International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

A hybrid Approach Using Rule Induction And Clustering Techniques In Terms Of Accuracy And Processing Time In Data Mining

Kapil sharma, Sheveta Vashisht, Richa Dhiman

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 using rule induction and clustering techniques in terms of accuracy and processing time in Data Mining we using induction algorithms and clustering 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- Rule induction, clustering, SOM algorithm, decision list induction, CN2

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 clustering 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. But sometimes it increases the time complexity so due to this we using clustering approach along with rule induction so we can reduce the time complexity.

Manuscript published on 28 February 2013.
***** Correspondence Author (s)
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,
Sheveta Vashisht, Assistant Professor in Department Of CSE, Lovely Professional University, Phagwara, Punjab, India,

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.

Hypothesis testing algorithms: A statistical hypothesis test is a method of making decisions using data, whether from a controlled experiment.

In statistics, a result is called statistically significant if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by Ronald Fisher.

Inductive logic programming: Inductive logic programming (ILP) is a subfield of machine learning which uses logic programming as a uniform representation for examples, background knowledge and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program which entails all the positive and none of the negative examples.

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

This research can improves the traditional algorithms like CART, SVM and PLS-DA, and presents supervised Rule Induction algorithms like CN2.

A. High coverage of data from training dataset with minimization of induced set of rules:

When we apply rule induction to mine the data from large dataset then we drives the number of rules to cover maximum of data with less induced set of rules. This obtains the best rule among all the induced set of rules from the training dataset. That particular rule covers the maximum dataset with less number of rules and high coverage of data.

B. Reduction in error rate from large dataset using induced set of rules in induction of rule and reduce time complexity:

When we drives induced set of rules in training dataset through induction of rule then we can obtain accurate result with low error rate. We choose the best rule among all the induced set of rules that covers maximum data .So that we obtain less error rate data from noisy data. We apply decision list induction along with clustering algorithm SOM then compare the result. It will reduce the time complexity and numbers of rules with accurate result.

C. Solution:

When we apply decision list induction to particular dataset then we check the result and then we use clustering algorithm SOM along with Decision List Induction algorithm or we can say that use decision list induction and SOM as hybrid approach then check the result. That method will give the best result with less numbers of rules and reduces the time complexity. This approach will be very beneficial in business domain and others domains in Data mining.

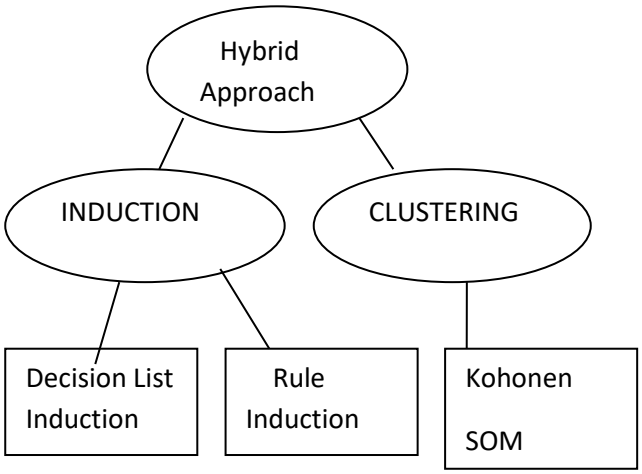

Figure 1 Hybrid approach

IV. CLUSTERING

SOM: SOM is a clustering method. Indeed, it organizes the data in clusters (cells of map) such as the instances in the same cell are similar, and the instances in different cells are different. In this point of view, SOM gives comparable results to state-of-the-art clustering algorithm such as K-Means.

SOM can be viewed also as a visualization technique. It allows us to visualize in a low dimensional representation space (2D) the original dataset.

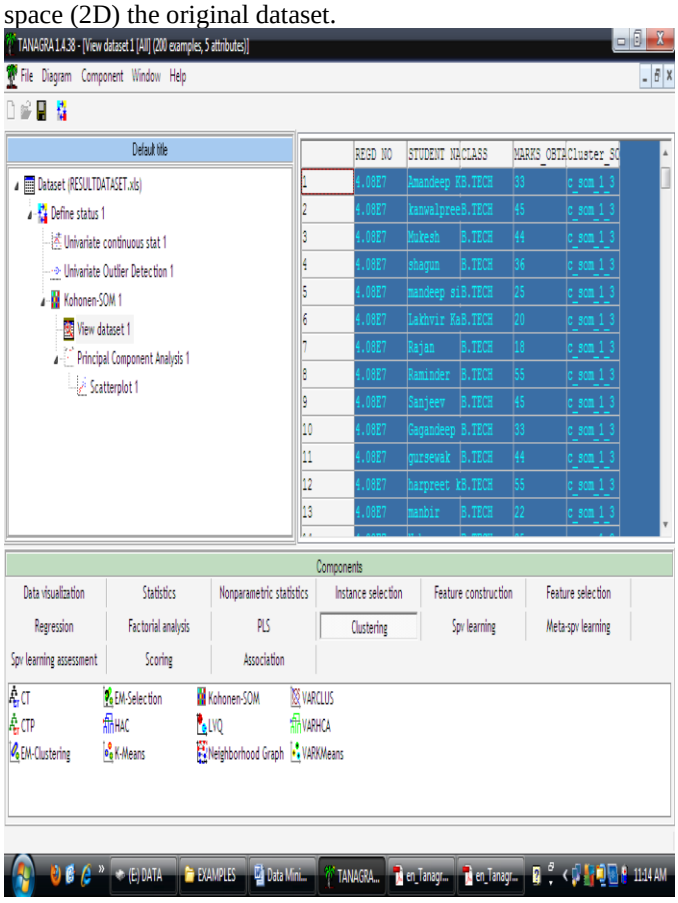

Figure 2 Graph using Kohnen-SOM

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

V. 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 Clustering 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 to mine the data using rule induction. 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 clustering algorithm SOM. 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.

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.