

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	The application of cluster analysis in geophysical data interpretation
Author(s)	Song, Yu-Chen; Meng, Hai-Dong; O'Grady, Michael J.; O'Hare, G. M. P. (Greg M. P.)
Publication Date	2010-03
Publication information	Computational Geosciences, 14 (2): 263-271
Publisher	Springer
This item's record/more information	http://hdl.handle.net/10197/1914

Downloaded 2012-05-16T20:41:35Z

Some rights reserved. For more information, please see the item record link above.



ORIGINAL PAPER

The Application of Cluster Analysis in Geophysical Data Interpretation

Yu-Chen Song^{1,*}, Hai-Dong Meng¹, M.J. O'Grady², G.M.P. O'Hare²

¹*Inner Mongolia University of Science and Technology, Baotou, China.*

²*School of Computer Science & Informatics, University College Dublin (UCD),
Belfield, Dublin 4, Ireland.*

Abstract: A clustering algorithm which is based on density and adaptive density-reachable is developed and presented for arbitrary data point distributions in some real world applications, especially in geophysical data interpretation. Through comparisons of the new algorithm and other algorithms, it is shown that the new algorithm can reduce the dependency of domain knowledge and the sensitivity of abnormal data points, that it can improve the effectiveness of clustering results in which data are distributed in different shapes and different density, and that it can get a better clustering efficiency. The application of the new clustering algorithm demonstrates that data mining techniques can be used in geophysical data interpretation and can get meaningful and useful results, and that the new clustering algorithm can be used in other real world applications.

Keywords: cluster analysis; geophysical data interpretation; data mining

*Corresponding author: Yu-Chen Song

Tex: +86 13947268432.

E-mail address: songyuchen@imust.edu.cn (Yu-Chen Song), bjsongyc@hotmail.com .

Address: Inner Mongolia University of Science and Technology,
7#, Aerding Street, Kunqu District, Baotou,
Inner Mongolia , China.

Post code: 014010

The Application of Cluster Analysis in Geophysical Data Interpretation

Abstract: A clustering algorithm which is based on density and adaptive density-reachable is developed and presented for arbitrary data point distributions in some real world applications, especially in geophysical data interpretation. Through comparisons of the new algorithm and other algorithms, it is shown that the new algorithm can reduce the dependency of domain knowledge and the sensitivity of abnormal data points, that it can improve the effectiveness of clustering results in which data are distributed in different shapes and different density, and that it can get a better clustering efficiency. The application of the new clustering algorithm demonstrates that data mining techniques can be used in geophysical data interpretation and can get meaningful and useful results, and that the new clustering algorithm can be used in other real world applications.

Keywords: cluster analysis; geophysical data interpretation; data mining

1. Introduction

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large amounts of data. It has also opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways. Clustering is one of data mining techniques. Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters. Cluster analysis has played an important role in a wide variety of fields. Many different clustering algorithms have been developed for meeting different applications, such as K-means algorithm.

In this paper we develop a cluster algorithm – CADD (Clustering Algorithm based on Density and adaptive Density-reachable) and try to use it to cluster a set

of geophysical data from Ningxia Autonomous Region in China. The two different algorithms, K-means algorithm and the CADD algorithm, are applied in the real word application (geophysical data interpretation) so that it will be shown that if the new algorithm is good for geophysical prospecting application. By comparing the two pictures, the original geomorphological map in Ningxia and the clustering result using the CADD algorithm, it will be shown that whether the new algorithm is effective in this particular area application.

The rest of the paper is organized as follows: In the section 2, Background and Related Research. In the section 3, the development of the new algorithm will be introduced after some main problems are analyzed. In the section 4 is the design and implementation of the algorithm which includes the concepts and description of the clustering algorithm. In the section 5, performances of the new algorithm will be tested so that the new algorithm would be suited for some real word applications, such as arbitrary data point distributions and time and space complexity. In the section 6, the new algorithm is applied to a real world application, geophysical prospecting. The conclusion and future works will be presented in section 7.

2. Background and Related Research

Cluster analysis is an active research area in data mining technology. In general, the major clustering algorithms can be divided into followings: the prototype-based clustering method, the grid-based clustering method, the partitioning method, the density-based clustering method, etc.

In the prototype-based method [34], a cluster is a set of objects in which an object is closer or similar more to the prototype that defines the cluster than to the prototype of any other cluster.

The grid-based clustering methods [16, 18, 25] first quantize the clustering space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. Cells that contain more than a certain number of points are treated as dense and the dense cells are connected to form the clusters.

Partitioning methods [34, 14, 21, 6] are divided into two major subcategories. One of the partitioning methods is the centroid algorithm [35, 14] which represents each cluster by using the gravity centre of the instances. The most well-known centroid algorithm is the K-means algorithm [19, 1, 13, 34] which partitions the data set into k subsets such that all points in a given subset are closest to the same centre. K-means algorithm randomly selects k of the instances to represent the clusters and if k cannot be known ahead of time, varies values of k can be evaluated until the most suitable one is found. K-means algorithm is efficient in processing large data sets [12] and it handles spherical shapes well [34]. But K-means has the weaknesses [23, 24] that it often terminates at a local optimum, it is sensitive to noise, it cannot handle non-globular clusters and it cannot detect outliers.

Density-based methods [4, 26], such as the DBSCAN algorithm [4, 26, 32, 2, 5, 36] and the DENCLUE algorithm [9, 28], cluster objects based on the notion of density. The DBSCAN algorithm locates regions of high density that are separated from one another by regions of low density. The DBSCAN algorithm is a typical and effective density-based clustering algorithm which can find different types of clusters, can identify outliers and noise, but it cannot handle the clusters of varying density. The DENCLUE algorithm has a solid theoretical foundation. It models the overall density of a set of points as the sum of influence functions associated with each point. The DENCLUE algorithm is good at handling noise

and outliers and it can find clusters of different shapes and size, but it has trouble with high-dimensional data and data that contains clusters of widely different densities, and it can be more computationally expensive than other density-based clustering techniques.

Different clustering algorithms are selected for different application areas [26, 29, 30, 20, 22, 7, 33, 10, 31, 11, 17, 40, 3]. A variety of factors need to be considered when deciding which type of clustering techniques to use. Our goal is that clustering algorithm can be appropriate for a particular clustering task. Generally, the task of choosing the proper clustering algorithm involves considering these issues such as type of clustering, type of cluster, number of data objects and number of attributes, and domain-specific issues as well.

3. The development of the new algorithm

According to the analysis above, some main problems in existing algorithms are as following:

3.1 Selection of cluster shapes

Generally, shapes of original clusters are divided into several different types [34, 15]: ①Well-separated clusters, each point is closer to all of the points in its cluster than to any point in another cluster. ②Centre-based clusters, each point is closer to the centre of its cluster than to the centre of any other cluster. ③Contiguity-based clusters, each point is closer to at least one point in its cluster than to any point in another cluster. ④Density-based clusters, clusters are regions of high density separated by regions of low density. Because the structures of data sets are complicated in some real world applications, and distributions of data

points are different and cluster shapes cannot be predicted, clustering algorithms are needed to handle different shapes of original clusters.

3.2 Dependency on domain knowledge

For some algorithms it is necessary to input the parameters of the number of clusters and the initial centroids of clusters. This is difficult for unsupervised data mining when there is lack of relevant domain knowledge [29, 30]. At the same time, different random initializations of numbers and centroids of clusters produce different clustering results, which have an effect on the stability of clustering methods.

3.3 Sensitivity to noise or outliers

There are large amounts of noise or outliers in some real word applications. Some algorithms are sensitive to noise or outliers, such as partitioning methods, prototype-based clustering methods, and grid-based methods. For example, if there are some maximum value existed, the data point distribution may be highly distorted. So a better clustering algorithm is needed to be less sensitive to noise or outliers, also it can handle noise or outliers effectively.

From the above analyses, we develop CADD algorithm, especially for some real world applications, such as geophysics or geochemistry. The aim is that it can promote the ability of finding clusters of arbitrary distribution and handling noise or outliers in some real world applications, that it has better performances and scalabilities in high dimensional data sets, and that it can automatically get some parameters and make such parameters less dependent on domain knowledge.

4. Design and implementation of the algorithm

4.1 The concepts of the clustering algorithm

Based on the notions of density and adaptive density-reachable, the CADD algorithm which is designed and implemented in this paper has ability to find clusters of arbitrary shapes and sizes, to handle clusters of varying densities, and to identify noise or outliers effectively. The concepts of the CADD algorithm are as follows:

①**The density of data points:** The density of data points models the overall density of a set of points in dataset D as the sum of influence function associated with each point:

$$Density(x_i) = \sum_{j=1}^n e^{-\frac{d(x_i, x_j)^2}{2\sigma^2}} \quad (1)$$

Where, Gaussian influence function $f_{Guass}(x_i, x_j) = e^{-\frac{d(x_i, x_j)^2}{2\sigma^2}}$ indicates the density influence of each data point to the density of point x_i , and σ is density adjustment parameter which is analogous to the standard deviation, and governs how quickly the influence of a point drops off.

②**Local density attractors:** Local density attractors are the data points at which the values of density function $Density(x)$ are locally maximum.

③**Density-reachable distance:** Density-reachable distance is used to determine a circular area of data point x , labeled as $\delta = \{x | 0 < d(x_i, x_j) \leq R\}$, the data points in which are belong a same cluster. The definition formula is:

$$R = \frac{mean(D)}{n^{coefR}} \quad (2)$$

Where, $mean(D)$ is the mean distance between all data points in dataset D , and $coefR$ ($0 < coefR < 1$) is named as the original adjustment coefficient of density-reachable distance.

④**Density-reachable:** Density-reachable means that if there is an object chain $p_1, p_2, \dots, p_n = q$, q is a local density attractor, and p_{n-1} is density-reachable from q , then for $p_i \in D, (1 \leq i < n-1)$ and $d(p_i, p_{i+1}) \leq R$, we define that object $p_i, (1 \leq i < n-1)$ is density-reachable from q .

⑤**Adaptive density-reachable distance:** In handling the clusters of varying densities, it is important to adjust the density-reachable distance R step by step during clustering. The adaptive adjustment is carried out through multiplying the original density-reachable distance R with an adjustment coefficient α :

$$R_{Adap} = \alpha R \quad (3)$$

Where, R_{Adap} is adaptive density-reachable distance, and α is defined as:

$$\alpha = \frac{Density(Attractor_{i-1})}{Density(Attractor_i)} \quad (4)$$

This is because that when the density value of local density attractor of a cluster is greater, the distance between objects in the cluster is smaller, and on the contrary, when the density value of local density attractor of a cluster is smaller, the distance between objects in the cluster is larger. When $i=1$, let $Density(Attractor_0) = Density(Attractor_1)$, and so $\alpha \geq 1$. It is necessary to note that adaptive adjustment coefficient α may be also other function.

4.2 Description of the clustering algorithm

Local density attractors are used to determine centroids of clusters, and adaptive density-reachable distance to locate data objects that belong to the cluster. The descriptions of CADD algorithm are as follows:

Algorithm Clustering Algorithm based on Density and adaptive Density-reachable

Input: Adjustment coefficient of density-reachable distance $CoefR$, Density adjustment parameter σ .

Output: Number of clusters, the members of each cluster, outliers or noise points.

- 1: Compute the densities of each data point.
 - 2: $i \leftarrow 1$
 - 3: **repeat**
 - 4: Seek the maximum density attractor $O_{DensityMaxi}$ in the original data set of clustering objects as the cluster centroid of C_i .
 - 5: Assign the data objects which are density reachable within adaptive density-reachable distance from $O_{DensityMaxi}$ to cluster C_i , and at the same time delete the clustered objects from original data set.
 - 6: $i \leftarrow i+1$
 - 7: **until** The original data set is empty.
 - 8: Assign the clusters which have fewer objects (such as less 5 or 10) into outlier or noise group.
-

5. The performances of the algorithm

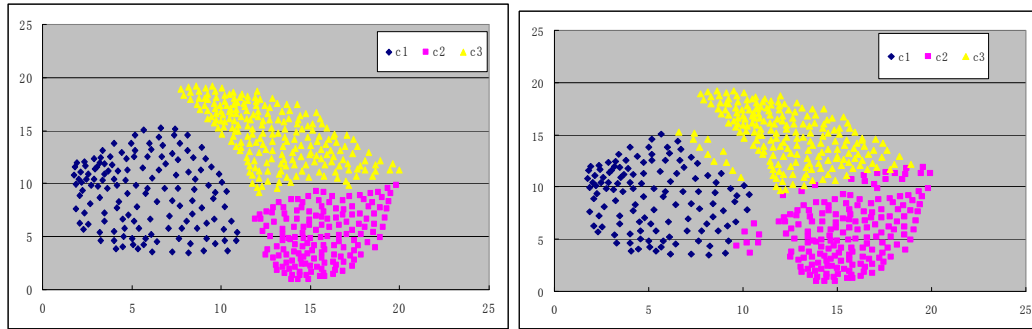
In this section, the performances of the CADD algorithm are tested by comparing the CADD algorithm and other algorithms. The testing of the performances include comparing of the clustering results between the clusters of

different data point distributions, the comparing time and space complexities using different algorithms.

5.1 Arbitrary data point distributions

According to data point distributions in some certain real world applications, there are different shapes of original clusters which are well-separated clusters, centre-based clusters, contiguity-based clusters and density-based clusters, etc. The following experiments will show that the CADD algorithm can handle arbitrary data point distributions, such as non-globular clusters of different shapes, different sizes and different densities, in some real world applications.

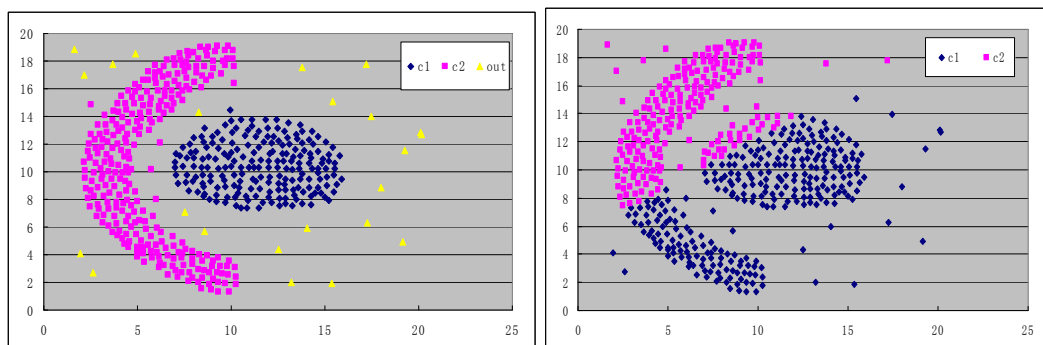
Figure 1 (a) and Figure 1 (b) show the clustering results in unequal density data point distributions using the CADD algorithm and K-means algorithm, respectively. In figure 1(a), the density of Cluster 1 is different from that of Cluster 2 and Cluster 3, and the density of Cluster 2 is different from that of Cluster 3. The shapes of the clustering results accord with the shapes of natural data point distributions. When data set is the equal density of data point distribution, K-means algorithm handles globular clusters well. But when data set is the unequal density data point distribution, K-means algorithm can not get good result even the shapes are globular. As figure 1(b) shows: K-means algorithm assigns some data objects of Cluster 1 to Cluster 2 and to Cluster 3 wrongly, and some data objects of Cluster 2 to Cluster 3 wrongly. It makes the shapes of clustering results differ from the shapes of natural data point distribution. Through this experiment we can clearly get that the CADD algorithm can handle clusters in unequal densities data point distribution, while K-means algorithm can not handle clusters in unequal densities data point distribution well.



(a) Clustering results by CADD algorithm (b) Clustering results by K-means algorithm

Figure 1 Comparing clustering results by different algorithms

Figure 2 shows the clustering results of non-globular clusters by using the CADD algorithm and using K-means algorithm, respectively. In Figure 2 (a) using the CADD algorithm, Cluster 1 is a globular cluster, Cluster 2 is a non-globular cluster, and the outliers scatter around the two clusters. The CADD algorithm handles non-globular clusters very well and can recognize outliers. In Figure 2 (b) using K-means algorithm, because one of the clusters is non-globular shape, so it makes the globular cluster divided into two parts and the non-globular cluster also into two parts. K-means cannot handle non-globular clusters and cannot recognize outliers. This experiment clearly shows that the CADD algorithm in handling non-globular clusters is better than K-means algorithm.



(a) Non-globular clusters by CADD algorithm (b) Non-globular clusters by K-means algorithm

Figure 2 Comparing Non-globular clusters by CADD and K-means

The two experiments show that the CADD algorithm has good property in handling both non-globular clusters and unequal density data distribution.

5.2 Time and space complexity

The two experiments of the time and space complexities are carried out so that the time and space complexity of the CADD algorithm can be compared with that of K-means algorithm. The curves of experimental results are shown in Figure 3.

Figure 3 shows the time and space complexities using the CADD algorithm and K-means algorithm. When the amount of data objects is less than 3000, the running time of the CADD algorithm is the same as that of K-means algorithm. As the amount of data objects increases greatly, more than 3000, the running time of the CADD algorithm is much less than that of K-means algorithm. It is mainly because that as the amount of data objects increases, the iteration of K-means algorithm increases, and the running time increases. Because the CADD algorithm only necessary to search the density-reachable objects in data set one time for each cluster, its running time is lower than that of K-means algorithm.

The time complexity of the CADD algorithm is $O(kn)$, while the time of complexity of K-means algorithm is $O(knt)$, where n is the number of data objects, k is the number of clusters, and t is the number of iterations required for convergence. The space complexity of the CADD algorithm is $O(n)$ and K-means algorithm is $O((k+n)m)$.

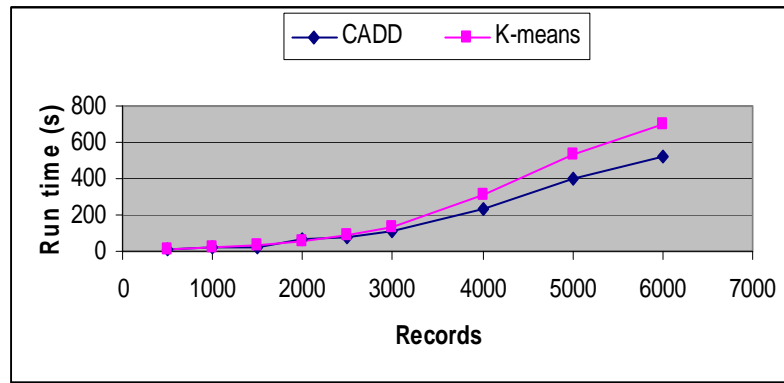


Figure 3 The time and space complexity of CADD and K-means

Through the previous experiment and comparing analysis, the performances of the CADD algorithm, which include the ability to handle the non-globular and unequal density of data distribution, the time and space efficiency, are better than that of K-means algorithm. Through the special design of the CADD algorithm, k value of the number of clusters and cluster centers are can be determined automatically. The CADD algorithm is insensitive to abnormal data and can be able to discover outliers. So we can conclude that the most performances of the CADD algorithm are good for some real world applications.

6. A real world application

The geophysical data are automatically measured and recorded by geophysical instruments. Generally, the amount of data is very large and relatively standard. It is suitable to be processed and be analyzed by data mining techniques. Using clustering algorithm of data mining, we can process some real word data which are measured in geophysical prospecting [8, 37, 27], such as Electrical Method, Gravity Exploration, and Magnetic Exploration, etc. We can also process the real word data which are obtained in different geophysical prospecting measurements

comprehensively. For example, we can store the geophysical prospecting data in spatial database, and every data point in the spatial database has some attributes associated with geophysical parameters measured in field. This is a very new approach of the real word application in geophysical data interpretation.

The apparent resistivity curves of electrical sounding are stored in spatial database, which were measured in the whole province region, Ningxia province of China (Figure 5(a)). We try to test and analyze the clustering results of the electrical sounding data by using the CADD algorithm. By comparing the clustering results obtained by using different clustering algorithms, we can know that ①whether the CADD algorithm is effective and reasonable in the real word application of geophysical prospecting; ②whether the clustering results can describe the characteristic of underground electrical distribution of the real word application objectively and accurately.

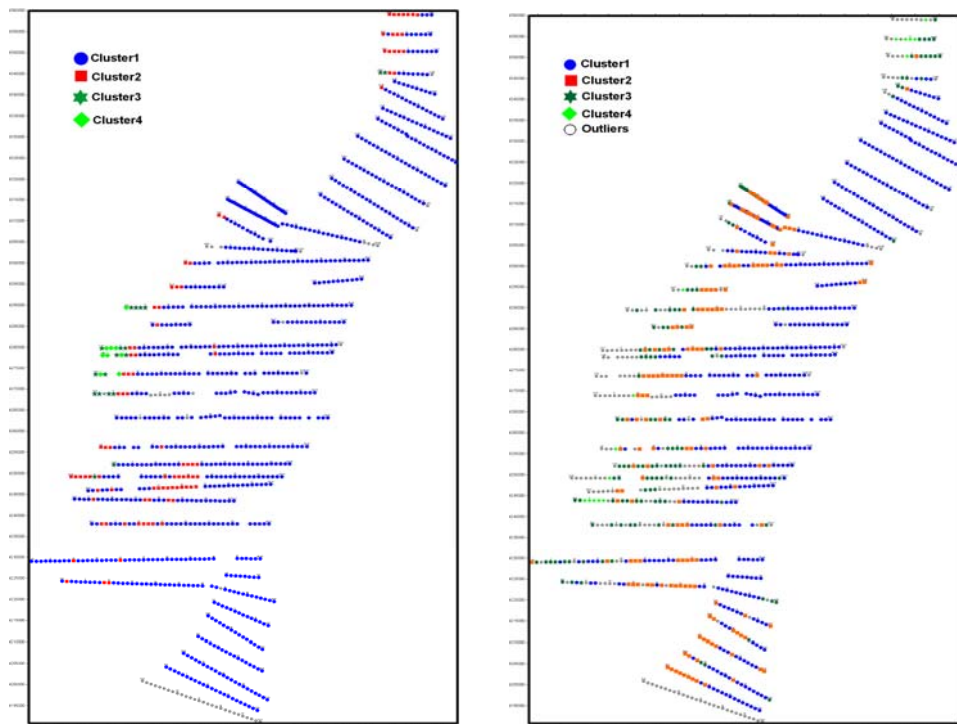
6.1 Data preparation

In the measuring region, a whole province region [39, 38], there are about 1100 electrical sounding points. At every sounding point, apparent resistivity is measured at 14 different $AB/2=\{3, 4.5, 7, 12, 20, 30, 45, 70, 120, 200, 300, 450, 700, 1000\text{m}\}$, and every sounding curve has 14 apparent resistivity values $\rho_s = \{\rho_s(3), \rho_s(4.5), \dots, \rho_s(1000)\}$. The sounding curves are stored in spatial database, and every electrical sounding curve is a data point (tuple) which has 14 attributes $\rho_s = \{\rho_s(3), \rho_s(4.5), \dots, \rho_s(1000)\}$ associated with 14 different $AB/2$.

The purpose of clustering analysis is to divide electrical sounding curves into different types at different part of the measuring regions, and the types of sounding curves reflect the geo-electrical features of the region.

6.2 The clustering results

Figure 4 (a) shows the clustering result distribution of the apparent resistivity curves of electrical sounding by using K-means algorithm. As can be seen, there are two main clusters in the clustering result. The blue area distributes widely and continuously, which is Cluster 1. The orange area distributes in small area and continuously partly, which is Cluster 2. There are two other clusters which include only less than 5 points, so they are shown in the same color as no-data sample points, and also there is no meaning in practice. It can be seen in Figure 4 (a) that the clustering result distributions by using K-means algorithm is simple and it cannot reflect the real distribution of the apparent resistivity curves of electrical method in the measured area.



(a) Clustering Result by K-means

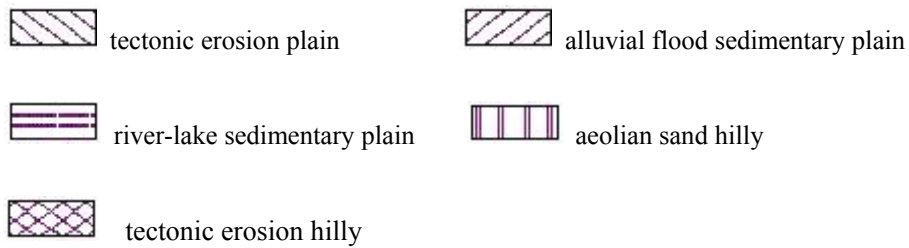
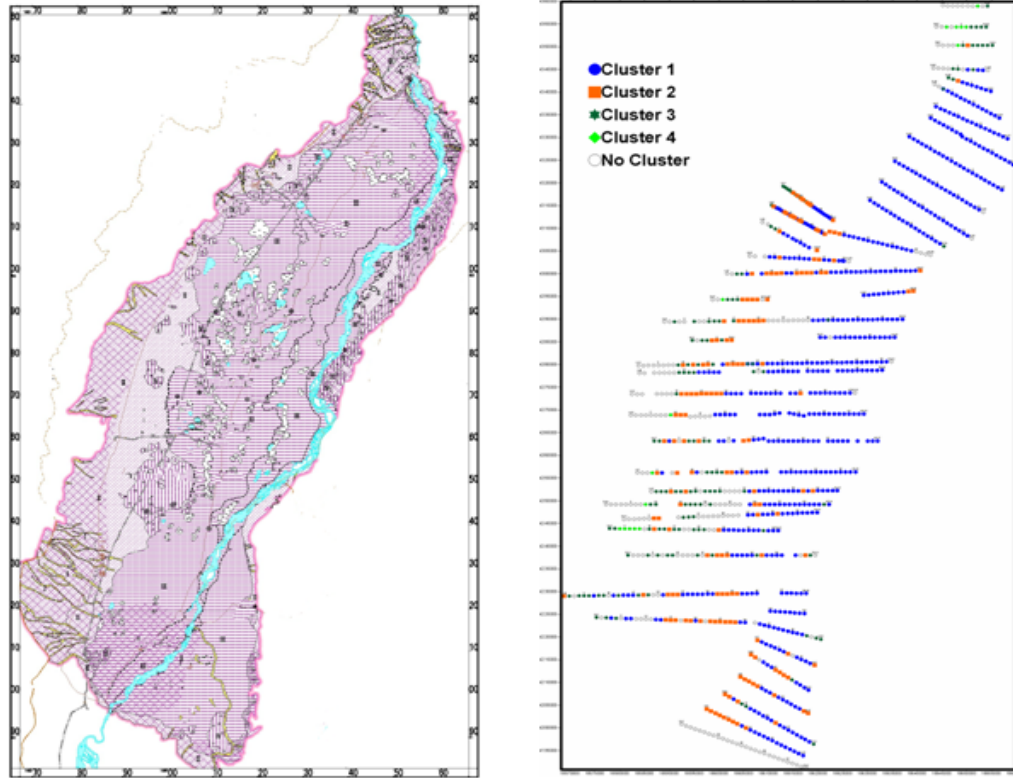
(b) Clustering Result by CADD

Figure 4 Clustering Results of the Apparent Resistivity Curves

Figure 4 (b) is the clustering result distribution of the apparent resistivity curves of electrical method by using the CADD algorithm. It can be seen from Figure 4

(b) that there are four clusters in the measured area. Cluster 1 is blue color which includes over six hundred measuring points. Cluster 2 is orange color which includes one hundred and seventy measuring points. Cluster 3 is dark green which includes about one hundred and thirty measuring points. Cluster 4 is light green color which includes about fifteen measuring points. The outliers are the points which the clusters include less than 5 measuring points. The no-data sample points and the outliers are shown in no cluster.

Comparing Figure 5 (a) with Figure 5 (b), it is obvious that each cluster (b) is fitted well with each region in Geomorphological Map of Ningxia Plain (a). The distribution area of sounding points (blue points (b)) in Cluster 1 is fitted with the region of river-lake sedimentary plain (a). The distribution area of sounding points (orange points (b)) in Cluster 2 is fitted with the region of alluvial flood sedimentary plain and tectonic erosion hilly (a). The distribution area of sounding points (dark green points (b)) in Cluster 3 is fitted with the region of aeolian sand hilly (a). The distribution area of sounding points (light green points (b)) in Cluster 4 is seldom scattered in the region of tectonic erosion plain and alluvial lake sedimentary plain (a). The distribution area of outliers (no cluster (b)) reflects the region of tectonic erosion plain and alluvial flood sedimentary plain in the piedmont (a), because the variational range of electric sounding data is too large to cluster which is caused by tectonic erosion. The clustering result reflects the electric distribution characteristics in Ningxia Plain, and the result is meaningful, useful and effective.



(a) Geomorphological Map

(b) Clustering Result by CADD

Figure 5 Geological-Geomorphological Map and the Clustering Result in Ningxia Plain

7. Conclusion and future works

In this paper, we have developed a new clustering algorithm for the real word application of geophysical prospecting. Firstly, the clustering analysis, one of the data mining techniques, has successfully been applied in geophysical data interpretation and the meaningful and useful results can be got by CADD algorithm. Secondly, the new algorithm was based on density and adaptive

density-reachable so that it could handle the data sets with arbitrary data point distributions. After it had been tested by using different data sets, the new algorithm could have many anticipative performances such as time and space complexity. Thirdly, the results for both test data sets and real data sets indicated that the new algorithm was effective and efficient and that it could eliminate the effect of abnormal data (noise or outliers), and suitable for large data sets and high-dimensional data sets.

Since the CADD algorithm is especially designed for the applications of geophysical prospecting, future research will have to consider the applications of different geophysical prospecting in many different ways. Also, we may try to extend it to other domains such as wireless sensor network or even some social data sets. At the same time, the performances of the CADD algorithm can be tested in many different domains and it can be improved further.

Acknowledgment

The material is based on work supported by National Natural Science Foundation of China under Grant No. 40764002. Michael O'Grady thanks the support of the Irish Research Council for Science, Engineering & Technology (IRCSET) through the Embark Initiative postdoctoral fellowship programme. Gregory O'Hare thanks the support of Science Foundation Ireland under Grant No. 03/IN.3/1361.

References

1. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York. (1973)
2. Birant, D., Kut, A.: ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*. 60(1), 208-221 (2007)
3. Chang, H., Yeung, D.Y.: Locally linear metric adaptation with application to semi-supervised clustering and image retrieval. *Pattern Recognition*. 39(7), 1253-1264 (2006)

4. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd int. Conf. on Knowledge Discovery and Data Mining. pp. 226–231. Portland, Oregon (1996)
5. Ester, M., Kriegel, H.P., Sander, J., Wimmer, M., Xu, X.: Incremental Clustering for Mining in a Data Warehousing Environment. Proceedings of the 24th VLDB Conference. New York, USA. 323-333 (1998)
6. Event, G., Naor, J., Rao, S., Schieber, B.: Fast Approximate Graph Partitioning Algorithms. SIAM Journal on Computing. 28(6), 2187-2214 (1999)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: towards a unifying framework. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. 82–88 (1996)
8. Fu, L.K.: Textbook of Electrical Imaging Surveys. The Geological Publishing House, Beijing, China (1983)
9. Hinneburg, A., Keim, D.A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In Proc. Of th 4th Intl. Conf. on Knowledge Discovery and Data Mining, New York City. pp. 58-65. AAAI Press (1998)
10. Horovitz, O., Krishnaswamy, S., Gaber, M.M.: A fuzzy approach for interpretation of ubiquitous data stream clustering and its application in road safety. Intelligent Data Analysis. 11(1), 89-108 (2007)
11. Hsieh, K.L., Tong, L.I., Wang, M.C.: The application of control chart for defects and defect clustering in IC manufacturing based on fuzzy theory. Expert Systems with Applications, 32(3), 765-776 (2007)
12. Huang, Z.: Extensions to the K-Means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery. pp. 283-304 (1998)
13. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall Advanced Reference Series, Prentice Hall. (1988)
14. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys. 31(3). (1999)
15. Kaufman, L., Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, New York (1990)

16. Kotsiantis, S., Pintelas, P.: Recent Advances in Clustering: A Brief Survey. WSEAS Transactions on Information Science and Applications. 1(1), 73-81 (2004)
17. Li, L.Q., Ji, H.B., Gao, X.B.: Maximum entropy fuzzy clustering with application to real-time target tracking. Signal Processing. 86(11), 3432-3447 (2006)
18. Ma Eden, W.M., Chow, T.W.S.: A new shifting grid clustering algorithm. Pattern Recognition. 37(3), 503-514 (2004)
19. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, pp. 281-297. University of California Press (1967)
20. Meng, H.D., Song, Y.C.: The implementation and application of data mining system based on campus network. Journal on communications, Beijing, China. 26,185-187 (2005)
21. Mitra, D., Ziedins, I.: Hierarchical virtual partitioning--algorithms for virtual private networking. Bell Labs Technical Journal. 2(2), 68-81 (1997)
22. Peng, W.: The computer processing of remote sensing data and geography information system. Beijing Normal School Publishing House, Beijing, China. (2006)
23. Pham, D.T., Dimov, S.S., Nguyen, C.D.: A two-phase K-means algorithm for large datasets. Proceedings of the Institution of Mechanical Engineers -- Part C -- Journal of Mechanical Engineering Science. vol. 218 Issue 10, pp.1269-1273. October (2004)
24. Pham, D.T., Dimov, S.S., Nguyen, C.D.: An Incremental K-means algorithm. Proceedings of the Institution of Mechanical Engineers -- Part C -- Journal of Mechanical Engineering Science. vol. 218 Issue 7, pp.783-795. July (2004)
25. Pilevar, A.H., Sukumar, M.: A grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern Recognition Letters. 26(7), 999-1010(2005)
26. Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery, Kluwer Academic Publishers. 2, 1-27 (1998)
27. Shen, Z.L.: Hydrogeology Textbook. Beijing Science and Technology Press, Beijing, China (1985)
28. Song, Y.C.: Application Research on Clustering Analysis of Data Mining in Geophysical-Geochemical Data Processing. Dissertation, China University of Geoscience, Beijing, China (2006)

29. Song, Y.C., Meng, H.D.: The design of expert system of market basket analysis based on data mining. *Market modernization J*, Beijing, China. 7,184-185 (2005)
30. Song, Y.C., Meng, H.D.: The design of expert system of market basket analysis based on data mining. *Market modernization J*, Beijing, China. 6,152-153 (2005)
31. Soto, J., Aguiar, M.I.V., Flores-Sintas, A.: A fuzzy clustering application to precise orbit determination. *Journal of Computational & Applied Mathematics*. 204(1), 137-143 (2007)
32. Stefanakis, E.: NET-DBSCAN: clustering the nodes of a dynamic linear network. *International Journal of Geographical Information Science*. 21(4), 427-442 (2007)
33. Su, M.C., Liu, Y.C.: A new approach to clustering data with arbitrary shapes. *Pattern Recognition*. 38(11), 1887-1901 (2005)
34. Tan, P.N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Post & Telecom Press of P.R.China (China Edition), Beijing. 359-371 (2006)
35. Trepalin, S., Osadchiy, N.: The centroidal algorithm in molecular similarity and diversity calculations on confidential datasets. *Computer-Aided Molecular Design J*. 19(9), 715-729 (2005)
36. Xu X., Jager J., Kriegel, H.P.: A Fast Parallel Clustering Algorithm for Large Spatial Data sets, *Data Mining and Knowledge Discovery*. 3, 263–290 (1999)
37. Yang, J.: *The Data-Processing and Interpretation Software System for IP Water Exploration and Its Application*. *Geophysical and Geochemical Exploration (China Edition)*. 23(5), 363-375 (1999)
38. Yang, J., Meng, H.D., Fu, L.K.: *The Data-Processing and Interpretation Software System for IP Water Exploration*. *Geophysical and Geochemical Exploration*. 22(3), 204-210 (1998)
39. Yin, B.X.: *Integrated Study of Groundwater Recharge and Water Quality Distribution in Yin Chuan lain*. Dissertation, China University of Geoscience, Beijing, China (2006)
40. Zhao, W.Z., Wu, C.Y., Yin, K., Young, T.Y., Ginsberg, M.D.: Pixel-based statistical analysis by a 3D clustering approach: Application to autoradiographic images. *Computer Methods & Programs in Biomedicine*. 83(1), 18-28 (2006)