ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Hybrid Approach of Data Mining clustering algorithms

Neetu Wadhwa*
Student of M.Tech (CSE)
Department of Computer Sc & Engg
Haryana College of Technology & Management,
Kaithal, India
its_neetu78@yahoo.co.in

Er. Alankrita Aggawal
Sr. Assistant Professor
Department of Computer Sc & Engg
Haryana College of Technology & Management,
Kaithal, India
alankrita.agg@gmail.com

Er. Anish Soni
Department of Computer Sc & Engg
Haryana College of Technology & Management,
Kaithal, India
soni_anish@yahoo.com

Abstract: Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets. It is a process of grouping data objects into disjoint clusters so that data in the same cluster are similar, and data belonging to different cluster are differ. Many algorithms have been developed for clustering. In this paper we are reviewing different clustering algorithms like K-Means , HAC , SOM and their comparison by applying hybrid approach on different datasets.

Keywords: Clustering Algorithms, Data Mininig, HAC, SOM, K-Means

I. INTRODUCTION

Recently many commercial data mining clustering techniques have been developed and their usage is increasing tremendously to achieve desired goal.

Data mining is a convenient way of extracting patterns, which represents knowledge implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. It can be viewed as an essential step in the process of knowledge discovery. Data are normally preprocessed through data cleaning, data integration, data selection, and data transformation and prepared for the mining task. Data mining can be performed on various types of databases and information repositories, but the kind of patterns to be found are specified by various data mining functionalities like class description, association, correlation analysis, classification, prediction, cluster analysis etc.

Clustering[3] is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets. It is a process of grouping data objects into disjoint clusters so that data in the same cluster are similar, and data belonging to different cluster are differ. Many algorithms have been developed for clustering.

A clustering algorithm[4] typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem. The simplest approach is dimension reduction techniques including principal component analysis (PCA)[7] and random projection. In these methods, dimension reduction is carried out as a preprocessing step.

Many diverse techniques have appeared in order to discover cohesive groups in large datasets.

A. *Basic Clustering Techniques:*

We distinguish two types of clustering techniques: *Partitional* and *Hierarchical*. Their definitions are as follows:

a. ***Partitional****:* Given a database of objects, a partitional clustering algorithm constructs partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the *sum of squared distance from the mean* within each cluster.

One of the issues with such algorithms is their high complexity, as some of them exhaustively enumerate all possible groupings and try to find the global optimum. Even for a small number of objects, the number of partitions is huge. That's why, common solutions start with an initial, usually random, partition and proceed with its refinement. A better practice would be to run the partitional algorithm for different sets of initial points (considered as representatives) and investigate whether all solutions lead to the same final partition.

Partitional Clustering algorithms try to locally improve a certain criterion. First, they compute the values of the similarity or distance, they order the results, and pick the one that optimizes the criterion.

Hence, the majority of them could be considered as greedy-like algorithms.

b. ***Hierarchical****:* Hierarchical algorithms create a hierarchical decomposition of the objects. They are either *agglomerative* (*bottom-up*) or *divisive* (*top-down*):

i. *Agglomerative* algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. The clustering may

stop when all objects are in a single group or at any other point the user wants.

These methods generally follow a greedy-like bottom-up merging.

ii. *Divisive* algorithms follow the opposite strategy. They start with one group of all objects and successively split groups into smaller ones, until each object falls in one cluster, or as desired.

Divisive approaches divide the data objects in disjoint groups at every step, and follow the same pattern until all objects fall into a separate cluster. This is similar to the approach followed by divide-and-conquer algorithms.

Most of the times, both approaches suffer from the fact that once a merge or a split is committed, it cannot be undone or refined.

Partitional and hierarchical methods can be integrated. This would mean that a result given by a hierarchical method can be improved via a partitional step, which refines the result via iterative relocation of points.

This paper is organized as follows. In Section 2, the concepts related to Data mining clustering techniques are introduced. In Section 3, we are reviewing different clustering algorithms used for data mining. In section 4, we are discussing hybrid approach . In section 5, we conclude the paper.

II. DATA MINING CLUSTERING TECHNIQUES

Apart from the two main categories of partitional and hierarchical clustering algorithms[2], many other methods have emerged in cluster analysis, and are mainly focused on specific problems or specific data sets available. These methods include :

A. Density-Based Clustering:

These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighbourhood of a data objects.

In these approaches a given cluster continues growing as long as the number of objects in the neighbourhood exceeds some parameter. This is considered to be different from the idea in partitional algorithms that use iterative relocation of points given a certain number of clusters.

B. Grid-Based Clustering:

The main focus of these algorithms is spatial data, *i.e.*, data that model the geometric structure of objects in space, their relationships, properties and operations. The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.

C. Model-Based Clustering:

These algorithms find good approximations of model parameters that best fit the data. They can be either partitional or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitionings. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

D. Categorical Data Clustering:

These algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied. In the literature, we find approaches close to both partitional and hierarchical methods.

For each category, there exists a plethora of sub-categories, *e.g.*, density-based clustering oriented towards geographical data, and algorithms for finding clusters. An exception to this is the class of categorical data approaches. Visualization of such data is not straightforward and there is no inherent geometrical structure in them, hence the approaches that have appeared in the literature mainly use concepts carried by the data, such as co-occurrences in tuples. On the other hand, categorical data sets are in abundance.

III. DIFFERENT CLUSTERING ALGORITHM USED FOR DATA MINING

K-means [9] is a prototype-based, simple partitional clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. The algorithm consist of two separate phases: the first phase is to select k centers randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, recalculating the average of the clusters.

This iterative process continues repeatedly until the criterion function becomes minimum. The k means algorithm works as follows:

a) Randomly select k data object from dataset D as initial cluster centers.
b) Repeat
a. Calculate the distance between each data object $d_i(1<=i<=n)$ and all k cluster centers $c_j(1<=j<=n)$ and assign data object d_i to the nearest cluster.
b. For each cluster $j(1<=j<=k)$, recalculate the cluster center.
c. Until no changing in the center of clusters.

The most widely used convergence criteria for the k-means algorithm is minimizing the SSE. The k-means algorithm always converges to a local minimum. The particular local minimum found depends on the starting cluster centroids. The k-means algorithm[12] updates cluster centroids till local minimum is found. Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset.

A. Hierarchal Agglomerative Clustering:

HAC[3] is a clustering method that produces "natural " groups of examples characterized by attributes. A tree, called dendrogram, where successive agglomerations are showed, starting from one example per cluster, until the whole dataset belong to one cluster, describes the clustering process.

a. Initialize the cluster set assuming each data point be a distinct cluster.
b. Compute the similarity between all pairs of clusters i.e evaluate the similarity matrix whose ij^{th} entry gives the similarity between the i^{th} and j^{th} clusters.
c. Merge the most similar (closest) two clusters.
d. Update the similarity matrix to reflect the pairwise similarity between the new cluster and the original (remaining clusters).
e. Repeat steps 3 and 4 untill only a single cluster remains

The main advantage of HAC is the user can guess the right partitioning by visualizing the tree, he usually prune the tree between nodes presenting an important variation. The main disadvantage is that requires the computation of distances between each example, which is very time consuming when the dataset size increases.

B. Self Organising Map Algorithm:

The SOM[5] is an algorithm used to visualize and interpret large high-dimensional data sets. Typical applications are visualization of process states or financial results by representing the central dependencies within the data on the map. The map consists of a regular grid of processing units, "neurons". A model of some multidimensional observation, eventually a vector consisting of features, is associated with each unit. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other.

a. Randomly choose an input vector x.
b. Determine the "winning" output node i, where w_i is the weight vector connecting the inputs to output node i. Note: the above equation is equivalent to $w_i \, x >= w_k \, x$ only if the weights are normalized.

$$\left|\omega_i - x\right| \le \left|\omega_k - x\right| \qquad \forall k$$

c. Given the winning node i, the weight update is

$$w_k(new) = w_k(old) + \mu \, \aleph(i,k)\,(x - w_k)$$

where $\aleph(i,k)$ is called the neighbourhood function that has value 1 when i=k and falls off with the distance $|r_k - r_i|$ between units i and k in the output array. Thus, units close to the winner as well as the winner itself, have their weights updated appreciably. Weights associated with far away output nodes do not change significantly. It is here that the toplogical information is supplied. Nearby units receive similar updates and thus end up responding to nearby input patterns.

IV. HYBRID APPROACH OF CLUSTERING TECHNIQUES

In the above section we have discussed various clustering techniques for data mining. But each technique has its advantage and disadvantages too. For example consider K-Means Algorithm[10] .

A. Advantages of k-Means Technique:

a. With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
b. K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

B. Disadvantages of K-Means Technique:

a. Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).
b. Fixed number of clusters can make it difficult to predict what K should be.
c. Does not work well with non-globular clusters.
d. Different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved.

C. Advantages of HAC Algorithm are:

a. The user can guess the right partitioning by visualizing the tree,
b. he usually prune the tree between nodes presenting an important variation.
c. It requires the computation of distances between each example, which is very time consuming when the dataset size increases.

D. Advantages of SOM Algorithm are:

SOM clustering is very useful in data visualization since the spacial representation of the grid, facilitated by its low dimensionality, reveals a great amount of information on the data.

E. Proposed Hybrid Algorithm:

There are two steps in the new algorithm:

a. First, a low-level clusters are built from fast clustering method such as K-MEANS, SOM;
b. HAC starts form these clusters and builds the dendogram.

V. CONCLUSION

In this paper, we dealt with algorithmic aspects of clustering in data mining. Mining clusters is one of the most used functions in data mining. clusters are of interest to both database researchers and data mining users. We have provided a survey of previous research in this area as well as provide a brief comparison of different.After studying various advantages and disadvantages of different algorithms we propsed a hybrid algorithm which will work on combination of two algorithms. In future, for mining the clusters different researchers are working on different algorithms and data structures in order to refine the database.

VI. REFERENCES

[1]. Tajunisha N., Saravanan V.,"An increased performance of clustering high dimensional data using Principal Component Analysis",Proceedings of the IEEE fist international

conference on integrated intelligent computing pp 17-21,(2010).

[2]. B.Chandra., "Hybrid Clustering Algorithm" , IEEE International Conference on Syatem, Man and cybernetics(Dec 2009).

[3]. Abu Abbas., "Comparison between Data Clustering algorithms" , International Arab Journal of IT Vol 5 (July 2008)

[4]. Rui Xu, "Survey of Clustering Algorithms", IEEE Transaction on Neural N/w (May 2005)

[5]. Sundareshan M.K, "Comparison of SOM with K-Means hierarichal clusetering for bioinformatic comparison" , IEEE international joint conference on Neural Network (2004)

[6]. Honda, K.Notsu, "Fuzzy PCA-Guided Robust K-Means Clustering" IEEE Transactions on fuzzy syatem(Feb 2010)

[7]. H.S Behara, "An Improved Hybridized K-Means Algorithm for high dimensional dataset and its performance analysis" , International Journal of Computer Sc & Enggineering(Mar 2010)

[8]. Bembiring, Zain & Embang, "Clustering high Dimensional data using subspace and projected clustering algorithms.", IJCSIT(Aug 2010).

[9]. Abdul Nazar, " Improving the accuracy and efficiency of the K-Means clustering algorithm" , Proceeding of world congress on engg Vol I (WCE 2009)

[10]. T. Kanungo, " An Efficient k-Means clustering algorithm: Analysis and Implementation", IEEE Transaction on Pattern Analysis and Machine Intelligence (July 2002).

[11]. Rajshree Dash, Debahuti Mishra, "A Hybridized K-Means clustering approach for high dimensional dataset", IJEST (2010)

[12]. Fahim A.M,Salem A.M, Torkey A and Ramadan M.A (2006) :An Efficient enchanced k-means clustering algorithm,Journal of Zhejiang University,10(7): 1626-1633,2006.

[13]. Fahim A.M,Salem A.M, Torkey F. A., Saake G and Ramadan M.A (2009): An Efficient k-means with good initial starting points, Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19,pp. 47-57.

[14]. Nazeer K. A., Abdul and Sebastian M.P. (2009): "Improving the accuracy and efficiency of the kmeans clustering algorithm", Proceedings of the World Congress on Engineering,Vol. 1, pp. 308-312.

[15]. Samarjeet Borah, Mrinal Kanti Ghose, "Performance Analysis of AIM-K-Means and K-means in Quality cluster generation," Journal of computing, Volume1, issue 1, December 2009.

[16]. Adam schenker, mark last, horst bunke, Abraham kandel: "Comparison of two noval algorithm for clustering using web documents, WDA", (2003)

[17]. Arthur D., vassilvitskii S.(2007): K-means++ the advantages of careful seeding, on discrete algorithms (SODA).

[18]. Jolliffe I.T. (2002): Principal Component Analysis, Springer, Second edition.

[19]. Babu G. and Murty M. "A near Optimal initial seed value selection in k-means algorithm using a genetic algorithm, Pattern Recognition Letters" Vol.14,1993, PP, 763-769. (2003)

[20]. Chris Ding and Xiaofeng He: " k-means Clustering via Principal component Analysis", In Proceedings of the 21st international conference on Machine Learning, Banff, Canada(2004).

