.cn

Keywords: Clustering; Compatible data; Spherical space

Abstract. In this paper, we discuss the problem of compatible clustering, we propose a new compatible clustering algorithm based on pNCompClu[9]. The new algorithm adopts spherical space approximation technique to replace the point neighborhood mechanism of pNCompClu. Experiments show that the proposed algorithm can get some consistent clustering results, and theory analysis also demonstrates that the proposed algorithm has higher clustering precision than pNCompClu has

Introduction

Cluster analysis is a very important tool in data analysis and data processing field, it has a wide range of applications in machine vision, statistics, machine learning and data mining. The aim of cluster analysis is to partition a data set into subsets (clusters) such that members of the same cluster are similar and members of distinct clusters are dissimilar, where the similarity of two data members is usually defined by a distance function[1]. While from the view of objects' relations, dissimilarity is just one kind measure for the objects relation analysis. Recently, a growing number of scholars are shown their interest on objects which have complex relationship. Liu J. et al. presented a cluster algorithm so called as Poclustering for ontology grouping[2,3]. This algorithm generates clusters from a pairwise dissimilarity matrix based on partially ordered relations among objects, and can preserve more available information than traditional clustering methods do during the whole clustering process. Socolovsky E.A. introduced a projecting method[4], it defines a complementary measure to form a similarity-dissimilarity measure pair by orthogonal components. Ng M.K. et al. described an extension to the k-modes algorithm for clustering categorical data[5], by modifying a simple matching dissimilarity measure for categorical objects, a heuristic approach was developed, which allows the use of the k-modes paradigm to obtain a cluster with strong intrasimilarity and to efficiently cluster large categorical data sets. To reduce the passive influence of noise, Hitchcock D.B.& Chen Z. proposed a method for cluster analysis of binary data based on "smoothed" dissimilarities[6]. The smoothing method presented borrows ideas from shrinkage estimation of cell probabilities. Rodriguez A. & Laio A. introduced an approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities[7]. Elhamifar E. & Vidal R. presented an algorithm to cluster high-dimensional data points that lie close to low-dimensional structures corresponding to several classes or categories to which the data belong[8]. A method to the automatic classification of objects based on an extension of the topological measure is presented[9].

Wan et al proposed an algorithm called as CompClustering[10], which is based on the compatible relation theory. As an extended work of CompClustering, CNclustering algorithm is proposed [1].

pNCompClu is proposed to cluster compatible objects with point neighborhood theory[11], it present every compatible subset with one special point "P" and its neighborhood. pNCompClu is an incremental clustering method, it scans the object set only once. The computational cost of pNCompClu is $O(kn)$, where n is the number of objects, k is the number of clusters.

But pNCompClu has a weakness, instead, the point P's neighborhood which present a compatible cluster may be too large in some cases.

In this article, we perfect the compatible clustering method via spherical space approximation. The rest of this paper is organized as follows: In section 2, we introduce some concepts of compatible relation and review pNCompClu. In section 3, we present a new method called as SSAclu. The experimental results and evaluation are reported in section 4. Finally, we draw our conclusions in section 5.

Compatible relation

Firstly, we give a sketch of compatible cluster, further details about these can be found from papers produced by Wan R. et al. (2009a, b).

Definition.1 Supposed D is the relation measure on an object set S, $\delta(\geq 0)$ is a threshold, if it satisfies:

D is reflexive (if and only if, for all objects $x \in S, D(x,x) \leq \delta$)

D is symmetric(if and only if, for all objects $x, y \in S$, if $D(x,y) \leq \delta$ then $D(y,x) \leq \delta$

Then we called D is a compatible relation, and S is a compatible set under D.

Definition.2 Let C be a subset of object set S, namely $C \subseteq S$, and supposed D is the relation measure on S, if C is a compatible set, then we call C is a compatible subset of S under D

Definiton.3 Let C be one compatible subset of S under relation measure D, if there does not exist another compatible subset C′ of S, such that C is one proper subset of C′, then we call C is a maximal compatible subset of S.

Compatible data clustering via spherical space approximation

NCompClu clusters compatible objects by point neighborhood[11], but the spherical space decided by a certain radius ρ is not well suitable to represent the cluster in some cases. For example, in figure 1, the real cluster is the circular space decided by points A, B, C and D, this cluster is only a subset of the circular space decided by ρ.

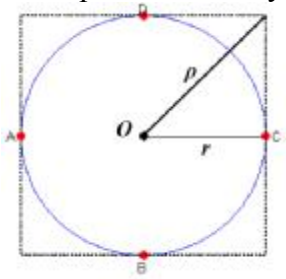

Figure 1. Exception 1 of pNCompClu

In the following, we try to improve the accuracy of clustering method with point neighborhood technique.

For convenience, we denote the maximal compatible subset under the threshold δ by $\delta-$MC. And if there is no special manifesto in the next discussions, D(X, Y) is the bigger of D(X, Y) and D(Y, X). In the following discussion, we suppose relation measure D to be Euclidean distance.

Proposition.1 Given a compatible cluster $\delta-$MC then $\delta \geq \max_{1\leq i\leq d}(x_i^{\max}-x_i^{\min})$, where $x=(x_1,x_2,...,x_d)$ is an arbitrary data point in $\delta-$MC .

Proof. Let x=($x_1,x_2,...,x_d$), y=($y_1,y_2,...,y_d$) be the two objects of "2d points" (Wan R. et al. 2009b) in the compatible cluster $\delta-$MC, and in the i0 dimensional space: $y_{i_0}-x_{i_0}=\max_{1\leq i\leq d}(x_i^{\max}-x_i^{\min})$.

$$\delta \geq D(y,x) = |yx| = \sqrt{\sum_{i=1}^{d}(y_i - x_i)^2}$$

$$= \sqrt{(y_1 - x_1)^2 + \ldots + (y_{i_0} - x_{i_0})^2 + \ldots + (y_d - x_d)^2}$$

$$\geq y_{i_0} - x_{i_0}$$

The proposition is thus demonstrated. ■

Since the threshold δ of a natural compatible cluster usually is unknown, and different clusters may have different cluster thresholds. Thus finding an appropriate cluster threshold is momentous to describe cluster.

In the process of clustering data points into compatible clusters, the "2d points" plays an important role, they determine not only the center of every cluster, but also the distributing area of their cluster. From proposition 1, we can see, in compatible cluster δ−MC, $r \leq \delta \leq \rho$, where $r = \max_{1 \leq i \leq d}(x_i^{\max} - x_i^{\min})$.

In this paper, we design an algorithm to describe compatible cluster via spherical space approximation. The overview of this algorithm is shown in figure 2.

Algorithm: SSAclu (Object set: S, Relation measure: D, Threshold of D: δ)

1: $SC \leftarrow \phi$; $Min\_Max \leftarrow \phi$; $C\_2d = \phi$; $SC\_2d = \phi$
2: for i 1:|S|
3: read data $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)})$ from S;
4: if i=1 then
5: $C^{(i)} = \{x^{(i)}\}$; $SC \leftarrow C^{(i)}$; $C\_2d^{(i)} = \{x^{(i)}\}$; $SC\_2d \leftarrow C\_2d^{(i)}$;
6: $d^{(i)} = \{[x_1^{\min(i)} - x_1^{(i)}, x_1^{\max(i)} - x_1^{(i)}], [x_2^{\min(i)} - x_2^{(i)}, x_2^{\max(i)} - x_2^{(i)}], \ldots, [x_d^{\min(i)} - x_d^{(i)}, x_d^{\max(i)} - x_d^{(i)}]\}$;
7: $Min\_Max \leftarrow d^{(i)}$;
8: end if
9: if $\exists k: 1 \leq k \leq |SC|$
10: if $\forall j (1 < j < d): x_j^{\min(k)} < x_j^{(i)} < x_j^{\max(k)}$
11: $C^{(k)} \leftarrow x^{(i)}$;
12: else
13: temp_C_2d ← Updating_center($x^{(i)}$, $C\_2d^{(k)}$);
14: calculate ρ in temp_C_2d by formula (2);
15: if $\rho \leq \delta/2$
16: $C^{(k)} \leftarrow x^{(i)}$;
17: else
18: $C^{(|SC|+1)} = \{x^{(i)}\}$;
19: $SC \leftarrow C^{(|SC|+1)}$;
20: end if
21: end if
22: else
23: $C^{(|SC|+1)} = \{x^{(i)}\}$;
24: $SC \leftarrow C^{(|SC|+1)}$;
25: end if
26: end for

Fig. 2. Algorithm: SSAclu

Algorithm: Updating_center($x^{(i)}$, $C\_2d^{(k)}$)

1: temp $C\_2d = C\_2d^{(k)}$
2: for j=1: d
3: if $x_j^{(i)} < x_j^{\min(k)}$
4: $x_j^{\min(k)} = x_j^{(i)}$;
5: else
6: $x_j^{\max(k)} = x_j^{(i)}$
7: end if
8: end for
9: return temp_C_2d;

Fig. 3. Algorithm: Updating_center

In the figure 2, SC is the set of compatible clusters, Min_Max is the set of pair vector, in which every pair vector $d^{(k)}$ is corresponding to compatible cluster $C^{(k)}$, and every pair array $[x_m^{\min}, x_m^{\max}]$ represents respectively the minimal value and maximal value of $C^{(k)}$ on the m-th dimension space. $C\_2d^{(k)}$ is spherical space of $C^{(k)}$, and SC_$2d$ is the set of all spherical spaces.

SSAclu uses approximately cubic spaces to store cluster information. The cubic space of corresponding cluster is fixed by those "2d points" of the cluster, instead of calculating the center and using neighborhood of this center to represent the cluster. Spherical space is updated on the line 9 of

Updating, in our method, we use linear discriminant function to produce and update the spherical space.

Empirical results and evaluation

In order to get more knowledge of the quality of our proposed algorithms, we used a synthetic data set SD (shown as figure 4) to test the clustering quality of the new algorithms. In our experiments, we chose the threshold δ as 0.12, 0.18, 0.23, and the clustering results are shown as follows:

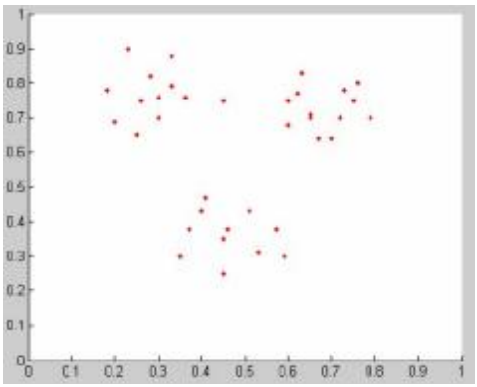

Fig.4. Data set SD

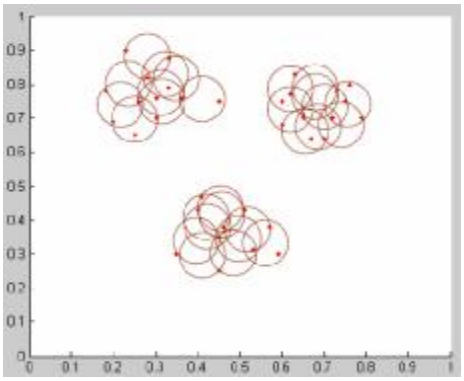

Fig.5. δ=0.12

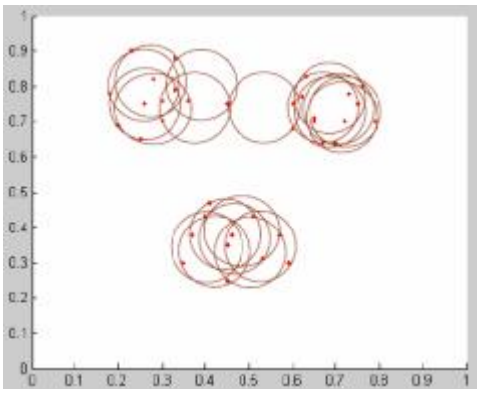

Fig.6. δ=0.18

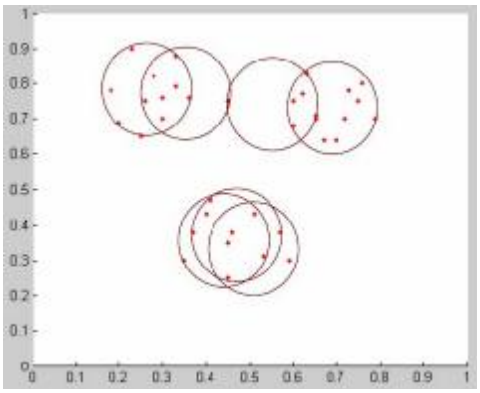

Fig.7. δ=0.23

We can see that our new algorithm can group objects into several clusters according to the given threshold, and has the compatible clustering effect as CompClusteing and pNCompClu do, namely, SSAclu allows overlaps between clusters.

The computational cost of CompClusteing is $O(kn^2)$, where k is the number of clusters. pNComClu is incremental clustering method, it scan the data set only once, pNCompClu has $O(kn)$ computational cost. SSAclu is also an incremental clustering method, it measure the distance between every object and "2d points", it will take a $O(2dn)$ computational consumption on every candidate cluster, thus, the total computational cost of SSAclu is $O(2kdn)$.

The computational cost of CompClusteing is $O(kn^2)$, where k is the number of clusters. pNComClu is incremental clustering method, it scan the data set only once, pNCompClu has $O(kn)$ computational cost. SSAclu is also an incremental clustering method, it measure the distance between

every object and "2d points", it will take a $O(2dn)$ computational consumption on every candidate cluster, thus, the total computational cost of SSAclu is. $O(2kdn)$.

Conclusions

In this paper, we explore clustering with compatible relation of objects, and propose a new algorithm called as SSAclu. SSAclu is an improved algorithm of pNCompClu. Theory analysis shows that the new algorithm has a linear cost of scale of the processing data set, and has a higher precision than pNCompClu has.

Future work will aim at more complex relationship between objects, and considering different clustering techniques to be embedded into our approach.

Acknowledgements

This research is financially supported by the National Natural Science Foundation of China (61163017, 61273304, 61102008, 61440044), the Research Fund for the Doctoral Program of Higher Education of China(20130072130004)

References

[1] R.Wan, L.Wang, Wang M., X. Su, CNclustering: Clustering with compatible nucleoids. The IEEE 2009 4th International Conference of Computer Science & Education.(2009)797-800.

[2] J.Liu, Q.Zhang, W.Wang, L.McMillan, J.Prins, Clustering pari-wise dissimilarity data into partially ordered sets. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.(2006)637-642.

[3]J.Liu, Q.Zhang, W.Wang, L.McMillan, J.Prins, PoClustering: lossless clustering of dissimilarity data. In Proceedings of the Seventh SIAM International Conference on Data Mining(2007).

[4] E.A.Socolovsky, A dissimilarity measure for clustering high and infinite dimensional data that satisfies the triangle inequality. NASA LaRC Technical Library Digital Repository.(2002)1-12.

[5]M.K.Ng, M.J.Li, J.Z.Huang, Z.He, On the impact of dissimilarity measure in k-Modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence.29(2007) 503-507.

[6] D.B.Hitchcock, Z.Chen, Smoothing dissimilarities to cluster binary data. Computational Statistics and Data Analysi. (2008).

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

[8] E.Elhamifar, R.Vidal, Spares subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (2013) 2765-2781.

[9] P.Valtchev, J.Euzena, Dissimilarity measure for collections of objects and values. Advances in Intelligent Data Analysis Reasoning about Data.(2006)259-272..

[10] R.Wan, L.Wang, Z.Liu, X.Su , Clustering on compatible relation. Application Research of Computers (in chinese).26(2009) 1301-1305.

[11] R.Wan, L.Wang, Z.Hao, Clustering compatible objects by point neighborhood. International Conference on Artificial Intelligence and Education.(2010)171-174.