A Hybrid Approach Based On Association Rule Mining and Rule Induction in Data Mining

Kapil Sharma, Sheveta Vashisht, Heena Sharma, Richa Dhiman, Jasreena Kaur Bains

Abstract--Data Mining: extracting useful insights from large and detailed collections of data. With the increased possibilities in modern society for companies and institutions to gather data cheaply and efficiently, this subject has become of increasing importance. This interest has inspired a rapidly maturing research field with developments both on a theoretical, as well as on a practical level with the availability of a range of commercial tools.

In this research work titled a hybrid approach based on Association Rule mining and Rule Induction in Data Mining we using induction algorithms and Association Rule mining algorithms as a hybrid approach to maximize the accurate result in fast processing time. This approach can obtain better result than previous work. This can also improves the traditional algorithms with good result. In the above section we will discuss how this approach results in a positive as compares to other approaches.

Keywords-- Association Rule mining, A priori algorithm, Rule Induction, Decision list induction, Data mining

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

A. Rule induction through data mining with Association Rule mining approach:

We use rule induction in data mining to obtain the accurate results with fast processing time. We using decision list induction algorithm to make order and unordered list of rules to coverage of maximum data from the data set. Using decision list induction we can generate number of rules for training dataset to achieve accurate result with less error rate. Through rule induction we can minimize the numbers of rules and maximize the coverage of data.

Manuscript receive on March, 2013.

Kapil Sharma, Research Scholar, Done B.TECH (CSE) from L.L.R.I.E.T, Moga. Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India.

Sheveta Vashisht, Assistant Professor in Department Of CSE, Lovely Professional University, Phagwara, Punjab, India.

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

Richa Dhiman, Research Scholar, Done MSC (IT) from DOABA COLLAGE Jalandhar, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India.

Jasreena Kaur Bains, Research Scholar, Done B.TECH (CSE) from Lovely Professional University, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India.

If we use rule induction along with association rule mining then it can generates less numbers of rules with more accurate result. Association Rules form a much applied data mining approach. Association Rules are derived from frequent item-sets.

B. Decision List induction algorithm

The CN2 induction algorithm is a learning algorithm for rule induction. It is designed to work even when the training data is imperfect. It is based on ideas from the AQ algorithm and the ID3 algorithm. As a consequence it creates a rule set like that created by AQ but is able to handle noisy data like ID3. The algorithm must be given a set of examples, Training Set, which have already been classified in order to generate a list of classification rules. A set of conditions, SimpleConditionSet, which can be applied, alone or in combination, to any set of examples is predefined to be used for the classification.

C. Rule induction

Rule induction [1] is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

Some major rule induction paradigms are:

Association rule algorithms: In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

Decision rule algorithms: Decision rules play an important role in the theory of statics and economics. In order to evaluate the usefulness of a decision rule, it is necessary to have a loss function detailing the outcome of each action under different states.

II. PREVIOUS WORKS

Khurram Shehzad (2012) represents a new discretization technique EDISC which utilizes the entropy-based principle but takes a class-tailored approach to [2]discretization. The technique is applicable in general to any covering algorithm, including those that use the class-per-class rule induction methodology such as CN2 as well as those that use a seed example during the learning phase, such as the RULES family.

Anil Rajput *et al*. (2012) they proposed [4] the rule based classification model of historical BSE stock data with data mining techniques. In this Paper we have used decision tree and rule induction method with the help of data mining software.

D T Pham *et al.* (2011) they represents a new hybrid pruning technique for rule induction, as well as an incremental post-pruning technique based on a misclassification [5] tolerance.

Alexander Borisov *et al.* (2011) a methodology based on association rule concepts is given for detecting fab tool commonality of affected lots. The performance of the methodology is then compared to several traditional methods such as ANOVA [6] and contingency tables using eight actual production cases.

III. PROPOSED WORK

Reduction in error rate from large dataset using association rule mining along with decision list induction:

When we apply induction algorithm to particular data set then we generates numbers of rules to cover maximum coverage of data. If we use association rule mining along with rule induction then it gives more accurate result with less error rate.

IV.SOLUTION

Suppose we applying rule induction algorithm to particular dataset then it gives the numbers of rules with maximum data coverage. Our target is to drives less number of rules with high coverage of data so we use A-priori algorithm along with induction algorithm to obtain high accurate result with less error rate. This also reduces the number of rules with fast processing time and high accuracy.

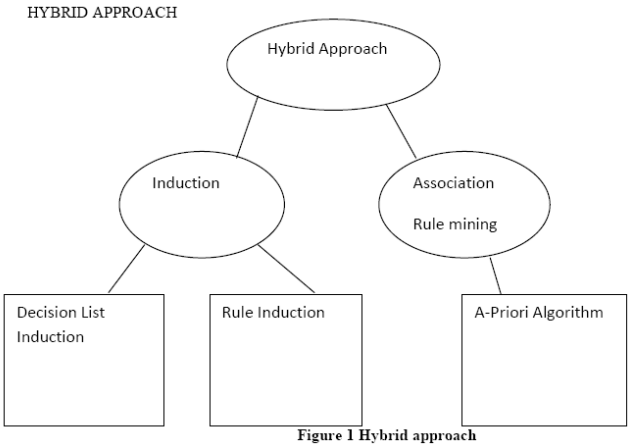

Figure 1 Hybrid approach

V. ASSOCIATION RULE MINING

A-priori Algorithm: Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase

milk." An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, and catalog design and store layout.

In Association Rule Mining, we will generate association rules and calculate support and confidence. Assume minimum support and minimum confidence. The rules satisfying both the criteria of minimum support & minimum confidence is true otherwise false. Rule induction technique retrieves all interesting patterns from the database. In rule induction systems the rule itself is of the simple form of "if this and this and this then this". In some cases accuracy is called the confidence and coverage is called the support. Accuracy refers to the probability that if the antecedent is true that the precedent will be true. High accuracy means that this is a rule that is highly dependable. Coverage refers to the number of records in the database that the rule applies to. High coverage means that the rule can be used very often and also that it is less likely to be a spurious artifact of the sampling technique or idiosyncrasies of the database. Assume minimum accuracy and minimum coverage. The rules satisfying both the criteria of minimum accuracy & minimum coverage is true otherwise false.

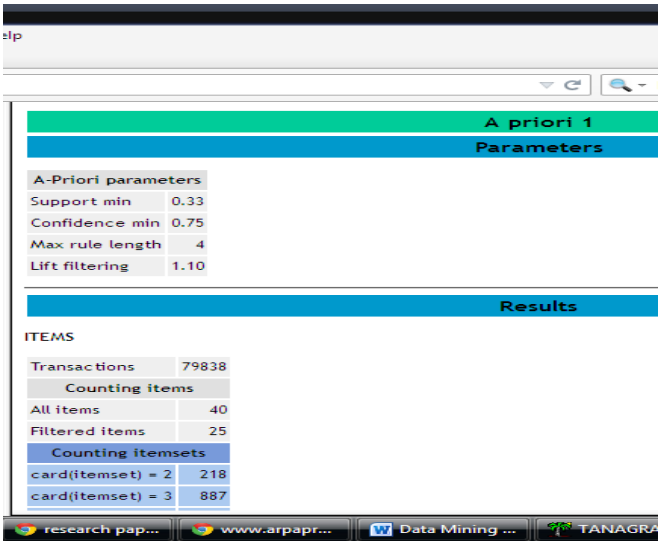

Figure 2 Association rule mining graph using TANAGRA tool

VI. CONCLUSION

Data mining is the biggest issue in every domain of research. It is a big task to mine the data with more accuracy and processing time. The research we develop using rule induction along with Association rule mining algorithm in data mining is beneficial in terms of accuracy and processing time. With the use of this we can minimize the number of rules with more data coverage. We can also reduce the error rate with fast processing time from the large dataset and reduces the time complexity with combine use of rule induction and Association algorithm A-Priori. During the study work we concentrate on learning WEKA or TANAGRA tool so that we can implement our theoretical idea to realization and see results.

REFERENCES

[1] http://en.wikipedia.org/wiki/Inductive_Logic_Programming.

[2] Khurram Shehzad(2012)" *EDISC: A Class-Tailored Discretization Technique for Rule-Based Classification*", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 8, AUGUST 2012.

[3] Ning Zhong, Yuefeng Li(2012)" *Effective Pattern Discovery for Text Mining",* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[4] Anil Rajput, S.P. Saxena(2012)**"** *Rule based Classification of BSE Stock Data with Data Mining***",** International Journal of Information Sciences and Application. ISSN 0974-2255 Volume 4, Number 1 (2012), pp. 1-9.
[5] K. Shehzad(2011)**"** *Simple Hybrid and Incremental Post-pruning Techniques for Rule Induction***",** IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.
[6] Alexander Borisov(2011)" *Rule Induction for Identifying Multilayer Tool Commonalities*", IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, VOL. 24, NO. 2, MAY 2011.
[7] Alexander Borisov(2011)" *Rule Induction for Identifying Multilayer Tool",*IEEE.
[8] Fernando E. B. Otero(2011)" *A New Sequential Covering Strategy for Inducing Classification Rules with Ant Colony Algorithms*",IEEE.
[9] Thomas R. Gabriel and Michael R. Berthold(2010)" *Missing Values in Fuzzy Rule Induction*", IEEE.
[10] Nick F Ryman-Tubb(2010)" *SOAR – Sparse Oracle-based Adaptive Rule Extraction: Knowledge extraction from large-scale datasets to detect credit card fraud***",** IEEE.
[11] Alberto Fern´andez(2010)" *Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study*", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 14, NO. 6, DECEMBER 2010.
[12] Jeremy Davis (2010)" *Methods of Information Hiding and Detection in File Systems",* 2010 Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering.
[13] Richard Jensen, Chris Cornelis(2009)" *Hybrid Fuzzy-Rough Rule Induction and Feature Selection*", R. Jensen and Q. Shen are with the Department of Computer Science, Aberystwyth University, UK.

Kapil Sharma, Research Scholar, Done B.TECH (CSE) from L.L.R.I.E.T, Moga. Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Data Mining.

Sheveta Vashisht, Assistant Professor in Department Of CSE, Lovely Professional University, Phagwara, Punjab, India, have done B.Tech, M.Tech from Lovely Professional University, Research area is Networking, Security, Data Mining.

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

Richa Dhiman, Research Scholar, Done MSC (IT) from DOABA COLLAGE Jalandhar, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Data Mining,

Jasreena Kaur Bains, Research Scholar, Done B.TECH (CSE) from Lovely Professional University, Now doing M.Tech(CSE) from Lovely Professional University, Phagwara, Punjab, India, Research area is Network security using Neural Network.