Quality Metrics for Non-Redundant Association Rules Mining

Muhammad Junaid[1] and Muhammad Nadeem[2]
mjunaididrees@yahoo.com
SZABIST
Karachi, Pakistan

Abstract*: In dealing with association rule mining we often come across with a lot of efforts have been given to efficiency and usefulness but rather a quality has been one of the core area that has not been focused by the researchers that is why very often we see using association rule mining so many rules can be exposed or derived and overcome the user but more significantly some of these rules could be unnecessary and redundant specially in the case of multi-level datasets and carry nothing new to the domain. In this paper we proposed an approach for measuring quality parameters and rather how the quality of a desire rule is to be measured to find out its usefulness, interestingness and importance.*

Keywords: *Association Rules Mining, Data Mining, Rule Induction Quality Parameters for Association Rules, Rules Redundancy.*

1. INTRODUCTION

Association rules mining is one of the widely-used data mining technique that specify regularities and similarities within set of items of the similar items inside set of transactions. It finds interesting associations or correlations interactions among large set of data items. Most classical and widely-used example of association rule mining is Market Basket Analysis. In supermarkets date are collected by means of bar-code scanners each record represent all the items purchased by a customer on a single purchase transaction. Managers might be interested to know the consistency in certain group of items frequently purchases together to conduct surveys, know the consumer behavior, cross selling and catalog design based on buying patterns.

This type of information is fetched using Association rules that provide it in the form of "if-then" statements. These rules are figure out from the data and are differ with the if-then rules of logic because they are probabilistic in nature.

The "if" part is said to be an antecedent and the other part is called a consequent (the "then" part), in association rule mining there are two numbers representing as a set that expresses the degree of uncertainty about the rule. These two the antecedent and consequent are sets of items (called item sets) that do not have any items common.

1.1 Inspiration and Rationale

When we discuss association rule mining it is repeatedly stated that Enormous works and efforts has been done on efficiency and effectiveness in tem of speed but on the other hand less efforts has been given to quality parameters for finding rules in tem of their accuracy and applicability in different environments and circumstances that most of the time a specific rule is either rejected or ignored due to its lack of information from the data source that if selected with its supporting parameters might have some sort of effectiveness. The reason behind this problem is that these rules could be redundant and do not possess new knowledge with them although some effort has been directed to deal with redundant rules in flat datasets [4, 5, 6]. However these datasets could have hierarchies and taxonomies or multiple perception levels and consequently redundancy in these datasets should be focused on.

This concern is one of the parts of this research but most important thing that is the main topic of this research is that when it comes to association rules mining is not just dealing with redundant rules but somewhat how and what quality parameters are setups to weigh a association rule to determine if it is interesting, useful and important and in what circumstances.

It is important to consider that how the How the quality of an association rule is find out but on the other hand there is no formal definition of quality and interestingness [2].

Presently there is a collection of various measures available which is to a certain extent to the traditional methods of support and confidence being considered insufficient [7]. With this reason and the fact that many of these quality measures have given considerations to association rules mainly to those derived from flat datasets, but here we suggest that for better measure for weighing and measuring the interestingness or quality of multi-level and cross-level association rules those derived from datasets having multiple perceptions levels is needed.

1.2 Data Mining

One of the best technology to extract the hidden analytical information from the databases is data mining that has enormous potential to assists organizations to put their considerations and move their focus towards the information lying in their large database for utilizing it in a better way that help business growth.

These data mining tools are used to get information as they predict and estimate future trends and activities helping organizations to make their businesses to take proactive and useful decisions. These tools can be used as to answer business problems that were traditionally complex and difficult to solve they not only offer businesses to solve their long time facing problems but also make it easy for them to get unseen patterns, predictive information that most of the time might miss because as it lies outside their expectations.

1.3 Data Mining Foundations

Long research and product developments make it possible for the data mining techniques to come out as a great technology for decision making, finding analytical information and hidden patterns.

When it was not possible to find out some sort of information from the large databases then the evaluation of data mining technology began. This evolutionary process of data mining is beyond the demonstration data access step towards the navigation to prospective and proactive information delivery.

1.4 Understanding Rule Induction System

One of the major forms of data mining is rule induction and it is perhaps one of the most common forms of knowledge discovery in an unsupervised learning systems. It is also termed as the form of data mining that somewhat be similar to the process whenever it comes in mind bout a data mining specially when dealing with large databases that has some sort of parameters and these parameters of a rule indicate the interestingness and focused on something about the hidden patterns and the information that were outside the expectations and sometimes were unaware of these hidden patterns that might be very helpful to businesses growth within the databases.

It is stated that Rule induction on a data base may be a very huge undertaking where in the first step all likely and expected patterns and samples are analytically fetched and pulled out of the data and then in the second step some sort of accuracy and significance in term of checks and constraints are employed to these patterns that finally let us know that how valuable and strong the pattern and the sample we are interested is and what are the chances of Its occurrences that it would appear again. Typically these above cited rules are comparatively very simple to understand likewise if we talk about the market basket database transactions usually termed as market basket analysis where items are normally scanned in a consumer market basket one might be very much interesting in finding interesting correlations and relations between different set of items sets within the transactional databases like wise Whenever potato is purchased then onion is also purchased with 70% of the time and this pattern occurs in 9% in all shopping baskets similarly whenever a charismas tree is purchased from a superstore then a light gift is also purchased 50% of the time and these two items are purchased mutually in 12% in all shopping baskets.

2. RELATED WORK

As its introduction in [1], association rule mining is one of the extensively used data mining technique and the aspire of this technique is to extract and find out repeated and frequent hidden patterns, interesting correlations relations and associations amongst the sets of items in lying in the vary large transactional databases lot of work in this field has focused and concentrated on finding out more and more proficient and resourceful ways to discover all of the rules possible. But as far as quality is concern less work has been done. Presently technique towards is to determine which rules are redundant i.e. unnecessary and remove them, thus reducing the number of rules a user has make sure that not reducing the content of information content .

These approaches explain many promise and work by Xu & Li [8] that shows that a decrease of over 80% for exact basis rules can be achieved. Recently on the basis of discovered approximate rules the work was extended that also include removing redundancy [9]. Although it was good approach but it is stick on datasets where all items are on the same concept level. Therefore they do not consider redundancy among the itemset that have hierarchies.

"These approaches have been adapted originally made for single level datasets into techniques usable on multi-level datasets. Han & Fu's work" [10] termed as one of the initial approaches that have proposed to find out repeated itemsets in multi-level datasets and then later it was revisited [11]. The primary focus of this work was on finding repeated itemsets at every level in the dataset but it did not include cross-level itemsets i.e. those itemsets that consist of items from two or more different levels.

Regardless of every thing the focus of all the work, was on finding the repeated and frequent itemsets efficiently, effectively and significantly but the issue of quality and redundancy in single level datasets was remain in its form. Some sort of work by Han & Fu [3] that discusses the removing rules from the discovered database rules which are hierarchically redundant, but it relies on the user giving an expected confidence.

3. PROBLEM STATEMENT

Besides redundancy within association rules there is no proper definition of quality and/or interestingness that tells how the quality of a rule is to be measured to determine whether it is useful, interesting, important etc. and at what extent and the bifurcation of the domain. This would eventually leads towards the development of redundant rules that bring no new knowledge and rules to the transaction databases and wastage of efforts and also unable to measure the quality in term of accuracy and its correctness for specific a domain.

4. SOLUTION TO PROBLEM STATEMENT (MATHEMATICAL APPRAOCH)

let suppose that data is being extracted from the transaction database that contain all the relevent information reagrding transactions.

Let $P=\{P_1, P_2, P_3.....P_n\}$ be a set of n binary attributes called items for products and $T=\{T_1, T_2, T_3.....T_n\}$ be a set of transactions which has a unique ID for each transaction and contain subset of items of the set P. A quality parameter namely Buoyancy is defined as a proposition of the form $a \Rightarrow b$, where (a , b) belongs to set I and $a \cap b$ <>0 and a <> b ; Buoyancy is defined as

1. Count occurrences of all items as support parameter i.e. $\text{Support}=\text{Support}\{a\}$
2. For each bill, consisting of items $P=\{P_1, P_2, P_3.....P_n\}$ get all pairs (a , b) in term of three tuples followed by two and store them.
3. Less all those transaction both in two and three tuples which do not meet the desired results.
   Buoyancy=

$$Confidence=\left(\text{Support}\{a,b\} \middle/ \text{Support}\{a\}\right)-\text{Support}\{a\}$$

4. At the end of the extract cycle, determine I1 the items with counts at least s i.e support
5. Finally, (a ; b) can be a candidate in the target table only if all of the following are true:
   i. a is in I.
   ii. b is in I
   iii. (a ; b) to a frequent rules
6. If a pair meets all three conditions, add to its count in a target table with relevant parameters, or create an entry for it if one does not yet exist..

4.1 Solution to Problem Statement (ETL scripting) using Star Schema Dimensional Model.

Extracting the quality oriented rules in a separate database from the transactional database ETL is one of the best way to adopt because using this we can setup a new staging data warehouse which will not only contain the non redundant rules but also be a performance oriented database due to its nature of denormalization and will not disturb the original and the source database. In the first phase of our approach we need to select dimension model and here we have Star schemas as our dimension model schema the star schema consists of fact tables and dimension tables. Usually quantitative and factual data about a business process is considered as a part of a Fact tables and this is the information that usually being queried. The data within the fact tables is often in numerical form and may have several columns and lots of rows while on the other hand Dimension tables are usually kept small and can consists only descriptive data that replicate the dimensions, or qualities, of a business. Simple and complex structure queries are used to joins the fact and dimension tables and imposed some sort of constraints and conditions on the required and desired data in order to get selected information as a result[3]

The fact table in our approach i.e. Nonredundent_rules_Facts would contain all the factual values and quantifiable parameters needed to identify non redundancy and accuracy of rules. The accuracy of a rule as defined above using mathematical approach i.e. buoyancy will be calculated in between ETL scripting. As it has already been defined time is one of the core angle and dimension to be considered as a major dimension value to view the quantifiable parameters in defining accuracy of association rules in our approach we have written a procedure for building a dimension table named as D_time that would automatically be updated on daily basis. In order to view the different rules to measure to the quality metrics in different times the table will consist of day, month and year to view a specific rule either on daily, monthly or yearly basis.

4.2 Proposed ETL Script:

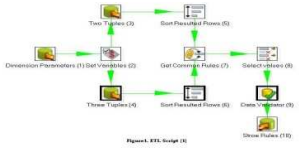

Figure 1

ETL defined as for extraction, transformation and loading, the processes that enable organizations to circulate their data from multiple sources in following the different phases of formatting, cleansing, and loading into another database, its analysis a data mart or a data warehouse, or on another required and desired system is often required to carry a business process efficiently and accurately.

5. CONCLUSIONS

Association Rule mining is being given much importance in the field of data mining but how to rank a rule in term of its accuracy and effectiveness less importance has been given, in this paper we have tried to find out the quality parameters and metrics to weigh a rule weather it is applicable, effective and accurate to be considered to implement, in the initial phase we focused on the quality issues and with help of mathematical representation provided a buoyancy factor that excluded all the transactions that results either null values or the less likelihood of occurrences i.e. if a rule has 20% support factor and 50% confidence factor we cannot stop here but also have to less all those transactions which are either results in null values or less likelihood of occurrences. Let say 20% are those transactions that results either in null values or other probabilistic values other than expectations and then less this 20% from confidence factor finally we will have 30% accuracy for a specific rule.

In the second phase we have written an ETL script to view a rule from different angle i.e. time dimension. But a lot of works are there to research and in future I will be focusing on more quality metrics of association rules and taking in to account more dimensions to view a rule facts in different dimensions.

ACKNOWLEDGMENT

First of all I would like to thank the ALMIGHTY ALLAH who helps me and enable me so that I have successfully completed my IS. I am very much thankful to Mr. Muhammad Nadeem for his guidance and encouragement towards this success.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in ACM SIGMOD International Conference on Management of Data (SIGMOD'93), Washington D.C., USA, 1993, pp. 207-216

[2] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey," ACM Computing Surveys (C SUR), vol. 38, 2006.

[3]. Software Information Center, website

[4] G. Shaw, Y. Xu, and S. Geva, "Eliminating Redundant Association Rules in Multi-level Datasets," in 4th International Conference on Data Mining (DMIN'08), Las Vegas, USA, 2008, p. To appear

[5] Y. Xu, Y. Li, and G. Shaw, "Concise Representations for Approximate Association Rules," in IEEE International Conference on Systems, Man & Cybernetics (SMC'08) Singapore: IEEE, 2008, p. To appear.

[6] Q. Zhao and S. S. Bhowmick, "Association Rule Mining: A Survey," Nanyang Technological University, Singapore, 2003.

[7] K. McGarry, "A Survey of Interestingness Measures for nowledge Discovery," The Knowledge Engineering Review, vol. 20, pp. 39-61, Mar 2005

[8] Y. Xu and Y. Li, "Generating Concise Association Rules," in 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07), Lisbon, Portugal, 2007, pp. 781-790.

[9] Y. Xu, Y. Li, and G. Shaw, "Concise Representations for Approximate Association Rules," in IEEE International Conference on Systems, Man &

[10] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in 21st International Conference on Very Large Databases, Zurich, Switzerland, 1995, pp. 420-431.

[11] J. Han and Y. Fu, "Mining Multiple-Level Association Rules in Large Databases," IEEE Transactions on Knowledge and Data Engineering, vol. 11, pp. 798 - 805, Sep/Oct 1999.