COMPARATIVE STUDY OF DATA MINING ALGORITHMS FOR HIGH DIMENSIONAL DATA ANALYSIS

Smitha .T[1], V. Sundaram[2]
[1] PhD-Research Scholar, Karpagam University, Coimbatore, India
(Asst. Prof., Deptt. of Computer Application, SNGIST, N. Paravoor, Kerala, India)
[2] Director-MCA, Karpagam College of Engineering, Coimbatore, India

ABSTRACT

The main objective of this research paper is to prove the effectiveness of high dimensional data analysis and different algorithm in the prediction process of Data mining. The approach made for this survey includes , an extensive literature search on published papers as well as text books in the application of Data mining in prediction. Many data tables were searched for this purpose and research was conducted during JAN 2009 - 2012 and I have retrieved many published articles on the usage of Data mining algorithm in prediction. I have retrieved those articles by searching the data bases with the usage of the keywords "data mining and algorithm". Titles of the articles were analyzed by usage of association rules that analyze the most frequently used words. The main algorithm which were includes in this survey are decision tree, k-means algorithm ,and association rules. I have Studied each algorithm with the help of high dimensional data set with UCI repository and find the advantages and disadvantages of each and made a comparative result for this.

KEYWORDS: *Algorithm, Clustering. Data mining, Decision Tree, High Dimensional analyzis.*

I. INTRODUCTION

Data mining is an interdisciplinary field of computer science. It is the process that results in the discovery of new patterns in large database. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. Besides the raw analysis step, it involves database and data management aspects, data processing model and inference considerations, interestingness metrics, complexity, considerations, post-processing of found structures, visualization, and online updating.[1] There are different techniques for multidimensional data analysis. The models included in the predictive data mining consists of clustering, decision tree, association rules, pattern matching, classification rules, statistical analysis etc. Through this paper we tried to analyse the advantages and disadvantages of main three algorithms such as association rules, clustering algorithm and classification and prediction methods. In clustering algorithm, we have selected the k-means clustering algorithm for sample study, similarly for classification and prediction methods, decision tree have used which is used c4.5 algorithm. Association rule algorithm also verified with sample data. Data set used for this study is UCI repository.

1.1 Multidimensional Analysis

Multidimensional data is a type of data that records facts related to variable entities called dimensions. Technologies such as Integrated Data Analysis & Simulation module(IDASM) provide an environment where multiple data sets can be integrated to conduct analysis across different cases Dimensions are entities which are used to analyze data and may represent the sides of a

multidimensional cube. Selecting proper dimensions in data analysis is indeed very crucial for multidimensional analysis and gaining greater insights from the data. Dimension modeling is very important in data analysis because queries can be satisfied only if dimensions are properly defined.
If one of the dimensional values changes, then output also will change. The multi dimensional analysis tools are capable of handling large data sets. These tools provide facilities to define convenient hierarchies and dimensions. But they are unable to predict trends and find the patterns and guess the hidden behaviour.[2]

1.2 Related Works in this Area

Many works related in this area have been going on. "In A New Approach for Evaluation of Data Mining Techniques" by Moawia Elfaki Yahia[19],the authors tried to put a new direction for the evaluation of some techniques for solving data mining tasks such as: Statistics, Visualization, Clustering, Decision Trees, Association Rules and Neural Networks. The article on " A study on effective mining of association rules from huge data base " by V.Umarani, [20] It aims at finding interesting patterns among the databases. This paper also provides an overview of techniques that are used to improvise the efficiency of Association Rule Mining (ARM) from huge databases. In another article " K-means v/s K-medoids: A Comparative Study" Shalini S Singh explained that portioned based clustering methods are suitable for spherical shaped clusters in medium sized datasets and also proved that K-means are not sensitive to noisy or outliers.[21]. In an article "Predicting School Failure Using Data Mining C". MÁRQUEZ-VERA explained the prediction methods and the application of classification rule in decision tree for predicting the school failures.[22].There are many research works carrying out related with data mining technology in prediction such as financial stock market forecast, rainfall forecasting, application of data mining technique in health care, base oils biodegradability predicting with data mining technique etc,[23].

II. THE ROLE OF DATA MINING IN HIGH DIMENSIONAL ANALYSIS

Due to the advancement in algorithm and changing scenario, new techniques have emerged in data analysis, which are used to predict and generate data patterns and to classify entities having multivariate attributes. These techniques are used to identify the pre-existing relationship in the data that are not readily available. Predictive Data mining deals with impact patterns of data.[4]

2.1 . Models used in Predictive Data Mining

The models mainly used in predictive data mining includes Regression, Time series, neural networks, statistical mining tools, pattern matching, association rules, clustering, classification trees etc[5]
Regression model is used to express relationship between dependent and independent variables using an expression. It is used when the relationship is linear in nature. If there is a non linear relationship, then it cannot be expressed using any expression, but the relationship can be built using neural networks. In time series models, historic data is used to generate trends for the future. Statistical mining models are used to determine the statistical validity of test parameters and can be utilized to test hypothesis undertake correlation studies and transform and prepare data for further analysis. Pattern matching are used to find hidden characteristics within data and the methods used to find patterns with the data includes association rules. [16]
Association rules allows the analysts to identify the behavior pattern with respect to a particular event where as frequent items are used to find how a group are segmented for a specific set. Clustering is used to find the similarity between entities having multiple attributes and grouping similar entities and classification rules are used to categorize data using multiple attributes.

III. C4.5

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan[1]. It is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. [3]This algorithm builds decision trees from a set of training data using the concept of information entropy.
It is handling both continuous and discrete attributes, handling training data with missing attribute values and also handling attributes with differing costs. In building a decision tree we can deal with

training sets that have records with unknown attribute values by evaluating the gain, or the gain ratio, for an attribute by considering only the records where that attribute is defined. In using a decision tree, we can classify records that have unknown attribute values by estimating the probability of the various possible results.[14]
Follows the algorithms employed in C4.5 using decision tree.

3. 1 Decision trees

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:[7]
If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.
Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets *S*1, *S*2, . . . according to the outcome for each case, and apply the same procedure recursively to each subset.
Use either information gain or gain ratio to rank the possible tests.
Check the estimation error. [12]

Table 1: Decision Tree-Advantages and Disadvantages

Advantages	Limitations
Error rate is less	Decision trees typically require certain knowledge of quantitative or statistical experience to complete the process accurately. Failing to accurately understand decision trees can lead to a garbled outcome of business opportunities or decision possibilities.
Decomposition is easier as compared with other techniques	It can also be difficult to include variables on the decision tree, exclude duplicate information or express information in a logical, consistent manner. The inability to complete the decision tree using only one set of information can be somewhat difficult.
Represent the knowledge in the form of IF-THEN rules. Rules are easier for humans to understand.	While incomplete information can create difficulties in the decision-tree process, too much information can also be an issue.

IV. CLUSTERING ALGORITHM

Clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for data analysis used in many fields including information retrieval. .Cluster analysis groups objects based on their similarity. The measure of similarity can be computed for various types of data. [5]
Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods, k-means algorithm, graph based model etc.

4.1 K means algorithm

In data mining K-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.
The k-means algorithm[also referred as Lloyd's algorithm**]** is a simple iterative method to partition a given dataset into a user specified number of clusters, *k*.[8]The algorithm operates on a set of *d*-dimensional vectors, $D = \{\mathbf{xi} \mid i = 1, \ldots, N\}$, where $\mathbf{xi} \in \_d$ denotes the *i*th data point. The algorithm is initialized by picking *k* points in *_d* as the initial *k* cluster representatives or "centroids".

Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. Then the algorithm iterates between two steps till convergence:[18]
Step 1: Data Assignment. Each data point is assigned to its *closest* centroid, with ties broken arbitrarily. This results in a partitioning of the data.
Step 2: Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.
The algorithm converges when the assignments (and hence the **cj** values) no longer change. The algorithm execution is visually depicted in Fig. 1. Note that each iteration needs $N \times k$ comparisons, which determines the time complexity of one iteration. The number of iterations required for convergence varies and may depend on *N*, but as a first cut, this algorithm can be considered linear in the dataset size.

4.2 Advantages and Limitations

Table 2: Advantages and Limitations of k-means algorithm

Advantages	Limitations
Relatively efficient and easy to implement.	Sensitive to initialization
Terminates at local optimum.	Limiting case of fixed data.
Apply even large data sets	Difficult to compare with different numbers of clusters
The clusters are non-hierarchical and they do not overlap	Needs to specify the number of clusters in advance.
With a large number of variables, K-Means may be computationally faster than hierarchical clustering	Unable to handle noisy data or outliers.
K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular	Not suitable to discover clusters with *non-convex shapes*

V. ASSOCIATION RULES ALGORITHM

Association rule mining searches for interesting relationships among items in a given set. Here the main rule interestingness are rule support and confidence which reflect the usefulness and certainty of discovered rules. Association rule mining algorithm is a two step process where we should have to find all the frequent item sets and generate strong association rules from the frequent item sets.[9]
If a rule concerns association between the presence or absence of items, it is a Boolean association rule. If a rule describes association between quantitative items or attributes, then it is known as quantitative association rules. Here the quantitative values for items are partitioned into intervals. The algorithm can be formed based on dimensions, based on level of abstractions involved in the rule set and also based on various extensions to association mining such as correlation analysis.[17]

5.1 Multi dimensional Association Rules:

In multi dimensional databases, each distinct predicate in a rule as a dimension. Association rule that involve two or more dimensions or predicates each of which occurs only once in the rule can be referred as multidimensional association rules.
Multi dimension association rules with no repeated predicates are called inter dimension association rules and may with repeated predicates which can contain multiple occurrences of some predicates are called hybrid dimension association rules.
For example

Age(X,50…..70)^ FAMILYHISTORY(X,DISEASE)=> DISEASEHIT(X,"TYPHOID").

Here the database attributes can be categorical or quantitative with no ordering among the values The basic definition of association rule states that, Let A={l1,l2,……..lm} be a set of items, and Let T, the

transaction database, be a set of transaction, where each transaction t is a set of items and thus t is a subset of A.

An association rule tells us about the association between two or more items. For example, If we are given a set of items where items can be referred as disease hit in an area and a large collection of patients who are subsets of some inhabitants in the area. . The task is to find relationship between the presences of disease hit within these group. In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule: Support *is* how often does the rule apply? and confidence is how often is the rule is correct. [19]

In fact association rule mining is a two-step process: Find all frequent item sets / disease hit - by definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, and then generate strong association rules from the frequent item sets by definition, these rules must satisfy minimum support and minimum confidence.

In this study predicting the chances of disease hit an area, by correlating the parameters or attributes such as climate, environmental condition, heredity, education with the inhabitants. And also finding how these parameters are associated with the chances of disease hit.

Table 3: Association Rules-Advantages and Disadvantages

Advantages	Limitations
Association rule algorithms can be formulated to look for sequential patterns.	Association rules do not show reasonable patterns with dependent variable and cannot reduce the number of independent variables by removing.
The methods of data acquisition and integration, and integrity checks are the most relevant to association rules.	Association rules cannot be useful if the information do not provide support and confidence of rule are correct.

VI. CONCLUSION

In this research work, I have made a study to make a comparison of the some of the existing data mining algorithm for high dimensional data clusters to estimate prediction in data mining technique. The main techniques included in the survey are decision tree, clustering algorithm, k-means algorithm and association rule algorithm. Studied each algorithm with the help of high dimensional data set with UCI repository and find the advantages and disadvantages of each. By comparing the advantages and disadvantages of each algorithm, I am trying to develop a hybrid algorithm for multidimensional data analysis. The efficiency was calculated on the basis of time complexity, space complexity, space requirements etc. The sample used in this study includes UCI repository. The efficiency of new algorithm can be checked with real time data.

VII. FUTURE ENHANCEMENT

We will be able to create a new hybrid algorithm by comparing the advantages and disadvantages of the existing ones. We can also take other techniques which are not included in this survey for comparison purpose and can find the best one by evaluating the advantages and limitations of the existing one.

REFERENCES

[1]. en.wikipedia.org/wiki/Data_mining
[2]. Multidimensional Data Analysis and data mining, Black Book, Arijay Chaudhry and Dr. P .S. Deshpande.
[3]. R. Agarwal, T. Imielinski and A. Swamy “Mining association Rules between Set of Items in Large Database”.In ACM SIGMO international conference on Management of Data .
[4]. Goulbourene G, Coenen F and Leng P, Algorithms for Computing Association Rules using a Partial support Tree” j. Knowledge Based System 13(2000)pp-141-149.
[5]. David Hand, Heikki Mannila, Padhraic Smyth,” principles of Data Mining”.
[6]. “Case study on High Dimensional Data Analysis using Decision Tree Model”,Smitha.T,Dr.V.Sundaram, IJCSI Vol9,Issue 3, May 2012.

[7]. "Classification Rules By Decision Tree for disease Prediction", Smitha.T, Dr. V. Sundaram, IJCA, vol-43, No-8, April 2012.
[8]. "Knowledge Discovery from Real Time Database using Data Mining Technique", Smitha.T, Dr. V. Sundaram, IJSRP vol 2, issue 4, April 2012.
[9]. "Another Look at Measures of Forecast Accuracy" Hyndman R and Koehler A(2005).
[10]. **"**Mining the structural knowledge of high-dimensional medical data using Isomap" S. Weng I C. Zhang I Z. Lin I X. Zhang 2
[11]. Bhattachariee, A., Richards, W. G., Staunton, J., Li, C., Monti,S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.,Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W.,Johnson, B. E., Golub, T. R., Sugarbaker, D. J., And Meyerson, M. (2001): 'Classification of human lung carcinomasby mRNA expression profiling reveals distinct adenocarcinoma sub classes', *Proc. Nat. Acad. Sci. USA,* 98, pp. 13790 13795 BLAKE, C. L. and Merz, C. J. (1998):
[12]. BORG, I., and GROENEN, P. (1997): 'Modern multidimensional scaling: theory and application' (Springer-Verlag, New York, Berlin, Heidelberg, 1997).
[13]. Adomavicius G,TuzhilinA2001 Expert-driven validation of rule-based user models in personalization applications. *Data Mining Knowledge Discovery* 5(1/2): 33–58.
[14]. Shekar B, Natarajan R 2004b A transaction-based neighbourhood-driven approach to quantifying interestingness of association rules. *Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM 2004)*(Washington, DC: IEEE Comput. Soc. Press) pp 194–201
[15]. Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and WeiLi. Parallel algorithms for discovery of association rules. Data Mining and Knowledge Discovery: An International Journal, special issue on Scalable High-Performance Computing for KDD, 1(4):343–373, December 2001.
[16]. Refaat, M. Data Preparation for Data Mining Using SAS,Elsevier, 2007.
[17]. El-taher, M. Evaluation of Data Mining Techniques, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan ,2009.
[18]. Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101, no 1, 2001, pp.41-46.
[19]. "A New Approach for Evaluation of Data Mining Techniques", Moawia Elfaki Yahia1, Murtada El-mukashfi El-taher2, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.
[20]. "A study on effective mining of association rules from huge database" V.Umarani et. al. / IJCSR International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.
[21]. " K-means v/s K-medoids: A Comparative Study" Shalini S Singh, National Conference on Recent Trends in Engineering & Technology, May 2011.
[22]. Predicting School Failure Using Data Mining"C. MÁRQUEZ-VERA
[23]. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" K.Srinivas et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

AUTHOR'S BIOGRAPHY

Smitha.T.: She has acquired her Post Graduate Degree in Computer Application and M.Phil in Computer science from M. K. University. Now doing PhD at Karpagam University under Dr. V. Sundaram. .She has 9 years of teaching experience and 4 years of industrial and research experience. She has attended many national and international conferences and workshops and presented many papers, regarding data mining,. She has also published many articles regarding data mining techniques in international journals with high impact factor. Now working as an Asst. Professor–MCA department of Sree Narayana Guru Institute of Science and Technology, N. Paravoor, Kerala. Her area of interest is Data mining and Data Warehousing.

Dr. V. Sundaram: He is a postgraduate in Mathematics with PhD in applied mathematics. He has 45 years of teaching experience in India and abroad and guiding more than 10 scholars in PhD and M.phil at Karpagam and Anna University. He has organized and presented more than 40 papers in national as well as international conferences and have many publications in international and national journals. He is now working as the Director of MCA Department of Karpagam Engineering College. He is a life member in many associations. His area of specialization includes fluid Mechanics, Applied Mathematics, Theoretical Computer Science, Data mining, and Networking etc.