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. 2, Issue. 4, April 2013, pg.197 – 204

RESEARCH ARTICLE

AN ADAPTIVE PARTITIONAL CLUSTERING METHOD FOR CATEGORICAL ATTRIBUTE USING K-MEDOID

A. Selvakumar[1]

[1]Assistant Professor of Computer Science, Dept. of Computer Science, Erode, Tamil Nadu, India

[1] *deesel@rediffmail.com*

Abstract— partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The operation is needed in a number of data mining tasks such as unsupervised classification and data summation as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed. Clustering is a popular approach used to implement this operation. Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters. More specifically, they attempt to determine an integer number of partitions that optimize as certain criterion function. The criterion function may emphasize the local or global structure of the data and its optimization is an iterative procedure. The intention to analyze the fact that partitional clustering algorithms performs efficiently for numerical attribute rather than categorical attribute. To analyze the algorithm best suits for a matrix data. They work with larger datasets with many attributes. For analysis the Iris dataset has been retrieved from UCI data repository and used in K-Medoid. The outcome of the algorithm is the partition of clusters which can also be visualized in graphical format. The cluster figures differentiate the cluster in various colors with the centroid measure distinctly. Finally it has been determined that K-Medoid is the better partitional algorithm.

I. INTRODUCTION

The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time, the users of these data are expecting more sophisticated information from them. Simple structured and query language queries are not adequate to support these increased demands for information. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

1.1 METHODOLOGY OF DATA MINING

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored

transaction data based on open-ended user queries. Several types of analytical software available are statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought as given below,

Classes Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns Data are mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

1.2 LEARNING IN DATA MINING

To do research work in data mining it is the task of researcher to make the system to understand the concepts as a human being do. To pull off this chore in point of fact learning in data mining theater a vital role. Learning is operationally defined as " An individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances or at different circumstances can carried out". If a researcher wants to implement this learning for computer systems then the question arises whether it is possible? Before answering the question a user should first determine whether the computer systems learn something or not. Of course it learns information to certain extent but not civilly. In practice, to enable computer systems to carry out the task correctly a computer must be able to write programs by it based on examples. This leads to new definition,

Machine Learning - A sub discipline of computer science that deals with the design and implementation of learning algorithms.

Concept Learning - A learning algorithm is that which has a good quality of good learning algorithms is that learns consistent and complete definitions. A definition of concept is complete if it recognizes all the instances of a concept.

1.3 ARCHITECTURE OF DATA MINING

The architecture of data mining system has the following major components such as,

Database, data ware house or other information repository - This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories.

Database or data warehouse server - The database or data warehouse server is responsible for fetching the relevant data based on the users' data mining request.

Knowledge Base - This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge includes concept hierarchies, used to organize attributes values in to different levels of abstraction.

Data Mining Engine - This consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis and evolution and deviation analysis.

Pattern Evaluation Module - This component typically employs measures and interacts with the data mining modules to focus the search towards interesting patterns.

Graphical User interface - This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data based on the intermediate data mining results.

1.4 APPLICATIONS OF DATA MINING

The discipline of data mining is driven in part by new applications that are not currently supplied by today's technology. These new applications can be naturally divided in to three broad categories,

Business and E-Commerce Data

This is a major source category of data for data mining applications. Back-office, front office and network applications produce large amounts of data about business processes. Using this data for effective decision-making remains a fundamental challenge.

Scientific, Engineering and Health Care Data

Scientific data and metadata tend to be more complex in structure than business data. These are applied in fields such as,

GENOMIC DATA
SENSOR DATA
SIMULATION DATA
HEALTH CARE DATA

Web Data

The data on the web is growing not only in volume but also in complexity and web data now includes not only text, audio and video material, but also streaming data and numerical data. The manuscript keeps its eye on the state of the art of Clustering techniques in Data Mining. The dissertation mainly focuses on the use of Partitional algorithm analysis. The eventual aspire is to calculate the efficiency of Partitional clustering algorithms.

1.5 CLUSTERING ALGORITHM CATEGORIES

Clustering algorithms can be classified to

- ♣ The type of data input to the algorithm
- ♣ The clustering criterion defining the similarity between data points.
- ♣ The theory and fundamental concepts on which cluster analysis techniques are based.

The algorithms can be classified into the following types,

1.5.1 Partitional Clustering

Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters. More specifically, they attempt to determine an integer number of partitions that optimize as certain criterion function. The criterion function may emphasize the local or global structure of the data and its optimization is an iterative procedure. Given a database of n objects or data tuples, a partitioning method constructs 'k' of the data, where each partition represents a cluster and k<= n. Thus it classifies the data into k groups, which together satisfy the following requirements

a. Each object must belong exactly one object and
b. Each object must belong to one group.

The applications adopt one of two popular heuristic methods,

i. k-Means algorithm, where each cluster is represented by the mean value of the objects in the cluster and

ii. k- Medoid algorithm, where each cluster is represented by one of the objects located near the center of the cluster. These clustering methods work well for finding spherical shaped clusters in small to medium sized databases.

1.5.2. Hierarchical Clustering

The hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of this algorithm is a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at desired level, a clustering of the data items into disjoint groups is obtained. A hierarchical method can be either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, called the bottom-up-approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all the groups are merged in to one. The divisive approach, also called the top-down approach, starts with all objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. There are two approaches to improve the quality of hierarchical clustering are,

- ✓ Perform careful analysis of object "linkages" at each hierarchical partitioning, such as CURE and CHAMELEON.
- ✓ Integrate hierarchical agglomeration and iterative relocation by first using a hierarchical agglomerative algorithm and refining the result using iterative relocation using BIRCH.

1.5.3 Density Based Clustering

The key idea of this type of clustering is to group neighboring objects of a data set into clusters based on density conditions. DBSCAN is a typical density-based method that grows cluster according to density threshold. OPTICS is a density-based method that computes an augmented clustering ordering for automatic and interactive cluster analysis.

1.5.4 Grid Based Clustering

This type of algorithm is mainly proposed for spatial data mining. Their main characteristics that they quantize the space into finite number of cells and then they do all operations on the quantize space. STING is a typical example of Grid –based method. CLIQUE and Wave cluster are two clustering algorithms.

II. LITERATURE SURVEY

Jain et al. [99] Clustering algorithms can be categorized into hierarchical clustering methods, partitioning clustering methods, density-based clustering methods, grid-based clustering methods, and model-based clustering methods.

Jain et al. [00] an excellent survey of clustering techniques for the statistical pattern recognition perspective, define cluster analysis as the organization of collection of patterns into clusters based on similarity.

Hastie et al. [01] define the goal of cluster analysis from his statistical perspective as a task to partition the observations into groups such that the pair wise dissimilarities between those assigned to the same cluster tend to be smaller than those in different clusters.

Tan et al. [02] states from data mining point of view that ”Cluster analysis divides data into groups that are meaningful, useful, or both.”. By meaningful they refer to clusters that capture the natural structure of a data set, whereas the useful clusters serve only as an initial setting for some other method, such as PCA (principal component analysis) or regression methods.

Kaufman et al. [03] propose six clustering algorithms (PAM, CLARA, FANNY, AGNES, DIANA and MONA) that they believe to cover a major part of the applications.

Anderberg et al., [03] states that there appears to be at least nine major elements in a cluster analysis study before the final results can be attained. Because the current real-world data sets contain missing values.

Dhillon et al., [04] describes Enhanced Word Clustering for Hierarchical Text Classification.

Sudipto Guha et al., [05] an efficient clustering algorithm for large CURE is a hierarchical clustering algorithm, that employs the features of both the centroid based algorithms and the all point algorithms.

Sanjay Goil, et al., [06] Efficient and Scalable Clustering for very large data sets.

Jain et al. [07] suggest that the strategies used in data collection, data representation, normalization and cluster validity is as important as the clustering strategy itself.

Richard O. Duda et al.,[08] Clustering algorithms developed in the literature can be classified into Partitional clustering and hierarchical clustering.

Anil K. Jain et al., [09] Hierarchical clustering algorithms, too, may be unsuitable for clustering data sets containing categorical attributes. For instance, consider the centroid-based agglomerative hierarchical clustering algorithm.

Eui-Hong Han [10] the authors address the problem of clustering related customer transactions in a market basket database. Frequent item sets used to generate association rules are used to construct a weighted hyper graph.

George Karypis et al., [11] a hyper graph partitioning algorithm from [KAKS97] is used to partition the items such that the sum of the weights of hyper edges that are cut due to the partitioning is minimized.

III. METHODOLOGY

3.1 EXISTING METHODOLOGY

3.1.1 CLUSTERING ALGORITHMS

There exist a large number of clustering algorithms in the literature. The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. Clustering can be performed on both numerical data and categorical data. For the clustering of numerical data, the inherent geometric properties can be used to define the distance between the points. But for clustering of categorical data, such a criterion does not exist and many data sets also consist of categorical attributes, on which distance functions are not naturally.

3.1.2 PARTITIONAL ALGORITHM

In this category, K-Means is commonly used algorithm. The aim of K-Means clustering is the optimization of an objective function that is described by the equation

$$\mathbf{E} = \sum_{i=1}^{c} \sum \mathbf{d(x,m_i)}$$

In the above equation, m_i is the center of cluster c_i while $d(x,m_i)$ is the Euclidean distance between point x and m_i . Thus, the criterion function E attempts to minimize the distance of each point from the center of the cluster to which the point belongs. More specifically, the algorithms begin by initializing a set of c cluster centers. Then, it assigns each object of the data set to the cluster whose center is the nearest, and re-computes the centers. The process continues until the centers of the clusters stop changing.

3.1.3 K-MEDIODS METHOD

The K-Means algorithms are sensitive to outliers since an object with an extremely large value may substantially distort the distribution of the data. The basic strategy of k-mediods clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoid, which is the most centrally located object in a cluster) for each cluster. Each remaining object is clustered with the medoid to which is the most similar. The strategy then iteratively replaces one of the mediods by one of the non-mediods as long as the quality of the resulting clustering is improved. This quality is estimated using a cost function that measures the average dissimilarity between an object and the medoid of its cluster.

3.1.4 PAM (PARTITIONING AROUND MEDIODS)

PAM was one of the first k-mediods algorithms introduced. It attempts to determine k partitions for n objects. After an initial random selection of k-mediods, the algorithm repeatedly tries to make a better choice of mediods. All of the possible pairs of objects are analyzed, where one object in each pair is considered a medoid and the other is not. The quality of the resulting clustering is calculated for each such combination. An object, o_j, is replaced with the object causing the greatest reduction in square error. The set of objects for each cluster in one iteration forms the mediods for the next iteration.

3.1.5 CLARA

The PAM works in effect for small data sets, but does not scale well for large data sets. To deal with larger data sets, a sampling based method; called CLARA (Clustering Large Applications) can be used. Instead of taking the whole set of data into consideration, a small portion of the actual data is chosen as representative of the data. Mediods are then from this sample using PAM. If the sample is selected in a fairly random manner, it should closely represent the original data set. The representative objects (mediods) will likely be similar to those that would have been chosen from the whole data set. CLARA draws multiple samples of the data set, applies PAM on each sample, and returns its best clustering as the output.

3.1.6 CLARANS (Clustering Large Applications based upon Randomized Search)

CLARANS was proposed that combines the sampling technique with PAM. However, unlike CLARA, CLARANS does not confine itself to any sample at any given time. CLARANS draws a sample with some randomness in each step of the search. The clustering process can be presented as searching a graph where every node is potential solution, that is, a set of k-mediods. The clustering obtained after replacing a single medoid is called the neighbour of the current clustering. The number of neighbours to be randomly tried is restricted by a user specified parameter. If a better neighbour is found, CLARANS moves to the neighbours node and the process starts again; otherwise the current clustering produces a local optimum. If the local optimum is found, CLARANS starts with new randomly selected nodes in search for a new local optimum. CLARANS has been experimentally shown to be more effective than both PAM and CLARA. It can be used to find the most "natural" number of clusters using a sithouette coefficient - a property of an object that specifies how much the object truly belongs to the cluster. The performance of CLARANS can be further improved by exploring spatial data structures, such as R^* -trees, and some focusing techniques.

3.2 COMPARISON OF CLUSTERING ALGORITHMS

Clustering is broadly recognized as a useful tool in many applications. Researchers of many disciplines have addressed the clustering problem. However, it is a difficult problem, which combines concepts of diverse scientific fields (such as database, machine learning, pattern recognition, and statistics). Thus, the differences in assumptions and context among different research Communities caused a number of clustering methodologies and algorithms to be defined. This slice offers an overview of the main characteristics of the clustering algorithms presented in a comparative way. Let us first consider the algorithms categorized in four groups based on their clustering method: Partitional, hierarchical, density-based and grid-based algorithms. More specifically comparison is based on the following features of the algorithms.

- ✓ The type of the data that the algorithm supports (numerical, categorical)
- ✓ The shape of the clusters
- ✓ Ability to handle noise
- ✓ Input parameters

The opening consortium of the clustering algorithms is the Partitional algorithm. Partitional algorithms are applicable mainly to numerical data sets. However there are some variants of K-Means such as K-mode, which handle categorical data. K-Mode is based on K-Means to discover clusters while it adopts new concepts in order to handle categorical data. Thus the cluster centers are replaced with "modes", a new dissimilarity measure used to deal with categorical objects. Another characteristics of Partitional algorithms is that they are unable to handle noised and they are not suitable to discover clusters with non-convex shapes. Moreover they are based on certain assumptions to partition a data set. Thus, they need to specify the number of clusters in advance except for CLARANS, which needs as input the maximum number of neighbours of a node as well as the number of local minima that will be found in order to partitioning of a dataset. The result of clustering proceeds in the set of representative points of the discovered clusters. These points may be the centers or the mediods (most centrally located object within a cluster) of the clusters depending on the algorithm.

A summarized view of the characteristics of hierarchical clustering methods is offered as subsequent kind in the manuscript. The algorithms of this category create a hierarchical decomposition of the database presented as the dendrogram. They are more efficient in handling noise than Partitional algorithms. However they break down due to their non-linear time complexity and huge I/O cost when the number of input data points is large. BIRCH tackles the problem using a hierarchical data structure called CF-tree for multiphase clustering. In BIRCH, a single scan of the dataset yields a good clustering and one or more additional scans can be used to improve the quality further. However, it handles only numerical data and it is order-sensitive. Also, BIRCH does not perform well when the clusters do not have uniform size and shape since it uses only the centroid of a cluster when redistributing the data points in the final phase. On the other hand, CURE employs a combination of random sampling and partitioning to handle large databases. It identifies clusters having non-spherical shapes and wide variances in size by representing each cluster by multiple points. Selecting well-scattered points from the cluster by multiple points generates the representative points of a cluster and shrinks them towards the center of the cluster by a specified fraction.

ROCK is a representative hierarchical clustering algorithm for categorical data. It introduces a novel concept called "link" in order to measure the similarity/complexity between a pair of data points. Thus, the ROCK clustering methods extends to non-metric similarity measures that are relevant to categorical data sets. It also exhibits good scalability properties in comparison with the traditional algorithms employing techniques of random sampling. Moreover, it seems to handle successfully data sets with significant differences in the size of the cluster. The density based clustering algorithms handle arbitrary shaped collections of points (like spiral, cylindrical etc.) as well as clusters of different sizes. Moreover these algorithms detect outliers or noises effectively. Two widely known algorithms of this category are DBSCAN and DENCLUE. DBSCAN requires the user to specify the radius of a neighborhood of a point and the minimum number of points in a neighborhood. Similarly DENCLUE requires careful selection of input parameters σ and ε, since such parameters may influence the quality of clustering results. However the major advantages of DENCLUE in comparison with other clustering algorithms are

i) It has a solid mathematical foundation and generalized other clustering methods such as Partitional and hierarchical.
ii) It has good clustering properties for data sets with large amount of noise
iii) It allows a compact mathematical description of arbitrary shaped clusters in high dimensional datasets
iv) It uses grid cells and only keeps information about the cells that actually contain points.

In general terms the complexity of density based algorithms is O (n logn). These types of algorithms do not perform any sorting on sampling, and thus they involve substantial I/O costs. And these algorithms fail to use random sampling to reduce input size, unless sample's size is large. This is because there may be substantial difference between the density in the sample's cluster and the clusters in the whole dataset. The last category refers to grid-based algorithms. The basic concept of these algorithms is that they define a grid for the data space and then do all the operations on the quantized space. These approaches are very efficient for large databases and are capable of finding arbitrary shape clusters and handling outliers. STING is one of the well-known grid-based algorithms. It divides the spatial area into rectangular cells while it stores the statistical parameters of the numerical features of the objects within cells. The grid structure facilitates parallel processing and incremental updating. Since STING goes through the database once to compute the statistical parameters of the cells, it is generally efficient method for generating clusters. Its time complexity is O(n). Sting does not consider spatial relationship between the children and their neighboring cells to construct the parent cell. The result is that all cluster boundaries are either horizontal or vertical and thus the quality of clusters is questionable.

3.3 CLUSTER VALIDITY

The procedure of evaluating the results of clustering algorithm is known under the term cluster validity. There are three approaches to investigate cluster validity.

i) External criteria – based on a predefined structure, which is imposed on a dataset and reflects our intuition about the clustering structure of the data set.
ii) Internal criteria - evaluation of results in terms of quantities that involve the vectors of the data set themselves.
iii) Relative criteria – Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameters.

3.4 PROPOSED SYSTEM

Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The operation is needed in a number of data mining tasks such as unsupervised classification and data summation as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed. Clustering is a popular approach used to implement this operation. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A *Cluster* is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in another cluster. Cluster analysis is an important human activity. Early in childhood one learns how to distinguish between cats and dog or between animals and plants by continuously improving subconscious clustering schemes. As a branch of statistics cluster analysis has been studied extensively for many years focusing mainly on distance based cluster analysis.

3.4.1 Simple K- Means Algorithm

The general idea is to start with a randomly chosen cluster points, then to reassign points so as to give the greatest increase (or decrease) in the score function, then to recalculate the updated cluster centers, to reassign points again , and so forth until there is no change in the score function or in the cluster memberships. The greedy approach has the virtue of being simple and guaranteeing at least a local a maximum (minimum) of the score function. Of course it suffers the usual drawback of greedy search algorithms in that we do not know how good the clustering c that it converges to is relative to the best possible clustering of the data (the global optimum for the score function being used). The number K of clusters is fixed before the algorithm is run (this is typical of many clustering algorithms). There are several variants of Simple K-Means algorithm. The basic version begins by randomly picking K cluster centers, assigning each point to the cluster whose mean is closest in the Euclidean distance sense, and then computing the mean vectors of the points assigned to each cluster, and using these as new centers iterative approach.

3.4.2 PROPOSED ALGORITHM

As an algorithm, the method is as follows: assuming we have n data points $D = \{x_1 \ldots\ldots, x_n\}$, our task is to find K clusters $\{C_1 \ldots\ldots C_k\}$;

Step1 : Input the number of clusters k and a database containing n objects.
Step2 : set of k clusters that minimizes the squared – error criterion.
Step3 : for K =l, ….K let r(K) be a randomly chosen point from D
while changes in clusters C_k happen do form clusters;
for k = 1,……k
do
$C_k = \{x \in D \ \ d(r_k,x) \leq d(r_j, x)$ for all j = l,…,K, $j \neq k\}$;
end.
Step4 : Compute new cluster centers,
r_k = the vector mean of the points c_k
Step5 : repeat step1.

3.4.3K-Medoid Algorithm

The k-medoid algorithm is a clustering algorithm related to the Simple K-Means algorithm and the medoid shift algorithm. Both the Simple K-Means and k-medoids algorithms are Partitional (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the Simple K-Means algorithm k-medoids chooses data points as centers. k-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the silhouette. It is more robust to noise and outliers as compared to Simple K-Means. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

3.4.4 PROPOSED ALGORITHM

Step1 : Initialize randomly select k of the n data points as the mediods
Step2 : Associate each data point to the closest medoid.
Step3 : for each mediod m
if each non-mediod data point o
swap m and o and compute the total cost of the configuration
Step4 : Select the configuration with the lowest cost.
Step5 : Repeat steps 2 to 5 until there is no change in the medoid.

IV. CONCLUSION

Data Mining has proved to be useful tool in exploring and discovering interesting and useful knowledge in very large datasets. However, as the datasets get larger the tools being used generally produce significantly more complex models. Successful applications of data mining are increasingly appearing as diverse applications. These are driven mainly by a glut in databases that have grown to surpass raw human processing ability. Driving the growth of this field are strong forces that are a product of the overloaded phenomenon. The thesis is designed to study about various clustering techniques in data mining. Cluster analysis is one of the major tasks in various research areas. The cluster aims at identifying and extracting significant groups in underlying data. Based on a certain criterion the data are grouped so that the data points in a cluster are more similar to each other than points in different clusters. Finally the clustering algorithms Simple k-means and K-Medoid for numerical data are used to calculate accurate clustering of transactional data has many potential applications in real industry, e-commerce intelligence etc. The effective clustering of transactional databases is extremely difficult because of the high dimensionality, sparsity and huge volumes often characterizing the databases. Distance based approaches like Simple k-means and K-Medoid are effective for low dimensional numerical data.

REFERENCES

[1] Adriaans.P and Zantige.D(1996), Data Mining, UK, Addision Wesley.
[2] Arun.K.Pujari , Data Mining Techniques, , Universities Press
[3] C.Apte ,Data Mining – An Industrial Research Perspective, IEEE Computational Science and Engineering April – June 1997.
[4] Catherine Bounsaythip and Esa Rinta- Runsala , Overview of Data Mining for Customer Behavior Modelling, Research Report TTE1 - 2001-18.
[5] David Hand, Heikki Mannila, Padhraic Symth, Principles of Data Mining, PHI 2004.
[6] Dharmendra S. Modha and W. Scott Spangler, Feature Weighting in K-means Clustering, Machine Learning,Vol.52, No.3, PP.217 - 237, 2003.
[7] George Karypis, Eui-Hong(Sam) Han Vipin Kumar, Chameleon: A Hierarchial Clustering Algorithm Using Dynamic Modeling ,Technical Report #99-007.
[8] Graham J.Williams and Zhexue Huang, (1997) Mining the knowledge Mine: The hot spot methodology for mining Large Real World databases, Advanced Topics in Artificial Intelligence, Lecture notes in Artificial Intelligence, Vol, 1342, pp340-348, Springer, Verlag,1997.
[9] Maria Halkidi, Yannis Batistakis,Michalis Vazirgiannis,On Clustering Validation Techniques.
[10] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Harcourt India Private Limited.
[11] Padhraic Smyth, Data mining : Data Analysis on a grand scale,Technical Report Kdnuggets. News 1:25,
[12] Pavel Berkhin, Survey of Clustering Data Mining Techniques.
[13] Periklis Andristos, Data Clustering Techniques,Technical Report, CSRG–443.
[14] Remzi Salih Ibrahim, Data Mining of Machine Learning Performance Data, Thesis report in 1999.
[15] Tian Zhuang, Raghu Ramakrishnan ,Miron Livny, BIRCH: An Efficient Data Clustering Method for very large Databases.

