Extraction of Interesting Frequent and Rare Association Rule using Multi Objective Approach

R. Das[1] & S.N.Singh[2]
[1]Assistant Professor, Department of Computer Science Engineering
National Institute of Technology-Mizoram, Aizwal, Mizoram, India
[2]Professor, Department of Electronics Engineering,
National Institute of Technology-Jamshedpur

ABSTRACT

Association rule mining is the process of data mining for finding some relationship among the attribute/attribute values of a huge database which will help in taking some decisions. Association rule mining can be of two types frequent association rule mining and rare association rule mining. Frequent association rule mining attempts to generate frequent rules, i.e. rules having higher support and confidence .The rare association rule mining generate rare rule which have lower support but higher confidence. However based on survey it has been observed that researchers have considered the problem of frequent and rare rule generation method separately[8,6]. That motivated to propose a method for generate frequent and rare rule using multiobjective approach. So association rule mining can be considered as a multi-objective problem rather than as a single objective one. Confidence, comprehensibility and interestingness measure used for evaluating a rule can and making it different objectives of association rule mining problem. Support count is the number of records, which satisfies all the conditions present in the rule. This objective gives the accuracy of the rules extracted from the database. Comprehensibility is measured by the number of attributes involved in the rule and tries to quantify the understandability of the rule. Interestingness measures how much interesting the rule is. Using these three measures as the objectives of frequent and rare rule mining problem, this paper uses a Pareto based non-dominated sorting for extracting some useful and interesting rules from any market-basket type database. Based on experimentation, the algorithm has been found suitable for large databases.

Index Terms: Association rule mining, Multi-objective association rule mining, Frequent rule generation, Rare rule generation.

1. INTRODUCTION

Association Rule Mining is one of the data mining techniques[1,2,12] that used to extract common pattern and rules from large databases, that can be used in decision making process. Following the original definition by Agrawal at al. the problem of association rule mining is defined as: Let I={i1,i2,….in} be a set of n binary attributes called items. Let D={t1,t2,…tm} be a set of transactions called database. Each transaction t in D has a unique transaction ID and contains a subset of the items in I. An association rules is defined as an implication of the form X=> Y where X, Y I and X∩Y=Φ, meaning that if find all of X in the market basket, then have a good chance of finding Y.To select interesting and useful rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best known constraints are minimum threshold on support and confidence. The support support(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. It is defined as,

Support(X) = Number of transactions in the database that contain itemset X/ Number of transaction in the entire database

Only those itemsets whose support is at least equal to a given minimum threshold are used for mining association rules. These itemsets with support at least equal to minimum support are called frequent itemsets.The confidence of rule is defined as

conf (X=>Y)= support(XUY)/support(X)

Only those rules with high confidence are used for association rule mining. A number of algorithms have been developed for searching these rules [1, 2]. Frequent association rule are all association rules in the database that have a support greater than the minimum support means rules are frequent and have a confidence greater than minimum confidence means rules are strong. Rare association rules correspond to rare, or infrequent itemsets. Rare association rules are those rules that have a support less than the minimum support means the rules are infrequent and have a confidence greater than minimum confidence means rules are strong. It has been observed that some of the rules generated by Frequent rule generation approach are not comprehensive (complex to understand), i.e. the rule length is higher. Specially if in antecedent part more number of attribute involved then it become complex for user to understand. If min support set very low then generated large number of rules and it is based on only one evaluation criteria i,e confidence or predictive accuracy. Even in case of rare rule generation approach also generates some rules which are not comprehensive means complex to understand and some of the interesting rule may be missed. To overcome this issue this paper introduced a multi-objective rule generation approach which should be generate minimized, comprehensive and interesting rules.

2. RELATED WORK

The Association rule mining problem divided in two ways one is frequent rule generation another is rare rule generation. Apriori algorithm which is frequent rule generation algorithm proposed by Agrawal et al in 1993[3]. .Most of the existing association rule mining algorithms are improvements to the Apriori algorithm. Among of these Srikant’s rule generation which is considered as faster rule generation approach, this algorithm eliminates the checking of many unnecessary candidate rules[2].This algorithm work in two phase the first phase is for frequent itemset generation and second phase rule generation on the basis of minimum confidence. So, confidence is used as a measure for generating rules.

In case of rare rule generation for generating the rare itemsets, it uses a modified version of Apriori algorithm, i.e. *Apriori-rare.* Finally, based on the rare itemsets generated, it applies the following rule generation technique for the rare rule generation. Here, the finding of minimal rare itemsets is based on a key technique, referred as *MRG-Exp[5]*, which attempts to find the frequent generators, but as a "side effect" it also explores the so-called minimal rare generators (mRGs). *MRG-Exp* retains these itemsets instead of pruning them. All minimal rare itemsets are identical to the set of minimal rare generators. Next, it finds the closures of the previously found minimal rare generators so as to obtain their equivalence classes. Finally, from the explored rare equivalence classes, it generates rare association rules in a way very similar way of finding (frequent) minimal non-redundant rules. These rare rules are referred here as "mRG rules" because their antecedents are minimal rare generators. In order to decide whether a candidate set *x* is a generator or not, one needs to compare its support with its immediate predecessor. It doesn't allow the predecessors to have the same support.

Existing Srikant's algorithms for association rule generation, try to measure the quality of generated rule by considering only one evaluation criterion, i.e., *confidence factor* or *predictive accuracy*. However, those rules may which not have expected confidence, but can be found to be interesting in a given context, will be missed by these algorithms. Usually in the medical diagnosis, it also demands for the generation of the hidden rules rather than only the obvious once. However, these two algorithms can be found to be incapable of finding those types of rules.Even in rare rule generation it has been observed that some of the rules generated by it, are not comprehensive i,e complex to understand and the rule length is higher. Specially if in antecedent part more number of attribute involved then it become complex for user to understand. To overcome this issue demanded multi-objective rule generation approach.

3. PROPOSED WORK

In the proposed work we tried to solve the frequent and rare rule generation using multi-objective association rule-mining technique with a Pareto based non-dominating approach.

3.1 Missing Value Prediction

The first task for this is to prepare the dataset for association mining. Today real world datasets contains missing values due to human, operational error, hardware malfunctioning and many other factors. The quality of knowledge extracted, learning and decision problems depend directly upon the quality of training data. Solving the problem of missing data is of a high priority in the field of data mining and knowledge discovery. The simple solution is to discard the data instances with some missing values. The algorithm that we have used to determine the missing attributes is given below:

1. Find an attribute with missing value (usually denoted by?)

2. Find out the class label of the transaction containing that missing attribute.

3. Find the most frequently occurred value (*m*) of that missing attribute for that class label.

4. Replace '?' by *m.*

5. Repeat step 1 through 4 until all the missing attributes are replaced

3.2 Conversion of Given Dataset to Market Basket Form

Real-life dataset contains continuous values. To generate frequent and rare itemset we convert the real life dataset to market basket dataset [3]. For that we adopt the strategy of discretization for converting numerical attribute to market basket form where each attribute is represented by 0 (absent) or 1(present). we kept discretization simple with equi-width binning with bin size or interval k. Simply, if a variable or attribute *p* is bounded by *pmin* and *pmax,* then bin width = (*pmax* – *pmin* / k). Then bin boundaries or subinterval range are *pmin* + i*(bin width) where i = 1, 2, .., k-1 i,e, we define subintervals for each attribute. Depending on their range, each attribute may have different number of sub intervals and for different attributes the range of these sub intervals may also be different. For each attribute, we assign 1 to the subinterval in which the attribute value falls and assign 0 to rest of the subintervals for that attribute.

3.3 Frequent and Rare Itemset Mining

In this work, we applied the KDD process [1] for mining multi-objective association rule. Algorithm used for frequent and rare pattern mining is Apriori[4] and Apriori-rare[5,6] respectively. Apriori algorithm used for frequent itemset mining. And apriori-rare, slight modification Of apriori algorithm used to generate rare itemset .

3.4 Rule Generation

Frequent and rare Association rules can be generated as follows: For each frequent and rare itemset *l*, generate all non-empty subset(s) of *l*.For every non-empty subset(s) of *l*, output the rule "*s* => (*l-s*)"

3.5 Multi-Objective Rule Generation and Optimization

Association rule mining problems can also be considered as a multi-objective problem [9] rather than as a single objective one. Measures like *support count*, *comprehensibility* and *interestingness [7]*, used in evaluating a rule can be thought of as different objectives of association rule mining problem. *Support count* is the number of records, which satisfies all the conditions present in the rule. This objective gives the accuracy of the rules extracted from the database. If a huge number of attributes are involved in a rule then the rule may be useless as it will be very difficult to understand. Involvement of fewer attributes makes a rule more understandable. *Comprehensibility* can be measured by the number of attributes involved in the rule and tries to quantify the understandability of the rule with the following expression:

$$Comprehensibility = \log(1+|Y|) / \log(1+|X\cup Y|) \quad \text{-----} \quad (i)$$

Here, $|Y|$ and $|X\cup Y|$ are the number of attributes involved in the consequent part and the total rule, respectively. Since association rule mining is a part of data mining process that extracts some hidden information, it should extract those rules that have a comparatively less occurrence in the entire database. *Interestingness* can be measured by how surprising the rule is. Such a surprising rule may be more interesting to the users. Interestingness of rule can be quantified with the following expression:

$$Interestingness = [SUP(X\cup Y)/SUP(X)] \times [SUP(X\cup Y)/SUP(Y)] \times [1-(SUP(X\cup Y)/|D|)] \quad \text{-----------} \quad (ii)$$

3.6 Pareto Optimality Principle

It is always difficult to find out a single solution for a multi-objective problem. So it is natural to find out a set of solutions depending on non-dominance criterion. According to non-dominace criteria a solution, say *x*, is said to be dominated by another solution, say *y*, if and only if the solution *y* is better or equal with respect to all the corresponding objectives of the solution *x*, and *y* is strictly better in at least one objective. So, the solution *y* is called a non-dominated solution. In order to sort a rule according to the level of non-domination, each rule must be compared with every other rule to find if it is dominated. When this process is continued to find the members of the first non dominated font or solution, all individuals in the first non-dominated front are found that is given rank1. In order to find the individuals in the next front, the solutions of the first front are temporarily discounted and the above procedure is repeated. The procedure is repeated to find the subsequent fronts. After that we will assign rank, which rules are in first font will be rank 1, second font will be rank2 subsequently others also. Here Lower rank has given higher priority. At the time of taking a decision, the solution that seems to fit better depending on the circumstances can be chosen from the set of these candidate solutions. Vilfredo Pareto suggested this approach of solving the multi-objective problem. Optimization techniques based on this approach are termed as Pareto optimization techniques. we adopt this concept for finding optimal solution.

4. EXPERIMENTAL EVALUATION

In our experiment purpose we use different dataset from UCI machine learning repository [10].Result comparing with srikant's rule generation method[2,12] and Szathmery[11] rare rule generation method. The experiment has been implemented using C as programming language in Linux platform and hardware used Processor: Intel Xeon: 3 GHz, RAM: 2GB Hard disk: 292 GB.

Table 1. Comparative result

Datasets	Srikant's second algorithm(frequent rule generation)	BtB algoritm(rare rule generation)	Proposed approach
Wisconsin breast cancer dataset	11,192(minsupp 40%)	351	11362
Glass	1522(minsupp 50%)	10	1532
Iris	88 (minsupp 40%)	3	91

4.1 Comparative Results

We observed that some of the best rules obtained using our propose method cannot be obtained using traditional Srikant's rule generation method. For example rules like UNIFORMITY_OF_CELL_SIZE=7,NORMAL_NUCLEOLI=1, MITOSES=1=>CLUMP_THICKNESS=8, SINGLE_EPITHELIAL_CELL_SIZE=4 having support 0.143% generated from Wisconsin dataset. However, the numbers of rules which are interesting cannot be generated using traditional rule mining algorithms.

4.2 Classification Result

Wisconsin Breast cancer dataset the objective of these identification techniques is to assign patients to either a benign group that does not have breast cancer or a 'malignant' group who has strong evidence of having breast cancer. Here presented some of high confident, high comprehensible and high interesting rule. Rules are useful for classification. Wisconsin cancer dataset having 699 instances among of them 599 used for training purpose and 100 used for testing purpose. First step finding the class of each rule after that in second step used that rule for classification of test dataset.

Table 2. Sample of best rule generated from Wisconsin breast cancer dataset by proposed approach

Rule	Attribute Present
R1	UNIFORMITY_OF_CELL_SIZE=1=>BARE_NUCLEI=1,NORMAL_NUCLEOLI=1,NORMAL_NUCLEOLI=10 conf=0.832278 : comp=0.861353 : int =0.465979 Rank=1
R2	UNIFORMITY_OF_CELL_SHAPE=1=>NORMAL_NUCLEOLI=1,UNIFORMITY_OF_CELL_SHAPE=1,UNIFORMITY_OF_CELL_SIZE=1 conf=0.900000 : comp=0.792481 : int =0.416970 Rank=1
R3	BLAND_CHROMATIN=7 =>MITOSES=1 conf=0.834483 : comp=0.630930 : int =0.490705
R4	UNIFORMITY_OF_CELL_SIZE=1=>MITOSES=1 conf=0.984177 : comp=0.630930 : int =0.343923 Rank=1
R5	BARE_NUCLEI=10 => NORMAL_NUCLEOLI=10 conf=0.853333 : comp=0.564575 : int =0.400124 Rank=1

Table 3. Training Dataset Classification

Rule	Benign	Malignant	Class
R1	100%	0%	Benign
R2	100%	0%	Benign
R3	4%	96%	Malignant
R4	100%	0%	Benign
R5	0%	100%	malignant

Table 4. Dataset classification

Rule	Classification Accuracy
R1	100%
R2	100%
R3	93% (M),7%(B)
R4	100%
R5	100%

5. CONCLUSION AND FUTURE WORK

The propose technique has been found to be capable of finding some of the rare association rules and found compact and scalable with a high ossification accuracy. However there are scopes for extending proposes work in the following direction Rough multi-objective rule generation and Classification accuracy. However there are scopes for extending proposes work in the following direction Rough multi-objective rule generation and Rough-fuzzy multi-objective rule generation.

6. REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concept and Techniques" 2nd Edition, Morgan Kaufmann.

[2] Rakesh Agrawal and Ramakrishnan Srikant, " Fast Algorithms for Mining Asociation rules", In Proceedings of the 20th Int.Conf. Very Large Data Bases,pp. 487-499.,1994.

[3] R. Agrawal, T. Imeilinski and A. Swami "Mining association rules between sets of items in large databases." Proceeding of ACM SIGMOD Conference on Management of data, pp. 207-216, 1993.

[4] M. Houtsman and A. Swami, 'Set-Oriented Mining for Association Rules in Relational Database'. Proc. Of the 11th IEEE Int. Conf. on Data Engineering, pp. 25-34, Taipei, Taiwan, March 1995.

[5] Laszlo Szathmary, Petko Valtchev, and Amedeo Napoli, "Generating Rare Association Rules Using the Minimal Rare Itemsets Family" Int J Software Informatics, Vol.4, No.3, September 2010, pp. 219–238

[6] R.U.Kiran and P.K.Reddy, "Mining Rare Association Rules in the Datasets with widely Varying Items Frequencies," The 15th International Conference on Database Systems for Advanced Applications Tsukuba,Japan,April 1-4,2010.

[7] A Ghosh and B. Nath, " Multi-objective Rule Mining using Genetic Algorithms," Information Sciences, vol 163, pp 123-133, 2004 .

[8] J. Hipp et. al. "Algorithms for association rule mining- a general survey and comparison," SIGKDD Explorations 2 (1), 2000.

[9] DEHURI, S., JAGADEV, A. K., GHOSH A. AND MALL R., " Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations," American Journal of Applied Sciences 3 (11): 2086-2095, 2006 ISSN 1546-9239, 2006.

[10] UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems, www.http://archive.ics.uci.edu.

[11] Szathmary L,Valtchev P, Napoli A."Finding Minimal Rare Itemsets and Rare Association Rules". Proc. of the 4th Intl. Conf. on Knowledge Science, Engineering & Management (KSEM '10),vol. 6291 of LNAI, pages 16–27, Belfast, Northern Ireland, UK, 2010. Springer, Berlin.

[12] Srikant R, Fast algorithms for mining association rules and sequential patterns, Phd thesis, University of Wisconsin-Madison, 1996.