CHAPTER 1

INTRODUCTION

With the development and improvement of data mining technology, data clustering algorithm are gradually applied to some fields. The definition of clustering in the academic community can be generalized as follows: first, the similarity of data objects. Data objects within the same cluster have great similarity, but data objects within the different cluster have great non-similarities. Second, the distance of data objects. Take entire data set as a test data object of the gathering, the distance between any pair of data objects within the same cluster size should not be greater than the distance between the different clusters of arbitrary data object. Third, the density of data objects. Take entire data set as a multi-dimensional space aggregation of the data object, a cluster is the spaces which contain the number of data object relatively high dimension cut-of by the space which contains the number of data object relatively low dimension. Thus form a relatively separated set of dimensional space.

For a long time, though many data cluster algorithm had been proposed, like K-means algorithm, Dbscan algorithm, and WaVecluster algorithm, etc, however, all these algorithm has big overhead, but low efficiency when mentioning data cluster of person data collection, this directly limited its application in related fields.

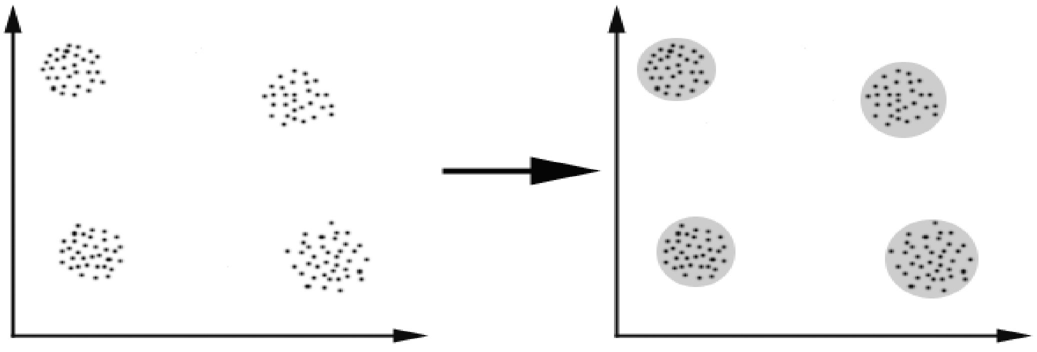

Fig1: Data Clustering

Among clustering algorithms, K-means clustering algorithm can be applied in many fields including image and audio data compression, pre-process of system modeling with radial basis function networks, and task decomposition of heterogeneous neural network structure,etc. k-means clustering is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid.

CHAPTER 2

OVERVIEW

Organizing data into *clusters* such that there is high intra-cluster similarity, low inter-cluster similarity. Informally, finding natural groupings among objects is called as clustering. We need clustering because organizing data into clusters shows internal structure of the data, for example Clusty and clustering genes. Sometimes the partitioning is the goal. For example Market segmentation. Prepare for other AI techniques .For example Summarize news (cluster and then find centroid) .Techniques for clustering is useful in knowledge discovery in data. For example underlying rules, reoccurring patterns, topics, etc.

2.1 Types of Clustering

2.1.1 Nesting

Hierarchical Clustering:

This separation is based on the characteristic of nesting clusters. **Hierarchical clustering** are nested by this we mean that it also clusters to exist within bigger clusters as shown in Figure 2.

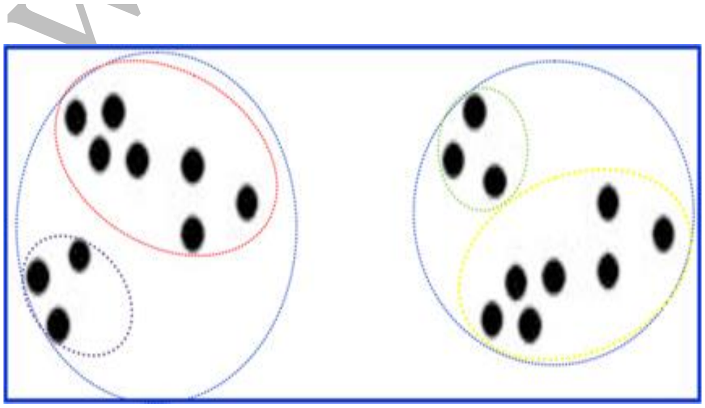

Fig2: Hierarchical Clustering

Partitional Clustering:

This separation is based on the characteristic of nesting clusters. **partitional clustering** prohibits subsets of cluster as shown in Figure 3 .

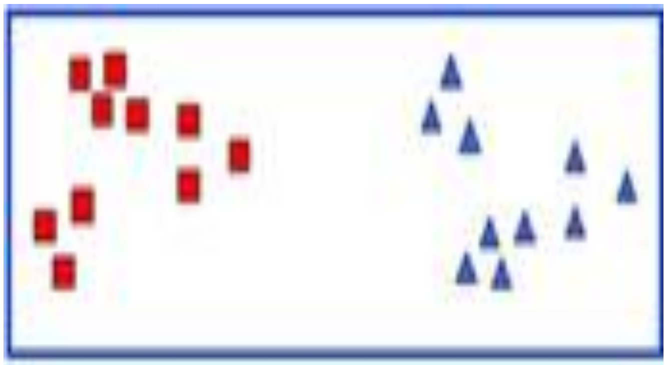

Fig3**:** partitional clustering

2.1.2 Exclusiveness:

This separation is based on the characteristic that allows a data object to exist in one or more than one clusters. **Exclusive clustering** is as the name suggests and stipulates that each data object can only exist in one cluster. Figure 4 is an example as each object is only a member of one cluster.

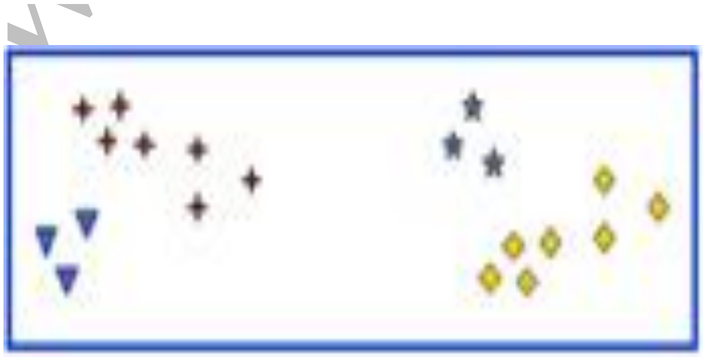

Fig4: **Exclusive clustering**

Overlapping clustering allows data objects to be grouped in 2 or more clusters as shown in figure 5. A real world example would be the breakdown of personnel at a school. Overlapping clustering would allow a student to also be grouped as an employee while exclusive clustering would demand that the person must choose the one that is more important.

In **Fuzzy clustering** every data object belongs to every cluster, I guess you can describe fuzzy clustering as an extreme version of overlapping, the major difference is that the data objects has a membership weight that is between 0 to 1 where 0 means it does not belong to a given cluster and 1 means it absolutely belongs to the cluster. Fuzzy clustering is also known as probabilistic clustering.

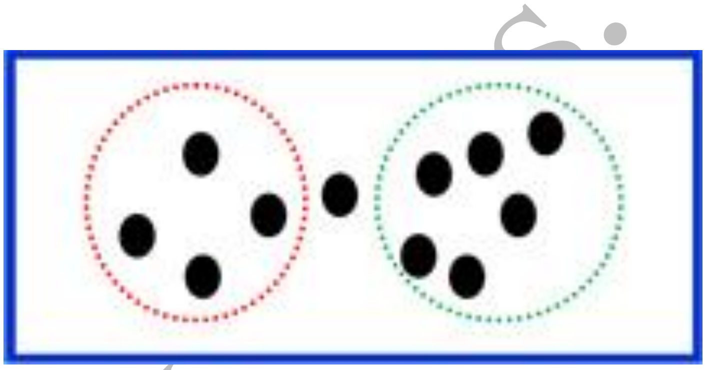

Fig5:**Overlapping clustering**

2.1.3 Completeness:

This separation is based on the characteristic that requires all data objects to be grouped. A **complete clustering** assigns every object to a cluster. All of the previous clustering figures are examples of complete clustering because in each one of them each data point is assigned to a cluster. P**artial clustering** as shown in Figure 6 on the other hand allows some data objects to left alone.

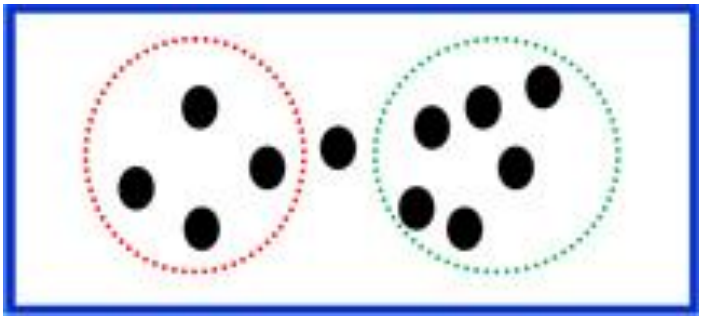

Fig6: P**artial clustering**

2.3 Desirable Properties of any Clustering Algorithm

- Scalability (in terms of both time and space)
- Ability to deal with different data types
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to the order of input records
- Incorporation of user-specified constraints
- Interpretability and usability.

CHAPTER 3

Traditional K-mean Algorithm

This is the most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm**; it is also referred to as **Loyd algorithm**, particularly in the computer science community.K-means cluster algorithm was proposed by J. B. MacQueen in 1967, which is used to deal with the problem of data clustering.

The algorithm is relatively simple, so generate a wide influence in the scientific field research and industrial applications. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

It is based on decomposition, using K as a parameter, divide n object into K relatively low similarity between clusters. And minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. The measurement of the similarity for the algorithm selection is by the reciprocal of the Euclidean distance.

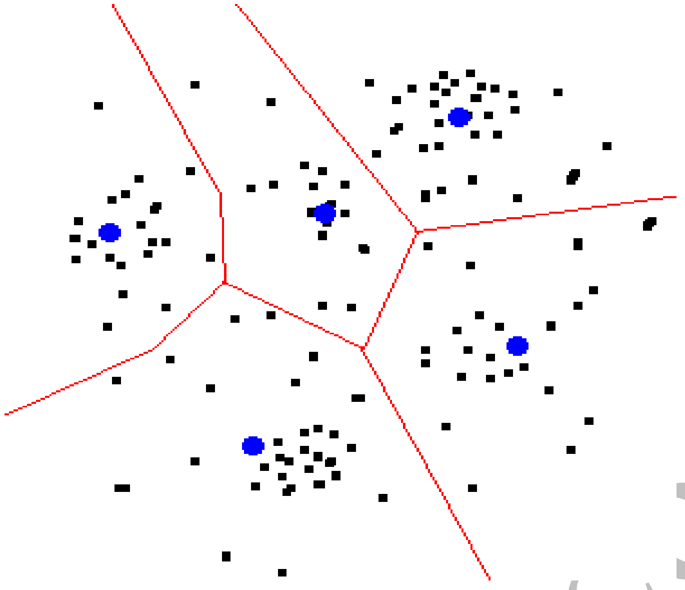

Fig7: k- means clustering

3.1 Procedure of K-means Algorithm

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The algorithm is as follows:

• Distribute all objects to K number of different cluster at random;

• Calculate the mean value of each cluster, and use this mean value to represent the cluster;

• Re-distribute the objects to the closest cluster according to its distance to the cluster center;

• Update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster;

• Calculate the criterion function E, until the criterion function converges.

Usually, the K-means algorithm criterion function adopts square error criterion, be defined as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

In which, J is total square error of all the objects in the data cluster, *xi* bellows to data object set, *Cj* is mean value of cluster (x and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect.

3.2 Analysis of the Performance of K-means Algorithm

3.2.1 Advantages:

1. K-mean value algorithm is a classic algorithm to resolve cluster problems; this algorithm is relatively simple and fast.
2. For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk). Among which, n is the times of iteration, k is the number of cluster, t is the times of iteration. Usually, k”n and t”n. The algorithm usually ends with local optimum.
3. It provides relatively good result for convex cluster.
4. Because the limitation of the Euclidean distance. It can only process the numerical value, with good geometrical and statistic meaning.

3.2.2 Disadvantages:

The inherent prosperities of the K-means clustering algorithm to determine its limitations, specific performance is as follows:

1. The K value is most important for K-means clustering algorithm. There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.
2. K-means clustering algorithm has a higher dependence of the initial cluster centers. If the initial cluster center is completely away from the cluster center of the data itself, the number of iterations tends to infinity, but also makes it easier for the final clustering results into local optimization, resulting in incorrect clustering results.
3. K-means clustering algorithm has a strong sensitivity to the noise data objects. If there is a certain amount of noise data in dataset, it will affect the final clustering results, leading to its error.

4. K-means clustering algorithm for the discovery of clusters of arbitrary shape is most difficult.

5. K-means clustering algorithm has man limitation on amount of data. In the iterative process, every time you need to adjust the cluster to which data object belongs and compute cluster center, so in case of large amount of data, the K-means clustering algorithm is not applicable.

CHAPTER 4

The Research Point of K-means Clustering Algorithm

The research on K-means clustering algorithm is mainly from the following two aspects: First, about the determination of K value. Through the above analysis, the K value of the initial cluster centers to determine the far-reaching impact throughout the clustering process and the final clustering results, while the K value in practical applications is very difficult to direct or one-time determination .Especially, if the amount of data tends to infinity which is pending, the K value of the K-means algorithm to determine will be very difficult.

At present, there are two clustering algorithms to determine the K value is relatively effective which is the cost function based on distance and propagation clustering algorithm based on nearest neighbors. The former find the minimum through using the cost function. Thus obtain the corresponding K value. The latter using nearest neighbor clustering algorithm to calculate the appropriate number of cluster center, the number of cluster center provides for the maximum K value of the K-means clustering algorithm to get the optimal value of K.

Second, about the choice of initial cluster centers. K-means clustering algorithm using the iterative method to solve the problem, except the first step, the clustering results of each step are improved to some extent; otherwise terminate the process of iteration. Traditional K-means clustering algorithm takes the cluster squares error and the criterion function value change or not as the iterative termination conditions. But the clustering results obtained from this criterion function easily fall into local minimum solution, the result is the clustering results of search are moving toward the direction of diminishing the criterion function value .

In this paper, the improvement of K-means algorithm is mainly reflected in the following two aspects:

- Optimize the initial cluster centers, to find a set of data to reflect the characteristics of data distribution as the initial cluster centers, to support the division of the data to the greatest extent.

• Optimize the calculation of cluster centers and data points to the cluster center distance, and make it more match with the goal of clustering.

CHAPTER 5

Improved K-means Clustering Algorithm

5.1 Related Concept:

Defnition 1: The distance between data points and the cluster center. The distance formula of data point xi and cluster center kj defined as following:

$$d_{j,i} = \sqrt{(x_{i_1} - k_{j_1})^2 + (x_{i_2} - k_{j_2})^2 + \cdots + (x_{i_w} - k_{j_w})^2} \qquad (2)$$

where w represents the number of attributes of the data points xi.

Defnition 2 :The density parameter τ . The number of data points which is contained by a scope defined as density parameter. The scope is a round which takes space point of not statistics xi as the center, γ as the radius. The greater the density of xi, the greater the value of the density parameter are.

Defnition 3: The core data points. If the γ -neighborhood of a data point contains at least PTS min number of data points, then the data point called the core data point.

Defnition 4: The cluster center. Differences from the traditional clustering adjustment, the improved clustering algorithm add the weight of data point to the cluster center. Data points near the center of the cluster weights, on the contrary, the value of data points away from the cluster center is less weight. The formula of cluster center defined as follow:

$$k = \frac{d_{j_h}}{D} x_{J_1} + \frac{d_{j_{(h-1)}}}{D} x_{J_2} + \cdots + \frac{d_{j_2}}{D} x_{J_{(h-2)}} + \frac{d_{j_1}}{D} x_{j_h} \qquad (3)$$

where j represents the jth cluster, h is the number of data points in the cluster, djh represents the distance between the hth data point which belongs to cluster c and cluster center. And with the restriction of $d_{j1} \le d_{j2} \le \ldots \le d_{jh}, \frac{d_{j1}}{D} + \frac{d_{j2}}{D} + \cdots + \frac{d_{jh}}{D} = 1.$

Defnition 5: The Euclidean distance between data points and the cluster center. The distance between data point and the cluster center determine the cluster which data point belongs to, the formula of Euclidean distance is defined as follows:

$$d_{ji} = \left(1 - \frac{\sigma_j}{\sigma}\right) d_{ji} \qquad (4)$$

where j represents the jth cluster cj , i represents the ith data point xi, dji is the Euclidean distance between data point xi and the cluster center cj , σ_i represents the squares error of the cluster cj , σ is the squares error sum of the K clusters c.

5.2 Improved K-means Algorithm Description

In computing the region of high-density data set D, set $2\gamma = 1/100$, the entire area is divided into 100 parts, set the density threshold *T*=*n*/100.

Algorithm 1: Improved K-means Algorithm

Input: data set x contains n data points; the number of cluster is k.
Output: k clusters of meet the criterion function convergence.

Program process:
Step 1. Initialize the cluster center.

Step 1.1. Select a data point *xi* from data set X, set the identified as statistics and compute the distance between *xi* and other data point in the data set X. If it meet the distance threshold, then identify the data points as statistics, the density value of the data point *xi* add 1.

Step 1.2. Select the data point which is not identified as statistics, set the identified as statistics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

Step 1.3. Select data point from data set which the density value is greater then the threshold and add it to the corresponding high-density area set D.

Step 1.4. Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster center set. Followed to find the k-1 data points, making the distance among k initial cluster centers are the largest.

Step 2. Assigned the n data points from data set X to the closest cluster.

Step 3. Adjust each cluster center K by the formula (3).

Step 4. Calculate the distance of various data objects from each cluster center by formula (4), and redistribute the n data points to corresponding cluster.

Step 5. Adjust each cluster center K by the formula (3).

Step 6. Calculate the criterion function E using formula (1), to determine whether the convergence, if convergence, then continue; otherwise, jump to Step 4.

CHAPTER 6

Simulation Experiment and Results Analysis

To further validate the effectiveness of the improved clustering algorithm, this paper uses a part of the UCI test data set IRIS and WINE data set for experiment, the data set name, the number of attributes, as well as the number of data objects contained in Table 1.

Table 1: UCI data set for experiment

Data set name	Propriety number	Data object number
IRIS	4	150
JWINE	13	178

UCI common data sets as a test of machine learning and data mining algorithms,all of its data objects have been assigned to the corresponding class in the data set in improved clustering algorithm .Clustering results can be intuitive and easy to get accuracy, which is the validity of improved clustering algorithm. Before running the algorithm, we set the value of K the number of categories of the standard data set, then turn to run the traditional K-means clustering algorithm and improved K-means clustering algorithm, the experimental results as shown for example in Table 2 and Table 3.

Table 2: Accuracy of the traditional and improved K-means clustering algorithm

Algorithm	**Data set**	**Accuracy of clustering**		
		Maximum	**Minimum**	**Average**
Traditional K-means	IRIS	89.33	82.00	87.30
	WINE	74.16	70.22	71.01
Improved K-means	IRIS	90.45	83.26	88.51
	WINE	76.29	71.54	74.23

Table 3: UCI data set for experiment

Algorithm	**Data set**	***Emax***	***Emin***	***Eavg***
Traditional K-means	IRIS	145.279	145.279	98.511
	WINE	2.647E6	2.371E6	2.424E6
Improved K-means	IRIS	78.483	75.731	76.190
	WINE	2.255E6	1.973E6	2.106E6

As seen from Table 2 and Table 3, compared to the traditional K-means clustering algorithm, the improved K-means clustering algorithm on the two data sets in IRIS and WINE clustering results accurate rate has improved significantly. Consider the K-means clustering algorithm, the basic idea is to make the data in the same cluster have high similarity, a relatively low similarity data will be divided into different clusters. And we can see from Table 3, improved clustering algorithm significantly decline in IRIS and WINE data clustering criterion function E 76.190 and 2.106E6, respectively, this indicates that the improved K-means clustering algorithm clustering result, each cluster is more compact.

Conclusion and Future Enhancements

As a key clustering algorithm, K-means cluster algorithm has already become one of the hotspots in the present. In this paper, through analysis the advantage and disadvantage of traditional K-means cluster algorithm, we elaborate two ways of improvement for K-means cluster algorithm, offer the improved algorithm. In the last, the simulation results show the improved clustering algorithm is not only the clustering process is more stable, at the same time, improved clustering algorithm to reduce or even avoid the impact of the noise data in the data set object to ensure that the final clustering result is more accurate and effective. However, researching on the improvement of K-means clustering algorithms are still not solved completely. And the further attempt and explore will be needed.

References

[1] Saifan Wang, Fang Dai, Bo Liang, A path-based clustering algorithm of partition, in: Information and Control, 40(1), 2011, 141-144

[2] Jiawei Wu, Xiongfei Li, Tao Sun, Wei Li, A density-based clustering algorithm concerning neighborhood balance, in: Journal of Computer Research and Development, 47(6), 2010, 1044-1052

[3] Zhaoxia Tang, K-means clustering algorithm based on improved genetic algorithm, in: Journal of Chengdu University (Natural Science Edition), 30(2), 2011, 162-164

[4] Yu Hu, JinZhi Bi, Optimized K-means clustering analysis based on genetic algorithm, in: Computer System Application, 19(6), 2010, 52-55

[5] Ke Sun, Jie Liu, Xueying Wang, K mean cluster algorithm with refined initial center point, in: Journal of Shenyang Normal University(Natural Science), 27(4), 2009, 448-451

[6] Ye Tao, Zhiyong Zeng, Jiankun Yu, Completeness proof and implementation of parallel K-means clustering algorithm, in: Computer Engineering, 36(22), 2010, 72-74

[7] Weiwei Ni, Geng Chen, Zhihui Sun, An efficient density-based clustering algorithm for vertically partitioned distributed datasets, in: Journal of Computer Research and Development, 44(9), 2007, 1612-1617