K-means Clustering Algorithm Characteristics Differences based on Distance Measurement

P.Indira Priya,
Assistant Professor,
Department of CSE,
Tagore Engineering College, Chennai, TN.

D.K.Ghosh, PhD.
Professor,
Department of CSE,
V.S.B. Engineering College, Karur, TN.

ABSTRACT
A distance measure for similarity estimation based on the differences is presented through our proposed algorithm. This kind of distance measurement is implemented in the K-means clustering algorithm. In this paper, a new Minkowski distance based K-means algorithm called Enhanced K-means Clustering algorithm (EKMCA) is proposed and also demonstrates the effectiveness of the distance measurement, the performance of this kind of distance and the Euclidian and Minkowski distances were compared by clustering KDD'99 Cup dataset. Experiment results show that the new distance measure can provide a more accurate feature model than the classical Euclidean and Manhattan distances.

Keywords - Clustering, Distance, K-means clustering algorithm, Enhanced K-Means Clustering Algorithm

1. INTRODUCTION

Cluster analysis is a set of different methodologies for clustering of data's into a number of groups using distance measure. Clustering is very important for many examining and finding tasks including machine learning, pattern recognition, and data mining. A number of research work has been done on building clustering algorithms and numerous clustering algorithms are proposed . Every approach has its own merits and demerits. Clustering are used to perform investigative data analysis technique, it attempts to partition a given data set in to dissimilar groups such that data patterns within a group are more similar to one another than those belonging to different groups[5].

Clustering techniques are classified into supervised and unsupervised methods. The unsupervised clustering method is used to detect the underlying structure in the data set for classification [6]. Supervised clustering method involved with the human interaction.
The unsupervised clustering techniques are most popular due to the minimal knowledge about the dataset. Similarity is fundamental to the definition of a cluster, and various clustering techniques use different similarity definitions and techniques. The very famous distance measure may be the Euclidean distance . But, it has taken much time to cluster the data's compare to other distance metrics like Minkowski.
In this paper, a new Minkowski distance based K-means clustering algorithm for clustering the data is proposed. In addition, the performance analysis is compared with our proposed algorithm. In this work, KDD Cup data set is used to perform the analysis of these two algorithms and the merits and demerits their distance metrics. The main advantage of this proposed algorithm is that the execution time compare to existing K-means algorithm is reduced to a considerable extent.

2. LITERATURE SURVEY

Tingting Cui and Fangshi Li [1] presented Weight Computing in Competitive K-Means Algorithm which is derived from Improved K-means method and subspace clustering. By adding weights to the objective function, the contributions from each feature of each clustering could simultaneously minimize the separations within clusters and maximize the separation between clusters. LI Han [2] focused on intrusion detection based on data mining.

Their aim is to improve the detection rate and decrease the false alarm rate, and the main research method is clustering analysis. A modified dynamic K-means algorithm called MDKM is proposed and corresponding simulation experiments are presented. Firstly, the MDKM algorithm filters the noise and isolated points on the data set to reduce the negative impact. Secondly by dynamic iterative process we find the *k* clustering center accurately, an anomaly detection model is presented and we get better detection effect.

A data clustering approach using modified K-Means algorithm based on the improvement of the sensitivity of initial center (seed point) of clusters. This algorithm partitions the whole space into different segments and calculates the frequency of data point in each segment. The segment which shows maximum frequency of data point will have the maximum probability to contain the centroid of cluster. The number of cluster's centroid (k) will be provided by the user in the same manner like the traditional K-mean algorithm and the number of division will be k*k ('k' vertically as well as 'k' horizontally). If the highest frequency of data point is same in different segments and the upper bound of segment crosses the threshold 'k' then merging of different segments become mandatory and then take the highest k segment for calculating the initial centroid (seed point) of clusters. In this paper we also define a threshold distance for each cluster's centroid to compare the distance between data point and cluster's centroid with this threshold distance through which we can minimize the computational effort during calculation of distance between data point and cluster's centroid [3].

Through research on K-means algorithm of text clustering and semantic-based vector space model, a semantic based K-means text clustering model is proposed to solve the problem on high-dimensional and sparse characteristics of text data set. The model reduces the semantic loss of the text data and improves the quality of text clustering [4].

Ge Xiufeng and Xing Changzheng [5] presented a new clustering method: KMCP algorithm. By the use of chromosome retreading and focus operators, the algorithm has

higher accuracy and convergence speed. Prepare a comparative test program, and repeatedly running test program in the analysis of large amounts of data. General Statistics proves that KMCP algorithm presented in this paper is a feasible and efficient clustering algorithm. Shi Na et al [6] proposes an improved k-means algorithm in order to solve this question, requiring a simple data structure to store some information in every iteration, which is to be used in the next interation. The improved method avoids computing the distance of each data object to the cluster centers repeatly, saving the running time.

3. PROPOSED WORK

Suppose a dataset X ={x1, x2 ,..., xn }include n data points and a feature set F ={f1 , f2 ,..., fd} comprise d features that describe the characteristics of each data point. A data point x= (xi1, xi2,..,xij ,..., xid)and *x* is the value of *j* th dimension of the *i* th point. The k-means algorithm partitions n data points into k clusters where the number of clusters k is pre-decided by users. Suppose C={C ,C ,,C }be a set of k clusters and c={c,c,..,c}is the set of the k corresponding cluster centers. A cluster center and is the value of j th dimension of their i^{th} cluster. Actually, the cluster centers are virtual data points and updated based on the assignments in the first step. cl= (cl1 ,cl2 ,...clj ,..., cld)

In k-means algorithm, the dissimilarity measure is the distance between a data point and a cluster center. The term determines the cluster membership to the cluster. The data point assign to the cluster if the is $d(x, c)$ is minimal.

$$U = \begin{cases} 1 & \text{if } l = arg\min d(x_i, c_t), t = 1,...,k \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Where

$$\sum u_{il} = 1 \quad (2)$$

for $1 \le i \le n$

U is a n ×k matrix that records the point-cluster membership and U€{1,0}is an element in U that represents the membership of data point x with the l th cluster c. U=1 indicates that the data point x belongs to the cluster c . Otherwise, indicates the data point x does not belong to the cluster c . If the distance measurements between one point and two cluster centers are equal, the point is assigned to the cluster with the smaller cluster index number.

Minkowski distance, which is the most used dissimilarity measure, is evaluated by summing the differences between two data points in terms of all features. The distance between the point x and cluster center c is written as d(x,c) d(x,c) with the Minkowski distance is written as below:

$$d = (\sum \text{mod}((M_k - N_k) * r))^{\wedge}(1/r)) \quad (3)$$

Where M & N are data objects r= Parameter
After an iteration of assignment is done, all the data points are assigned into k clusters. Then compute the geometric center of each cluster. The center of l th cluster C l is as follows:

$$C = \sum ux \ / \ \sum u \quad (4)$$

for and 1≤i≤n and *1*≤l≤ k.

The objective of k-means algorithm is to minimize the sum of the dissimilarity between all data points and their corresponding cluster centers, which is shown as below:

$$\text{Min } E = \Sigma\Sigma\, u_{il}\, d(x_i, c_l) \quad (5)$$

In most cases, the k-means algorithm minimizes the following Mean Square Error (MSE) function based on Minkowski distance

$$\text{Min } E = \sum\sum \text{umod}(x - c)^{\wedge} 2 \quad (6)$$

The Enhanced K-means Clustering algorithm is described below in three steps:

Step1: Randomly choose k number of points as the initial centers of k clusters.

Step2: Generate k number of new clusters by assigning each point to its closest cluster center by (1).

Step3: Calculate new cluster centers by (4).

Keep repeating Step 2 and Step 3 until the cluster centers are stable or the Mean Square Error (MSE) function converges to a threshold value.

4. RESULTS AND DISCUSSION

In order to evaluate *EKMCA*, the algorithm is tested on a benchmark dataset, the network traffic data from the KDD Cup 1999 Dataset [16]. KDD Cup data set is usually used as a standard dataset to evaluate the performance of clustering algorithm. Network data set includes 100 samples and can be divided into many groups. The KDD dataset includes a wide variety of intrusions together with normal activities simulated in a military network environment. The simulated attacks fall in one of four major categories: DOS (denial of service), R2L (unauthorized access from remote machine), U2R (unauthorized access to local super user privilege) and Probing (surveillance and other probing). In addition, the proposed algorithm evaluates by using the recall and precision. Precision and recall defined as follows,

$$\text{Precision} = [TP / (TP+FN)] * 100 \quad (7)$$

$$\text{Recall} = [TP / (TP+FP)] * 100 \quad (8)$$

Table 1 shows the comparison of recall and precision values of the proposed algorithm with the existing algorithms. It is inferred that the proposed algorithm improves the performance of existing algorithms.

Table 1 Performance evaluation of the clustering algorithms

Datasets	KM (%)		EKM (%)	
	Recall	Precision	Recall	Precision
WWW	85.77	85.14	94.58	93.27
Mail	84.32	85.63	94.26	94.01

Database	81.46	82.51	87.35	88.29
Media	80.54	82.29	86.37	84.62
Game	79.37	80.25	83.08	82.11
Average	**82.292**	**83.164**	**89.128**	**88.46**

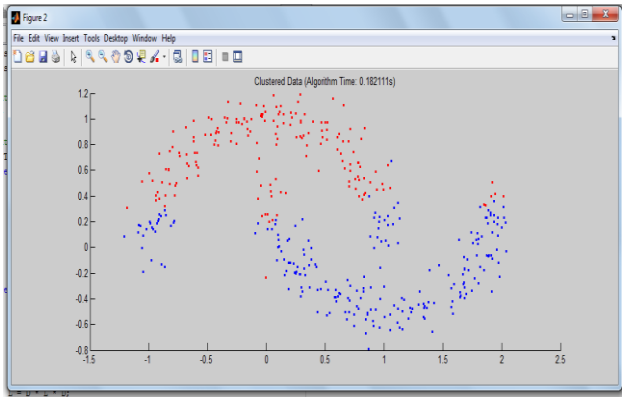

Figure 1 Clustered Data by Proposed Algorithm

Figure 1 shows the clustered data using the proposed Enhanced K-Means Clustering algorithm with time taken. From figure1, it can be observed that the proposed algorithm taken less number of milliseconds clustering the data efficiently.

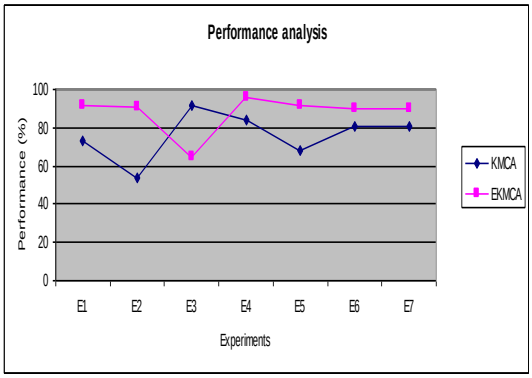

Figure 2 Performance analysis of proposed algorithm

Figure 2 shows the performance of the proposed algorithm for the same data set. From the five experiments conducted, it has been observed that the proposed EKMCA provides high performance when it is compared with K-Means Clustering algorithm. From the figure2, it can be observed that the proposed EKMCA provides better performance when it is compared with the existing methods.

Table 2 Time taken for clustering the random number of data

Experiments	Time taken for Clustering (sec)	
	KMCA	EKMCA
E1	1.25	0.09
E2	0.57	0.11
E3	0.50	0.11
E4	0.55	0.11
E5	0.55	0.09

Table 2 shows that the comparison of time taken for the random number of data clustering in existing K-Means clustering algorithm and proposed Enhanced K-Means Clustering Algorithm. From table 2 can observe that the proposed K-Means clustering algorithm taken less number of time for clustering the data.

5. CONCLUSION

In this paper, we proposed a new clustering algorithm that uses Enhanced K-means Clustering algorithm for achieve better performance in clustering the data. This proposed algorithm achieves the high performance when compared with K – means clustering algorithm which is used Euclidean distance measurement. In this same clustering algorithm provides high performance when we used the Minkowski distance measurement. The experimental results show that the proposed algorithm achieved high performance and less number of times only taken for clustering the data. The main advantage of this algorithm is that it helps to reduce the execution time.

REFERENCES

[1] Tingting Cui, Fangshi Li, "Weight Computing in Competitive K-Means Algorithm", IEEE, pp. 430-435, 2012.

[2] LI Han, "Using A Dynamic K-means Algorithm to Detect Anomaly Activities", Seventh International Conference on Computational Intelligence and Security, pp. 1049-1052, 2011.

[3] Ran Vijay Singh, M.P.S Bhatia, "Data Clustering with Modified K-means Algorithm", IEEE-International Conference on Recent Trends in Information Technology, pp.717-721, 2011.

[4] Yufang Liu, Shibin Xiao, Xueqiang Lv, Shuicai Shi, "Research on K-Means Text Clustering Algorithm Based on Semantic ", International Conference on Computing, Control and Industrial Engineering, pp.124-127, 2010.

[5] Ge Xiufeng, Xing Changzheng, "K-means Multiple Clustering Research Based on Pseudo Parallel Genetic Algorithm", International Forum on Information Technology and Applications, pp.30-33, 2010.

[6] Shi Na, Liu Xumin, Guan yong, "Research on k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics, pp.63-67, 2010.