International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-6 Issue-2, May 2017

A Review Paper: A Comparative Analysis on Association Rule Mining Algorithms

Gurpreet Singh, Sonia Jassi

Abstract: ***Data mining is a process which finds useful patterns from large amount of data. The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. [1] Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. But here we are going to discuss Association rules mining.***

Keywords: ***Data, Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm,***

I. INTRODUCTION

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.

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.

Key Concepts:

1. Frequent Itemsets: The sets of item which has minimum support (denoted by L_i for i^{th}-Itemset).
2. Apriori Property: Any subset of frequent itemset must be frequent.
3. Join Operation: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself.
4. Find the frequent itemsets: the sets of items that have minimum support – A subset of a frequent itemset must also be a frequent itemset
   a. if {AB} is a frequent itemset, both { A} and { B} should be a frequent itemset
   b. Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
5. Use the frequent itemsets to generate association rules.

The Apriori Algorithm : Pseudo code

- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code:

```
Ck: Candidate itemset of size k
Lk : frequent itemset of size k

L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
    Ck+1 = candidates generated from Lk;
   for each transaction t in database do
        increment the count of all candidates in Ck+1
      that are contained in t
   Lk+1  = candidates in Ck+1 with min_support
   end
return ∪k Lk;
```

The whole point of the algorithm (and data mining, in general) is to extract useful information from large amounts of data. For example, the information that a customer who purchases a keyboard also tends to buy a mouse at the same time is acquired from the association rule below:

Support: The percentage of task-relevant data transactions for which the pattern is true.

Support (Keyboard -> Mouse) =

$$\frac{\text{No. of transactions containing both Keyboard and Mouse}}{\text{No. of total transactions}}$$

Confidence: The measure of certainty or trustworthiness associated with each discovered pattern.

Confidence (Keyboard -> Mouse) =

$$\frac{\text{No. of transactions containing both Keyboard and Mouse}}{\text{No. of transactions containing (Keyboard)}}$$

The algorithm aims to find the rules which satisfy both a minimum support threshold and a minimum confidence threshold (Strong Rules).

A PRIORI MR (molecular replacement) component ASSOCIATION tab is an experimental tool which supplies several interesting measures for evaluating rules. There are widespread measures such as confidence, support, etc; there are also less known measures such as those based on the test value principle.

Revised Manuscript Received on 30 May 2017.
***** Correspondence Author
Dr. Gurpreet Singh, Professor & Head, Department of Computer Science & Engineering, St. Soldier Institute of Engineering Technology, Jalandhar (Punjab), India. E-mail: saini3077@gmail.com
Er. Sonia Jassi, M.Tech Scholar, Department of Computer Science & Engineering, St. Soldier Institute of Engineering Technology, Jalandhar (Punjab), India. E-mail: sonajas815@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

In this experiment I will show to implement the A PRIORI MR component, how to set the parameters in order to obtain more or less rules, and how to read the results.

Dataset

I use a modified version of the Banking CREDIT dataset. It depicts the characteristics of customers.

II. THE A PRIORI MR COMPONENT

Creating a diagram and importing the dataset Tanagra can load directly an Excel file format (XLS) even if the Excel software is not available on our computer. There are two restrictions for handling the data file: it must not be currently opened in other tool; the dataset must be in the first worksheet Into Tanagra, I click on the FILE / NEW menu; I select the CREDIT_ASSOC.XLS data file.

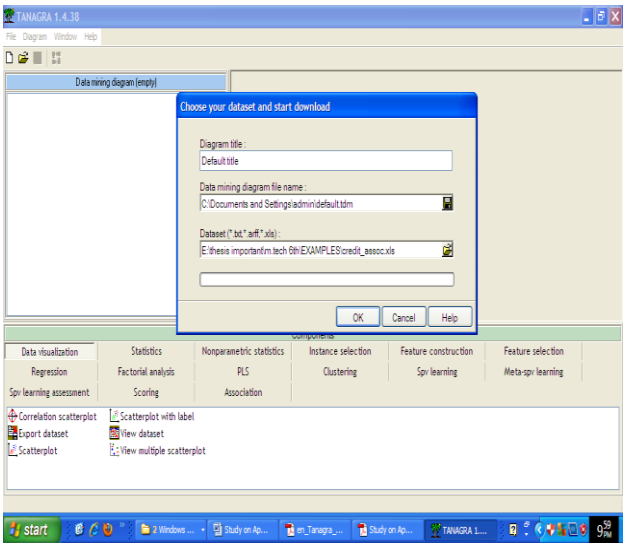

Figure 1

17 attributes and 1000 individuals from the DATASET sheet are now available for the analysis.

A. A Priori MR

I insert the DEFINE STATUS component into the diagram, using the shortcut into the tool bar. I set all the variables as INPUT.

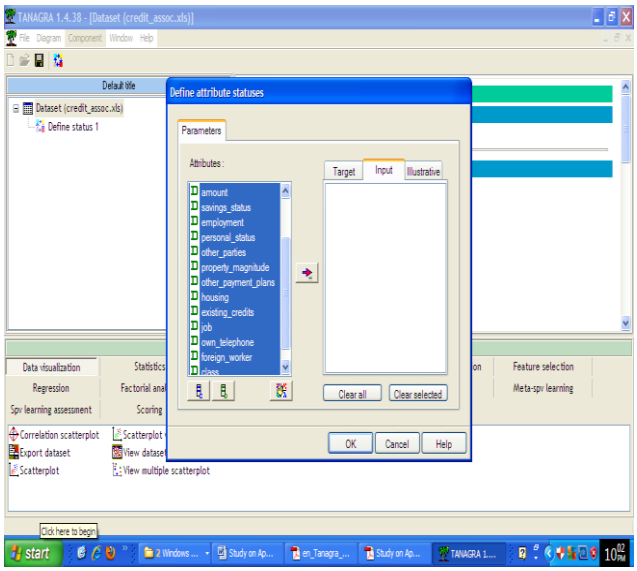

Figure 2

Then, I add the A PRIORI MR component (ASSOCIATION tab). I click on the VIEW menu. I obtain a first result with the default settings.

Table 1

A-Priori parameters	
Support min	0.33
Confidence min	0.75
Max rule length	4
Lift filtering	1.10
Learning set ratio	1.00
Repetition	1

- SUPPORT MIN states the minimum support of the mined rules;
- CONFIDENCE MIN states the minimum confidence;
- MAX RULE LENGTH states the maximum number of items (attribute = value) of rules ;
- LIFT FILTERING states the minimum of LIFT.

These settings allow to restrict the number of mined rules. The first 3 parameters determine the computation time and the memory occupation; the last one (LIFT) filters only the displayed rules.

Learning Set Ratio states the proportion of the dataset used for the learning phase. Indeed, Tanagra can subdivide the dataset into two parts: the first is used to extract the rules; the second is used to asses them. This framework is often used in the supervised learning task. Here, it allows evaluating the reliability of the numerical indicator associated to the rules. The default value is 1 i.e. all the observations are used in the learning phase.

Repetition is the number of the replication of the Monte-Carlo procedure for the computation of the VT-100.

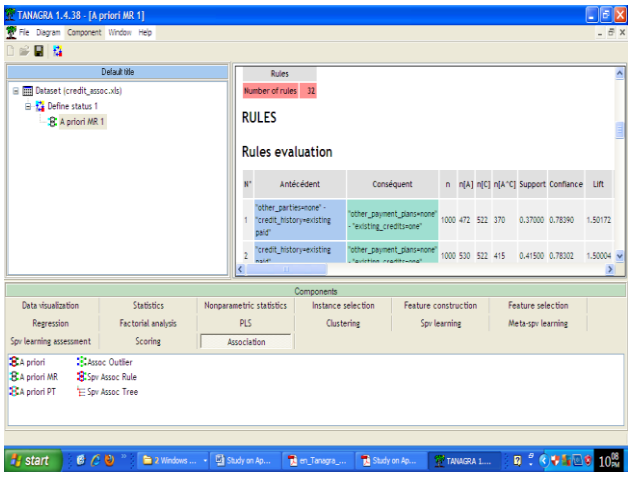

Figure 3

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

32 rules are generated. They are sorted in a decreasing order according the LIFT measure

The results supplied by the A PRIORI MR Component the ITEMS part describes the number of mined frequent item sets (those of which the support is higher than SUPPORT MIN), gathered by cardinality.

The total number of items is 66, 19 of them are frequent. For the item sets of cardinality equal to two, we have 68; etc.

B. A PRIORI PT (Pass –Through)

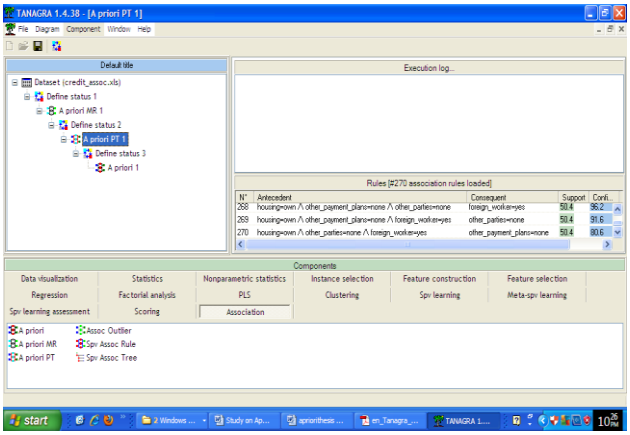

Figure 4

I obtain 270 association rules

III. EXPLORING THE RULES

Filtering the rules is an interesting functionality. But, I must know how to set appropriately the Parameters. It is not obvious. Tanagra supplies an option which allows to deeply exploring the extracted rules. I can copy the results in a spreadsheet, then I can use the abilities of Excel to organize (sort) the rules in different ways, according various interestingness measures.

I click on the COMPONENT / UNFORMATTED COPY in the main menu. Then, I launch EXCEL.

I paste the results. With the various tools supplied by Excel, I can explore deeply the rule base. In the screenshot below, I have sorted the rules according the LEVERAGE criterion.

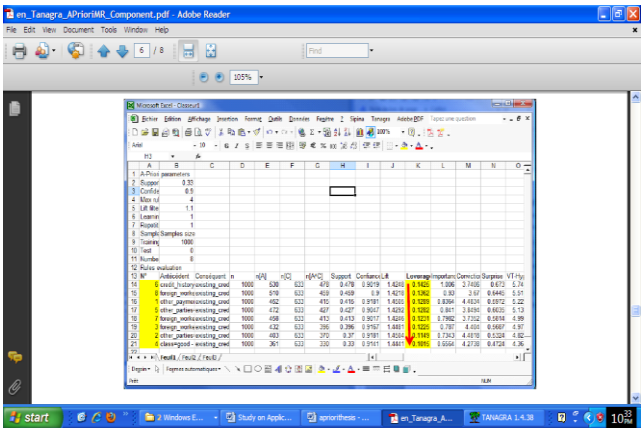

Figure 5

Subdividing the dataset into training and test sets I can subdivide the dataset into two parts for computing the rules (train set) and assessing them (test set). I obtain thus an honest estimation of the interestingness measure associated to the rule.

I click on the PARAMETERS menu. I set the LEARNING RATIO to 0.66 i.e. 660 observations on 1000 are used in the training phase; 340 in the testing phase.

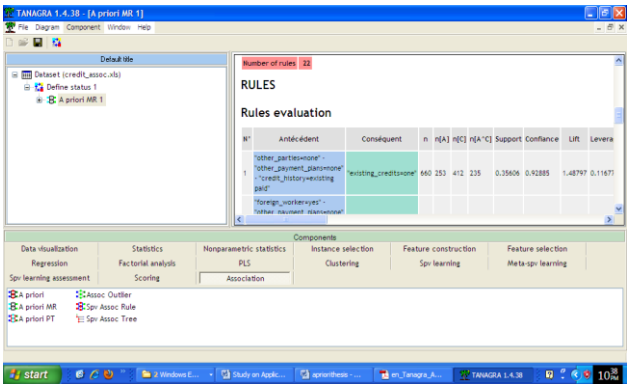

Figure 5

Tanagra extracts 22 rules. In the right part of the table enumerating the rules, after the TEST column, I have the interestingness measures computed on the test set. The comparison of the values obtained on the train and the test set allows assessing the stability of the rules.

IV. CONCLUSION

The A PRIORI MR component of Tanagra extracts rules from data using the A PRIORI algorithm. It differentiates oneself from other by offering additional tools for exploring and assessing the mined rules: original measures based on the "test value" principle allow to evaluate differently the rules; the ability to copy the results into a spreadsheet allows a more detailed exploration of the rule base; by subdividing the dataset into train and test sets, we obtain a more reliable values of the interesting measures of rules.

REFERENCES

1. Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
2. YanxiLiu, "Study on Application of Apriori Algorithm in Data Mining", IEEE Second International Conference on Computer Modeling and Simulation, 978-0-7695-3941-6/10, 2010
3. Wangjie Sun, ZhigaoZheng "An Advanced Design of Data Mining Algorithms", IEEE 978-1-4244-6943-7/10, 2010
4. S.P.Latha"Algorithm for Efficient Data Mining", IEEEInternational Conference on Computational Intelligence and Multimedia Applications, 7695-3050-8/07, 2007
5. XU Chi, ZHANG Wen Fang "Review of Association Rule Mining Algorithm in Data Mining"978-1-61284-486-2/11/2011 IEEE.
6. AmolGhoting, Gregory Buehrer "Knowledge and Cache Conscious Algorithm Design and Systems Support for Data Mining Algorithms", 1-4244-0910-1/07/2007 IEEE.
7. Wan Muhammad Zulhafizsyam Wan Ahmad, ShahidaSulaiman, UmiKalsomYusof "Data Mining Technique for Expertise Search in a Special Interest Group Knowledge Portal", 2011 3rd Conference on Data Mining and Optimization (DMO), 28-29 June 2011
8. Mining Association Rules Based on Cloud Model and Application in Credit Card Marketing, Yan-Li Zhu, Yu-Fen Wang, Shun-Ping Wang, Xiao-juanGuo, and 2010 Asia-Pacific Conference on Wearable.
9. Richard Lackes, Chris B6rgermann, Erik Frank, "Determining Factors of Online Auction Prices Analysis with Data Mining Algorithms on the basis of Decision Trees", 978-1-4244-8845-2/10/ 2010 IEEE
10. Chyouhwa Chen, Shi-Jinn Horng, Chin-Pin Huang, "Locality sensitive hashing for sampling-based algorithms in association rule mining", 2011 Elsevier Ltd.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.