A New Algorithm to Mine Most Interesting Association Rules

Deepika Aggarwal, Kritika Dixit, Dharmveer Singh Rajpoot

Department of Computer Science and Engineering
Jaypee Institute of Information Technology, Noida, India

Abstract: **Association rule mining algorithms generally mine a lot of rules quite a lot of which are often not useful in a lot of applications. These rules are called uninteresting rules. This paper discusses a new algorithm to mine the most interesting rules out of millions discovered. Also, a comparison with the existing algorithms has been presented.**

***Keywords*: Frequent, Pattern, Rule Mining, Correlation etc.**

I. INTRODUCTION

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

An association consists of two parts, one is antecedent and the other is consequent. Antecedent is defined as an item which is found in the data under observation. A consequent is an item that is found combined with the antecedent.

Association rules are generated by analyzing data for frequent if/then patterns and using the two criteria of *support* and *confidence*

For identify the most important relationships. *Support* tells how frequently the items are found in the database. *Confidence* tells about the number of times the if/then statements are found to be true.

So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are frequently bought together. For example, peanut butter and jelly are often bought together because a lot of people like to make sandwiches.

Application

Application of association rules in data mining are useful for analyzing and predicting customer nature. Programmers use association rules to made programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

The main applications of association rule mining:

- Basket data analysis - is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.
- Cross marketing - is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
- Catalog design - the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Algorithms for Association Rule Mining:

Various algorithms for association rule mining have been presented over time.The most well known among them and then a few new have been discussed here.

1) Apriori algorithm:

- Find the frequent item-sets: the sets of items that have minimum support.A subset of a frequent itemset must also be a frequent itemset .Iteratively find frequent itemsets with from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules.

2) F-P Tree algorithm

This algorithm generates frequent item sets by looking into the database only twice with the help of a tree structure design.Firstly we generate the frequent pattern tree and then we generate frequent patterns from the FP Tree.This algorithm does away with the

limitations of apriori algorithm such as complex candidate generation process and too many scans of the databse.

3) ECLAT algorithm(Equivalence Class Clustering and Bottom up Lattice Tarverser):

The algorithm has the following steps:

Step 1: get the transaction id list for each item.

Step 2: apply minimum support condition criteria and if it is satisfied then go to step 3 else stop

Step 3: intersect the transaction id list of {a} with all other items to get {a,b}, {a,c},.....so on

Step 4: Repeat steps 1 to 3.

4) Continuous association rule mining algorithm

Its an entirely new algorithm to compute large itemsetsonline.Atmost two scans of the database are needed to produce all large itemsetshere.It has 2 phases. Phase 1 constructs a lattice of all possible large itemsets.Phase 2 involves removing itemsets which have support value less than the minimum specified.Then this phase determines the exact number of occurrences of each of the remaining itemsets and continuously remove the small itemsets.

5) Rapid Association Rule Mining:

It makes use of an efficient tree structure to represent the databse and does not generate the candidate itemsetsanymore.It involves preprocessing through trieItemset(TrieIT).So, this algorithm eliminates second tym searching of the databse and generates 1-itemsets and 2-itemsets quickly through support oriented trieItemset(SOTrieIT).

Major Problems:

The major problems in the association rule mining algorithms are:

1) The existing association rule mining algorithms usually generate a huge quantity of rules without guaranteeing that all of them will be relavant . In short , not much attention has been paid towards the interstingness of the rules generated which is an important issue in association rule mining. Objective measures of interestingness have been propose like support and confidence of a rule but these are not enough . Subjective measure are becoming incresingly important.

2) In F-P tree algorithm , the construction of F-P tree takes a lot of time especially for a large database and it is also very expensive to build . It takes a lot of memory.Also it is difficult to use in interactive mining systems and is not suitable for incremental mining also.

3) Another important factor is to be able to determine the comprehensibility of an extracted rule i.e. the poorly understandable rules should be eliminated.This aspect of rule quality is often ignored by people beacuse of the subjective nature of comprehensibility.Past experience and the person's knowledge of the concerned domain is an important factor in deciding comprehensibilty.

4) Most of the association rule mining algorithms require the user to set thresholds –the minimum support and the minimum confidence.Therefore in order to find the right values of support and confidence , the user must possess some level of expertise , to find the best rules. In general, most of the data mining tools are too complex (have too many parameters)for common , everyday users like educators in e-learning who are non experts in data mining.More emphasis is given on power and flexibility of the algorithms rather than simplicity.

Chosen Research Gap:

Out of the various research gaps examined above , our team would like to work upon resolving the problem of efficiently finding ways to more rigorously examining the rules for their interstingness and hence reducing the output set of the data mining algorithms so that the rules obtained are for more useful.

II. LITERATURE REVIEW

In the previous years , finding out most interesting association rules has become very important as the number of rules generated by association rule mining algorithms are quite huge. The literature analyzed has several ways to do this some of which we have discussed in our paper.

One such approach was Hierarchical interestingness measure. We compared 35 flat and hierarchical interestingness measures to find out the ones which are most successful. Later , an entirely new class of interestingness measures was developed to solve the problem.

Advantage: The hierarchical methods are advantageous because they do not suffer from the need of parameters and also the steps required after processing are not costly. These methods are also beneficial because they produce rules according to their rank i.e. importance and arrange them rank-wise.

Disadvantage: These hierarchical methods have disadvantages too. While some like Jac and Acnf produced very good results , Cnf on the other hand produced dissatisfactory ones.

The other algorithms we studied include a multi objective evolutionary method to find rules with a very low computation cost. This algorithm is an extension of already existing evolutionary approach based on decomposition of the association rules obtained.

Advantage: The rules obtained by this method have very few attributes which makes comprehensibility feasible. Also, the trade-off obtained between number of rules, support and coverage is worth considering. This algorithm has a linear growth with the increase in the number of attributes in the data set whereas as far as the classical association rule learning methods are concerned , they have exponential growth when the number of attributes is greater than 10.

Performance Measure: Three objectives are maximized for this problem: interestingness, comprehensibility, and performance. Performance represents the attempt to improve the coverage of the dataset in order to extract more interesting knowledge from it. Performance is the product of support and CF , which allows to mine a set of accurate rules with a good trade-off between local and general rules.

Advantage: This DISJOINT algorithm has better performance when compared to random sampling. It selects the right tables that can be displayed to experts in the domain. This is because it selects table with two special characteristics. One is selecting tables that differ with each other in relative ranking. The other is selecting tables that generate conflicts in ranking.

One important finding we made is that the methods which have been proposed to find interesting rules use the dataset as an entirely complete entity and the doing the process of filtering of discovered rules in various manners but this is generally not enough. These techniques are unable to solve the issue of trust to be invested in these rules. Will it be possible to trust the rules obtained?

III. PROPOSED METHOD

Discovered rules with the given confidence and support threshold are large in number. All these rules are not useful, since they are heavily redundant in information such as method based on clustering technique.

In our method, rules are clustered on the basis of rule consequent information .So group of rules are in the form. X(i)□ y for i=1,2,3….n

That means we have different rule antecedents collected in a group for a same rule consequences y.

The next step involves selecting a representative subgroups RS(i) of each of these clusters base on measure of support as well as confidence.

Support of association rule:- X□Y= p(X,Y)

no of customers who bought x and y/#Total no of customers

Confidence of association rule:-X□= p(Y/X)

no of customers buying Xand Y/# customer buying X

All the rules which satisfy both minimum support as well as confidence criterion are grouped into the representative.

In this way we can obtain few rules for each item .

We can further prune each representative as:-

For i=123…n ,ifRs(i) has two rule X(i) □Y and X(j)□ Y for particular item y

Then if (X(i)int Y(i)) belongs to X(i) then ignore X(j)□Y

Or if (X(i) int Y(i))belongs to X(j) then ignore X(j)□Y (int = intersection)

In this way we obtain reduced representative sets Rs (i) for each item where i=1….n

Our final reduced set of association rules F, will be union of all the representative sets

i.e F=Rs(1) U Rs(2) U Rs(3)……Rs(n) for n items.

The proposed approach has the following steps:

1. 1.Partitioning the dataset: The original dataset D is first partitioned vertically into a number of blocks (or sub-datasets) DI, D2, ..., D, according to the time periods, TI, T2, ..., T,, in which they were collected, e.g., years or months.

2. Mining rules from sub-datasets: We then mine rules from each sub-dataset. Pruning is also performed to remove those insignificant rules.

3. Analyzing rules over time: After all the required information about support and confidence is obtained , these rules are then analyzed to ensure

that the user of the system gets different varieties of interesting rules.

IV. QUALITATIVE ANALYSIS OF RESULTS

Using these statistical tests, it is possible to identify stable rules and trend rules, and at the same time remove those unreliable (unstable) rules. Experiment results showed that the proposed technique is very effective and efficient.

We also studied an important rule mining algorithm named row enumeration algorithm to mine rules which are high in confidence value. This has been discovered to do mining in dense data sets. The main disadvantage of this approach is that quite a lot of rules which are low in support value but high in the confidence one get eliminated.

Advantage: The basic disadvantage of this approach has been overcome by another approach suggested called MAXCONF used for mining high confidence rules from dense data sets.

We came across a new technique has been developed to help find interesting rules from a set of discovered association rules. This interestingness analysis system (IAS) increases the user's present domain knowledge to analyze discovered associations and then rank those rules according to various criteria for interestingness, such as conformity.

This Interestingness Analysis System is very interactive and iterative in nature. In each round it asks the user of the system to specify existing domain knowledge. Then it starts discovering the rules. The discovered rules are then analyzed according to the criterion specified.

Working Procedure :

Do looping until the user decides to end:

- The user of the system specifies the knowledge already existing in the system.
- the system analyzes the rules found by the measures of conformity and unexpectedness.
- the user analyzes results generated, saves the interesting rules found and eliminates the unwanted rules.

Disadvantage: We notice that here subjective interestingness has been used which does not give an idea of all the interesting rules. Objective evaluation is missing here.

Advantage: However , this technique has been proved quite successful in filtering out the most interesting rules in quite a lot of practical situations.

This visualization system also allows the user of the system to save interesting rules discovered in an incremental fashion and remove unwanted rules.

V. CONCLUSION

Finding the most interesting association rules can be quite complex task. One of the problems is that many measures of interestingness do not work effectively for all datasets and are difficult to understand properly by the users.

To extract some different sets of more representative rules, subjective measures can be used too. This gives different type of information or knowledge for the same clusters. In addition, for increasing the quantity of top rules, some other measures such as cosine, coherence, jacard, all confidence, and recall/sensitivity give considerable number of representative rules. Especially, all-confidence, and recall/sensitivity give better results as observed by us.

REFERENCES

[1]. Fernando Benites and Elena Sapozhnikova "Evaluation of Hierarchical Interestingness Measures for Mining Pairwise Generalized Association Rules" .
[2]. Diana Mart´ın, Alejandro Rosete, Jes´usAlcal´a-Fdez, Member, IEEE, and Francisco Herrera, Member, IEEE "A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules.
[3]. Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma, Analyzing the Subjective Interestingness of Association Rules by National University of Singapore
[4]. Bing Liu, Yiming Ma , Selecting the Right Interestingness Measure for Association Pattern by Analyzing the Interestingness of Association Rules from thTemporal Dimension by School of Computing National University of Singapore
[5]. Tara McIntosh and Sanjay Chawla, High-Confidence Rule Mining for Microarray Analysis.