ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Approach for Rule Pruning in Association Rule Mining for Removing Redundancy

Ashwini Batbarai[1], Devishree Naidu[2]

P.G. Student, Department of Computer science and engineering, Ramdeobaba College of engineering and management, Nagpur, India[1].

Assistant Professor, Department of Computer science and engineering, Ramdeobaba College of engineering and management, Nagpur, India[2]

ABSTRACT: In Data Mining Association rule mining is an important component. It is used for prediction or decision making. Numbers of method or algorithm exist for generating association rules. These Methods generates a huge number of association rules. Some are redundant rules. Many algorithms have been proposed with the objective of solving the obstacles presented in the generation of association rules. In this paper we have given the approach for removing redundancy based on frequent closed itemset mining (FCI), and using lift as the interesting measure for gating the interesting rule and forming the non-redundant rule set based on completeness and tightness properties of rule set..

KEYWORDS: Association Rule Mining, Frequent Closed Itemset, Non-redundant rule, Redundant rule**.**

I. INTRODUCTION

Association rule is an association relationship among the attributes in the relevant data or transaction data. It gives the result in the form of rules between the different set of items on the basis of metrics like support and confidence i.e. joint and conditional probability respectively of antecedent and consequent.

Association rule mining required the two important constraints support and confidence [1]. It is finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

Association Rules**: Antecedent →Consequent [**support, confidence**].**

Let **I={i1,i2…..in}** A set of items and D={t1,t2, …, tn} be the set of Transactions where tj ⊆ I is represent the set of items purchased by customer, then association rule is an Implication: X ➔ Y where X,Y⊆ I and X ∩Y= ∅; Support of AR (s) X ➔ Y: Probability that a transaction contains X∪Y. Confidence of AR (a) X ➔ Y: Conditional probability that a transaction having X also contains *Y* [11].

There are two definitions of association rules. :

1st an association rule **r** is an implication of two frequent itemset X, Y⊆I of the form X→Y where X∩Y=Ø and support and confidence of rule **r** are defined as support(X) = support(Y) and confidence(**r**) = support(Y)/support(X) [11].

2nd an association rule **r** is an implication of two frequent itemset X, Y⊆I of the form X→Y where X∩Y≠Ø and support and confidence of rule **r** are defined as support(X) = support(Y) and confidence(**r**) =support(Y)/support(X) [2].

Here we are considering the 1st definition of AR for our approach.

Example: I = {1 2 3 4 5}

Transaction id	items
t1	1 2 3
t2	1 3
t3	1 4 3
t4	5 1
t5	5 4

Table 1: Transaction dataset

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Rule X→ Y	support X→ Y	Confidence X→ Y
1→3	(3/5)=60%	(3/5)*(4/5)=75%
3→1	60%	(3/5)*(3/5)=100%
2→4	0%	0%
2→3	(1/5)=20%	(1/5)*(1/5)=100%

Table 2: Rules generated from dataset

In this way rule are formed from the itemset using support and confidence measures.

There are two types of association rule 1st is exact association rule in this rule having confidence equal to 100% then is called exact rule.2nd is approximate association rule which having confidence less than 100%. We consider both in the approach.[3]

There are two phases of generating association rules 1st one is Extracting Frequent Itemset and 2nd Generating association rules from Frequent Itemsets. There are two types of Itemsets generation. One is Frequent Itemset generation which mining frequent itemset using Apriori algorithm (candidate set generation), FP-Growth algorithm (without candidate set generation). Second is Closed Frequent Itemset generation using CHARM ALGORITM, BIDE ALGORITHM, TOP-K ALGO-RITHM (TFP), and CLOSET ALGORITHM.

Generating association rule has different algorithm like Apriori, Apriori (TID) but these algorithm present some drawback like multiple database scanning. This traditional association rule mining algorithm presents some obstacles when generating association rule .These obstacles are Complexity of data, Time required to mining is more, Space required to store the rules are more, Cost required to mining rules is also more, Obtaining non interesting rules. Number of algorithms has been proposed for solving the obstacles present in generation of association rules [1][2][3][4].

Generating rules using frequent itemset gives large number of rules and in dense dataset generating the rules using frequent itemset is impossible where generating the rules from closed frequent itemset is possible. Forming association rule with lower frequency means minimum support level gives large number of rule and increasing the frequency i.e. maximum support level gives the rules but interesting rules pruned.

These methods produce large number of rules many of which having same meaning; also some rule only change the items in the antecedents and consequence side which not having different meaning than the existing one; and Many rules are valid rules but having same meaning, this all types of rules called redundant rule.

Redundant Association Rules: Definition: Let X →Y and X'→ Y' be two rules with confidence cf and cf', respectively. X→ Y is said a redundant rule to X' →Y' if X belong to X'; Y' belong to Y, and cf ≤ cf' [4].

Here five type of redundancy are given:

1: If rule X→YZ is redundant when the rules such as XY→Z, XZ→Y, *X→Y,* and *X→Z* are satisfy the minimum support and confidence. This is because the support and confidence values *X→YZ* are less than the support and confidence values for the rules *XY→Z, XZ→Y, X→Y,* and *X→Z.*

2: Check for combination of rules

A rule r in R is said to be redundant if and only if a rule or a set of rules S where S in R, possess the same intrinsic meaning of r. For example, consider a rule set R has three rules such as milk→tea, sugar→tea, and milk, sugar→tea. If we know the first two rules i.e. milk→tea and sugar→tea, then the third rule milk, sugar→tea becomes redundant, because it is a simple combination of the first two rules and as a result it does not convey any extra information especially when the first two rules are present.

3: Interchange the antecedent and consequence

Swapping the antecedent item set with consequence item set of a rule will not give us any extra information or knowledge.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

4: Redundant Rules with Fixed Consequence Rules [2]
Let us apply this theorem to a rule set R that has three rules such as {AB→X, AB→Y and AB→XY}. Consider the rule AB→XY has s% support and c% confidence. Then, the rules such as AB→X and AB→Y will also have at least s% support and c% confidence because X→XY and Y→XY. Since AB→X and AB→Y dominate AB→XY both in support and confidence, for this reason AB→XY is redundant.

5: Redundant Rules with Fixed Antecedent Rules [2]
Let us apply this theorem to a rule set R that has three rules such as {XY→Z, X→Z and Y→Z}. Suppose rule XY→Z has s% support and c% confidence. If n (i.e. number of items in the antecedent) number of rules such as X→Z and Y→Z also satisfy s and c then, the rule XY→Z is redundant because it does not convey any extra information if rule X→Z and Y→Z are present. All the above rules cannot be considered for dealing redundant rules in one domain.
Complete and tight ruleset is non-redundant ruleset. The ruleset having all the rules and having the rule which infer the other rule called the complete ruleset.

Ruleset which does not contain the redundant rule called tight ruleset.

II. RELATED WORK

Association rule mining has different application in data mining like analysis of market data, purchase histories, web log. This type of application has large data if we use the traditional algorithm for mining association rule it give large amount of association rule. From these number of rule many rule has same meaning so there is need for reducing the rule which having the same mining. Different approaches present for removing redundant rule. That all approaches gives rules is in the form of implication or functional dependency form in which rules are formed by using Armstrong axiom scheme namely Reflexivity (X→Y where Y⊆X) Augmentation (if X →Y and X' → Y' then XX' →YY', where juxtaposition denotes union) and Transitivity (if X →Y and Y → Z then X → Z).

For partial rules, the Armstrong schemes are not valid anymore. Reflexivity does hold, but Transitivity takes a different form that affects the confidence of the rules: if the rule A → B (or A → AB, which is equivalent) and the rule B →C both hold with confidence at least α (α is minimum threshold level), we still know nothing about the confidence of A → C .If it is less than the minimum confidence threshold and by using transitivity rule we take it as interesting or important rule it affect the accuracy of rule. So we cannot use the transitivity rule here also reflexivity not hold here [6].

Bastide et al.[1] gives the algorithm which using the semantic for extraction of association rule based on closure of Galois connection, the generic basis for exact association rule and informative basis for approximate association rule. It constructed using frequent closed itemset and generators .They is using A **Close** algorithm for frequent closed itemset mining and their generators. It generate the rule which having minimal antecedents and maximal consequents. It gives the user the set of rules covering all the attributes of dataset i.e. containing the rule where union of antecedent (resp. consequents) is equal to the union of antecedent (resp. consequents) of all association rules valid in context.
Ashrafi et al [2] the proposed methods not only remove redundant rules generated from frequent itemset but also remove redundant rules generated from the frequent closed itemset**.** The proposed methods are not based on any bias assumptions. It verifies all rules that have one or more items in the consequence. Therefore, it has the ability to eliminate redundant rules that contain single or multiple items in the consequence.

David Lo et al [3] in proposed method several rule sets based on composition of various types of pattern sets namely generators, projected-database generators, closed patterns and projected-database closed patterns. This set evaluated based on the two criteria of completeness and tightness and used as a composite filter , replacing a full set of rules with non redundant subset of rules dose not impact the accuracy of filter.

Zaki et al [4] present the framework for generating non-redundant association rule. They are using frequent closed itemset mining using the charm algorithm and generating the rule; then reducing the redundancy using transitivity rule. E.g. {TW→A, TW→AC, CTW→A} in this set all rule having same confidence as 1then put rule TW→A as interesting rule and remove other as redundant rule, because its only adding the items in the antecedent and consequent side items.{A→W,A→CW,AC→W} in this set all rule having same confidence hence put A→W and remove other as redundant rule, but rule A→W is not fully characterized the knowledge of rule AC→W so in this we are removing the interesting rule so we are not getting accurate rule set. .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Philippe Fournier-Viger et al [5] this gives the TNR algorithm for removing redundancy. In real time selecting the parameters to generate a desired amount of rules is usually difficult and time-consuming and generate large amount of redundancy in the results. These problems are addressed using TNR algorithm but it costly.

Redundancy can be removed by interesting measures. It measures how strongly one attribute implies other based on available data. Many objective measures are present like lift, Convention which can use for getting interesting rule depending on application [6],[7],[8].

III. METHODOLOGY

As traditional method produces large no of redundant rule which affect on the decision making problem to overcome this problem in this paper we are giving approach based on frequent closed item set mining then generating the rule set and then removing the redundancy using interesting measure and giving resultant rule set based on completeness and tightness

FLOW OF WORK:

I: Generate the synthetic dataset by using IBM dataset generator. Take the retail and mushroom dataset.

II: Generate the frequent closed itemset by using charm algorithm with minimum support value [9]. Generating the frequent closed itemset than frequent itemset required less time. Also generating rule from frequent itemset required more time and give large number of rule which is not efficient so we are using FCI method.

III: Generate the ruleset from the frequent closed itemset.

IV: Here we find the generator itemset and closed itemset then take the generator of closed itemset as antecedent and remaining items are consequence. And generate the rule. Here rule with minimum antecedent and maximum consequence generated

A frequent itemset P is considered to be a generator in DB if there exist no proper sub-sequence itemset of P having the same support as P in DB

Still redundancy present use transitivity rule for removing redundancy.

Transitivity rule: X→Y having s% support and c% confidence, Y→Z having s% support and c% confidence , then X→Z having s% support and c% confidence.

Use the lift as another interesting measure for removing redundancy and getting the interesting rule

$$\text{Lift } (X \rightarrow Y) = \text{conf}(X \rightarrow Y)/\text{supp}(Y) = P(X \text{ and } Y)/ (P(X)P(Y)).$$

Lift measures how many times more often X and Y occurs together than expected if they where statistically independent. Lift is not down-ward closed and does not suffer from the rare item problem. Also lift is susceptible to noise in small databases.

If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

If the lift is positive, like we know the degree to which those two occurrences are dependent on one another, and make those rules potentially useful for predicting the consequent in future data sets.

If rules generated from single closed itemset having same support and same confidence; in this redundancy present and there is need to get the interesting rule from the entire rule set. Lift is the interesting measure which gives the interesting rule from all these rules which having same support and same confidence.

Eg.36 79 94 ==> 90 #SUP: 4744 #CONF: 1 #LIFT:1
36 79 90 ==> 94 #SUP: 4744 #CONF: 1 #LIFT: 1.02434

In this second rule is interesting rule so there is need to remove first rule and put second rule which is most interesting rule.

E.g. 88 94 ==> 36 90 #SUP: 4640 #CONF: 1 #LIFT: 1.02634
36 88 ==> 90 94 #SUP: 4640 #CONF: 1 #LIFT: 1.02434
88 90 ==> 36 94 #SUP: 4640 #CONF: 1 #LIFT: 1.02734
88 ==> 36 90 94 #SUP: 4640 #CONF: 1 #LIFT: 1.02734

Here third rule and forth is interesting rule and other need to remove.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

In this way lift gives interesting rule. Lift also gives the interesting rule from approximate rule set.
V: getting the rule set as non-redundant rule base on completeness and tightness.

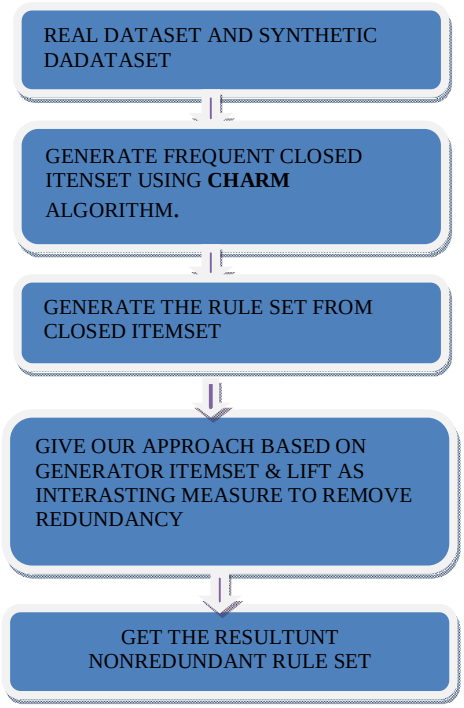

Fig1. Flow of approach

IV. RESULT

STUDY RESULT:

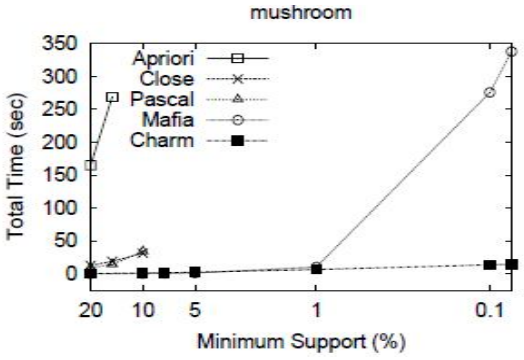

Fig2. Performance of charm verses Apriori, Close, Pascal and Mafia.

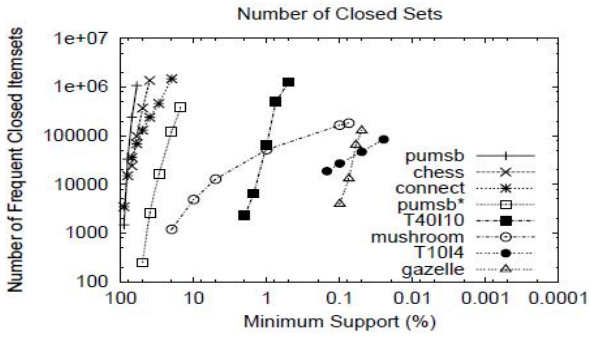

Fig3. Number of frequent closed itemset distribution by length

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

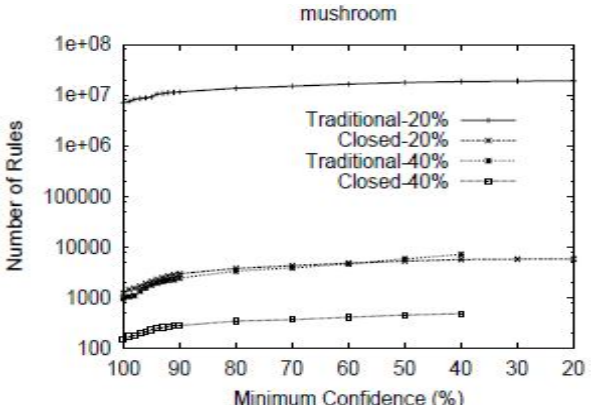

Fig4. Number of rules from traditional verses frequent closed itemset.

Here the studied result has shown [9]. In 2nd fig. time required for generating frequent closed itemset are less than other algorithms. In 3rd fig we are getting more no of frequent closed itemset as support threshold as decreased. In this result are shown on different dataset. We may use only mushroom and retail dataset. As the no of frequent closed itemset are less than other algorithm it give the reduce rule than other so we have selected the charm algorithm for generating closed itemset. In this only the closed frequent itemset are generated by using this when generating rules it give less rule than traditional algorithm fig 4th [4]. In fig 4th numbers of rules are generated as confidence threshold decreased. Still in that many no of rule for reducing that we are using our approach.

V. CONCLUSION

In this paper we have study obstacle come in association rule mining due to redundancy. We have given the approach for removing redundancy based on frequent closed itemset mining and generator. It gives the rule set which is smaller in size than the traditional approach and reduces the redundancy and from this we get good prediction

REFERENCES

1. Bastide Y., Pasquier N., Taouil R., Stumme G., Lakhal, L., "Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets", 1st International Conference on Computational Logic, pp. 972 – 986,2000,.
2. Mafruz Zaman Ashrafi , David Taniar , Kate Smith "New approach of Elim inating Redundant Association Rules'' 15th International Conference, DEXA , Zaragoza, Spain, Proceedings pp 465-474 2004.
3. David Lo , Siau-Cheng Khoo , Limsoon Wong , "Non-redundant Sequential Rule Theory and Algorithm '' School of Information System, Singapore Management University, 2008.
4. Mohammed J. Zaki, "Generating Non-Redundant association ruleKDD'00 Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM New York, NY, USA ,Pages 34-43 ©2000 .
5. Philippe Fournier-Viger, Vincent S. Tseng "Mining Top-K Non-Redundant Association Rules" 20th International Symposium, ISMIS 2012, Macau, China, December 4-7, 2012. Proceedings.
6. Pang -Ning Tan, Vipin Kumar, Jaideep Srivastava "Selecting the Right Interestingness Measure for Association Patterns"KDD'02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining ACM New York, NY, USA Pages 32-41 ©2002 .
7. P. D. McNicholas, T. B. Murphy, M. O'Regan "standardising the lift of an association Rule " computational statistics and Data Analysis, volume 52,Issue 10, pages 4712-4721,15 june 2008.
8. Paulo J. Azevedo, and Alıpio M. Jorge2 "Comparing Rule Measures for Predictive Association Rules" 18th European Conference on Machine Learning, Warsaw, Poland, pp 510-517 Proceedings 2007.
9. Mohammed J. Zaki ,Ching-Jui Hsiao "CHARM: An Efficient Algorithm for Closed Itemset Mining" Computer Science Department, Rensselaer Polytechnic Institute, Troy NY 12180.2002.
10. Qiankun Zhao, Sourav S. Bhowmick "Association Rule Mining: A Survey" Nanyang Technological University, Singapore.2003.
11. Book (Second Edition) Data Mining, Concept and techniques by Jiawei Han and Micheline Kamber .
12. J. Pei, J. Han, and R. Mao. "Closet: An efficient algorithm for mining frequent closed itemset". In SIGMOD Int'l Workshop on Data Mining and Knowledge Discovery, May 2000.
13. Kunkle D, Zhang DH, Cooperman G. "Mining frequent generalized itemsets and generalized association rules without Redundancy", Journal of Computer Science and Technology 23(1): 77 -102, Jan 2008.
14. Jose L. Balcazar "Redundancy, Deduction Schemes, And Minimum-Size Bases For Association Rules". Logical Methods in Computer Science Vol. 6 (2:3), pp. 1–33, 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

15. Bay Vo and Bac Le. "A Frequent Closed Itemsets Lattice-based Approach for Mining Minimal Non-Redundant Association Rules". International Journal of Database Theory and Application Vol. 4, No. 2, June 2011.

BIOGRAPHY

Ashwini Batbarai has received B.E. degree in Information Technology from K.I.T.S., Ramtek Nagpur University in 2011. She is pursuing M.Tech. degree in computer science and engineering from Ramdeobaba college of engineering and management (Autonomous) Nagpur her research interest include association rule mining, Datamining.

Devishree Naidu,has received B.E degree in Computer science and M.E. Specialization in Wireless Communication and Computing from G.H.Raisoni College of Engineering, Rashtrasant Tukdoji Maharaj University (University of Nagpur) INDIA in year2009. She has around Seven years of Teaching Experience, Currently working as a Assistant Professor at Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, INDIA. She has guided undergraduate and post graduate student for research project work in the field of mobile computing, and wireless sensor network. She is currently carrying her research work in sensor networks. Her previous research work was in Header compression Related to TCP/IP Protocol. Also she has interest in association rule mining, datamining.

.