CIMCA 2008, IAWTIC 2008, and ISE 2008

A Clustering Algorithm Incorporating Density and Direction

Yu-Chen Song[1], M.J. O'Grady[2], G.M.P. O'Hare[2], Wei Wang[1]
[1] Inner Mongolia University of Science and Technology, Baotou, China.
songyuchen@imust.edu.cn
[2]Adaptive Information Cluster (AIC), School of Computer Science & Informatics,
University College Dublin (UCD), Belfield, Dublin 4, Ireland.
{ michael.j.ogrady, gregory.ohare }@ucd.ie

Abstract

This paper analyses the advantages and disadvantages of the K-means algorithm and the DENCLUE algorithm. In order to realise the automation of clustering analysis and eliminate human factors, both partitioning and density-based methods were adopted, resulting in a new algorithm – Clustering Algorithm based on object Density and Direction (CADD). This paper discusses the theory and algorithm design of the CADD algorithm. As an illustration of its applicability, CADD was used to cluster real world data from the geochemistry domain.

1. Introduction

Clustering analysis is a very important and very active research area within data mining [1], [2]. As a function of data mining, clustering analysis can be used as an independent tool for accessing the data distribution, observing the characteristics of each cluster and enabling further analysis of specific clusters. Clustering facilities the identification of intensive and sparse regions; and enables the discovery of overall distribution patterns together with relations among the data attributes [3]. A high quality clustering algorithm should be capable of identifying:

- Strongest similarity of intra-class data or objects;
- Weakest similarity of inter-class data or objects.;

The quality of clustering usually depends on the similarity measurement adopted and the implementation of the algorithm. It also depends on whether the algorithm can identify some or all hidden patterns.

K-means algorithm is well established and mature, and is the most frequently used clustering method. Many methods and algorithms have been developed that harness the K-means algorithm .Within this paper we describe the CADD algorithm which fuses the particular strengths of both the K-means algorithm and the DENCLUE algorithm, resulting in performance enhancements.

This paper is structured as follows: Section 2 reviews both K-means and DENCLUE. Section 3 introduces some essential definitions while Section 4 presents the algorithmic design and data structure . A real-world application is illustrated in Section 5 after which the paper is concluded.

2. A Brief Review of K-Means & DENCLUE

A key precondition of the K-means algorithm [4], [5] is that the user must determine the number K of clusters before hand. As the clustering result is very sensitive to the value of K, and different K values can often result in completely different results, a user-determined K value can make the clustering result very unsatisfactory. Thus users need domain knowledge to estimate a good K value. This reduces the applicability and automation level of K-means. So far, there is no simple and universally applicable solution to the initial points' selection problem. K-means algorithm is also very sensitive to abnormal data. If some maximum value exists, the data distribution may be highly distorted.

In summary, it is desirable to improve K-means as follows:

- eliminate the need for a user determined K, the number of result clusters;
- enable initial cluster centres be chosen reasonablely;
- reduce the sensitivity to abnormal data.

DENCLUE [6], [7] adopts a density-based clustering approach that models the overall density of a set points as the sum of influence functions associated

978-0-7695-3514-2/08 $25.00 © 2008 IEEE
DOI 10.1109/CIMCA.2008.34

with each point. The resulting overall density function will have local peaks, i.e., local density maxima, and these local peaks can be used to define clusters in a natural way. But DENCLUE can be more computationally expensive than other density-based clustering techniques and it is one of the limitations of DENCLUE.

DENCLUE [8] has a solid theoretical foundation and it can be used to determine the K value and the initial cluster centre points based on a density function, without a need for human intervention. In this way, the K-means algorithm will not be influenced by user determined K value and random initial cluster centre points, and analysts no longer need to have domain expertise. So, a select combination of the both should theoretically improve clustering performance.

3. Basic concepts

Before outlining CADD, some key concepts must be explained.

Definition 1, Object Density: given space $\Omega \in Fd$ consists of a data set of n objects $D=\{x_1, x_2, \ldots, x_n\}$ in which the density of x_i, density(x_i), is the value of the influence function of the object in space.

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

where the 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 object to the density of object x_i, and σ is the density adjustment parameter which is analogous to be the standard deviation, and governs how quickly the influence of a object drops off.

Definition 2, Object Neighbourhood: for any object x and distance R in space, a circular region with centre x and radius R is called the neighbourhood of object x, defined as $\delta = \{x | 0 < d(x, x_i) \le R\}$, in which $d(x, x_i)$ is the distance between object x and x_i.

Definition 3, Neighbourhood radius R: for any object x and distance R in space, a circular region with centre x and radius R is defined as the neighbourhood of object x, marked as $\delta = \{x | 0 < d(x,x_i) \le R\}$. The radius of the defined neighbourhood can be calculated as follows:

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

where mean(D) is the mean distance among all objects, and coefR is the coefficient of neighbourhood radius adjustment.

4. Algorithm design and Data structure

4.1 Algorithm design of CADD

CADD is an incremental algorithm. It uses the nearby data points to increase the initial clusters. If a candidate data point meets any principle of any cluster, the point can be clustered to a certain cluster. For every single clustering event, CADD can start from any object in the cluster. This does not affect the clustering result. It is important to note that only if an object is a density attractor, the point is a cluster centre. Only a cluster centre can cluster its nearby points. Thus, there are three main steps of the algorithm:

(1) Read the data in database to memory and construct the necessary data set structures.

(2) Cluster using CADD. During the process, keep record of the identification of the clusters of every object. This process has two sub procedures:

• Find density attractors to be the cluster centres and produce candidate objects.

• Examine the remaining objects.

A candidate object is such a point that does not yet belong to the current cluster but needs to be clustered. The following rule is used to produce candidate objects: for every new member S of the current cluster C, a circle region with a suitable radius R is used to examine and find out new candidate objects.

Algorithm Clustering Algorithm based on on object Density and Direction（CADD）

Input： *Data set, coefR (Coefficient of neighbourhood radius adjustment).*

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

Method：

1：Compute the densities of each data objects in the data set.

2：i←1

3：repeat

4：Seek the maximum density attractor ODensityMaxi in the original data set of clustering objects as the first cluster center of Ci.

5：Assign the objects in the data set which are in the direction from the objects to ODensityMaxi to

cluster Ci, and at the same time delete the clustered objects form original data set.

6: i←i+1

*7: **until** The original data set is empty.*

8: Mark clusters which have a few objects (such as less 5) as outliers or noise points.

---

(3) Examine the candidate samples (data objects). Check whether the distance between each object and the cluster centre is less or equal to the neighbourhood radius R and if it does, the object belongs to the cluster otherwise it does not. The examination method in CADD algorithm implies the dependency of the distance between objects but is independent of the examination sequence. In reality, in order to reduce the dependency of the distances, CADD also combines the following features:

• After an object is clustered, the object does not participant in the next clustering process, and is removed from the original data set.

• After all clusters having been completed, all remaining samples (data objects) will be examined based on the definition of object direction.

• The clusters which have only a few objects are marked as outliers or noise points.

4.2 Data structure of CADD

The neighbourhood radius ensures adjacent relations are formed among data sets objects, in much the same way as us found in graphs. Adjacent list is a data structure that is frequently used for representing graphs [9]. Thus, an adjacent list was identified as a suitable data structure for CADD. In addition, adjacent lists are conceptually straightforward, easy to manipulate and frequently used. An adjacent list consists of two parts – a list header and a list node, and an example may be seen in Figure 1.

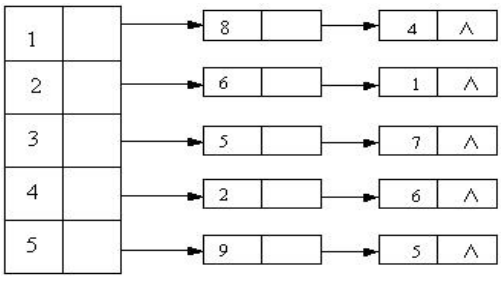

Figure 1. Sample of adjacent list

Since clustering results have a similar structure to data sets, besides using Adjacent List for data sets, such list together with linked list can also used for clustering results. Hence, we can specify a data structure for the clustering results:

```
struct DoubleNode
{
    DataNode*  object;
    DoubleNode* next;
    int        IdNode;
    int        flag;
};
```

• The data for every sample point needs to be stored. Since the sample points are dynamic, it is necessary to have a linked list, DataNode, in the structure to store the data.

• A pointer next needs to be defined to point to the connecting subsequent nodes.

• In order to uniquely identify each sample point, every sample point must have a unique identification number for use in processing. This identification number, IdNode, is set to integer.

• It is frequently necessary to check whether a sample point is clustered or not. In order to eliminate unnecessary repeatitive operations, an integer variable flag needs to be defined using 0 and 1 to represent whether a sample is clustered or not.

5. A real-world application

In this section, we present a real-world application of geochemistry. Firstly, we show the two plots: one is a geological plot and the other is produced by CADD algorithm. Figure 2 is the result of a regional geological survey. Different colours represent the distribution of different rock lithology. It can be seen from Figure 2 that there is an obvious heterogeneity of the combination of the chemical elements in the sampling region. The changing of the heterogeneity follows a certain rule which is represented by the band shape changing from the top left corner to the bottom right corner in the sampling region. It reflects the migration and proliferation characteristics of the chemical elements.

Figure 3 is the clustering result distribution using CADD algorithm. The clustering was performed automatically. Some outliers and non-data sample points can be seen, but there are four clusters that reflect the chemical element distribution. The four clusters are distributed in a similarthe same band shape as the original geological plot (Figure 2) from

the top left corner to the bottom right corner in the sampling region. The CADD result broadly reflects shows good function in the migration and proliferation characteristics of the chemical elements. It reflects the chemical element distribution characteristics in more detail and thus has a better result than that with two clusters that is produces by by K-means algorithm (because the limitation of pages, the result using K-means is not represented here).

Figure 2. Results of regional geological survey

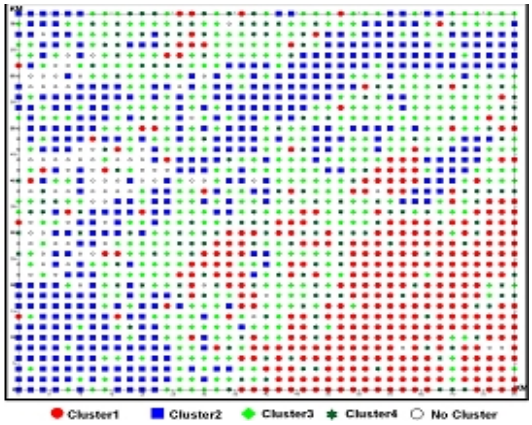

Figure 3. Clustering using CADD

6. Conclusion

In this paper, we described the CADD algorithm - a combination of K-means and DENCLUE algorithms respectively. It was successfully applied to a real world domain, namely geochemistry. CADD was demonstrated to be robust in that it automatically determined the number K of clusters, and is capable of identifying clusters of multiple shapes and sizes.

At present, a range of new algorithms for clustering analysis are being produced. It is hoped that these may prove effective in different application domains. In this spirit, we have conducted a number of exploratory studies in other areas, such as shopping carts [10], [11], and a campus network [12]. It is intended to continue evaluating the algorithm in otherdifferent domains, for example, geophysics and wireless sensor networks.

7. Acknowledgment

This material is based upon works supported by:

(1) The National Natural Science Foundation of China (No. 40764002).

(2) Science Foundation Ireland under Grant No. 03/IN.3/I361.

(3) China Scholarship Council.

(4) Science Foundation of Inner Mongolia University of Science and Technology.

8. References

[1] P.N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining,” Addison Wesley, pp. 487-568, 2005

[2] R. Xu, D. Wunsch II, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, , vol.16, no.3, pp. 645-678, May 2005

[3] L. Kaufman, P.J. Rousseeuw, “Finding Groups in Data: An Introduction to Cluster Analysis,” John Wiley and Sons, pp.114-138, 1990.

[4] S.Z. Selim，M.A. Ismail, “K-Means-Type Algorithms: A Generalized Convergence Theorem and characterization of Local Optimality,” IEEE Trans Pattern Analysis and Machine Intelligence, pp. 81-87, 1984.

[5] D.T. Pham, S.S. Dimov, C.D. Nguyen, “An Incremental K-means Algorithm”, Proceedings of the Institution of Mechanical Engineers, Journal of Mechanical Engineering Science, vol. 218, Issue 7, pp.783-795, 2004.

[6] A. Hinneburg and D.A. Keim, “An Efficient Approach to Clustering in Large Multimedia Databases with Noise,” In Proc. Of th 4th Intl. Conf. on Knowledge Discovery and Data Mining, pp. 58-65. 1998.

[7] A. Hinneburg and D.A. Keim, “A General Approach to Clustering in Large Databases with Noise,” Knowledge and Information Systems (KAIS), vol. 5, no. 4, pp. 387-415, 2003.

[8] P.N. Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining,” Post & Telecom Press of P.R.China (China Edition), pp. 377-379, 2006. (Chinese version)

[10] Y.C. Song and H.D. Meng, “The knowledge base building of expert system of market basket analysis based on data miming,” Market modernization, Beijing, China, no.7, pp. 184-185, 2005.

[11] Y.C. Song and H.D. Meng, “The design of expert system of market basket analysis based on data mining,” Market modernization, Beijing, China, no.6, pp. 152-153, 2005.

[12] H.D. Meng and Y.C. Song, “The implementation and application of data mining system based on campus network’”. Journal on communications, Beijing, China, vol.26, no.1A, pp.185-187, 2005.