Data Mining with Improved and Efficient Mechanism in Clustering Analysis and Decision Tree as a Hybrid Approach

Heena Sharma, Navdeep Kaur Kaler

Abstract-In this research, we are using clustering and decision tree methods to mine the data by using hybrid algorithms K-MEANS, SOM and HAC algorithms from clustering and CHAID and C4.5 algorithms from decision tree and it can produce the better results than the traditional algorithms. It also performs the comparative study of these algorithms to obtain high accuracy. Clustering method will use for make the clusters of similar groups to extract the easily features or properties and decision tree method will use for choose to decide the optimal decision to extract the valuable information.This comparison is able to find clusters in large high dimensional spaces efficiently. It is suitable for clustering in the full dimensional space as well as in subspaces. Experiments on both synthetic data and real-life data show that the technique is effective and also scales well for large high dimensional datasets.

Index terms- Clustering, Decision tree, HAC, SOM, C4.5, Data Mining, K-Means

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. To

mine the data we use these [1] important steps or tasks: Classification use to classify the data items into the predefined classes and find the model to analysis. Regression identifies real valued variables. Clustering Use to describe the data and categories into similar objects in groups,Find the dependencies between variables, Mine the data using tools. Clustering and decision tree are two of the mostly used methods of data mining which provide us much more convenience in researching information data. Cluster analysis groups objects based on the information found in the data describing the objects or their relationships.

The goal is that the objects in a group will be similar to one other and different from the objects in other groups.The greater the similarity or homogeneity within a group and the greater the difference between groups, the "better" or more distinct the clustering. Clustering is a tool for data analysis, which solves classification problems. Its object is to distribute cases into groups, so that the degree of association to be strong between members of the same cluster and weak between members of different clusters. This way each cluster describes, in terms of data collected, the class to which its members belong.

Manuscript received on April, 2013

Heena Sharma, Research Scholar, Done B.Tech. (CSE) from L.L.R.I.E.T Moga (P.T.U), now doing M.Tech (CSE) from L.L.R.I.E.T Moga (P.T.U), Punjab, India.

Navdeep Kaur Kaler, Assistant Professor in Department Of CSE, L.L.R.I.E.T, Moga, Punjab, India.

Classification is an important task in data mining. It belongs to directed learning and the main methods include decision tree, neural network and genetic algorithm. Decision tree build its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Algorithms include ID3, C4.5, CART and SPRINT etc.

II. PREVIOUS WORK

Hong Yu *et al*. [2] 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*. [3] 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.

Mohamed El far et al. [4] comparing between data mining algorithms: "Close+, Apriori and CHARM" and "K-means classification algorithm" and applying them on 3D object indexing. Three-dimensional models are more and more used in applications in which the necessity to visualize realistic objects is felt (CAD/CAO, medical simulations, games, virtual reality etc.).

Wangjie Sun *et al.* [5] implement an advanced design of data mining algorithms. In order to save the computer data effectively, we should not only check the integrity for the data , but also check storage system to recover data in a timely manner to reduce losses to a minimum, to prevent the recover fails when the fault occurred.

S.P.Latha [6] presents algorithm for efficient data mining. Over the years, a variety of algorithms for finding frequent item sets in very large transaction databases have been developed. Data mining algorithms are used extensively to analyze business, commerce, scientific, engineering, and security data and dramatically improve the effectiveness of applications in areas such as marketing, predictive modeling, life sciences, information retrieval, and engineering.

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 [7]. In most of clustering algorithms, the criterion that is used to measure the quality of resulting clusters is defined as in [8] equation(1) which is known as minimizing sum of squared error:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2,$$

Usually, similarity and dissimilarity between objects are expressed through some distance functions. The most common distance function is the Euclidean distance.

IV. DECISION TREE ANALYSIS

Decision tree is a popular classification method that results in a flow-chart like tree structure where each node denotes a test on an attribute value [9] and each branch represents an outcome of the test. Decision tree is a supervised data mining technique. It can be used to partition a large collection of data in to smaller sets by recursively applying two-way and /or multi way. Using the data, the decision tree method generates a tree that consists of nodes that are rules. Each [10] leaf node represents a classification or a decision. The training process that generates the tree is called induction. It has been shown in various studies that employing pruning methods can improve the generalization performance of a decision tree, especially in noisy domains According to this methodology; a loosely stopping criterion is used, letting the decision tree to overfit the training set. Then the over-fitted tree is cut back into a smaller tree by removing sub–branches that are not contributing to the generalization accuracy.

V. HYBRID APPROACH

We are using Hybrid techniques of clustering and decision tree method for large dimensional dataset. Clustering analysis is an important and popular data analysis technique that is large variety of fields**.** Clustering and decision tree are the mostly used methods of data mining. Clustering can be used for describing and decision tree can be applied to analyzing. After combining these two methods effectively we compare the effectiveness of clustering data mining algorithms HAC and SOM with the traditional algorithms with using decision tree algorithms C4.5 and CHAID by applying them to data sets. After using the hybridization, algorithms produce the best classification accuracy, and also show the higher robustness and generalization ability compared to the other algorithms.

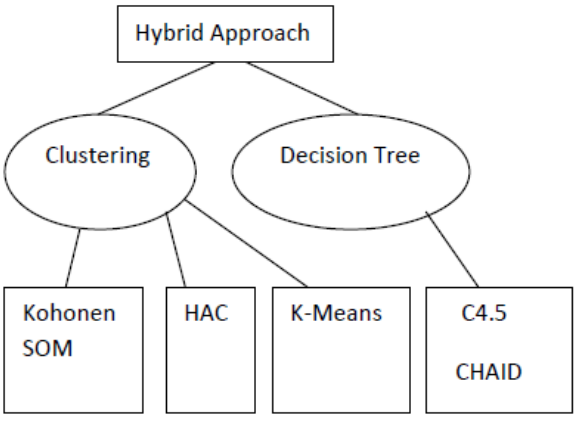

Fig 1 Hybrid Approach

VI. PROPOSED WORK

We compare the effectiveness of two stage clustering and decision tree data mining algorithms by applying them to data sets. Experiment results will show that like two stage algorithms produce the best classification accuracy, and also show the higher robustness and generalization ability compared to the other traditional algorithms. Our approach can remove the shortcoming of hybridization of algorithms (clustering and decision tree algorithms) and improve the results on applying them to data sets. Our approach gives us effective results, better performance and reduces the error rate than the traditional algorithms of clustering and classification in data mining.

A. K- Means Method

1. Select k points as the initial centroids in a random way.
2. (Re) Assign all objects to the closest centroid.
3. Recalculate the centroid of each cluster.
4. Repeat steps 2 and 3 until a termination criterion is met.
5. Pass the solution to the next stage.

B. SOM (Self Organization Map)

The self-organising maps (SOM) introduced by Teuvo Kohonen [11] are deemed as being highly effective as a sophisticated visualization tool for visualizing high dimensional, complex data with inherent relationships between the various features comprising the data. The SOM's output emphasises the salient features of the data and subsequently lead to the automatic formation of clusters of similar data items. We argue that this particular characteristic of SOMs alone qualifies them as a potential candidate for data mining tasks that involve classification and clustering of data items.

C. HAC (Hierarchical Agglomerative Clustering)

1. Compute the proximity matrix containing the distance between [12] each pair of patterns. Treat each pattern as a cluster.
2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.
3. If all patterns are in one cluster, stop. Otherwise, go to step 2.

D. CHAID (Chi-square–Automatic–Interaction–Detection)

Starting from the early seventies, researchers in applied statistics developed procedures for generating decision trees, such as: AID, MAID (Gillo, 1972), THAID (Morgan and Messenger, 1973) and CHAID. It was originally designed to handle nominal attributes only. This procedure also stops when one of the following conditions is fulfilled:

1. Maximum tree depth is reached.
2. Minimum number of cases in node for being a parent is reached, so it cannot be split any further.
3. Minimum number of cases in node for being a child node is reached.

CHAID handles missing values by treating them all as a single valid category. CHAID does not perform pruning.

E. C4.5

C4.5 is an evolution of ID3, presented by the same author (Quinlan, 1993). It uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. Error–based pruning is performed after the growing phase. C4.5 can handle numeric attributes. It can induce from a training set that incorporates missing values by using corrected gain ratio criteria.

VII. CONCLUSION

This research work can improve the performance of traditional algorithms Like K-Means and presents a hybrid approach like algorithm SOM, HAC and C4.5 and CHAID for mining large-scale high dimensional datasets. The mostly used algorithm is K-MEANS which can deal with small convex datasets preferably. It reduces the error rate and achieves accuracy. This research compares the effectiveness of three clustering data mining algorithms - K-means, Kohonen-SOM and HAC by applying them to data sets. Experiment results will show that the Kohonen-SOM and HAC algorithms produce the best classification accuracy, and also show the higher robustness and generalization ability compared to the other algorithms.

REFERENCES

1. Tipawan Silwattananusarn, Dr. KulthidaTuamsuk “Data Mining and Its Applications for Knowledge Management -A Literature Review from 2007 to 2012” International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012 pp 13-24.
2. Hong Yu, Xiaolei Huang, Xiaorong Hu, Hengwen Cai (2010) “A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation”, International Conference on Management of e-Commerce and e-Government.
3. Ji Dan, Qiu Jianlin (2010) “A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree”, 10th IEEE International Conference on Computer and Information Technology, CIT.
4. Mohamed El far, Lahcen Moumoun, Mohamed Chahhou, Taoufiq Gadi (2010) “Comparing between data mining algorithms: "Close+, Apriori and CHARM" and “K-Means classification algorithm” and applying them on 3D object indexing”, 10th IEEE International Conference on Computer and Information Technology, CIT.
5. S.P.Latha (2007) “Algorithm for Efficient Data Mining”, International Conference on Computational Intelligence and Multimedia Applications, Kavaraipettai.
6. Wangjie Sun, Zhigao Zheng (2010) “An Advanced Design of Data Mining Algorithms”, IEEE.
7. Abdolreza Hatamlo and Salwani Abdullah “A Two-Stage Algorithm for Data Clustering” Int Conf. Data Mining DMIN 2011 pp-135-139.
8. http://en.wikipedia.org/wiki/CURE_data_clustering_Algorithm
9. S.Balaji and Dr.S.K.Srivatsa” Decision Tree induction based classification for mining Life Insurance Data bases” International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3, June 2012 pp-699-703.
10. Lior Rokach and Oded Maimon” Top-Down Induction of Decision Trees Classifiers- A Survey” IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002 pp-1-12.
11. T. Kohonen”The Self-Organizing Map” Proceedings of the IEEE, 78(9):1464-1480, 1990.
12. Lior Rokach and Oded Maimon” Top-Down Induction of Decision Trees Classifiers- A Survey” IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: PART C, VOL. 1, NO. 11, NOVEMBER 2002 pp-1-12.

AUTHORS PROFILE

Heena Sharma, Research Scholar, Done B.Tech. (CSE) from L.L.R.I.E.T Moga (P.T.U), now doing M.Tech (CSE) from L.L.R.I.E.T Moga (P.T.U), Punjab, India, Research area is Data Mining.

Navdeep Kaur Kaler, Assistant Professor in Department Of CSE, L.L.R.I.E.T, Moga, Punjab, India, have done B.Tech. (CSE) from Guru Nanak Dev Engineering College, Ludhiana and have done M.Tech. (CSE) from Punjab Agriculture University (PAU), Ludhiana, Research area is Software Engineering, Data Mining