I J C T A, 9(28) 2016, pp. 57-63

AN EFFICIENT AND MODIFIED FP-GROWTH BASED POSITIVE ASSOCIATION RULE MINING TECHNIQUE FOR VERY LARGE DATASETS

R. Z. Inamul Hussain* and S. K. Srivatsa**

Abstract: Association Rules Mining is one of the standard methods used in data mining. To make decision and for correlation analysis positive association rules are very much useful. So in our paper, we have proposed a model which contains of preprocessing, mining patterns and assigning weight to discover highly positive association rules. To find the frequently occurring sets of data in databases, we have implemented algorithm to mine the frequent pattern. This algorithm is more memory efficient and leads to run extremely fast on large databases. The result displays that the projected model can mine association rules with high correlation.

Keywords: Association rule mining, large datasets, sampling, pruning, Frequent Item Set

1. INTRODUCTION

In this world of fast information communication, massive amount of data is generated and stored in computer database systems. Association rule mining (ARM) is the most popular knowledge discovery technique used in several areas of applications. In ARM, large number of Association rules or patterns or knowledge is generated from the large volume of dataset. But most of the association rules have redundant information and thus all of them cannot be used directly for an application. So pruning or grouping rules by some means is necessary to get very important rules or knowledge. One way of selecting very interesting rules is using interestingness measures to rank and select a small set of rules of different characteristics.[4]

Association rule mining was initially established to support "market basket analysis" to examine products that buyers tend to buy at the same period. For instance, an association rule milk; butter→bread means that consumers who buy milk and butter are also likely to buy bread. Mining Association rules has been used recently to provide recommender systems in various domains like intelligent web applications and e-commerce. Association rules can be educated from the product-by-feature matrix, which is used to build recommendations. For instance, consider a rule stating that e-mail spam detector, virus description update cookies management and web account: then, a product profile containing an e-mail spam indicator and a virus description update could result in a recommendation to add a feature to support web account and cookies management. In common, when a partial profile is matched beside the antecedent of a exposed rule, the attributes on the right-hand side of the similar rules are arranged conferring to the confidence standards for the rule, and the highest ranked items from this list form the recommendation set. The finding of association rules includes two key parts: the finding of frequent item sets (i.e., item sets which fulfill a least support threshold) and the finding of association rules from these frequent element sets which satisfy a minimum confidence threshold. Numerous algorithms exist for determining the frequent item

* Research Scholar Sri Chandrashekhendra Saraswathi Viswa maha Vidyalaya University, Enathur,Kanchipuram-631561 India Email: inamulhasan.rz@gmail.com

** Retired Anna University Senior Professor Anna University, Chennai -600 025 Email: profsks@rediffmail.com

sets and association rules between which we have selected is FP-growth algorithm as it is exposed to be fairly memory-efficient, and therefore suitable for the extent of our data set. After creating the association rules, different types can be recommended to an early product profile by finding all the matching rules[1]

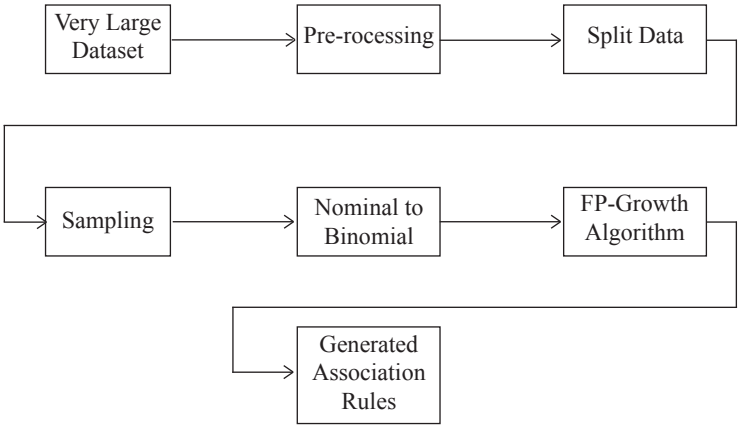

Figure 1. Various steps to generate Positive Association Rules.

2. ASSOCIATION RULES [2]

The purpose of association rule mining is the elicitation of exciting rules from which information can be derived. Those rules describe original, important, surprising, nontrivial and even actionable relationships among different features or attributes. Association rule mining is normally detailed as follows: Assume $I = \{I_1, I_2, \ldots, I_n\}$ be a set of *items*, and D be a set of transactions. Every transaction comprises of a subset of items in I.

An association rule, is explained as an effect of the form $A \rightarrow B$ where $A, B \subseteq I$ and $A \cap B = \Phi$. Support, confidence, and the minimum improvement constraint between there might be measured as events to assess the excellence of the extracted rules. Support decides how frequently a rule is applicable to a certain data set of an attribute and it represents the probability that a transaction contains the rule. The confidence of a rule $A \rightarrow B$ denotes the probability that B arises if A have already occurred P(B/A); then it estimates how frequently items B appear in transactions that contain A. Finally the minimum improvement constraint amount not only specifies the strength of a rule but it prunes any rule that does not provide a important predictive advantage over its proper sub-rules.

In this work, in the process for obtaining rules, we consider the FP-Growth that is supported on frequent item sets. For the purpose of this effort items will be measured as features and transaction as PMs and the outcome of this couple wise is what we call the binary features matrix.

Step 2: *Running Association Rules FP-GrowthAlgorithm*. Once the matrix is constructed, we have the datas to apply the association rule data mining tool, that allow us to discover the relationships, dependencies and also to handle a large amount of data in an best way. But, such algorithms developed are occasionally restricted to the memory since of its size and calculus that they do.

In fact the most composite task of the entire association rule mining process is the generation of frequent itemsets. Numerous different mixtures of features and rules have to be discovered which can be a very computation-intensive task, especially in huge databases. With this we can study only single relations between features to avoid those computation difficulties. Often, a cooperation has to be made between learning more complex rules and calculation time.

To screen those rules that might be not appreciated, it is vital to calculate its support. As we have previously seen, the support governs how frequent the rule is appropriate to the product P. This value

related with the minimum support believed by an expert (min support threshold), prunes the uninteresting rules. To estimate the interestingness and pertinence; that is it the dependability of the implication made by a rule, it is valuable to calculate its confidence. The job is now to produce all possible rules in the frequent item set and then equate their confidence value with the minimum confidence. All rules that happen this condition are observed as interesting. All the last revealed associations with their support and confidence values, hence, may be presented to stakeholders.

Furthermore the calculation of other measures is relevant to refine the process of selecting the appropriate association rule. The minimum development constraint portion not only gives us an impression about the strength but also prunes any rule that does not offer a important predictive benefit over its proper sub-rules. This increases efficiency of the algorithm.

Eliminating all association rules that do not fulfill the minimum development constraint, suggests us the most related and important rules available for the study. It is clear that those relations that are always present in all the product models may be considered as compulsory. Now, if some vague information is existing in the database and this one is not consistent at λ%, in order to obtain compulsory relationships, the analyst may create as a least confidence threshold the value (100-λ)%. The rules whose confidence is superior than the (100-λ)% may be measured as compulsory relationships. Bidirectional rules such as X1→ X2 and X2→ X1 may be also measured as compulsory relationships.

The association is classified as compulsory if at least one of the two things mentioned before (high frequent features and bidirectional rules) happens and, of course, the relations belong to a parent child.

3. SAMPLE DESIGN [3]

There are three techniques of gathering data such that the information collected can be used to draw implications about the target universe. These are:

- Gathering of data from all enterprises. This is a expensive and long procedure unless the target dataset is small;
- Gathering of data from a model of units that have been chosen from the target dataset with the intention that they should be representative of that dataset A sample of this type is mentioned to as a *purposive (*or sometimes *judgmental)* sample. In order to draw inferences about the target dataset using a purposive sample, a number of `conventions have to be completed about the representativeness of the data collected and of the writing units and, in general, there are restrictions to the implications that can be drawn from purposive samples when the probability of choice is not known;
- Gathering of data from a random sample of items which have been selected with recognized probabilities of choice from between all units in the target dataset. In this case no expectations about representativeness are wanted in estimating totals or averages for the target dataset and, in addition, there are well known methods for determining the precision of these estimations. This said, estimations based on random samples will only be balanced if the business registers from which they are drawn are complete and up to date.

Various algorithms are used to extract positive rules from large datasets such as Apriori, FP-Growth etc. They are described as follow

4. APRIORI [5]

The general idea of the Apriori algorithm is to create frequent itemsets for a certain dataset and then scan those frequent itemset to differentiate most frequent items in this dataset. The process is iterative. Since

generated frequent itemsets from a phase can construct another itemsets by combining with previous frequent itemsets. It is a confidence-based Association Rule Mining algorithm. The Confidence is purely accuracy to estimate the rules, created by the algorithm. The rules are classified according to the confidence rate. If two or more rules share the identical confidence then they are primarily ordered using their support and then the time of discovery. Support is the percentage of a specific record in a data set. General steps for rule generation by Apriori are:

- Create frequent itemsets of length 1
- Repeat until count of newly generated frequent itemsets are zero(0)
- From *m* frequent itemsets, produce *m*+1 candidate itemsets.
- Prune infrequent candidate of size m.
- Calculate the support of each candidate by scanning the entire database.
- Recollecting the frequent candidate, eliminate the mfrequent one.
- Yield Apriori rules based on support and confidence.

5. FREQUENT PATTERN (FP) GROWTH METHOD [6]

Apriori requires m+1 scans, where m is the length of the lengthiest pattern, we can have frequent pattern (FP) growth technique to reduce the amount of scans of the entire database, d to find the frequent itemsets using only two scans of database.

5.1 Process FP-growth

1. IF Tree comprises a single path G THEN
2. FOR all combination (indicated as β) to the nodes in the path G DO
3. Produce patter β_α with support=minimumsupport of nodes inβ;
4. ELSE FOR all α*i* in the pass of Tree DO BEGIN
5. Produce patter β = α*i*_α with support = α*i*.support;
6. build restricted pattern base and produce Treeβ;
7. IF Tree β ≠ Φ THEN CALL FP-growth (Tree β, β);
8. END

the FP growth technique is faster than Apriori Algorithm since FP scans the entire database at most twice, while in Apriori this is not known in advance and may be quite big. FP based method is quite effective in reducing memory.

It has advanced mining efficiency in execution time, usage of memory and utilization of CPU than most current ones like Apriori.

6. DATA MINING TOOLS [7]

There are numerous data mining tools existing. This division gives a brief report of some of them.

6.1 Microsoft SQL Server

Microsoft SQL Server is a “complete, united data society and investigation software that permits organizations to reliably achieve mission-critical material and positively run today’s gradually complex

business presentations. (Microsoft, 2007) ” SQL Server 2005 is the platform frontrunner in a variety of zones containing business intelligence and database management systems.

6.2 Rapid Miner

“Rapid Miner (previously YALE) is the world foremost open-source association for discovery of knowledge and data mining (Rapid-I, 2008)”. This tool is not only accessible as a stand-alone application for data analysis but also as a data mining machine that can be united into your own products.

6.3 Weka

Weka (Waikato Environment for Knowledge Analysis) is a standard collection of machine learning software that is written in Java. The software was established at the University of Waikato and is open source under the GNU General Public License (Wikipedia, 2008).

6.4 Oracle Data Mining

Oracle Data Mining (ODM) is an choice of Oracle Database 10*g* Enterprise Edition. This tool offers the skill to “yield actionable predictive information and shape integrated business intelligence applications (Oracle, 2007)”.

7. PROBLEM DEFINITION

As stated previously the accuracy of predictive modeling algorithms is very important. The results of these models contain information that can be very valuable and useful in aiding decision making and driving change. Ensemble methods (e.g., bagging, boosting and stacking) have been proven to produce a model that will outperform single models. However, these methods only deal with combining classification algorithms.

This research investigates the idea of using clustering and predictive modeling as an ensemble. The hypothesis is that by clustering the data set then applying a predictive model to each cluster and combining the results of each cluster’s predictive model we will produce a model that will have a higher accuracy than the predictive model alone.

8. EXPERIMENT

A preliminary experiment was conducted to determine positive association rules using a FP-Growth modeling approach. It could possibly produce a model or set of rules that has a interesting rules.

8.1 Data

The dataset was given by Tom Brijs and comprises the (anonymized) retail market basket data from an anonymous Belgian retail store and it is found in frequent item set data repository and file name is retail (.gz). It is a huge dataset.It contains nearly 90,000 datasets. This is mainly used for association rule mining.

8.2 Materials/Tools

Rapid miner was used for this experiment which is mainly used mining large amount of datas in a small amount of time.

8.3 Procedure

Initially we take large datasets and it is preprocessed to give that as input to rapid miner. Filtering is done to remove unwanted data's and to be suitable for mining tool. Then the data is splitted. Since the dataset is huge. Each sample of 1000 records are taken. It is applied to frequent item set mining algorithm and the output is added with pervious output. Sampling is done since our system is of 4GB memory and the intermediate results are stored in memory. It occupies much memory size.

9. RESULTS

Figure 2 depicts the positive association rules generated from the training data. Result shows that when the customer buys the product number 39 and 41 and he is likely to purchase product number 32. When the customer buys the product number 39 alone he is likely to purchase product number 41 This tells the seller to how to improve the product sales.

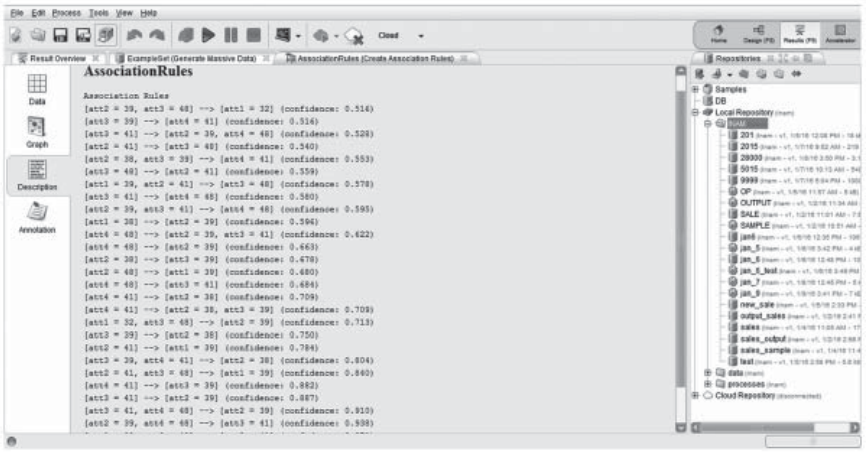

Figure 2. Number of association rules generated from sale data set using Rapid Miner Tool.

Figures 3 shows the graph with this we can easily find the product and their relevant rules.Graph shows thatproduct 48 is related to the rule14,12 Figure 4 show that we can find the positive results of each and every product. In the given output it shows that product 39 has association with product 41 and 48.With this we can find the associations of each and every product with other products which makes the promotion in sales as well as it is mainly used as recommendation system.

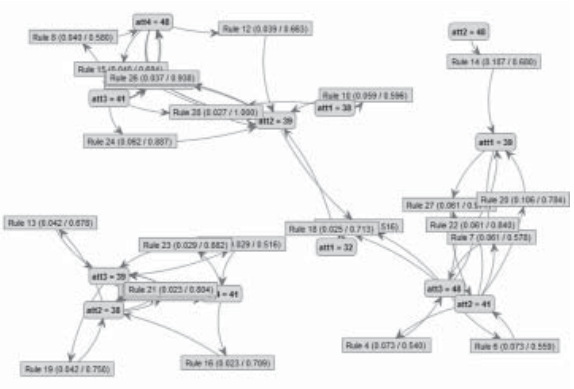

Figure 3. Generated graph for every product

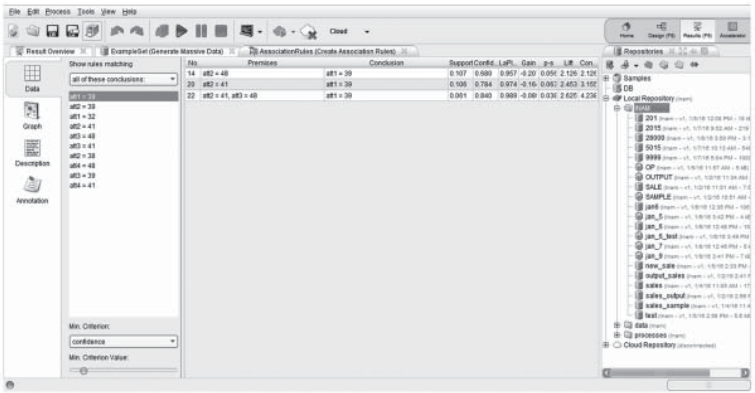

Figure 4. Rules generated for particular item 39

The below figure shows the graph that for attribute 39 three rules are generated. This gives the clear view.

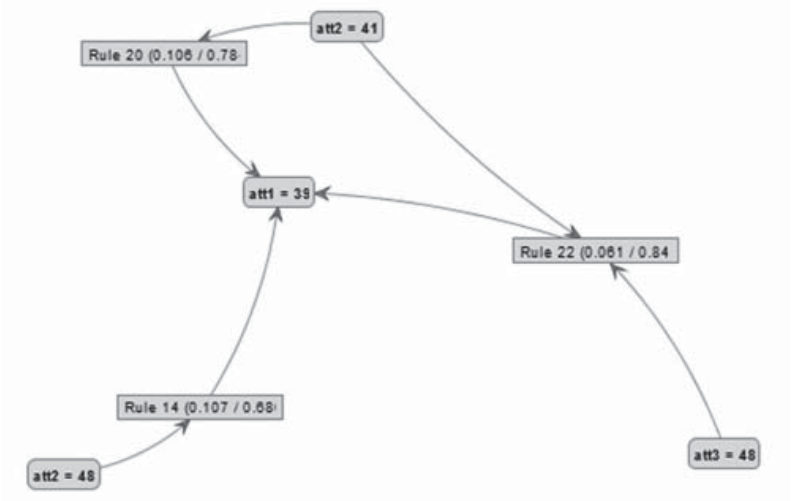

Figure 5. Graph for particular item number 38

10. CONCLUSION

Data mining is a dominant and valuable tool that can be used to extract related data's from large data sets. These methods are used in a variety of domains to achieve an array of dissimilar tasks. However, these techniques focus on mining positive association rules for large datasets. The goal of this experiment was to extract rules from the idea of using a sampling and modified FP-growth approach. The results suggest that sampling and modified FP-growth modeling approach can possibly produce a model that has a higher accuracy versus a predictive model alone.

11. FUTURE WORK

This research focused on the idea of using sampling and modified FP-growth approach. The dataset used in this experiment was very large. We would like to do more tests with data sets changing in size and domain. Also, there is presently no open source tool existing to do this type of analysis.

References

1. Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings "Negar Hariri, Carlos Castro-Herrera, Member, IEEE, Mehdi Mirakhorli, Student Member, IEEE, Jane Cleland-Huang, Member, IEEE, and Bamshad Mobasher, Member, IEEE
2. Alberto Lora-Michiels, Camille Salinesi, Ra_ul Mazo,2010 "A Method based on Association Rules to Construct Product Line Model", 4th International Workshop on Variability Modelling of Software-intensive Systems (VaMos), Jan 2010, Linz, Austria. pp.50. <hal-00707527>
3. Sample Design,*Business Tendency Survey Handbook*,STATISTICS DIRECTORATE, OECD, 2003
4. S.Kannan1 and R.Bhaskaran,2009," Association Rule Pruning based on Interestingness Measures with Clustering" IJCSI International Journal of Computer Science Issues, Vol. 6, No. 1, 2009
5. Mohammed M Mazid, A B M Shawkat Ali, Kevin S Tickle,2009, "A Comparison Between Rule Based and Association Rule Mining Algorithms" 2009 Third International Conference on Network and System Security, 978-0-7695-3838-9/09 © 2009 IEEE DOI 10.1109/NSS.2009.81
6. Wei Zhang, Hongzhi Liao, Na Zhao, Research on the FP Growth Algorithm about Association Rule Mining, 2008 International Seminar on Business and Information Management, 978-0-7695-3560-9/08 $25.00 © 2008 IEEE DOI 10.1109/ISBIM.2008.177
7. Philicity K. Williams, "CLUSTERING AND PREDICTIVE MODELING: AN ENSEMBLE APPROACH" Thesis, Auburn University Auburn, Alabama August9,2008