International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

A CLUSTER ANALYSIS AND DECISION TREE HYBRID APPROACH IN DATA MINING TO DESCRIBING TAX AUDIT

Richa Dhiman
[1]Department of computer science and engineering, Lovely Professional University, Phagwara
Richadhiman58@gmail.com
Sheveta Vashisht
[2]Department of computer science and engineering, Lovely Professional University, Phagwara
sheveta.16856@lpu.co.in
Kapil Sharma
[3]Department of computer science and engineering, Lovely Professional University, Phagwara
kapilsharma701@gmail.com

ABSTRACT
In this research, we use clustering and classification methods to mine the data of tax and extract the information about tax audit by using hybrid algorithms K-MEANS, SOM and HAC algorithms from clustering and CHAID and C4.5 algorithms from decision tree and it produce the better results than the traditional algorithms and compare it by applying on tax dataset. 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 from samples of tax datasets? 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.
Keywords- Clustering, Decision tree, HAC, SOM, C4.5.

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.
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.

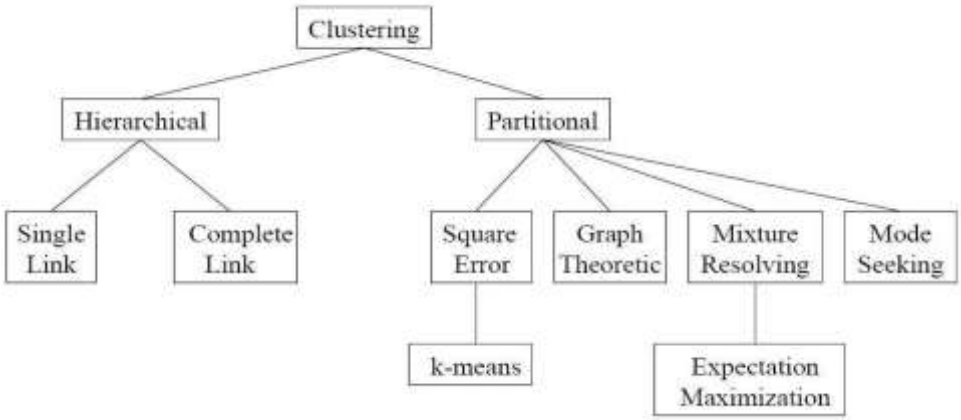
II. BACKGROUND
Ji Dan *et al* (2010) they presented a new synthesized data mining algorithm[3] named CA which improves the original methods of CURE and C4.5. CA introduces principle component analysis (PCA), [2] grid partition and parallel processing which can achieve feature Reduction and scale reduction for large-scale datasets. This paper applies CA algorithm to maize seed breeding and the results of experiments show that our approach is better than Original methods. They introduces feature reduction, scale reduction and classification analysis to handle large and high dimensional dataset By applying CA algorithm in maize seed breeding and find out the important features which will influence breeding tremendously and obtain the classification model of whole maize samples. They conclude that efficiency of CA is higher not only in clustering but also in decision tree. CA is sensitive to some parameters like the clustering number, shrink factors and the threshold etc. C4.5 only can deal with the dataset which has the classification feature. The dataset we treated is a little small which will impact the final output of algorithms.
Guojun Mao *et al* (2011) micro-cluster based[3] classification problems in distributed data streams, and proposes an approach for mining data streams in the distributed environments with both labelled and unlabeled data. For each local site, a local micro-cluster based ensemble is used and its updating algorithms are designed. Making use of the time-based sliding window techniques, the local models in a fixed time-span are transferred to a central site after being generated in all

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

local sites, and then the global patterns related to this time-span can be mined at the central site. They proposed a micro-cluster based ensemble Classification approach for distributed data streams. At local sites, they designed the micro-cluster structure and built "Local MC-Ensembles"; At the central site, they designed a ensemble named "Central MC-Ensemble" and presented an novel method for updating global patterns. The experimental results have shown that it is an applicable approach for its effectiveness and efficiency on maintaining global patterns in distributed environments.
In Reference [4] presented k-attractors, a partitioned clustering Algorithm tailored to numeric data analysis. They uses maximal frequent item-set discovery and partitioning to define the number of clusters k and the initial cluster attractors. This algorithm uses a distance measure, which is adapted with high precision to the way initial attractors are determined. They applied k-attractors as well as k-means, EM, and Farthest First clustering algorithms to several datasets and compared results. In reference [5] presented three different algorithms for data clustering. These are the Self-Organizing Map (SOM), Neural Gas (NG) and Fuzzy Means (FCM) algorithms. SOM and NG algorithm are based on competitive learning. In references [6] a decision tree algorithm was developed to classify the surface soil frozen/thawed status based on the microwave emission and scattering characteristics of the frozen/thawed soil, and the cluster analysis of samples from the frozen soil, the thawed soil, the desert and the snow.

III. CLUSTERING METHOD

Different approaches to clustering data can be described with the help of the hierarchy shown in Figure1 (other taxonometric representations of clustering methodology are possible; ours is based on the discussion in Jain and Dubes [1988]). At the top level, there is a distinction between hierarchical and partitional approaches (hierarchical methods produce a nested series of partitions, while partitional methods produce only one)[7].



The taxonomy of clustering [7] shown in Fig1

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, 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, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset

Hierarchical clustering builds a cluster hierarchy or in other words, a tree of clusters also known as a dendrogram [8]. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [Jain & Dubes 1988; Kaufman & Rousseeuw 1990]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

IV. 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 [9]. In most of clustering algorithms, the criterion that is used to measure the quality of resulting clusters is defined as in [10] equation (1) which is known as minimizing sum of squared error:

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

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

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

V. DECISION TREE METHOD

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 [11] 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 splits in figure 2 [12].

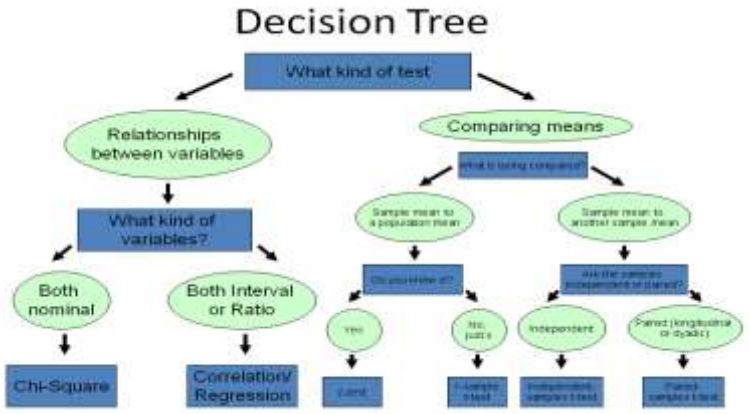


Fig 2 example of [12] Decision tree method

Using the data, the decision tree method generates a tree that consists of nodes that are rules. Each [13] 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.

VI. HYBRID ALGORITHMS

We use Hybrid techniques of clustering and decision tree applying for large dimensional dataset figure 3. 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 compares 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.

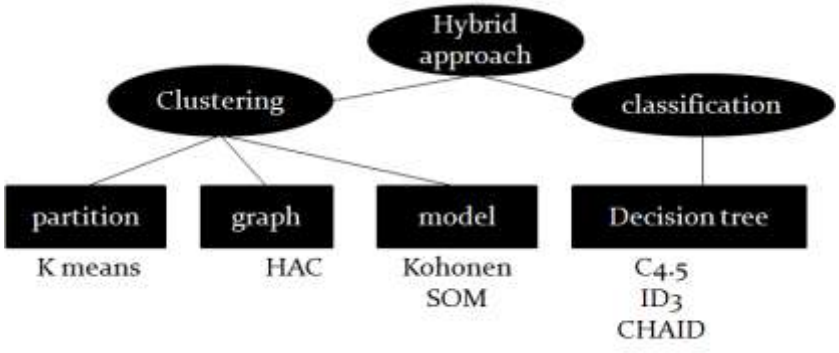


Fig 3 Hybrid Approach

1. 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.

2. SOM (Self Organization Map)

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

The self-organising maps (SOM) introduced by Teuvo Kohonen [14] 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.

3. HAC (Hierarchical Agglomerative Clustering)

1. Compute the proximity matrix containing the distance between [15] 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.

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

Starting from the early seventies, researchers [16] 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.

5. 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.

V. PROPOSED METHODOLOGY

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. We hybridization techniques clustering and decision tree in Figure 4, clustering and classification used to improve the terms of quality, performance, accuracy and error rate. 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.

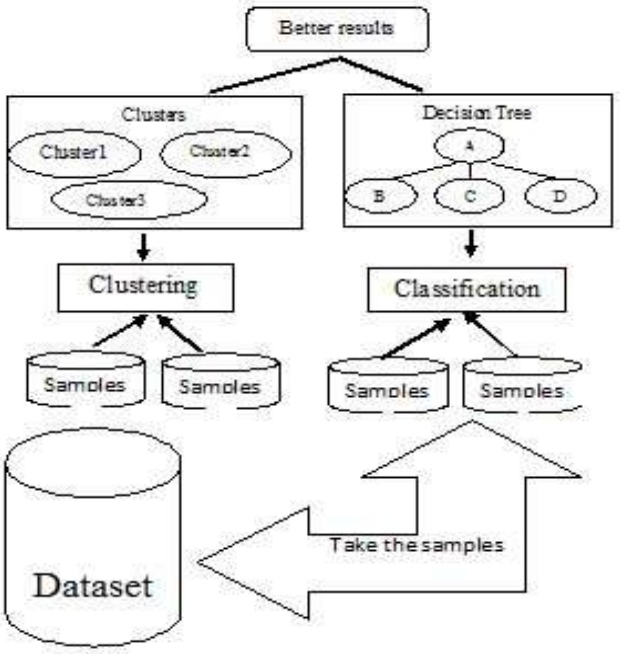
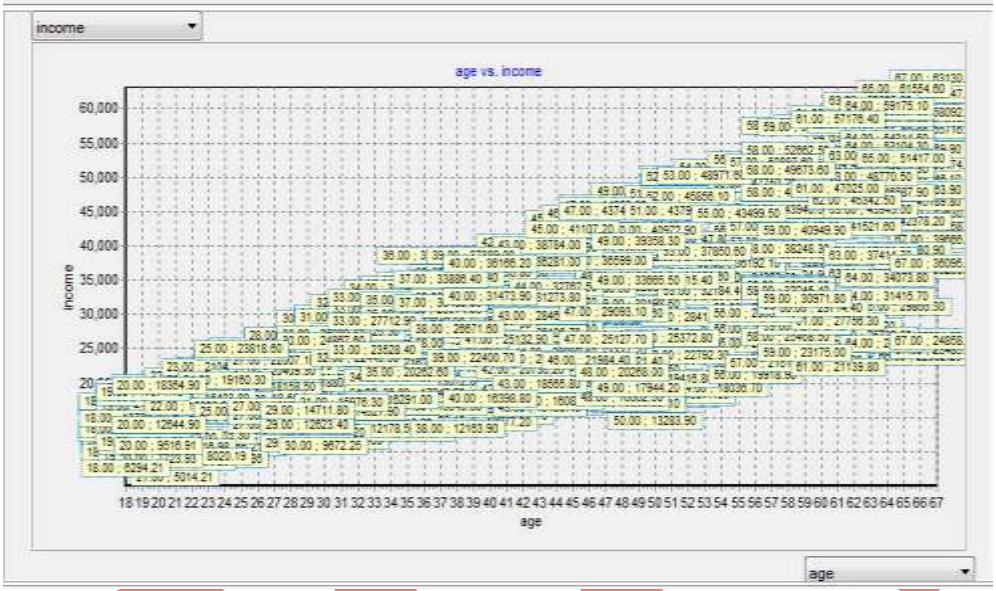


Fig 4 Two stage clustering and decision tree data mining algorithms.

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013 ISSN 2277-3061

VII. DATASET AND TOOLS

We will use Tax data to mine the data with 26 attributes of tax returns like source of income of any person, tax deduction, and total tax paid etc. We will use Tanagra tool for clustering and Spinia tool for decision tree and mine the tax data then compare the result with traditional algorithms and find the tax audits for example in which year tax agencies had been top 50% tax from population under tax rules and comparison between income and age attributes for tax Figure 5.



In Tanagra, comparison between income and age attributes for tax Fig 5.

VIII. CONCLUSION

This research can improve the performance of traditional algorithms K-Means and presents a synthesized 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. But it also exist some shortcomings. For example, it can only deal with numeric data, find convex or spherical sharps be sensitive to the input and noise and can't deal with large datasets. increase the degree of association between the members of the same cluster, increase cluster quality, PCA used for add more features and to overcome the limitations such as optimal searching samples, To reduce the sensitivity to outliers or noises, outfit problems for classifying samples perfectly. This research combine clustering technique which is based on a supervised learning technique called decision tree construction, find clusters in large high dimensional spaces efficiently and improve the results with clusters quality and performance. It reduce the error rate and achieve accuracy.

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] Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, He Peng" A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree" 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010) 978-0-7695-4108-2/10 © 2010 IEEE ,pp 2722-2728.

[3] Guojun Mao ,Yi Yang" A micro-Cluster based Ensemble Approach for Classifying Distributed Data Streams" 2011 23rd IEEE International Conference on Tools with Artificial Intelligence pp-753-759

[4] Y. Kanellopoulos, P. Antonellis, C. Tjortjis, C. Makris, N. Tsirakis" k-attractors a partitional clustering algorithm for numeric data analysis" Applied Artificial Intelligence, 25:97–115, 2011 Copyright 2011 Taylor & Francis Group, LLC pp 97-115.

[5] Terje Kristensen, Vemund Jakobsen "Three Different Paradigams for Interactive Data Clustering "Int Conf. Data Mining DMIN 2011 pp-3-9.

[6] Rui Jin, Xin Li" A Decision Tree Algorithm Freeze /Thaw Classification Of Surface Soil Using SSM/I" 978-1-4244-2808-3/08/$25.00 ©2008 IEEE pp-742-744.

[7] A.K. Jain, M.N. Murty, P.J. Flynn" Data Clustering"

[8] Pavel Berkhin "Survey of Clustering Data Mining Techniques

[9] Abdolreza Hatamlo and Salwani Abdullah "A Two-Stage Algorithm for Data Clustering" Int Conf. Data Mining DMIN 2011 pp-135-139.

[10] http://en.wikipedia.org/wiki/CURE_data_clustering_Algorithm.

[11] 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.

[12] http://my.ilstu.edu/~wjschne/138/Psychology138 Exam3StudyGuide.html.

[13] 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.

[14] T. Kohonen"The Self-Organizing Map" Proceedings of the IEEE, 78(9):1464-1480, 1990.

[15] 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.

CIR