International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

Negative Associative Classification Rules; An approach to Generalized Negative Rules

[1]**Dr. M. Kalpana Devi**, [2]**Prof. M. Usha Rani**

[1]*Sr. Asst. Professor, Dept. of MCA, Srinivasa Institute of Technology and Management Studies, Chittoor, Andhra Pradesh, India*

[2]*Professor, Dept. of Computer Science, Sri Padmavathi Mahila Viswavidhyalayam, Tirupati, Andhra Pradesh, India*

Abstract* - Association rule mining finds frequent itemsets and from which it extracts all the association rules. Classification rule mining aims to discover small set of rules to form an accurate classifier. Associative Classification is a combination of the above, which provides a special set of association rules whose right hand side is restricted to the classification class attribute and is referred to as Class Association Rules(CAR). Associative Classifier provide greater accuracy than the other classifiers and rules produced may be both Positive and Negative. Frequent itemsets produce positive rules and infrequent itemsets produce negative rules. Negative Association Rules also provide useful information in devising marketing strategies and many applications would benefit from negative associative rules. There are two types of negative association rules, the Confined and the Generalized negative rules. Many authors proposed confined negative rule generation but there are very few papers to deal with the generalized negative rules. In this paper, we discuss the problem of generating generalized negative associative classification rules and its importance.

***Keywords:* Association rule mining, Classification rule mining, Confined and Generalized negative associative classification rules.**

I. INTRODUCTION

The definition of an association rule is the implication of the form X => Y and X ∩ Y=φ, where X and Y are frequent item sets in a transactional database. In the rule X => Y , where X is called as an antecedent and Y as consequent of the rule. The rule has a *Support* 's' and *Confidence* 'c'. In the transaction database D if s% of transactions contain X U Y, and if c% of transactions in D that contain X also contains Y. The Association rules which satisfy minimum support and minimum confidence thresholds are deemed as strong rules and identify the positive relationship or co-occurrence of frequent itemsets.

Several algorithms have been proposed for mining positive rules but a few are there for mining both positive and negative rules. It would be necessary not only to consider all items in a transaction, but also all possible items absent from the transaction. A negative item is defined as ¬A, means item A is absent from a transaction and support of ¬A is sup(¬A)=1- sup(A). Negative association rule is of the form ¬X=>Y and the itemsets X and Y does not coexist in any transactions. There could be a considerable exponential growth in the candidate generation phase. This is especially true in datasets with highly correlated attributes.

According to John Tsiligaridis [1], there are two types of negative associative rules, the Confined negative association rule and the Generalized negative association rule. *The Confined negative association rule* is one of the following: ¬X=>Y, X=>¬Y or ¬X=>¬Y where the entire antecedent or consequent must be a conjunction of negated or non-negated attributes. *Generalized negative association rule* is a rule for which its antecedent or its consequent can be formed by a conjunction of presence or absence of terms. An example for such association would be: A∧¬B∧C∧¬D=> Y. Therefore negative association rule is either a negative relationship between two positive itemsets or a rule that contains at least one negative item in the antecedent or consequent itemsets.

Associative Classification rule mining is a hybrid approach, which provide supervised learning process. Class Associations Rules (CAR) may be both Positive and Negative, and by generating large number of both positive and negative rules, effective classifier can be built. Rule based classifiers play an important role in the classification problems in many application domains. In this paper we propose an associative classifier algorithm which generates both positive and negative rules, negative rules in the sense both confined and generalized negative rules.

II. LITERATURE SURVEY

The concept of negative relationships was mentioned first time by Brin et.al [2]. To verify the independence between two variables, they used statistical test. To verify the positive or negative relationship, a correlation metric was used. Their model is chi-squared based. The chi-squared test rests on the normal approximation to the binomial distribution. This approximation breaks down when the expected values are small.

Interesting negative rules can be produced with Generalized negative association rules (GNAR) [3], and this approach can speedup execution time efficiently through the domain taxonomy. Advantage of taxonomy tree is to eliminate unnecessary rules. But, if the domain does not have any taxonomy tree then this algorithm will not be sufficient.

Dong, X et al. [4] proposed an algorithm called PNAR_IMLMS, which produces valid association rules based on correlation coefficient but the drawback of this algorithm is, negative rules extract from uninteresting pattern sets which is useless. Optimized association rule mining with genetic algorithm produces more reliable interesting rules.

Old techniques based on support-confidence generate large scale of redundant rules therefore Xiufen Piao et al.[5] proposed a new algorithm. This algorithm performs two functions, correlation and dual confidence. Algorithm's first function is finding correlation value and according to this value, association rules divide into positive and negative rules and second function used dual confidence to reduce unwanted negative rules and mined useful negative rules.

An accurate classifiers is better than decision tree classifier. Accurate classifiers have added a new direction to the ongoing research. In some cases, It is found that the accurate classifier contain the incorrect rules. The idea is to put accurate negative rules in place of inaccurate positive rules. Generate enough number of negative rules efficiently so that classification accuracy is enhanced. Associative classifier with negative rules (ACN) was proposed by Kundu,G et al.[6] which is time-efficient method and it can be used for classification perspective.

A novel approach was proposed by Wu. X et al. [7]. In this, mining both positive and negative association rules of interest can be decomposed into the following two sub problems, (1) generate the set of frequent itemsets of interest and the set of infrequent itemsets of interest (2) extract positive rules PL, and negative rules NL.

Wei-Guang et. al.[8] proposed an algorithm called SRM to discover subset of negative association rules of the type X=> ⌐ Y which are used to find items that are substitutes for others in market basket analysis.

Fuzzy Weighted Associative Classification based on Positive and Negative Rules algorithm proposed in [9] generated only Confined Negative rules and have shown that the efficiency of the algorithm has been increased by adding negative rules to the positive rules.

III. NEGATIVE CLASS ASSOCIATIVE RULES

Class Associative Rule (CAR) is an implication of the form X => C, where X is an itemset and C is a Classification class attribute. Classification algorithms are used to build a model from a set of training data whose target class labels are known and then this model is used to classify unseen instances. Associative classifier will provide improved performance when negative rules were also employed in the training and classification process. Most common framework in the association rule generation is the *Support- Confidence.* As given in the last section negative rules are of two types, confined and generalized negative rules and if the support of a rule is greater than or equal to user defined minimum support *(minsup)* threshold and minimum confidence *(minconf)* threshold then the itemset is called frequent itemset otherwise as infrequent itemset.

Confined negative class associative rules are of the form ⌐X => C, X => ⌐C, ⌐X => ⌐C

Support of these rules can be calculated as

- $sup(\neg X \Rightarrow C) = sup(X) - sup(X \cup C)$ ---------(1)
- $sup(X \Rightarrow \neg C) = sup(C) - sup(X \cup C)$----------(2)
- $sup(\neg X \Rightarrow \neg C) = 1 - sup(X) - sup(C) + sup(X \cup C)$ ---------------------(3)

Confidence of these rules are

- $conf(X \Rightarrow \neg C) = \frac{sup(X) - sup(X \cup C)}{sup(X)}$ ----------(4)
- $conf(\neg X \Rightarrow C) = \frac{sup(C) - sup(X \cup C)}{1 - sup(X)}$ -----------(5)
- $conf(\neg X \Rightarrow \neg C) = \frac{1 - sup(X) - sup(C) + sup(X \cup C)}{1 - sup(X)}$ -----------------------(6)

Generalized negative class association rules are as follows

For example let A,B,C are attributes and Z is a class attribute. Let the domains of these attributes are {a1,a2,a3}, {b1,b2},{c1,c2,c3,c4} and {u, v} then few frequent rule items of length 3 are

$a1 \wedge \neg b1 \wedge c1 \Rightarrow u$
$a2 \wedge b2 \wedge \neg c2 \Rightarrow u$
$a3 \wedge \neg b3 \wedge \neg c3 \Rightarrow v$.

Support of the rule $a1 \wedge \neg b1 \wedge c1 \Rightarrow u$ can be calculated as

- $sup(a1 \wedge \neg b1 \wedge c1 \Rightarrow u) = sup(a1 \wedge c1 \Rightarrow u) - sup(a1 \wedge b1 \wedge c1 \Rightarrow u)$ -----------------(7)
- $sup(a1 \wedge \neg b1 \wedge c1) = sup(a1 \wedge c1) - sup(a1 \wedge b1 \wedge c1)$ ------------------(8)

Confidence of the rule is

- $conf(a1 \wedge \neg b1 \wedge c1 \Rightarrow u) = \frac{sup(a1 \wedge \neg b1 \wedge c1 \Rightarrow u)}{sup(a1 \wedge \neg b1 \wedge c1)}$ ---------------------(9)

Many algorithms have been proposed for mining negative association rules and illustrated some weaknesses in the Support- Confidence framework. Several measures have been proposed for discovering interesting rules such as Correlation measure, Chi Square measure, mininterest etc., Most of the authors have proposed generation of only Confined negative association rules, because the generation of Generalized negative association rules is a complicated process. In the discovery of generalized negative association rules, there will be a considerable exponential growth in candidate generation, moreover memory requirement is also one of the problem and the existing pruning techniques are not sufficient.

IV. PROPOSED METHOD

Associative Classification produces less number of Negative class association rules and Association rule mining produces more number of Negative association rules. Associative Classification technique can be used as a prediction technique[9].

We propose a method to discover highly interesting generalized negative class association rules. Frequent and infrequent itemsets are generated using support, confidence measures and correlation measure.

Correlation coefficient measures the strength of the linear relationship between a pair of two variables. For two variables X and Y, the correlation coefficient is given by the following formula:

$$\rho = \frac{Cov(X,Y)}{\sigma x \sigma y} \quad \text{----------------------(10)}$$

Cov(X, Y) represents the covariance of the two variables and σx stands for the standard deviation. The range of values for ρ is between -1 and +1. If the two variables are independent then ρ equals 0. When ρ = +1 the variables considered are perfectly positive correlated. Similarly, When ρ= -1 the variables considered are perfectly negative correlated.

Using positively correlated itemsets, positive class association rules are generated and with negatively correlated itemsets Confined negative class association rules are generated. Generalized negative class association rules are generated by the computation of contingency table for each attribute. If the observed frequency is greater than the expected frequency, the model of no effect is rejected and high frequent items can be found.

Formula for calculating expected frequency is,

$$EF_{ij} = \frac{Ti \; X \; Tj}{N} \quad \text{---------------------- (11)}$$

Where EF_{ij} is the expected frequency for the cell in i^{th} row and j^{th} column. T_i is the total number of subjects in the i^{th} row, T_j is the total number of subjects in the j^{th} row and N is the total number of subjects in the table.

If $OF_{ij} > EF_{ij}$ then the subject in the concerned cell is highly frequent, where OF_{ij} is observed frequency of the cell in i^{th} row and j^{th} column.

All combinations of attributes are generated i.e., 'n' attributes produce 2^n combinations, and less frequent items with negation in conjunction with high frequent items with non-negated. The algorithm is as follows,

ALGORITHM : Positive and Negative Class Association Rule generation
Input : TD, minsup, minconf, mincorr
Output : Positive and Negative Class Association Rule

1. Generate Classification Association Rules using frequent itemset generation and is represented as X =>c, where X is an itemset and c is a class label. Examples of such rules are {(Age, 'Young'), (Income, 'High')} (Student, 'Yes')} =>(Buy-Computer, 'Yes').
2. Compute Correlation (corr) between itemset and class attribute.
3. If corr >= mincorr
4. Compute Support (S) and Confidence (C) for each rule.
5. If S >= minsup && C>= minconf
6. Store these rules into Positive rule base
7. If corr < - mincorr
8. Generate Confined Negative rules such as {X=>¬c}, {¬X=>c} and {¬X=>¬c}
9. Compute support S and confidence C for each rule.
10. If S >= minsup && C >= minconf
11. Store these rules into Negative rule base.
12. Generate Contingency table by calculating Observed Frequency(OF) and Expected Frequency (EF) of each attrubute
13. If OF > EF then Bit Vector BV = 1
14. else BV = 0
15. Generate Generalized Negative Rule using BV If BV == 1 then item is positive Else the item is negative such as ⌐A (y) ∧ S (Y) ∧ ⌐ I (M) => BC
17 Repeat step (4) to (16) until maximum candidate itemset length is obtained.
18. Keep all the Positive and Negative rules into main rule base

In the above algorithm binary vector BV, is an array of one's and zero's which is used to generate generalized class association rules

For example consider buying of a computer using transaction data base and domain of attributes such as Age (Youth, Middle-aged, Senior), Student (Yes, No), Income (Low, Medium, High) and class attribute Buy-Computer(Yes, No). Contingency tables for two attributes are shown in the Table.4.1 and Table.4.2, Observed Frequency and Expected Frequency are calculated, the cells with Observed frequency greater than Expected frequency are denoted with * symbol. The cells with * symbol are having high frequency and remaining cells are less frequency attributes. To represent this in the rule, a bit vector array is generated. Binary Vector for an attribute *Age and Income* generated will be as in the Table 4.3 and Table. 4.4. If BV=0 then the attribute is an infrequent item. In this example *Youth* will not buy computers, that is A(Y) is infrequent and will be designated with negative ⌐A (Y). Generalized class association rules are generated by considering less frequency attributes with negative symbol. Few rules shown in the Table 4.5.

AGE	Buy-Computer	Not Buy-Computer	Total
Youth Exp Freq	25 36.2	*33 21.8	**58**
Middle aged Exp Freq	*41 28.7	5 17.29	**46**
Senior Exp Freq	32 33.08	*21 19.92	**53**
Total	**98**	**59**	**157**

Table. 4.1

INCOME	Buy-Computer	Not Buy-Computer	Total
Low Exp Freq	*34 29.34	13 17.66	**47**
Moderate Exp Freq	*41 39.32	22 23.67	**63**
High Exp Freq	23 29.34	*24 17.66	**47**
Total	**98**	**59**	**157**

Table. 4.2

AGE	Buy-Computer	Not Buy-Computer
Youth	0	1
Middle Aged	1	0
Senior	0	1

Table. 4.3

INCOME	Buy-Computer	Not Buy-Computer
Low	1	0
Moderate	1	0
High	0	1

Table. 4.4

GNCAR	Confidence
¬A (Y) ∧ S (Y) ∧ I (M) => BC	1
¬A (Y) ∧ S (Y) ∧¬ I (H) => BC	0.86
¬A (Y) ∧ ¬ S (N) ∧ I (L) => BC	0.75
¬A (Y) ∧ ¬ S (N) ∧ I(M) => BC	1
A(M) ∧ ¬S(N) ∧ I(L) => BC	0.8
A(M) ∧ S(Y) ∧ ¬I(H) => BC	1
A(M) ∧ ¬S(N) ∧ I(M) => BC	1
¬A (S) ∧ S (Y) ∧ I (L) => BC	1
¬A (S) ∧ S (Y) ∧ I (M) => BC	1
¬A (S) ∧ ¬ S (N) ∧ I (L) => BC	0.75
¬A (S) ∧ ¬ S (N) ∧ I (M) => BC	1

Table. 4.5

Contingency tables are used to test hypothesis in order to decide whether or not effects are present. Effects in a contingency table are defined as relationship between rows and column variables. The association between row and column attributes are found and less correlated attributes are denoted with negation. A number of algorithms have been proposed recently to generate negative rules but produce large number of negative rules and also suffer from multi scan problem. In this process pruning will also be done while generating the rules.

V. CONCLUSIONS

Mining of negative association rules have got a greater demand and various algorithms being added to the literature. Negative rules are combination of both Confined and Generalized rules. Most of the algorithms have proposed for mining Confined Negative rules. Generalized negative rule generation is a complicated process, since it is well known that the itemset generation process itself is an expensive one. Associative Classification is gaining more importance, it associates attributes to class labels. In this paper we proposed a method to generates both positive and negative CARs, which will discover both Confined and Generalized negative class association rules. The proposed method is very simple and easy to implement and generate strong negative rules.

REFERENCES

[1] John Tsiligaridis, "Mining Positive and Negative Association Rules: An Approach for Binary Trees", (IJNCAA) 3(3): 81-87 The Society of Digital Information and Wireless Communications.

[2] Brin,S., Motwani,R. and Silverstein,C., " *Beyond Market* Baskets: Generalizing Association Rules to *Correlations*," Proc. ACM SIGMOD Conf., pp.265-276, May 1997.

[3] Li-Min T sai, Shu-Jing Lin and Don-Lin Yang: "Efficient Mining of Generalized Negative Association Rules" in International Conference on Granular Computing 978-0-7695-4161-7/10, IEEE(2010).

[4] Dong, X., Niu, Z., Shi, X., Zhu, D.: "Mining both Positive and Negative Association Rules from Frequent and Infrequent itemsets", ADMA, LNAI 4632, Springer-Verlag Berlin Heidelberg(2007).

[5] Xiufend Piao,Zhanlong Wang,Gang Liu "Reasearch on mining positive and negative association rules based on dual confidence"978-1-4244-9954-0/11 @31 @2011.

[6] Kudu, G.Islam,M.M; Munir,S.; Bari ,M.F . "An associative classifier with negative rules"978-1-4244-2172-5 @31 @ 2008

[7] Wu, X., Zhang, C., Zhang, S.: Efficient m*ining both positive and negative association rules*. ACM Transactions on Information Systems, Vol. 22, No.3, July 2004,Pages 381-405.

[8] Wei-Guang Teng, Ming-Jyh Hsieh, and Ming-Syan Chen. A statistical framework for mining substitution rules. Knowl. Inf. Syst., 7(2):158{178}, 2005.

[9] M. Kalpana Devi, M. Usharani, "Fuzzy Weighted Associative Classifier based on Positive and Negative rules", IJCSIT, Vol. 5 (4) , 5400-5403, 2014.

[10] B. Kavitha Rani, K. Srinivas, B. Ramasubbareddy,: " Mining Negative Association Rules" , IJET, Vol.3 (2), 100-105, 2011.

[11] Chandrasekar Ravi, Neelu Khare and N. Deepa," NP-FARM: Negative and Positive Fuzzy Association Rule Mining in Transaction Dataset", IJST, *Vol 7(11), 1774–1779, November 2014*

[12] Nikky Suryawanshi Rai, Susheel Jain, Anurag Jain, " Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm (MIPNAR_GA)", *IJACSA, Vol. 5, No. 1, 2014.*