Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.81 – 88

RESEARCH ARTICLE

A HYBRID APPROACH FOR DATA CLUSTERING USING DATA MINING TECHNIQUES

K.PRABHA[*]
Research Scholar in Computer Science
Vivekanandha College for Women,
Unjanai, Tiruchengode, India
prabha.bca2007@gmail.com

K.RAJESWARI, M.Sc., M.Phil,
Assistant Professor in Computer Science
Vivekanandha College for Women,
Unjanai, Tiruchengode, India

ABSTRACT

Data clustering is a process of arranging similar data into groups. Data clustering is a common technique for data analysis and is used in many fields, including data mining, pattern recognition and image analysis. In this paper a hybrid clustering algorithm based on K-mean is described. K-means clustering is a common and simple approach for data clustering but this method has some limitation such as local optimal convergence and initial point sensibility. The algorithm then extended to use k-means clustering to refined centroids and clusters. The experimental results showed the accuracy and capability of proposed algorithm to data clustering.

Keywords: Data clustering, K-means, Data mining, Hybrid algorithm

I. INTRODUCTION

Data mining is the important step for discover the knowledge in knowledge discovery process in data set. Data mining provide us useful pattern or model to discovering important and useful data from whole database. We used different algorithms to extract the valuable data.

Data Base Management System (DBMS) and Data Mining (DM) are two emerging technologies in this information world. Knowledge is obtained through the collection of

information. Information is enriched in today's business world. In order to maintain the information, a new systematic way has been used such as database. In this database, there are collection of data organized in the form of tuples and attributes. In order to obtain knowledge from a collection of data, business intelligence methods are used. Data Mining is the powerful new technology with great potential that help the business environments to focus on only the essential information in their data warehouse. Using the data mining technology, it is easy for decision making by improving the business intelligence.

Cluster Analysis is an effective method of analyzing and finding useful information in terms of grouping of objects from large amount of data. To group the data into clusters, many algorithms have been proposed such as k-means algorithm, Fuzzy C means, Evolutionary Algorithm and EM Method. These clustering algorithms groups the data into classes or clusters so that object within a cluster exhibit same similarity and dissimilar to other clusters. Thus based on the similar-ity and dissimilarity, the objects are grouped into clusters.

In this paper, a new hybrid algorithm is for business intelligence recommender system based on knowledge of users and frequent items. This algorithms works in three phases namely preprocessing, modelling and obtaining intelligence. First, the users are filtered based on thuser's Profile and knowledge such as needs and preferences defined in the form of rules. This poses selection of features and data reduction from dataset. Second, these filtered users are then clustered using k-means clustering algorithm as a modelling phase. Third, identifies nearest neighbour for active users and generates recommendations by finding most frequent items from identified cluster of users. This algorithm is experimentally tested with e-commerce application for better decision making by recommending top n products to the active users.

II. RELATED WORKS

Alexandre et al, presented a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. They analyzed the nature of privacy breaches and proposed a class of randomization operators that are much more effective than uniform randomization in limiting the breaches.

Jiaqi Wang et al, stated that Support vector machines (SVM) have been applied to build classifiers, which can help users make well-informed business decisions. The paper speeds up the response of SVM classifiers by reducing the number of support vectors. It was done by the Kmeans SVM (KMSVM) algorithm proposed in the paper. The KMSVM algorithm combines the K-means clustering technique with SVM and requires one more input parameter to be determined: the number of clusters.

M. H. Marghny et al, stated that Clustering analysis plays an important role in scientific research and commercial application. In the article, they proposed a technique to handle large scale data, which can select initial clustering center purposefully using Genetic algorithms (GAs), reduce the sensitivity to isolated point, avoid dissevering big cluster, and overcome

deflexion of data in some degree that caused by the disproportion in data partitioning owing to adoption of multi-sampling.

Ravindra Jain, explained that data clustering was a process of arranging similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group was better than among groups. In the paper a hybrid clustering algorithm based on K-mean and K-harmonic mean (KHM) was described. The result obtained from proposed hybrid algorithm was much better than the traditional K-mean & KHM algorithm.

David et al, described a clustering method for unsupervised classification of objects in large data sets. The new methodology combines the mixture likelihood approach with a sampling and sub sampling strategy in order to cluster large data sets efficiently. The method was quick and reliable and produces classifications comparable to previous work on these data using supervised clustering.

Hong Yu et al. performed comparative study on data mining for individual credit risk evaluation. The researcher referred Credit risk is referred to as the risk of loss when a debtor does not fulfill its debt contract and is of natural interest to practitioners in bank as well as to regulators.

Ji Dan et al. performed synthesized data mining algorithm based on clustering and decision tree. At present, they have accumulated abundant agriculture information data for the vast territory and diversity of crop resources. However, we just can visit a small quantity of data for lack of useful tools.

III. CLUSTERING ANALYSIS

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. Clustering is the process of partitioning a set of objects into a finite number of k clusters so that the objects within each cluster are similar, while objects in different clusters are dissimilar.

Cluster analysis is an effective tool in scientific or managerial inquiry. The k-means clustering method is one of the simplest unsupervised learning algorithms for solving the well-known clustering problem. The goal is to divide the data points in a data set into *K* clusters fixed a priori such that some metric relative to the centroids of the clusters (called the fitness function) is minimized. The algorithm consists of two stages: an initial stage and an iterative stage. The initial stage involves defining *K* initial centroids, one for each cluster. These centroids must be selected carefully because of differing initial centroids causes differing results. One policy for selecting the initial centroids is to place them as far as possible from each other. The second, iterative stage repeats the assignment of each data point to the nearest centroid and *K* new centroids are recalculated according to the new assignments. This iteration stops when a certain criterion is met, for example, when there is no further change in the assignment of the data points. Given a set of n data samples, suppose that we want to classify the data into *K* groups, the algorithm aims to minimize a fitness function, such as a squared error function defined as:

$$F = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \qquad (1)$$

where $||xi^{(j)}-cj||^2$ is a chosen distance measure between the i^{th} data point of the data sample $xi^{(j)}$ which was classified into the j^{th} group) and the j^{th} cluster center *cj, 1 = i= n, 1 = j= K* and is an indicator of the distance of the n data samples from their respective cluster centroids. The k-means clustering algorithm is summarized in the following steps :

• Place *K* points into the space represented by the objects that are being clustered. These points represent the initial group centroids.

• Assign each object to the group that has the closest centroid.

• When all objects have been assigned, recalculate the positions of the K centroids.

• Repeat steps 2 and 3 until a certain criterion is met, such as the centroids no longer moving or a preset number of iterations have been performed. This results in the separation of objects into groups for which the score of the fitness function is minimized.

IV. HYBRID ALGORITHM

In this business world, there exists a lot of information. It is necessary to maintain the information for decision making in business environment. The decision making consists of two kinds of data such as Online Analytical Processing (OLAP) and Online Transactional Processing (OLTP).The former contains historical data about the business from the beginning itself and the later contains only day-to-day transactions on business. Based on these two kinds of data, decision making process can be carried out by means of a new hybrid algorithm based on frequent itemsets mining and clustering using k-means algorithm and knowledge of users in order to improve the business intelligence.

Algorithm: Hybrid Algorithm

Input:

The number of clusters *k.*

Dataset D with *n* objects.

Output: A set of clusters Ck.

Begin

Identify the dataset $D = \sum (A) = \{a_1, a_2 \ldots a_n\}$ attributes/objects.

Outline the Consideration Column (CC) from D.

$CC = D` = \sum (A') \{a'_1, a'_2 \ldots a'_m\}$

Repeat

Formulate the rules for identifying the similar objects.

$\sum (R) = \sum(R) \bowtie \sigma (a_{ij} (A')/D')$, where i=1to n, j=1 to m.

S = f(X) / D, where S is the sample set containing identified column

FIS = value(S) > (SUP(X) and/or (SUP(X U Y) / SUP (X))),

Where FIS is the frequent itemsets identified.

C_n (a_{ij} (A`)) > T from FIS, where T specifies the threshold value.

Generate the Resultant Dataset, D``

Until no further partition is possible in CC.

Identify the *k* initial mean vectors (centroids) from the objects of D``.

Repeat

Compute the distance, □ between the object a_i and the centroids c_j.

Assign objects to cluster with min {∂ (c_{jk})} of all clusters

Recalculate the *k* new centroids c_j from the new cluster formed

Until reaching convergence

Find the nearest neighbour of active object

Generate recommendation from most frequent items of nearest neighbour

End

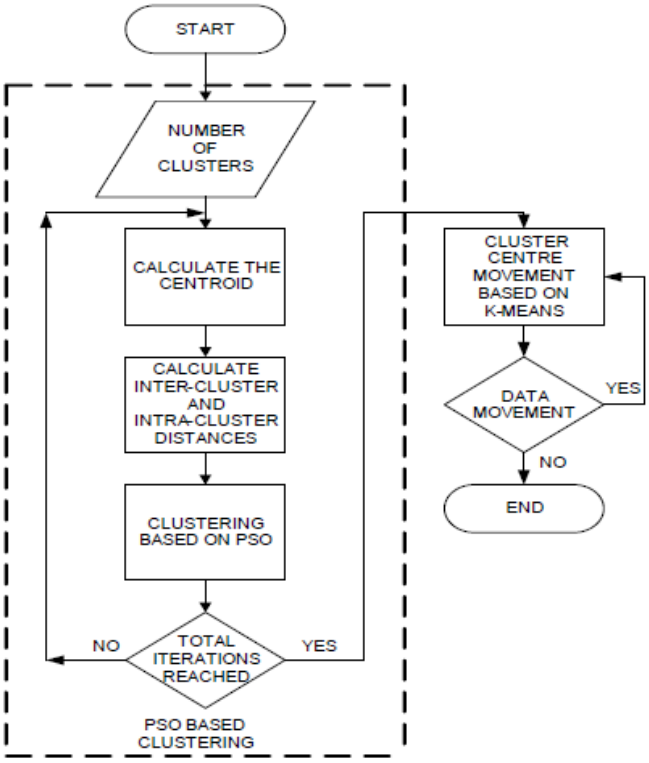

Figure.1. Flow chart of Hybrid Algorithm

V. PROPOSED WORK

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The result of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Steps of K-Means Clustering

K-means algorithm was developed by Macqueen which aims to find the cluster centers, (c1, . . . , cK), in order to minimize the sum of the squared distances (Distortion, D) of each data point (xi) to its nearest cluster centre (ck), as shown in Equation below where d is some distance function. Typically, d is chosen as the Euclidean distance. The steps of K-means algorithm are shown as follows:

(1) Initialize K centre locations ($c_1, \ldots, c_K$).

(2) Assign each x_i to its nearest cluster centre c_k.

(3) update each cluster centre ck as the mean of all x_i that have been assigned as closest to it.

(4) Calculate $D=\sum^{n}_{i=1}[min_{k=(1\ldots k)}\, d\,(xi, ci)]^{2}$

(5) If the value of D has converged, then return (c1, . . . , cK); else go to Step 2.

Main advantages:

1. K-means clustering is very Fast, robust and easily understandable. If the data set is well separated from each other data set, then it gives best results.
2. The clusters do not having overlapping character and are also non-hierarchical in nature.

Main disadvantages:

1. In this algorithm, complexity is more as compared to others
2. Need of predefined cluster centers.
3. Handling any of empty Clusters: One more problems with K-means clustering is that empty clusters are generated during execution, if in case no data points are allocated to a cluster under consideration during the assignment phase. The experimental results demonstrated that the proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

VI. EXPERIMENTAL RESULTS

Experiments have been performed on five data sets including Iris, WDBC, Sonar, Glass and Wine that were selected from standard data set UCI. Which the characteristics of each of them are described in the following:

Iris (fisher's iris plants database): This data set is according to the Iris flowers recognition that has three different classes and each class consists of 50 samples. Every sample has four attributes.

WDBC (Wisconsin diagnostic breast cancer): this data set is about breast cancer that is collected at the University of Wisconsin. That has two different classes including 357 and 212 samples. In this data set, each sample has 30 features.

Sonar: this data set is about sonar signals of submarine that totally has 208 samples. In this data set, Sonar signals divided in two classes including 111 and 97 samples with 60 features.

Glass (glass identification database): this data set is about several types of glass that has totally 214 samples in 6 classes. These classes are about building_windows_float_processed,

vehicle_windows_float_processe, containers‹, ableware, building_windows_non_float_processed and headlamps and each data has 9 attributes.

Wine (wine recognition data): This data set is regarding to drinks recognition that totally has 178 samples classified into three different classes including 59, 71 and 48 samples, respectively. In this data set, each sample has 13 attributes.

Table 1: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR IRIS DATA SET.

Algorithm	**Best**	**Mean**	**Std.Dev**
k-means	97.32	102.57	11.34

Table 2: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR WDBC DATA.

Algorithm	**Best**	**Mean**	**Std.Dev**
k-means	152647.25	179794.25	55222.17

Table 3: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR SONAR DATA.

Algorithm	**Best**	**Mean**	**Std.Dev**
k-means	234.77	235.06	0.15

Table4: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR GLASS DATA.

Algorithm	**Best**	**Mean**	**Std.Dev**
k-means	213.42	241.03	25.32

Table 5: COMPARISON OF INTRA-CLUSTER DISTANCE BETWEEN DIFFERENT METHODS FOR WINE DATA.

Algorithm	**Best**	**Mean**	**Std.Dev**
k-means	16555.68	17662.73	1878.07

VII. CONCLUSION

In this paper, a new hybridizes method based on k-means clustering method proposed to cluster data. In the proposed method, to find optimal cluster centers and then initialized the k-means algorithm with this centers to refine the centers. This method applies to 5 dataset. Experimental results for optimizing fitness function related to intra-cluster distance showed that the proposed obtained results that are relatively stable in different performance.

REFERENCES

[1] K. C. Wong and G. C. L. Li, "Simultaneous Pattern and Data Clustering for Pattern Cluster Analysis'', in IEEE Transaction on Knowledge and Data Engineering, Vol. 20, pp. 911-923, Los Angeles, USA, June 2008.

[2] D. W. van der Merwe and A. P. Engelbrecht, ''Data Clustering Using Particle Swarm Optimization'', in the 2003 Congress on Evolutionary Computation, Vol. 1, pp. 215-220, December 2003.

[3] Jiaqi Wang, Xindong Wu, Chengqi Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems", Int. J. Business Intelligence and Data Mining, Vol. 1, No. 1, 2005.

[4] Chonghui GUO, Li PENG, "A Hybrid Clustering Algorithm Based on Dimensional Reduction and K-Harmonic Means," IEEE 2008.

[5] Tsai C. Y. and Chiu C. C., 2008. Developing a feature weight self-adjustment mechanism for a K-Means clustering algorithm. *Computational Statistics and Data Analysis*, Vol. 52, pp. 4658-4672.

[6] Carlos Ordonez," Integrating K-Means Clustering with a Relational DBMS Using SQL", *IEEE Trans. Knowl. Data Eng.*, VOL. 18, NO. 2, PP. 188-201, 2006.

[7] Kao, Y.T., E. Zahara and I.W. Kao, 2008. A hybridized approach to data clustering, Expert Systems with Applications, 34 (3): 1754-1762.

[8] Ji Dan, Qiu Jianlin (2010) "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", 10th IEEE International CIT.

