ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Efficient Mechanism for Data Mining with Clustering and Classification Analysis as a Hybrid Approach

M. Nageshu[1]
Research Scholar,
Department of Computer Science,
Sri Venkateswara University,
Tirupati, India

Dr. E. Kesavulu Reddy[2]
Assistant Professor,
Department of Computer Science,
Sri Venkateswara University,
Tirupati, India

Abstract: ***Appearance of modern techniques for scientific records collection has resulted in big scale accumulation of data pertaining to various fields. Predictable database querying methods are insufficient to extract functional information from enormous data banks. In this research, we are using clustering with classification and decision tree methods to mine the data by using hybrid algorithms like EM, K-MEANS and HAC algorithms from clustering, J48 and C4.5 algorithms from decision making and it can generate the improved outcome than the conventional algorithms. It also performs the proportional study of these algorithms to acquire elevated accuracy. This contrast is capable to discover clusters in huge high dimensional spaces proficiently. It is appropriate for clustering in the complete dimensional space as fit as in subspaces. Experiments on both synthetic data and real-life data demonstrate that the method is successful and also balance well for huge high dimensional datasets.***

Keywords: HAC, EM, J48, C4.5, K-Means, Data Mining, Clustering, Decision tree.

I. INTRODUCTION

Today Information Technology plays a very important function in every aspects of the human life. It is extremely necessary to collect data from dissimilar sources. This data can be stored and maintained to create information and knowledge. Data mining is the non trivial procedure of identifying suitable, original, potentially helpful and eventually understandable patterns in data. With the extensive use of databases and the volatile development in their sizes, organizations are faced with the problem of information burden. The problem of effectively utilize these huge volumes of data is becoming a foremost problem or all enterprise. We used special algorithms to mine the precious data. To extract the data we use these significant steps or tasks: Classification use to classify the data items into the predefined classes and discover the replica to analysis. Regression identifies real valued variables. Clustering use to illustrate the data and categories into comparable objects in groups, get the dependencies between variables, extract the data via tools. Cluster analysis groups objects based on the information found in the data relating the objects or their relationships. Clustering is a tool for data analysis, which solves classification problems. Its object is to allocate cases into groups, so that the quantity of association to be strong between members of the identical cluster and weak between members of dissimilar clusters. This way every cluster describes, in requisites of data collected, the class to which its members belong. Classification is an essential task in data mining. It belongs to intended for learning and the core methods include decision tree, neural network and genetic algorithm. Decision tree build its finest tree form by selecting essential relationship features. Special decision making methods will adopt different technologies to resolve these problems.

II. BACK GROUND

Data mining is the study and analysis of big data sets, in order to find out meaningful pattern and rules. The key scheme is to discover useful way to merge the computer's power to process the data with the human eye's ability to identify patterns. The idea of data mining is to plan and work professionally with large data sets. Data mining is the component of wider procedure

called knowledge discovery from database. Data Mining is the procedure of analyze data from dissimilar perspectives the results as positive information. It has been defined as "the nontrivial process of identifying suitable, original, potentially useful, and eventually reasonable patterns in data" The meaning of data mining is directly associated to one more usually used word knowledge discovery. Data mining is an interdisciplinary, incorporated database, artificial intelligence, machine learning, statistics, etc. Several areas of theory and technology in present era are databases, artificial intelligence, data mining and statistics is a study of three strong large expertise pillars. Data mining is a multi-step process; require accessing and preparing data for a mining the data, data mining algorithm, analyzing consequences and taking suitable action. The data, which is accessed can be stored in one or further operational databases. In data mining the information can be mined by passing different processes.

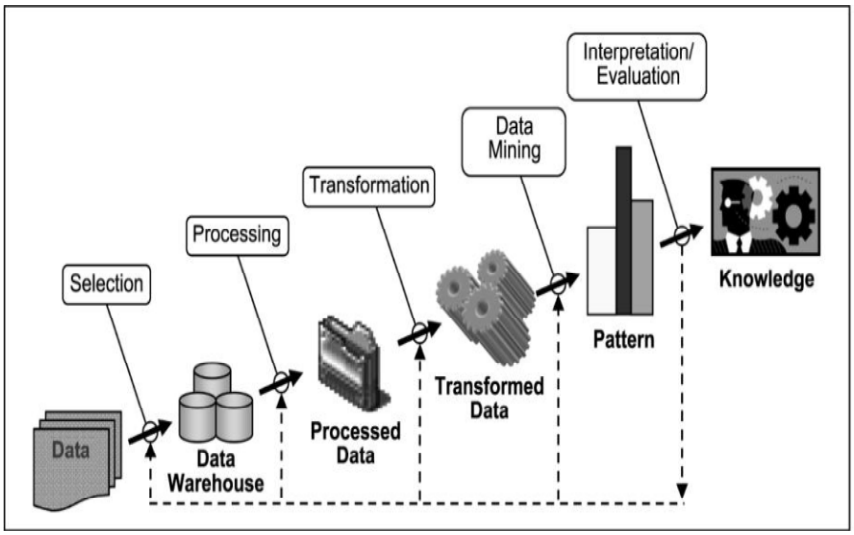

Figure 1: Process of Knowledge Discovery

A. Data Mining Application Areas

Data mining is determined in part by latest applications which necessitate new capabilities that are not at present being supplied by today's technology. These new applications can be logically into the following categories.

» Business and E-Commerce

» Scientific, Engineering and Health Care Data

» Telecommunication Industry

» Biological Data Analysis

» Intrusion Detection

» Other Scientific Applications

B. Data Mining Tasks

Data mining tasks are principally classified into the following wide categories:

» Classification/Prediction

» Estimation

» Association

» Clustering

» Visualization

C. Issues In Data Mining

Data mining has evolved into a significant and dynamic area of research because of the hypothetical challenges and realistic applications connected with the problem of discovering exciting and previously unknown knowledge from real-world databases. The major challenge to the data mining and the equivalent consideration in scheming the algorithms are as follows:

1. Huge datasets and high dimensionality.

2. Over fitting and assessing the statistical implication.

3. Understandability of patterns.

4. Non-standard unfinished data and data integration.

5. Diverse changing and redundant data.

III. Decision Tree Analysis

Decision tree/making is a popular classification method that results in a flow chart like tree arrangement where every node denotes a test on an attribute value and every branch represents an outcome of the test. Decision tree is a supervised data mining technique. It can be used to partition a big collection of data in to smaller sets by recursively applying two-way and /or multi way. Using the data, the decision tree technique generates a tree that consists of nodes that are rules. Every leaf node represents a classification or a decision. The training process that generates the tree is called induction. It has been shown in different studies that employing pruning methods can develop the overview performance of a decision tree, particularly in noisy domains according to this methodology; a loosely stopping criterion is used, letting the decision tree to over fit the training set. Then the over-fitted tree is cut back into a minor tree by removing sub branches that are not contributing to the simplification accuracy.

IV. Hybrid Approach

We are using Hybrid techniques of clustering and decision tree method for large dimensional dataset. Clustering analysis is an essential and popular data analysis technique that is great mixture of fields**.** Clustering and decision tree are the frequently used methods of data mining. Clustering can be used for describing and decision tree can be applied to analyzing. After combining these two methods successfully we evaluate the usefulness of clustering data mining algorithms HAC and EM with the conventional algorithms with using decision tree algorithms C4.5 and J48 by applying them to data sets. After using the hybridization, algorithms construct the greatest classification accuracy and also show the advanced robustness and generalization capability compared to the further algorithms.

V. Proposed Work

We contrast the efficiency of two stage clustering and decision tree data mining algorithms by applying them to data sets. Testing results will show that like two stage algorithms create the greatest classification accuracy and also demonstrate the higher robustness and simplification capability compared to the other traditional algorithms. Our approach can eliminate the shortcoming of hybridization of algorithms (clustering and decision tree algorithms) and develop the results on applying them to data sets. Our approach gives us efficient results, improved performance and reduces the error rate than the traditional algorithms of clustering and classification in data mining.

We consider the absolute value equation (AVE):

$Ax-|x|=b$

Where $A \in R^{n\times n}$ and $b \in R^n$ are given, and $|\cdot|$ denotes absolute value. A slightly more common form of the AVE, $Ax + B|x| = b$ was introduced for investigated algorithmically in a novel general context. Our current hybrid approach consists of an iterative process whereby we first solve a system of linear equations based on AVE.

A. K- Means Method

1. Select k points as the initial centroids in a random way.

2. (Re) allocate all objects to the closest centroid.

3. Recalculate the centroid of every cluster.

4. Repeat steps 2 and 3 until a termination criterion is met.

5. Pass the clarification to the next stage.

B. HAC (Hierarchical Agglomerative Clustering)

1. Calculate the proximity matrix containing the distance between every pair of patterns. Treat every pattern as a cluster.

2. Find the majority comparable pair of clusters using the proximity matrix. Combine these two clusters into one cluster. Modernize the proximity matrix to reflect this merge operation.

3. If all patterns are in one cluster, stop. Otherwise, go to step 2.

C. EM (Expectation Maximization)

Although the fact that EM can infrequently get stuck in a limited maximum as you guess the parameters by maximizing the log likelihood of the experimental data, in my mind there are three things that construct it magical:

» The capability to concurrently optimize a huge number of variables.

» The ability to discover good estimates for any absent information in your data at the same time.

» In the situation of clustering multidimensional data that lends itself to modeling by a Gaussian mixture, the ability to generate both the traditional “hard” clusters and not-so- traditional “soft” clusters.

D. J48

Classification is the procedure of build a model of classes from a set of records that hold class labels. Decision Tree Algorithm is to discover exposed the way the attributes-vector behaves for a number of instances. Furthermore on the bases of the training instances the classes for the recently generated instances is being originated. This algorithm generates the rules for the prediction of the objective variable. With facilitate of tree classification algorithm the significant distribution of the data is easily understandable. J48 creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is useful to each tuple in the database and results in classification for that tuple.

D. C4.5

C4.5 is a development 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 mounting phase. C4.5 can hold numeric attributes. It can induce from a training set that incorporates missing values by using corrected gain ratio criteria.

VI. Performance Evaluation

This section has shown the evaluation of the dissimilar data mining algorithms applied on several data sets with data mining toolkit. The formula to estimate accuracy is:

$$TA = \frac{(TP+TN)}{TP+TN+FP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$RA = \frac{(TP+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{(TOTAL*TOTAL)} \ldots(2)$$

In the equation (1) TA represents Total Accuracy, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. In equation (2) RA represents Random Accuracy.

Table 1: Prediction Performance Comparison of Algorithms

ALGORITHM	ACCURACY	ERROR
K-Means	99.8700	0.1300
HAC	68.099	31.901
EM	76.3021	23.6979
J48	73.8281	20.1719
C4.5	74.0885	25.9115

VII. CONCLUSION

This research work can improve the performance of traditional algorithms Like K-Means and presents a hybrid approach like algorithm HAC, EM, J48 and C4.5 and for mining large-scale high dimensional datasets. The frequently 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 efficiency of these clustering and classification data mining algorithms by applying them to data sets. Experiment results will show that the K-Means and J48 algorithms generate the best classification accuracy and also show the high robustness and generalization capacity compared to the other algorithms.

References

1. Usama Fayyad, G. Paitetsky-Shapiro, and Padhrais Smith, “knowledge discovery and data mining: Towards a unifying framework”, proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 82-22, 1996.
2. ZhaoHui Tang, Jamie MacLennan, “Data Mining with SQL server 2005”, John Wiley & Sons, 2005.
3. Yasogha, P; M. Kannan ,“Analysis of a Population of Diabetics Patients Databases in Weka Tool”, International Journal of Science & Engineering Research, Vol. 2, Issue 5, May 2011.
4. Jinxin Gao, David B. Hitchcock, “James-Stein Shrinkage to Improve K-means Cluster Analysis”, University of South Carolina,Department of Statistics November 30, 2009.
5. A. P. Dempster; N. M. Laird; D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. Pp.1-38, 1977.
6. M. and Heckerman, D., “An experimental comparison of several clustering and initialization method”, Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA, February, 1998.
7. Mrs. Bharati M. Ramageri, “Data Mining Techniques and Applications,” Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp.301-305, 2010.
8. Karimella Vikram and Niraj Upadhayaya, “Data Mining Tools and Techniques: a review,” Computer Engineering and Intelligent Systems, Vol 2, No.8, pp.31-39, 2011.
9. Fayyad, U., “Mining Databases: Towards Algorithms for Knowledge Discovery”, IEEE Bulletin of the Technical Committee on Data Engineering, 21, 1, pp.41-48, 1998.
10. Zhong, N.; Zhou, L., “Methodologies for Knowledge Discovery and Data Mining”, the Third Pacific-Asia Conference, Pakdd-99, Beijing, China, April 26-28, 1999; Proceedings, Springer Verlag, ,1999.
11. Abdolreza Hatamlo and Salwani Abdullah “A Two-Stage Algorithm for Data Clustering” Int Conf. Data Mining DMIN, pp-135-139, 2011.
12. Ji Dan, Qiu Jianlin, “A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree”, 10th IEEE International Conference on Computer and Information Technology, CIT, 2010.
13. Ng, R.T. and Han, J., “Efficient and effective clustering methods for spatial data mining. In Proc. 20th Int. Conf. on Very Large Data Bases, Santiago, Chile, pp. 144–155,1994.
14. Agrawal, R., Imielinski, T., and Swami, A., “Database mining: A performance perspective”, IEEE Transactions on Knowledge and Data Engineering, 5.6, pp.914–925, 1993.

