2008 International Conference on Computer Science and Software Engineering

# Research and Application of Clustering Algorithm for Arbitrary Data Set

Yu-Chen Song
*Inner Mongolia University of Science and Technology, Baotou, China.*
*bjsongyc@hotmail.com*
*songyuchen@imust.edu.cn*

M.J. O'Grady
*School of Computer Science & Informatics, University College Dublin, Belfield, Dublin 4, Ireland.*
*michael.j.ogrady@ucd.ie*

G.M.P. O'Hare
*School of Computer Science & Informatics, University College Dublin, Belfield, Dublin 4, Ireland.*
*gregory.ohare@ucd.ie*

## Abstract

*This paper discusses the theory and algorithmic design of the CADD (Clustering Algorithm based on object Density and Direction) algorithm. This algorithm seeks to harness the respective advantages of the K-means and DENCLUE algorithms. Clustering results are illustrated using both a simple data set and one from the geological domain. Results indicate that CADD is robust in that automatically determines the number K of clusters, and is capable of identifying clusters of multiple shapes and sizes.*

## 1. Introduction

Cluster analysis groups data objects based on only information found in the data that describes the objects and their relationships. The goal is that objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between the groups are, the better or more distinct the clustering. A large number of clustering algorithms have been developed in a variety of domains for different types of applications [1]. None of these algorithms is suitable for all types of applications. In fact, it seems that there is always a need for a new clustering algorithm that is more efficient for a particular type of applications.

In this paper, we briefly review two algorithms - , K-means and DENCLUE. We then present a new algorithm that incorporates the respective advantages of K-means and DENCLUE. This algorithm - CADD (Clustering Algorithm based on object Density and Direction) is then compared with k-means using a standard dataset and one from the geochemistry domain. The paper is then concluded.

## 2. Review of K-means & DENCLUE

### 2.1 K-means – a classic algorithm

K-means [2] [3] is based on the principle which minimizes the clustering property indicator. A commonly used clustering criteria function is the minimized variance of all the sample points (data or objects) to the central point in the cluster. Strengths of K-means include:

- simplicity and applicability for a wide variety of data types. It is also quite efficient, even though multiple runs are often performed.
- good results when the cluster is intensive and the distinction between clusters is obvious.
- with large data sets, it is relatively efficient and scalable.

Weaknesses of K-means include:

- a requirement for a known K value before clustering is carried out, which is difficult for inexperienced users.
- the choice of initial centers of the clusters have a great effect on the clustering result. The algorithm is not suitable for non-convex shape clusters nor cluster sizes that are highly variable.
- A sensitivity to noise points, marginal points and isolated points.

These disadvantages would greatly weaken the efficiency and utilities of K-means in unsupervised clustering in data mining. It will normally get the local optimal result rather than the global optimal result.

978-0-7695-3336-0/08 $25.00 © 2008 IEEE
DOI 10.1109/CSSE.2008.415

### 2.2 Density-based DENCLUE algorithm

DENCLUE (DENsity-based CLUstEring), designed by Hinneburg and Keim, is based on the density distribution function. The basic tenets of the DENCLUE algorithm are:

- the influence of each data point can be formally simulated by a mathematical function, called the influence function [4], which describes the influence of a point in its neighbourhood.
- the overall density of a data space can be modelled to the sum of influence functions of all the data points.
- clusters can be determined by a density attractor.

Characteristics of DENCLUE include the following [5]:

- it has a solid mathematical foundation which summarises all other clustering methods including partitioning, hierarchical and grid-based methods.
- it produces good cluster characteristics from noisy data sets.
- it provides a simple mathematical description for the clusters of arbitrary shapes in high dimensional data sets.
- the data structure, based on a grid-cell makes the algorithm capable of handling large, high-dimensional data sets. But because the choice of density and noise threshold parameters can greatly influence the quality of the clustering results, DENCLUE highly depends on these parameters.

DENCLUE 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. K-means algorithm is simple and also quite efficient for large data sets. Thus a select combination of the both should theoretically improve clustering performance.

## 3. Algorithm design

Before presenting the CADD algorithm, it is necessary to construct some definitions.

### 3.1 Relevant definitions

**Definition 1:** Object Density: given space $\Omega \in F^d$ 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. Thus,

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

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 the standard deviation, and governs how quickly the influence of an 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 $\delta = \{x|0<d(x, x_i) \leq R\}$, 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 $\delta = \{x|0<d(x,xi)\leq R\}$. The radius of the defined neighbourhood can be calculated as follows:

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

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

**Definition 4:** Object Direction: based on the definition of clustering, a characteristic of an object at the boundary of a cluster is that there are more neighbour points in one direction (normally is the direction pointing to the cluster centre) of the object, but only very few neighbour points in the opposite direction. Thus object direction can be defined as follows: in a data set D, there exists $d(x,y)=\min(|x-y|)\leq R$. If $y \in cj$, $j \in \{1,2,\ldots,k\}$, then $x \in cj$ which is the direction from object x to the cluster centre cj, defined as $x \rightarrow cj$.

**Definition 5:** Density Attractor: the local maximum data point of the global density function in space, that is if the density of an object is greater than its surrounding object, the object is a density attractor.

### 3.2 CADD model

There are four parts in CADD algorithm:

1. Constructing a dissimilarity matrix: CADD is a clustering method based on dissimilarity matrix.
2. DENCLUE clustering method: determining the number and positions of density attractors through

a hill-climbing algorithm (number of clusters and the positions of the clusters centre points) and obtaining the input parameters for K-means. Thus CADD achieves automation of clustering analysis and elimination of the human factor.

3. Partitioning data points: Assign the data objects which are in a circular region with centre of density attractor and radius R to each cluster, and at the same time delete the clustered objects from the original data set. Because the cluster centre points are pre-determined, it is not necessary to use iterations to optimise cluster centres. Rather, one iteration will produce the optimised result.
4. Clustering based on the definition of object direction: direction based clustering of the remaining objects. If no remaining objects, all samples were clustered. If some remaining, cluster using the definition of object direction until all objects meeting requirements are clustered.

## 4. Comparison of clustering results

Two clustering results are presented as follows: one is an arbitrary data set with two-dimension graphs, and the second is a real-world geochemical application.

### 4.1 An arbitrary data set

In order to compare the clustering result of K-means and CADD, the clustering result analysis graphs of two-dimensional, arbitrary objects distribution are shown in Figure 1 and Figure 2.

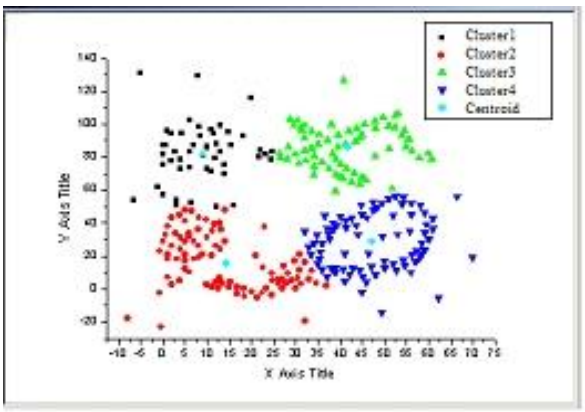


**Figure 1. Clustering using K-means**

It can be seen from Figure 1, using K-means, some objects which naturally belong to Cluster2 or Cluster3 are assigned to Cluster1; a number of objects which naturally belong to Cluster4 are assigned to Cluster2; some outliers are unrecognized; and non-globular distribution of data is not found.

It can be seen from Figure 2 that CADD is capable of handling arbitrary shape data sets very well, eliminating the influence of noise and outliers, and accurately reflecting the spatial distribution characteristics of the original data sets.

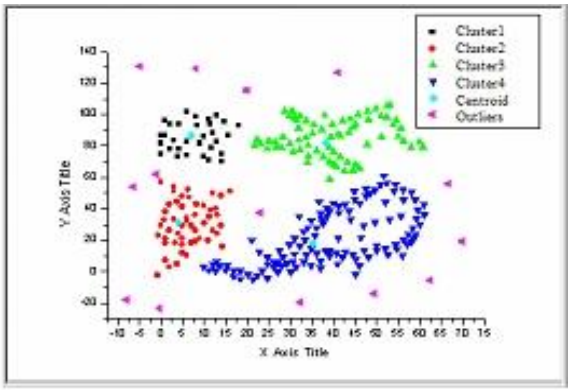


**Figure 2. Clustering using CADD**

### 4.2 A real-world application

**Figure 3. Results of regional geological survey**

In this section, we present a real-world application from geochemistry. Figure 3 is the result of a regional geological survey. Different colors represent the distribution of different rock lithology. It can be seen from Figure 3 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 4 is the clustering result distribution using CADD. 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

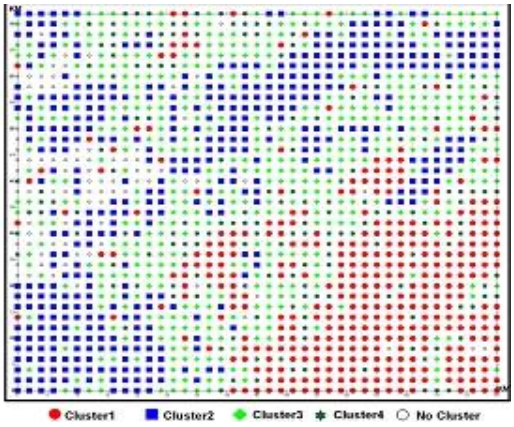


**Figure 4. Clustering using CADD**

in a similar band shape as the original geological plot from the top left corner to the bottom right corner in the sampling region. The CADD result broadly reflects 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 produced by K-means (Figure 5).

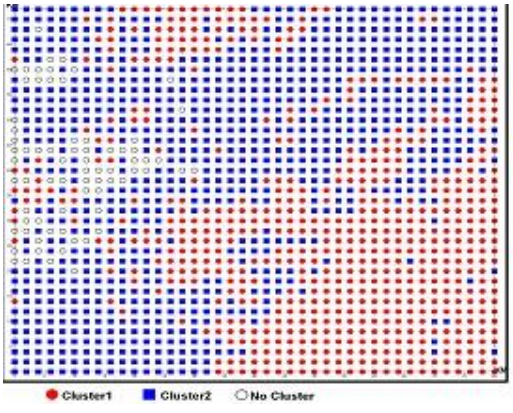


**Figure 5. Clustering using K-means**

## 5. 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 [6] [7] and a campus network [8]. It is intended to continue evaluating the algorithm in other domains, for example, geophysics and wireless sensor networks.

## Acknowledgments

This material is based upon works supported by the National Natural Science Foundation of China (No. 40764002 ), Science Foundation Ireland under Grant No. 03/IN.3/I361; and the China Scholarship Council.

## References

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

[2] 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.

[3] 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.

[4] 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)

[5] 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.

[6] Y.C. Song, H.D. Meng, "The design of expert system of market basket analysis based on data miming," Market modernization, Beijing, China, no.7, pp. 184-185, 2005.

[7] Y.C. Song, 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.

[8] H.D. Meng, 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.