MINING POSITIVE AND NEGATIVE ASSOCIATION RULE FROM FREQUENT AND INFREQUENT PATTERN BASED ON IMLMS_GA

Nikky Rai
PG Research Scholar
Department of CSE
RITS, Bhopal, India

Susheel Jain
Assistant Professor
Department of CSE
RITS, Bhopal, India

Anurag Jain
Head of the Department
Department of CSE
RITS, Bhopal, India

ABSTRACT
Association rule mining is one of the most significant tasks in data mining. The essential concept of association rule is to mine the positive patterns from transaction database. But mining the negative patterns has also received the interest of publishers in this region. This paper shows an efficient algorithm (IMLMS-GA) for mining both positive and negative association rules in transaction databases. The goal of this study is to build up a new model for mining negative and positive (PR & NR) association rules out of transaction data sets. The proposed model is based on two models, the MLMS model and the Interesting Multiple Level Minimum Supports (IMLMS) model. This paper proposes a new approach (IMLMS-GA) for mining both negative and positive association rules. The interesting frequent patterns and infrequent patterns mined by the IMLMS-GA algorithm. This algorithm is accomplished in two phase: a. First phase find all frequent patterns & infrequent patterns b. Second phase efficiently generate positive and negative association rule by using useful frequent pattern set. The experimental results prove that the IMLMS-GA can remove the scale of uninteresting association rules and generates better results than the previous positive and negative association rule mining algorithm.

Keywords
Positive and Negative rules, Correlation coefficient, Frequent pattern set and Infrequent pattern set.

1. INTRODUCTION
Association rules was initially introduced by R. Agrawal et al [1], which is an important method in data mining. Initially association rules, for mining user transaction database association between pattern sets, that is the form X→Y, interesting rule, strong relation, which is defined as positive association rule, is a strong association of the correspondent model, and there are number of mining algorithm[2-9]. In fact, a lot of algorithms which is based on this technology cannot extract the unseen patterns from the transaction databases, these unseen patterns are called the negative association rule, which has week relation, less frequency, meaningful association and the characteristic of the meaningful association in the transaction databases which is difficult to extract. These rules describe that which data pattern has less frequency and it can define a week relation but hold very precious information, such as: X→¬Y、¬X→Y、¬X→¬Y. So this mining unseen pattern is very important.

Presently used for the negative association rules mining algorithm is not too much, such as [3] proposed a positive and negative association rules mining algorithm based on minimum interestingness, a positive and negative association rules mining algorithm based on support, confidence and correlation coefficient in [4], and a negative association rules mining algorithm based on Dual confidence in [5].

Negative association rules look for all infrequent pattern sets, and the infrequent pattern sets in the database is exponential, which is also the main reason of studying negative association rules more difficult than positive association rules. While obtained the support of all infrequent pattern set is meaningless. For example, for the negative association rules X→¬Y, X∪Y is the infrequent pattern set, but both X and Y are frequent pattern sets, Therefore, a negative association rules mining can be divided into two parts: a: Find all the frequent pattern sets from the transaction database; b: Determine negative association rules based on frequent pattern sets[6]. There are many algorithms can be used directly to solve the first part, such as the Apriori algorithm, FP-growth algorithm Improved Apriori. This paper describes an association rule mining algorithm to extract meaningful association or interesting patterns in massive data. Positive association rules are used to discover the interesting relationship between the pattern sets and other in the large data, generating the meaningful association between these patterns.

2. BASIC CONCEPTS

2.1 Positive and Negative rules
A strong positive association is referred to as a positive relation between two data sets. Negative relation implies a negative rule between the two pattern sets. However, strong negative associations disclose only the existence of negative rules in an unseen representation and do not give the real negative rules .Unlike existing mining techniques, this paper extends conventional associations to include association rules of forms A→¬B, ¬A→B, and ¬A→¬B, which indicate negative associations between pattern sets. Lets illustrate rules of the form A→B positive rules, and rules of the another forms negative rules [7]. While positive association rules are useful in marketing analysis, negative association rules also play key roles in marketing analysis.

For example: there are four types of products each product has profit gain P=20%, Q=30%, R=60%, S=80%.The goal of the market analysis team is to consider only those product which has high profit gain, positive rules help in determine which product can be adopt, negative association rules help in determine which product can be ignored, The goal of market analysis team is to ensure a fair and efficient trading for all suppliers through a signal system assume that each piece of

evidence P, Q, R and S, can cause a signal of profit and loss, if having rules in the form P→¬X, Q→¬ X, R →X and S→X ,the team can make the decision of profit gain when P,Q occurs, in other words ,signal caused by P ,Q can be ignored.

2.2 Pruning Strategy

The above study suggest that, there can be a huge number of infrequent pattern set in transaction database, and only some of them are interesting for mining strong association rules .However, pruning strategy is to well-organized search for interesting frequent pattern set. In case of mining positive association rules, adopt minimum support and minimum confidence threshold values to improve the usability of the rules. Through the outcome analysis, found that the association rules are generated in the form X→Y and X→¬Y .In particular situation the association rule of the form ¬X→¬Y is very large, but these negative association rules are little use in real application [8].

Let's assume that the database in a electronic shop contain t transactions .Now lets concern the sale of Laptop (L) and computer(c), and came to mine the rule of the form ¬L→¬C, which shows contradict relation user cannot purchase both items at same time. This rule is not helpful to analyze the market basket. So consider a pruning strategy that, not to adopt the part of association rules of the form ¬L→¬C to reduce search space and improved efficiency.

2.3 Genetic Algorithm

Genetic Algorithm (GA) is an artificial intelligence method. It is based on the theory of natural selection and development. A genetic algorithm (GA) is a heuristic scan the process of natural development. This heuristic scan is normally used to generate useful keys for optimization and search problems. Genetic algorithm is an iterative or level wise procedure that is appropriate for optimization problems. In order to use the genetic algorithm, the following points must be measured: Fitness value, Selection, Crossover and Mutation. Genetic algorithms place in to the big class of evolutionary algorithms (EA), which produces an optimize solution by using natural development, such as mutation, selection, and crossover, heredity relation [9].

This paper describes IMLMS_GA is applied on small data instances to discover an optimized association rule. Initially extract the sample of records from the transaction database. The algorithm learning starts as follows. A primary step population is created by randomly generated transactions. This proposed algorithm is based on extracting frequent pattern set repeatedly transforms the population by executing the following moves: (1) Fitness value: The fitness value is calculated for each individual. (2) Selection procedure: Individuals are select from the current population as parents to be concerned in reproduction. (3)Reproduction: New individuals are produced from the parents by applying genetic operators such as crossover and mutation. (4) Replacement: Some of the individuals are replaced with some another individuals (usually with their parents). One complete rotation of transforming a population is called as generation.

3. ALGORITHM DESIGN

3.1 Problem statement

Place L= {i1, i2....in} that contains a group of n items. Give a transaction database DB, where each transaction T is a group pattern set of L. If X is a subset of L with X is subset of T ,this proofs that a transaction T contains X. X negative association rules are an implication of the form ¬X→Y(or X→¬Y ,or ¬X→¬Y) ,Where X, Y is subset of T and X ∩ Y= Ø . Given support S and confidence C. If DB has (100 × sup) % of the transaction contains Y but does not contain X, the support of negative association rules ¬X→Y is S, denoted as S (¬X→Y) =sup. If the transaction does not contain X, there are (100 × conf) % of the transaction contains Y, the confidence of negative association rules ¬X→Y is C, denoted as C (¬X→Y) =conf .So the support and confidence of negative association rules X→¬Y、¬X→¬Y can also be defined.

Negative association rule find seeks rules of the form rules X→¬Y、¬X→Y, value of support and confidence is greater or equal to user-defined minimum support (minsup) and minimum confidence (minconf) thresholds correspondingly, where X and Y are frequent patterns. In this paper, introducing the minimum correlation theory, so it can reduce the production of the meaningless association rule.

3.2 Correlation Coefficient (CRC)

The algorithm uses the correlation coefficient (CRC) between pattern sets to extract positive and negative association rules. The method of finding those rules whose support and confidence meet some user specified minimum support (minsupp) and minimum confidence (minconf) is called association rule mining, but some algorithm defines this condition is not sufficient for strong rule. So correlation coefficient is needed [10].

According to IMLMS_GA, if supp (X∪Y) = supp (X) ∗ supp (Y), then the rule X→Y is not Strong rule. The rule is strong only if meet the term: supp(X∪Y)−supp(X)*supp(Y) ≥mininterest, but this term is not considered negative association rules, if extended this term negative association rules are generate easily . An extended term is that: |supp (X∪Y)−supp(X)*supp(Y)| ≥ mininterest ,this correction not only meet positive and negative association rules but also check the correlation coefficient of association rules.

There are three kinds of CRC x, y:

If CRC x, y >1, then itemsets X and Y is positive correlation;

If CRC x, y <1, then itemsets X and Y is negative correlation;

If CRC x, y =1, then itemsets X and Y is independent of each other.

3.3 Interesting multilevel minimum support

Interesting multilevel minimum support with genetic algorithm aims to extract positive and negative association rule .The IMLMS_GA algorithm assigns various minimum supports to dataset with different number of length ms(i) is the different minimum support of i-itemset (i=1 to m) which are the threshold for frequent item sets ,ms(0) is threshold for infrequent item set for any pattern set X ,ifs(x)≥ms(i) then X is frequent pattern set and s(x)≤ms(i) then X is infrequent pattern set. However, the number of frequent pattern set is required for extracting to positive rule or strong relation and number of infrequent pattern is required for generation of Negative rule. Fitness function plays an important role in the search of meaningful association and positive rule.

The pruning strategy is customized to be suitable to the MLMS model. Therefore, the MLMS model with the customized pruning strategy is called IMLMS [11]. A variety of methods have been used with the customized pruning

strategy to decrease the number of frequent pattern set and infrequent set by pruning the uninteresting patterns.

The following principle is utilized to prune uninteresting frequent pattern set: P is measured a frequent pattern set of interest(int) if s(P)≥ms(len(P)) and f(X,Y, ms(len(X∪Y)), mi)= 1, where len(X) is the number of items in an pattern set X and F(•) is a fitness function regarding the support and interest of the rule A→B, where int(X,Y) = |supp(X∪Y)−supp(X)*supp(Y)| and mi is the minimum threshold resolute by users , this threshold value is used to control the number of frequent pattern set and infrequent pattern set.

3.4 Roulette Wheel Selection

The simplest selection scheme is a roulette wheel selection, also called probability sampling with replacement .This technique is similar to a roulette wheel ,generally a section of the wheel is assigned to each of the possible individuals based on their fitness value ,which individuals has high fitness is selected. Each individual proportional in size to fitness the individual are mapped to contagious segment of a line, its such that each individual segment is equal in size to fitness a random number is selected and the individual whose segment spans is high is selected the process is repeated until the desire number of individual is obtained. Select a set of number(X>1) those pattern satisfy this condition is considered as a fittest member and discard the rest, ifs(x) ≥ms (i) then X is frequent pattern set and s(x) ≤ms (i) then X is infrequent pattern set.

3.5 Genetic operators

Genetic operators perform the evolutionary process of a population in genetic algorithm .After a new population is formed by selection procedure some individuals of the new population undergo transform by means of genetic operators to form new solution. Presentation of algorithm is depending on genetic operators, mutation and crossover are two primary operators of GA.

3.5.1 Crossover

Customized the standard crossover operator, and generate rules in massive data. A rule is measured as week if it consists less data instances i.e. when less data instances satisfy equally the antecedent and the consequent of the rule. Comparing a rule is considered as strong rule when it covers a lot of data instances. Apply bitwise OR and the bitwise AND for rule generalization. There are number of crossover operators that have been used on binary and real coded GA: single point crossover, multipoint cross over and uniform cross-over. With the help of single point cross over to create a new individual it will generate a single cut point and recombines the first part of the first parent with the second part of the second parent. Two point crossovers is similar to one point cross over except that two point is used in place of one, uniform crossover value of the first parent gene is assigned to the first offspring and the value of second parent gene is to the second offspring with probability with 0.2.

3.5.2 Mutation

Mutation operator is used to maintain genetic diversity. Mutation is a reproduction operator that changes one or more value in a new offspring from its primary state. So the outcome new gene value is added to the gene pool with this gene value, the GA may be able to arrive at better result. Mutation is performed in single individuals and used unary operator, mutation depends on the encoding as well as the crossover, mutation simulates the errors that happen with low probability during duplication.

3.6 Replacement:

A replacement approach is used to ensure that more fit genotypes are always introduced into the population. Uniqueness testing: The advantage of the genetic operators on the parent population may result in identical genotypes in the population. The algorithm first tests to ensure the new offspring does not reproduce any existing member of the population. Replacement strategy with genotype uniqueness enforced preserves genetic diversity inside the population [12] this testing permits the algorithm to discover very useful rule.

4. IMLMS_GA

This paper describes an algorithm which is applied on small data instances to discover the optimized association rules. IMLMS_GA utilizes improved apriori algorithm, this algorithm work in two stages:

a. The first stage produces all useful frequent pattern sets (FPS) by using, roulette wheel method
b. The second stage generates positive and negative rules.

INPUT

No. Of Length

Minimum support =MS

Minimum Confidence=MC

The algorithm consists following steps:

1. Initialize PR←Ø, NR←Ø
2. Find FPS← set of frequent pattern set 1-pattern set, ifs(x) ≥ms (i) then X is frequent pattern set and s(x) ≤ms (i) then X is infrequent pattern set.
   //Selections by roulette wheel method//
3. (For (i=2; fi-1! =Ø; i++)
4. {
5. Ji=Fi-1&& Fi-1
   //Apply pruning used fitness function//
6. For each t ε Ci any sub item set of t is not in fi-1 then J i= Ji − {t}
7. For each t ε Ci fined support (t)
8. For each X, T (X∪T=t)
9. {
10. O X, Y=Association Rule (X, T)
11. O>1
12. If (supp(X∪T)−(sup X)*sup(T))≥ MS&&conf (X→T) ≥MC) then
13. PR←PA∪ (X→T)
14. If O<1
15. {
16. If ((X∪T)−(sup X)*sup(T)) ≥ MS&&conf (X→¬T) ≥MC) then
17. NR←NA∪ (X→¬T)
18. AR=PR∪NR
19. }
20. }
21. }
22. If O=1 then
23. Contradiction Rule CR ←(¬X→¬T)
24. //Replacement (Remove) by using unique testing //
25. Repeat 3 step until all optimized rule is not generated.

O/P-Optimize positive and negative association rule.

5. EXPERIMENTAL RESULTS

These experimental results are used to demonstrate the performance of the IMLMS_GA, for mining both negative and positive rules. Experiments are performed on a computer Intel dual core processor with 2.10 GHz of Control processing Unit, running on a window 7, 64 bit OS and 4 GB memory. All codes are implemented under the MATLAB 7.8.0. Demonstrate the performance of this method on four databases from UCI which involve Lenses, Fertility, Iris and Seeds all information about databases shown in Table 1.

TABLE 1. Characteristics of Database

Date	No. of Instance	Attributes	Classes
Lenses	24	4	3
Fertility	100	10	2
Iris	150	4	3
Seeds	210	7	3

Because IMLMS_GA is designed to mine positive association rule and negative association rule from frequent pattern set with different user defined values: minimum support, minimum confidence and number of length. The results are represent in table 2- 5 .Where L is expressed as total number of length, MS is expressed as minimum support, MC is expressed as minimum confidence and Cr is expressed as contradict rule.

TABLE 2. No. of Rules for database Lenses

Threshold Values	MLMS	IMLMS	IMLMS_GA
MS=30% MC=40% L=4	10	8	2
MS=20% MC=30% L=5	8	5	0
Total Rules	18	14	2

TABLE 3. No. of Rules for database Fertility

Threshold Values	MLMS	IMLMS	IMLMS_GA
MS=30% MC=40% L=4	48	44	12
MS=20% MC=30% L=5	50	47	13
Total Rules	98	91	25

TABLE 4. No. of Rules for Iris database

Threshold Values	MLMS	IMLMS	IMLMS_GA
MS=30% MC=40% L=4	100	106	67
MS=20% MC=30% L=5	104	103	61
Total Rules	204	209	128

TABLE 5. No. of Rules for Seeds database

Threshold Values	MLMS	IMLMS	IMLMS_GA
MS=30% MC=40% L=4	223	180	124
MS=20% MC=30% L=5	127	134	108
Total Rules	350	314	232

The experiments defined in table 2 to 5 displays the number of association rules generated from frequent pattern sets with different support and confidence values. These association rules are mined with three algorithms, MLMS, IMLMS and IMLMS_GA at different threshold values. From above results, IMLMS_GA algorithm can successfully generate less number of association rules than MLMS,IMLMS .For example table 2 through 5 the number of association rules mined by MLMS=18 to 350 (minimum 18-maximum 350),IMLMS=14 to 314 whereas the total rules by IMLMS_GA= 2 to 232.

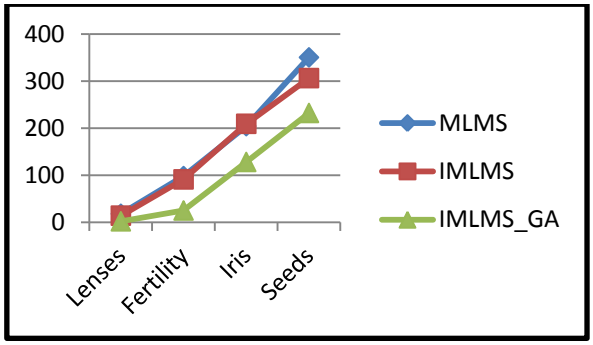

Figure 1: Positive (PR) and Negative rules (NR) are generated by MLMS, IMLMS and IMLMS_GA

6. CONCLUSION AND FUTURE WORK

This paper proposes a new algorithm for mining positive and negative association rules in datasets. This method is quite different from exiting methods. The algorithm is called IMLMS_GA and is suitable for mining positive and negative rule from useful frequent and infrequent pattern sets. The algorithm is based on implemented pruning technique for reducing large number of association rules and improving the performance of algorithm and has used the Genetic fitness function to check which form positive and negative rule should be mined. The only obstacle with the implemented method is that, this mining algorithm does not work very well for large instances of database. In the future, the study is still going on to modify this approach to achieve two goals:

1. An interesting measure is added to this approach for working with cloud computing environment (large instances of databases).
2. To improve the performance of the algorithm

7. ACKNOWLEDGEMENT

I would like to thank Prof. Anurag Jain (HOD, CSE) and Assistant Prof. Susheel Jain, for accepting me to work under his valuable guidance. They closely supervise the work over the past few months and advised many innovative ideas, helpful suggestion, valuable advice and support.

8. REFERENCES

[1] Agrawal R, Imielinski T, Swami A (1993)"Mining Association Rules between Sets of Items in Large Databases" .In Proceeding of the ACM SIGMOD International conference on Management of Data, Washington DC,ACM ,pp207-216.

[2] Yanguang Shen, Jie Liu, Zhiyong yang (2009) "Research on Positive and Negative Association Rules based on Interest Support confidence framework". Published in computational intelligence and software engineering CiSE 2009, IEEE, pp 1- 4

[3] Shi-ju Shang, Xiang –jun Dong ,Jie Li,Yuan-yuan Zhao (2008) "Mining Positive and Negative Association Rules in multi-database based on Minimum Interestingness".Published in Intelligent Computation Technology and Automation (ICICTA) 'International Conference Volume 1', IEEE, pp 791-794.

[4] He Jiang, Yuanyuan Zhao, Chunhua Yang, Xiangjun Dong (2008) "Mining both Positive and Negative Weighted Association Rules with Multiple Minimum Supports".Published in International conference volume 4, Computer science & software Engineering CiSE, IEEE, pp407-410.

[5] By Xiufend Piao, Zhan long Wang, Gang Liu (2011) "Research on Mining Positive and Negative Association Rules based on dual confidence". Internet computing for Science & Engineering ICICSE, IEEE,pp102-105.

[6] Wu, X., Zhang, C., Zhang, S (2004) "Efficient Mining of both Positive and Negative Association Rules". ACM Transactions on Information Systems 22(3): pp 381–405.

[7] By Idheba Mohamad Ali O,Swesi ,Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir(2012) "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets". 9th international conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE pp 650-655

[8] By Sandeep Kumar, K.Srinivas, Peddi Kishor, and T.Bhaskar (2011) "An Alternative Approach to mine Association Rules". Electronics Computer Technology (ICECT),3rd international conference volume 6, IEEE pp 420-424.

[9] By Chun-Hao-Chen, tzung-Pei Hong and Yeong Chy Lee (2011)"A Multiple level genetic Fuzzy Mining Algorithm". International Conference on Fuzzy System (FUZZ), IEEE, pp 278-282.

[10] By Xiangjun Dong (2011) "Mining Interesting frequent and frequent itemsets based on minimum correlation strength" @ Springer- Verlag Berlin Hedelberg.

[11] By Xiangjun Dong, Zhiyun Zheng ,Zhendong Niu, Donghua Zhu,Qiuting Jia(2008) "Mining Interesting Infrequent and frequent itemsets based on MLMS model ".The fourth international conference on advanced data mining and application ,ADMA 5139.444-451, IEEE pp 444-451 .

[12]C.F.Lima,M.Pelikan,D.Goldberg.K.S.O.G.Lobo(2008)"In fluence of Selection and Replacement strategies on linkage learning in bao. CEC 2007 IEEE pp 1083-1090