Graduate Research in Engineering and Technology (GRET)

Volume 1 | Issue 2 Article 3

October 2013

AN EFFICIENT APPROACH USING RULE INDUCTION AND ASSOCIATION RULE MINING ALGORITHMS IN DATA MINING

KAPIL SHARMA
Computer Science & Engineering, Lovely Professional University, Punjab, India,
kapilsharma701@gmail.com

SHEVETA VASHISHT
Computer Science & Engineering, Lovely Professional University, Punjab, India, sheveta.16856@lpu.co.in

AN EFFICIENT APPROACH USING RULE INDUCTION AND ASSOCIATION RULE MINING ALGORITHMS IN DATA MINING

KAPIL SHARMA, SHEVETA VASHISHT

Computer Science & Engineering, Lovely Professional University, Punjab, India
Email: kapilsharma701@gmail.com, sheveta.16856@lpu.co.in

Abstract- In this research work 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 induction rule via association rule mining we can generate number of rules for training dataset to achieve accurate result with less error rate. We also use induction rule algorithms like confidence static and Shannon entropy to obtain the high rate of accurate results from the large dataset. This can also improves the traditional algorithms with good result.

Keywords*: rule induction, association rule mining, decision list induction, Shannon entropy, data mining, confidence static*

I. INTRODUCTION

Data mining techniques are the result of a long process of research and product development. This evolution began when Business data was first stored on computers, continued with improvements in data access and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

A Separate and Conquer paradigm:

Among the rule induction methods, the "separate and conquer" approaches are very popular during the 90's. The goal is to learn a prediction rule from data If Premise Then Conclusion « Premise » is a set of conditions « attribute – Relational Operator – Value ». For instance, Age > 45 and Profession = Workman In the supervised learning framework, the attribute into the conclusion part is of course the target attribute. A rule is related to only one value of the target attribute. But one value of the target attribute may be concerned by several rules.

B Compared to classification tree algorithms:

Which are based on the divide and conquer paradigm, their representation bias is more powerful because it is not constrained by the arbores cent structure. It needs sometimes a very complicated tree to get an equivalent of a simple rule based system. Some splitting sequences are replicated into the tree. It is known as the "replication problem".

C Compared to the predictive association rule algorithms:

They do not suffer of the redundancy of the induced rules. The idea is even to produce the minimal set of rules which allows classifying accurately a new instance. It enables to handle the problem of collision about rules, when an instance activates two or several rules which lead to inconsistent conclusions.

We describe first two separate and conquer algorithms for the rule induction process. Then, we show the behaviour of the classification rules algorithms implemented by a tool.

Separate and Conquer algorithms

- **Induction of ordered rules(Decision list induction)**
- **Induction of unordered rules**

Induction of ordered rules (Decision list induction)

The induction process is based on the top down separate and conquers approach. We have nested procedures that are intended to create the set of rules from the target attribute, the input variables and the instances.

The rule based system has the following structure:

IF Condition 1 Then Conclusion 1
Else If Condition 2 Then Conclusion 2
Else If
Else If (Default rule) Conclusion M

Decision list induction algorithm:

Decision List (target, inputs, and instances)
Ruleset = ∞
Repeat
Rule = Specialize (target, inputs, instances)
If (Rule != NULL) Then
Ruleset = Ruleset + {Rule}
Instances = Instances – {Instances covered by the rule}
End if
Until (Rule = NULL)
Ruleset = Ruleset + {Default rule (instances)}
Return (Ruleset)

Induction of unordered rules: Ordered set of rules, when we read the i-th rule, we must consider the (i-1) preceding rules. It is impracticable when we have a large number of rules.

The classifier is now outlined as the following:

If Condition 1 Then Conclusion 1
If Condition 2 Then Conclusion 2
…
(Default rule) Conclusion M
(Ruleset)

II. BACKGROUND

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. IMPLEMENTED WORK

In this research work we used rule induction and association rule mining algorithms to maximize the accurate results with fast processing time. With these techniques we also minimize number of rules with high coverage of data. This research is beneficial in various business domains to mine useful data.

We take CREDIT-G.TXT dataset. In this we define the list of persons with their details so that we mine the data in the way that we provide them loan if their status will be good in various fields.

We applied rule induction and decision list induction with their various measure like confidence static, J-measure, misclassification, and Shannon entropy to minimize the numbers of rules with less error rate and high accuracy data.

We implemented the work using TANAGRA data mining tool. In which loaded the dataset. And performed different operations to get accurate results we applied firstly rule induction algorithm and calculated results as error rate, number of rules, with their respective accuracy and processing time. Similarly we applied decision list induction algorithm and calculate the numbers of rules and error rate with respective accuracy and processing time. We also calculate the performance measures of rule induction and decision list induction to maximize the accurate results with less number of rules and high coverage of data.

In this we also used Association rule mining algorithm A-priori that is used to induce number of rules from the data set. In this we also calculate lift and confidence so that we can induce numbers of rules according to item sets.

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.

The above figure is implemented work that is done in TANAGRA data mining tool.

The above figure describe the implemented work with corresponding algorithms rule induction, decision list induction and A-priori.

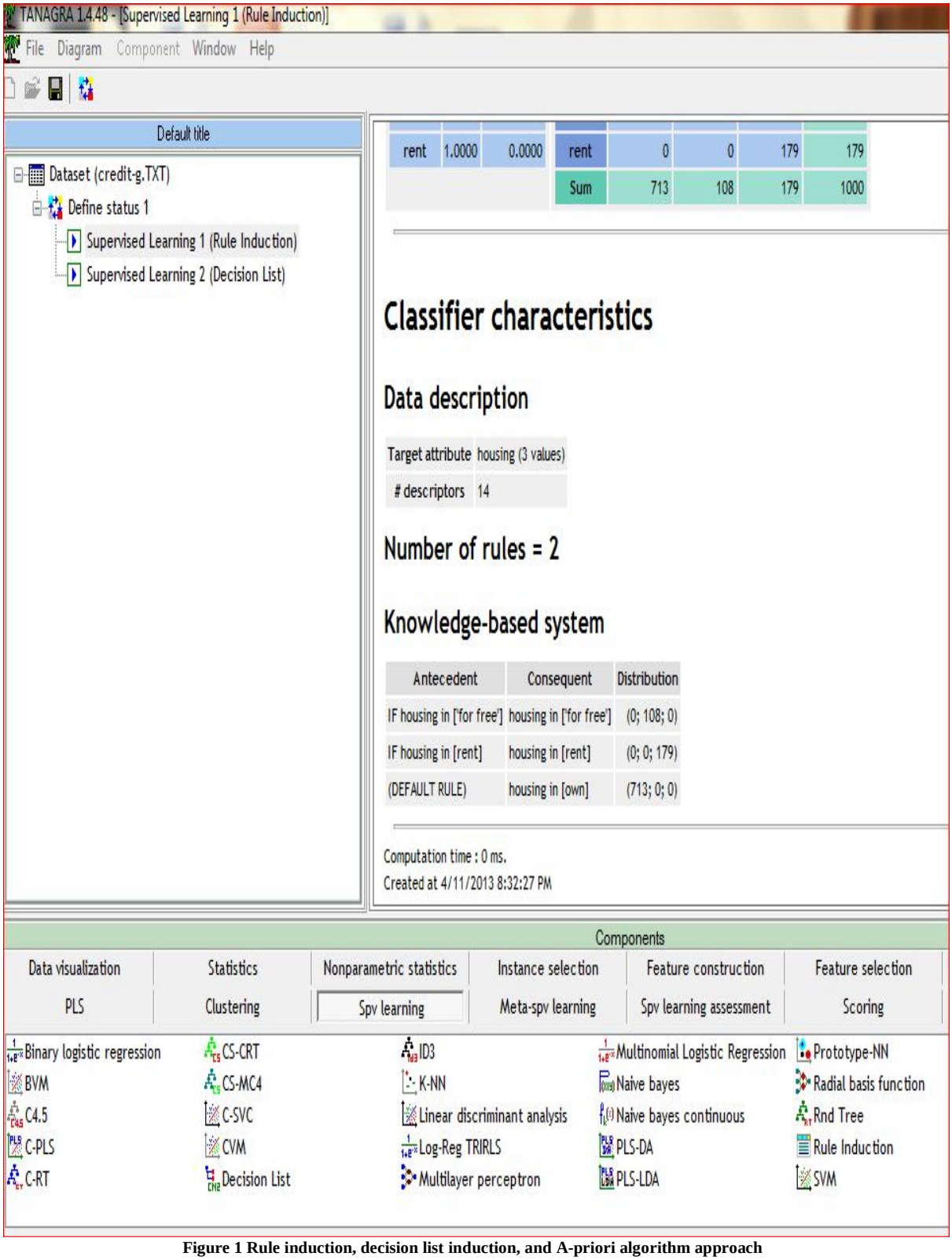

Figure 1 Rule induction, decision list induction, and A-priori algorithm approach

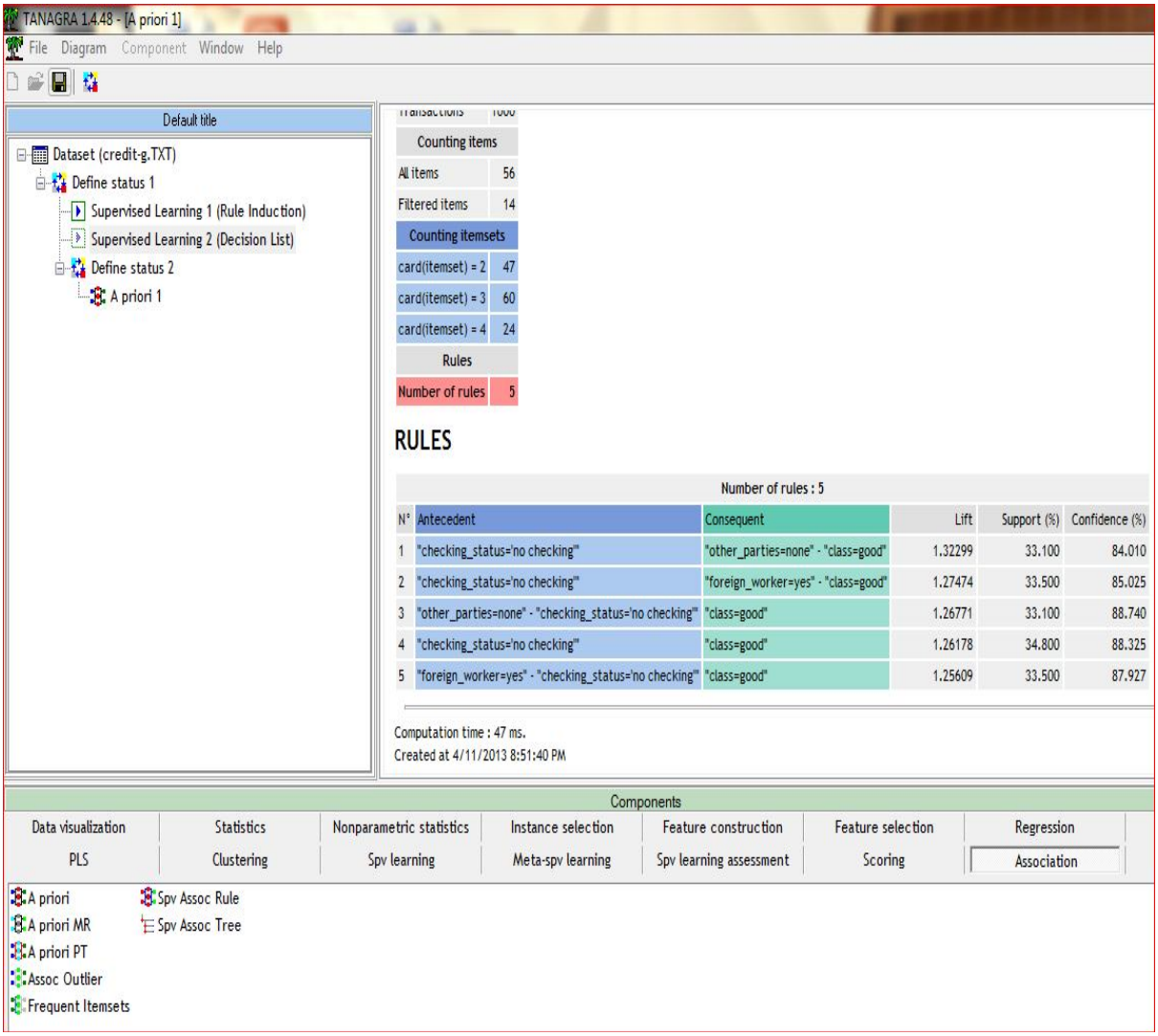

Figure 2 Numbers of rules using TANAGRA tool

IV. RESULTS AND DISCUSSION

- ***A. Error Rate:*** In Decision list induction, Shannon Entropy supervised algorithms is a best as compare to other algorithms because it has a minimum error rate 24.76%.
- ***B. No. of Rules:*** In rule induction, Misclassification Rate Static supervised algorithm is a best as compare to other algorithms because it has a minimum number of rules is 01.
- ***C. Computation Time:*** In rule induction, confidence static supervised algorithm is a best as compare to other algorithms because it has a minimum computation time is 176ms.
- ***D. Using Association Rule Mining:***
  Error Rate: 33%
  No. of Rules: 5
  Computation Time: 47 ms

The goal of this research is to reduce number of rules and maximize the data coverage to get more accurate results with less error rate. In this work we demonstrate the different measure of rule induction and decision list algorithm to reduces number of rules and maximize accurate date. This research can work on different business domains to obtain accurate results. In this work we implemented different algorithms to mine the data with less number of rules and less error rate.

V. CONCLUSION

In this Research paper, we wanted to highlight the approaches for the induction of prediction rules. They are mainly available into academic tools from the machine learning community. We note that they are an alternative quite credible to decision trees and predictive association rules, both in terms of accuracy than in terms of processing time. After analysis Order Rule Induction algorithm is more suitable to find accurate and consuming less access time to mine data with minimum error rate 24.76%. In this we also reduce the number of rules up to 1 with high coverage of data with less processing time. This research is beneficial in different business domains to get accurate result.

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. Fernando E. B. Otero(2011)" *A New Sequential Covering Strategy for Inducing Classification Rules with Ant Colony Algorithms*",IEEE.