Gathering Large Databases Using EK-Means & PCA Approach

R. Maruthaveni
Assistant Professor
Department of Computer Science

S. Sangamithra
Research Scholar
Department of Computer Science

Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore

ABSTRACT

Discovery of such clusters of data is must in illuminating main links in categorical data regulatory networks. There are lot of complications exists in the earlier clustering methods particularly while data clustering with mixed data types. This experiment analyzes those previous approaches and proposes with the new method for clustering the mixed data items. There are lot of methods occur for clustering the parallel data type, whereas only very few methods exists for clustering mixed data items and it leads to the need of better clustering technique for classification of mixed data. In this paper, we propose an innovative data clustering method for mining enormous databases. This proposed PCA with EK-means method results better performance better than the earlier method.

Keywords

Data clustering, Neural Network (NN), Self Organizing Map, EK-means, PCA, Iris data set.

1. INTRODUCTION

In KDD, a basis data mining technique used is data clustering. Most widely used methods in data mining is clustering and its core is grouping the whole data based on its parallel measures that based on some distance measure. The clustering problem has become more important in recent years. Data mining offers a suite of algorithms, each linking a different task and in the process elucidating a unique facet of the data. Among all the facets of data mining, we are specifically interested in clustering problems, i.e. the process of finding resemblance in the data and then clustering similar data into identifiable clusters.

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more resemble to each other than to those in other groups. Most techniques of clustering contain document grouping, scientific data analysis and survey segmentation. Usually, clustering take in the classification of the yielding data that includes n points in m dimension into k clusters. The clustering must be such that the data in the corresponding cluster that should be highly resemble to one another. The clustering problem has been focused in many perspectives and by researchers in many areas; this imitate its broad appeal and usefulness as one of the process in exploratory data analysis.

It should be highlighted that this research's aim is not to find an ideal clustering for the data but to get good understanding into the data cluster structure for data mining needs and hence, the clustering method should be quick, robust, and effective.

It is a important task of experimental data mining, and a usual method for mathematical data study, used in many areas, such as machine learning, image analysis, information retrieval and pattern recognition. Our proposed approach aims to works well with the enormous data set with high dimensional data.

2. REVIEW OF LITERATURE

Efficient collaborative method for mixed numeric and categorical data is recommended by earlier researchers. Most of the existing clustering algorithms focus on numerical data whose inbuilt geometric characteristics can be put-upon obviously to outline distance functions between data points and a massive or enormous data exist in the databases is categorical, where attribute values will not be reasonably arranged as numerical values. Since the differences in the features of the categories of data, force to build up standard functions for mixed data was not successful.

The problems exists in the clustering techniques are: Identifying the similar measure to find the similarity among various data, it is complex to find out the suitable methods for identifying the identical data in unsupervised way and originate a description that can differentiate the data of a cluster in an efficient manner.

In previous method NN-SOM method is applied to the issue. In the beginning, the real large database is segmented with different sub-datasets and next, accessible well predictable clustering method developed for various types of datasets is applied to produce equivalent clusters. Finally, the clustering results on the categorical and numeric dataset are combined as a enormous dataset, on which that clustering method is used to produce the final result. The vital contribution in this experimental study is to present an algorithm for the mixed features clustering complications, in which previous clustering algorithms like ant colony as shown in fig 1. can be effortlessly incorporated.

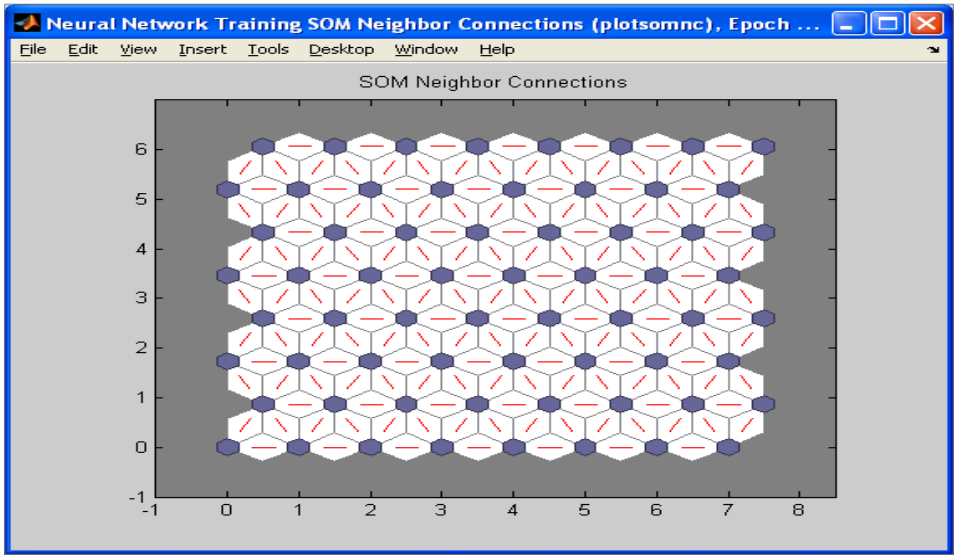

Fig 1:SOM based Neural Network

Ant colony clustering is a popular basic method for enhancement to different modifications. It depends on object function input from swarm intelligence. To analyze the algorithm result is to compare resemblance between objects while ants move continuously. It takes more time to cluster similar objects into one cell and partition different objects into different cells.

The majority of existing clustering algorithms encounter serious scalability and/or accuracy related problems when used on databases with a large number of records and/or attributes. This NN-SOM method detects all possible ways and find out the shortest path to measure the distance which takes more time to find the distance. The clustering of this dataset aims at finding of heavy linked groups of data points which can be divided in such a way that they can be used further. But with vectors, it is impossible to group the similarities even the distances between the points. As a result, to solve the most clustering problems NN-SOM algorithm has to be modified. The performance of the NN-SOM is not sensitive to the accurate points of the neighborhoods.

3. METHODOLOGY

The objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy. In existing, when the data sets have cluster sets size is increased, information cannot be retrieved by user efficiently. But, in proposed data set have cluster set is increased means the clustering process is done very quickly and the information will be retrieved very fast by user.

Different measures have been defined to measure the distance between two features. Euclidean distance has been widely applied on the numerical data. However, it is not applicable for categorical data. For any pair of categorical features, the data can be displayed as a contingency table, a table whose rows are labeled by the values of one categorical feature, whose columns are labeled by the values of the other categorical feature, and whose entries are nonnegative integers giving the number of observed events for each combination of row and column.

The problem with all the above mentioned algorithms is that they mostly deal with numerical data sets that are those databases having attributes with numeric domains. The basic reason for dealing with numerical attributes is that these are very easy to handle and also it is easy to define resemblance on them. But large dataset have multi-valued attributes.

The main goal of ensembles has been to improve the accuracy and robustness of a given classification or regression task, and spectacular improvements have been obtained for a wide variety of data sets.

3.1 EK-means

EK-means(Extended K-means) is simplest unsupervised learning algorithms which solve the clustering complications This approach is a simple and easy way to classify a given data set through a number of clusters.

The main idea is to define k centroids, one for each cluster because different location causes different result. The next process is to take each node related to a given data set and link it to the nearest centroids. EK-means is for dividing the sample data set into k clusters so as to reduce the sum of the squared distances to the cluster centers.

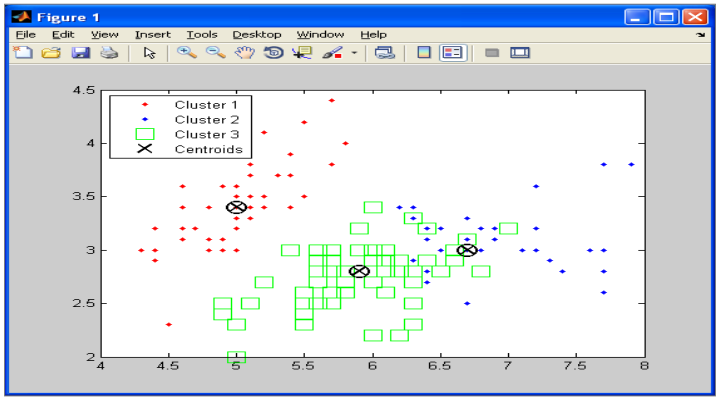

Fig 2. K means based data clustering:

3.2 PCA Method

Principal component analysis (PCA) [4][5] is a essential method in the framework of the various analysis methods. It is successfully used in many areas such as pattern recognition, process monitoring, data mining and feature extraction and image processing. It is due to reason of its easiness and capability in processing massive amount of process data, PCA is identified as a dominant tool of statistical process care and widely used in the area for fault detection. The advantage of PCA is that once it have found these nearby patterns in the data, then it can be compressed with the data, ie. by minimizing the number of sizes, with no information loss. Principal Component Analysis is a way of analyzing related data patterns and expressing the data with their resemblance and variance.

3.3 IRIS data set

In our approach we are performing incremental clustering on IRIS dataset for optimal clusters apart from the traditional approaches. The use of this data set in cluster analysis however is uncommon, since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used.

5.1	2.5	3.0	1.1
5.7	2.8	4.1	1.3
6.3	3.3	6.0	2.5
5.8	2.7	5.1	1.9
7.1	3.0	5.9	2.1
6.3	2.9	5.6	1.8
6.5	3.0	5.8	2.2
7.6	3.0	6.6	2.1
4.9	2.5	4.5	1.7
7.3	2.9	6.3	1.8
6.7	2.5	5.8	1.8
7.2	3.6	6.1	2.5
6.5	3.2	5.1	2.0
6.4	2.7	5.3	1.9
6.8	3.0	5.5	2.1
5.7	2.5	5.0	2.0
5.8	2.8	5.1	2.4
6.4	3.2	5.3	2.3
6.5	3.0	5.5	1.8
7.7	3.8	6.7	2.2
7.7	2.6	6.9	2.3
6.0	2.2	5.0	1.5
6.9	3.2	5.7	2.3
5.6	2.8	4.9	2.0
7.7	2.8	6.7	2.0
6.3	2.7	4.9	1.8
6.7	3.3	5.7	2.1
7.2	3.2	6.0	1.8
6.2	2.8	4.8	1.8
6.1	3.0	4.9	1.8

Fig 3. Data set (iris data

This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.

3.4 EXPERIMENT RESULTS

The investigation study of the proposed technique was accepted with the support of Iris Data Set. The number of clusters answered for the proposed method is lesser when compared to the other approaches and the outliers were not found by using the proposed method.

Owing to Iris data set with 2 attribute data which composed of the width of petal, and the length of petal. This experiment made data reduction with the Principal Component Analysis approach which was the implement to investigate high dimensional data.

The data consists of 150 samples from various types of iris flower as like 150 samples are taken for experiment. In each sample, two features were measured: the length and width of the petal, in centimeters. Due to this data clusters reasonably well into 3 clusters, it has become a prominent aspect of the power of a clustering algorithm.

The Proposed algorithm utilizes an active sampling mechanism and compels at most a single examine through the data. The capability to generalize our identifications to other instances is reduced in many approaches. EK-means method used to cluster the similarity data identified in the operational definition of length and width of the petal used in this study as in fig 4.

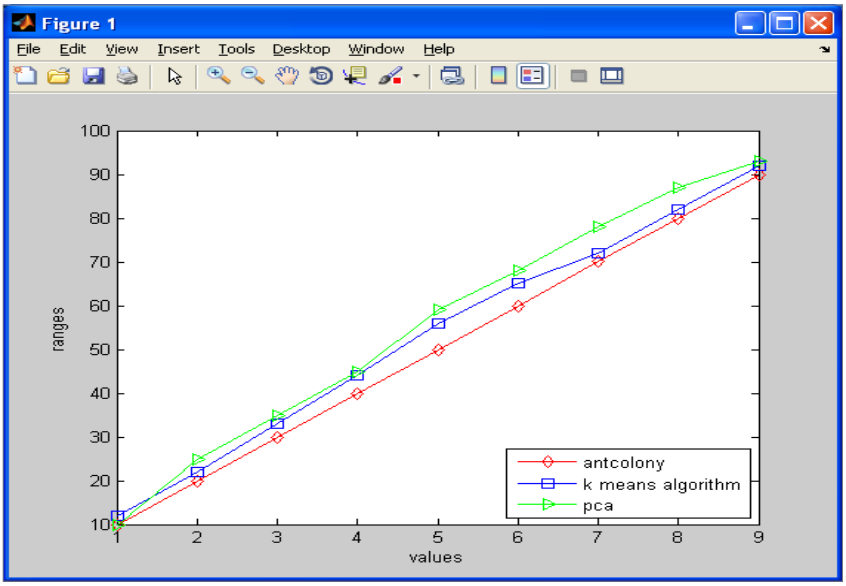

Fig 4.Performance results for the Iris data set.

We determine the high quality of the clustering solutions that was found, their explanatory power, and proposed model's good scalability. PCA method has good accuracy, uses a single tunable parameter, and can successfully function with partial memory resources.

4. CONCLUSION

This study focus on effective clustering approach for large database. In clustering analysis, there are many problems, responsible for the suitable number of clusters for a specific dataset and picture the performance of these clusters. This study have made a proposed new method with Iris Data Set.

A new approach has been proposed which combines the dimensionality reduction through PCA and a Kmeans algorithm which uses the basic EK-means algorithm. The main objective of applying PCA on original data before clustering is to obtain accurate results. But the clustering results depend on the initialization of centroid. This experiment shows a substantial improvement in running time and accuracy of the clustering results by reducing the dimension and initial centroid selection using PCA. Though it produces better result than the EK-means but when number of document increases, the intra-cluster similarity decreases. A method or algorithm can be taken for further research.

The results in all the data sets prove that the PCA combine with EK-means can achieve 95.8% accuracy. The experimental study shows that the PCA& EK-means model is very effective in terms of both evaluation time and estimating performance.

5. REFERENCE

[1] D. Pyle, *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann, 1999

[2] G. J. McLahlan and K. E. Basford, Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker, 1987, vol. 84..

[3] A. Varfis and C. Versino, "Clustering of socio-economic data with Kohonen maps," *Neural Network World*, vol. 2, no. 6, pp. 813–834, 1992.

[4] Jonathon Shlens. A Tutorial on Principal Component Analysis. University of California, San Diego, 2005.

[5] Sandro Saitta, Combining PCA and K-means March 26, 2007 by Filed under: PCA, k-means

[6] Huang, Z. (1998). Extensions to the EK-means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, 2, p. 283-304 (Pubitemid 128695480)

[7] D. Pollard, ªA Centeral Limit Theorem for EK-means Clustering, Annals of Probability, vol. 10, pp. 919-926, 1982.