RESEARCH ARTICLE OPEN ACCESS

An Optimized method for Generating All Positive and Negative Association Rules

Shivakeshi.C [1], Mahantesh.H.M [2], Naveen Kumar.B [3], Pampapathi.B.M [4]

Assistant Professor [1], [2], [3] & [4]
Department of Information Science and Engineering [1] & [2]
Department of Computer Science and Engineering [3] & [4]
Rap Bahadur Y. Mahabaleswarappa Engineering College
VTU, Ballari
Karnataka - India.

ABSTRACT

Association Rule play very important role in recent scenario of data mining but have only generated positive rule, negative rule also useful in today's data mining task. In this proposed model "An optimized method for generating all positive and negative Association Rules" (NRGA).NRGA generates all association rules which are hidden and have been applied an Apriori Algorithm. To represent the Negative Rules the new names are given as like: CNR, ANR, and ACNR. In this project the Correlation coefficient (CRC) equation is also modified, so all generated results are very promising. Firstly the proposed model applies the Apriori Algorithm for frequent itemset generation and that will also generate positive rules, after on generating frequent itemset it applies NRGA algorithm for all negative rules generation and optimize generated rules using Genetic Algorithm.

Keywords::- Association Rule, Data Mining, Genetic Algorithm, Negative Rule Generating Algorithm (NRGA).

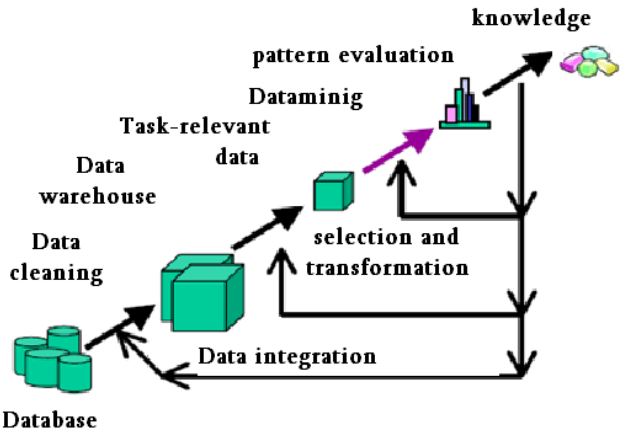
I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes)



.Figure 1: knowledge discovery process.

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of

data stored in multiple data sources such as file systems, databases, data warehouses and etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual.

Data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports for business users to make decisions. Data is now stored in database and/or data warehouse system so data mining system should be designed to decouple or couple with these systems. It leads to four possible architectures of a data mining system.

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database, those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … , Ik}, association rules with this item sets are generated in the following way: the first rule is {I1, I2, … , Ik-1}⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem.

.

The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation process and frequent itemsets generation process. It call those item sets whose support exceed the support threshold as large or frequent item- GESTS International Transactions on Computer Science and Engineering, sets, those itemsets that are expected or have the hope to be large or frequent are called candidate item sets

In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only “interesting” rules, generating only “non redundant” rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

II. OVERVIEW OF ASSOCIATION RULE

A-Apriori Algorithm

In data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. Apriori is the most popular and effective algorithm to find all the frequent itemsets in dataset. It is proposed by Agrawal and Srikant in 1994.

Let I = {I1, I2……, Ik} be a set of k distinct attributes, also called literals. Ai = s is an item, where s is a domain value is attributing, Ai in a relation, R (A 1… An). A is an itemset if it is a subset of I. DT =. {ti, ti+1… tn) is a set of transactions, called the transaction (tid, itemset).A transaction t contains an itemset A if and only if, for all items IA, i is in t-itemset. An itemset A in a transaction database DT has a support, denoted as Supp (A) (we also use p (A) to stand for Supp (A)), that is the ratio of transactions in DT contain A. Supp (A) = |A (t)| / |DT|, Where A(t) = {t in DT/t contains A}. An itemset A in a transaction database DT is called a large (frequent) itemset if its support is equal to, or greater than, a threshold of minimal support (minsupp), which is given by users or experts. An association rule is an expression of the form IF A THEN B (or A → B), A∩B = Ø, where A and B are sets of items. The meaning of this expression is that transactions of the databases, which contain A, tend to contain B. Each association rule has two quality measurements: support and confidence, defined as:

(1) The support of a rule A->B is the support of A∪B, where A ∪B means both A and B occur at the same time in same transaction.

(2) The confidence or predictive accuracy [2] of a rule A→B is conf (A→B) as the ratio:| (A ∪B)(t)| / |A(t) or Supp(A ∪B) / Supp(A).

That is, support = frequencies of occurring patterns, confidence = strength of implication. Support-confidence framework [5][11]: Let I be the set of items in database D, A, B I be itemset, A ∩B = f, p(A) is not zero and p(B) is not zero. Minimal support minsupp) and minimal confidence (minconf) are given by users or experts

Then A→B is a valid rule if

1. Supp (A∪B) is greater or equal to minsupp,

2. Conf (A→B) is greater or equal to minconf.

Mining association rules can be broken down into the following two sub-problems [5]:

1. Generating all itemsets that have support greater than, or equal to, the user specified minimal support. That is, generating all large itemsets.

2. Generating all the rules that have minimum confidence.

B-Negative Association Rule

Typical association rules consider only items enumerated in transactions. Such rules are referred to as positive association rules. Negative association rules also consider the same items, but in addition consider negated items (i.e. absent from transactions). Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. Mining negative association rules is a difficult task, due to the fact that there are essential differences between positive and negative association rule mining. The researchers attack two key problems in negative association rule mining: (i) how to effectively search for interesting itemsets, and (ii) how to effectively identify negative association rules of interest.

Brin et. al [8] mentioned for the first time in the literature the notion of negative relationships. Their model is chi-square based. They use the statistical test to verify the independence between two variables. To determine the nature (positive or negative) of the relationship, a correlation metric was used. This proposed model presents a new idea to mine strong negative rules. They combine positive frequent itemsets with domain knowledge in the form of taxonomy to mine negative associations. However, their algorithm is hard to generalize since it is domain dependant and requires a predefined taxonomy. Wu et derived a new algorithm for generating both positive and negative association rules. They add on top of the support-confidence framework another measure called mininterest for a better pruning of the frequent itemsets generated. In this project use only negative associations of the type X ⇒ ¬Y to substitute items in market basket analysis.

.

The negation of an itemset A is represented by ¬A, which means the absence of the itemset A. We call a rule of the form A=> B a positive association rule, and rules of the other forms (A=>¬B, ¬A=>B and ¬A=> ¬B) negative association rules. The support and confidence of the negative association rules can make use of those of the positive association rules [10].In this proposed model have create a meaning for these type rule like:

Positive Rule (PR) = A=>B

Consequent Negative Rule (CNR) = A=> ¬B

Antecedent Negative Rule (ANR) = ¬A=>B

Antecedent and Consequent Negative (ACNR) = ¬A=> ¬B

The support and Confidence for CNR, ANR.and ACNR rule is given by the following formulas:

Consequent Negative Rule (CNR):

Supp (A=>¬B) =supp (A)-supp (AUB)

Conf (A=>¬B) =supp (A)-supp (A ∪B)

Supp (A)

II- Antecedent Negative Rule (ANR):

Supp (¬A=>B) =supp (B) - supp (A ∪B)

Conf (¬A=>B) = supp (B)-supp (A ∪B)

1-supp (A)

III- Antecedent and Consequent Negative (ACNR):

Supp (¬A=>¬B) =1-supp (A)-supp (B) +supp (A ∪B)

Conf (¬A=>¬B) = 1-supp (A)-supp (B) +supp (A ∪B)

1-supp (A)

The negative association rules discovery seeks rules of the three forms with their support and confidence greater than, lesser then or equal to, user-specified minsupp and minconf thresholds respectively. These rules are referred to as an interesting negative association rule. The algorithm uses the correlation coefficient (CRC) between itemsets to find negative association rules. The correlation coefficient (CRC) between itemsets can be defined as:

$$CRC=\{Supp(AUB)|Supp(A) * Supp(B)\}$$

A and B are itemsets.

When CRC (A, B) = 1, A and B are independent.

When CRC (A, ⌐ B) <1, A and B have negative correlation.

When CRC (⌐ A, B) <1, A and B have negative correlation.

By using Mod1 in eq.7, CRC values not exceed more then 1.it is providing benefit in negative rule generation.

III. Negative Rule Generating Algorithm(NRGA)

NRGA mining is algorithm for generated all Association rule like CNR, ANR and ACN. Let DB is database that contain training dataset T, minsupp, minconf given by the user. This algorithm for extracting both positive and negative association rules as follows:

Algorithm

Input: A training dataset T, minsupp, minconf.

Output: frequent itemsets (FI), CNR, ANR, ACNR.

1) Initialize FI=NULL and CNR=NULL, ANR=NULL, ACNR=NULL.

2) Generate frequent itemsets from T.

FI ἐ T

3) For (any frequent itemset A and ⌐B in FI)

Calculate supports value of A => ⌐B.

3.2 Calculate confidence value of A => ⌐B

3.3 if (supp (A=> ⌐B) ⩾ minsupp and conf (A=> ⌐B) ⩾ minconf)

if (CRC (A, ⌐B) <1)

{

CNR=CNR Ụ (A=> ⌐B).

}

4) For (any frequent itemset ⌐A and B in FI)

4.1 Calculate supports value of ⌐ A => B

4.2 Calculate confidence value of ⌐A => B

4.3 if (supp (⌐A=> B) ⩾ minsupp and conf (⌐A=> B) ⩾ minconf)

if (CRC (⌐A, B) <1)

{

ANR=ANR Ụ (⌐A=> B).

}

5) For (any frequent itemset ⌐A and ⌐B in FI)

5.1 Calculate supports value of ⌐ A => ⌐B

5.2 Calculate confidence value of ⌐A => ⌐B

5.3 if (supp (⌐A=> ⌐B) ⩾ minsupp and conf (⌐A=> ⌐B) ⩾ minconf)

If (CRC (⌐A, ⌐B) <1

{

ACNR=ACNR Ụ (⌐A=>⌐ B).

}

6) Return CNR, ANR, and ACNR.

NRGA generate all positive and negative rules

IV. GENETIC ALGORITHM

Genetic Algorithm (GA) is general purpose search algorithm which use principles inspired by natural genetic populations to evolve solutions to problems [8]. All GAs typically starts from a set, called population, of random solutions (candidate). These solutions are evolved by the repeated selection and variations of more fit solutions, following the principle of survival of the fittest. The elements of the population are called individuals or chromosomes, which represent candidate solutions. Chromosomes are typically selected according to the quality of solutions they represent. To measure the quality of a solution, fitness function is assigned to each chromosome in the population. Hence, the better the fitness of a chromosome, the more possibility the chromosome has of being selected for reproduction and the more parts of its genetic material will be passed on to the next generations. Genetic Algorithms are very easy to develop and validate, which makes them highly attractive, if they applied. The algorithm is parallel, it can be applied to large populations efficiently, so if it begins with a poor original solution it can rapidly progress to good solutions. Use of mutation makes the method capable of identifying global optimal, even in very difficult problem domains. The technique does not need prior knowledge about the distribution of the data, this way Gas can efficiently explore the space of possible solutions. This space is called search space, and it contains all the possible solutions that can be encoded [4]

V. OPTIMIZATION OF ASSOCIATION RULE USING GA

In this proposed model describes the GA algorithm for optimization of association rule associated .First, explanation of how GA algorithm represents the rule individually and encodes scheme and the chromosome structure (Representation of rule) shown. After that, description of genetic operators and fitness function assignment and selection criteria are listed. Finally, the algorithmic structure is given.

A. Representation of Individually in rule and Encoding scheme

Representation of generated rule in GA is play very important role. Mainly two Methods are mostly based on how rules are encoded in the population of individuals (“Chromosomes”) as discussed Michigan and Pittsburgh, In the Michigan Approach each individual encodes a single prediction rule, whereas in the Pittsburgh approach each individual encodes a set of prediction rules. In this project are only interested to generate single rule so, here we are using Michigan approach. GA use various encoding scheme like tree encoding, permutation

encoding, binary encoding etc., Consider following example

If *paper* ***and*** *pencil* ***then*** *eraser not Ink*

Now, following Michigan's approach and binary encoding, for simplicity usage, this rule can be represented as 001 111 010 111 011 111 100 000 where, the bold tri-digits are used as attribute id, like 001 for paper, 010 for pencil, 011 for eraser and 100 the normal tri-digits are 000 or 111 which shows absence or presence respectively. Now this rule is ready for further computations.

B. Chromosomes Structures (Representation of attribute)

GA algorithms a fixed length chromosome structure. In this proposed project using three bit binary encoding for representation Table.3.1 show the attribute representation and Table.3.2 show the Presence and absence of rule , in this project are only interested to take 6 attribute like for example, A,B,C,D,E, and F.

A	B	C	D	E	F
001	010	011	100	101	110

Figure 2; Representation of Attribute in binary encoding.

Presence of Attribute	Absence of Attribute
111	000

Figure 3: Presence and Absence of attribute

C. Genetic operator

Genetic Algorithm uses genetic operators to generate the offspring of the existing population. This proposed model describes three operators of Genetic Algorithms that were used in GA algorithm: selection, crossover and mutation.

1) Selection:

The selection operator chooses a chromosome in the current population according to the fitness function and copies it without changes into the new population.GA algorithm used route wheel selection where the fittest members of each generation are more chance to select.

2) Crossover:

The crossover operator, according to a certain probability, produces two new chromosomes from two selected chromosomes by swapping segments of genes. GA algorithm used single-point crossover operation with probability 0.1; chromosomes can be created as in Fig.4.

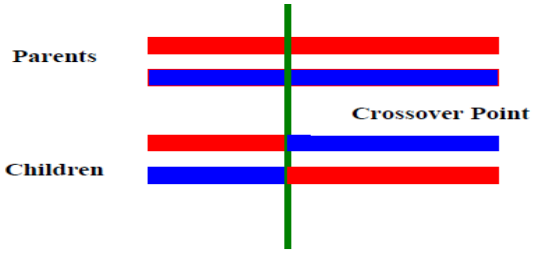


Figure 4: Single point Crossover

3) Mutation: The mutation operator is used for maintaining diversity. During the mutation phase and according to mutation probability, 0.005 in GA algorithm, value of each gene in each selected chromosome is changed.

D. Fitness Function

Ideally the discovered rules should: (a) have a high predictive accuracy ; (b) be comprehensible; and (c) be interesting. The fitness function should be customized to the specific search spaces, thus choice of Fitness function [6] is very important to get the desired results. The population is ranked with the help of fitness function. We apply genetic algorithm on the selected population from the database and compute the fitness function after each step until the genetic algorithm is terminated. Rules generally define as:

IF A THEN B

Where A is the antecedent and C is the consequent. The rules performance can be shown in Fig.5 by a 2x2 matrix, which is called confusion matrix.

Predicted/actual class	Item set A	Not Item set A
Item set B	TP	FP
Not item set B	FN	TN

Figure 5: Confusion matrix for a rule

It is known that higher the values of TP and TN and lower the values of FP and FN, the better is the rule.

Confidence Factor, CF = {TP/(TP+FN)} Mod1

We also introduce another factor completeness measure for computing the fitness function.

Comp = {TP/(TP+FP)}Mod1

Fitness = (CF*Comp) Mod1

The fitness function shows that how much we near the generate the rule. In this fitness function we are using Mod operation with 1 in order to insure that it will not exceed the range of fitness function, which is [0...1]. The fitness function shows that how much we near to generate the rule.

E. Algorithm Structure and Methodology

In this paper presenting algorithm structure and also describes the genetic algorithm is applied over the rules fetched from Apriori association Rule mining. The proposed method for generating association rule by genetic algorithm is as follows:

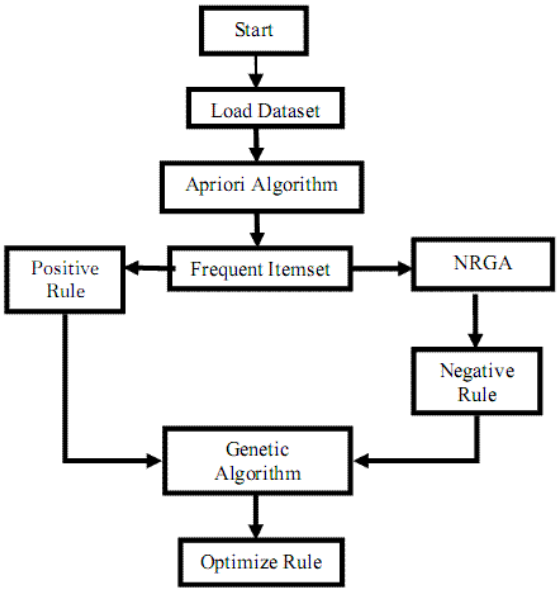


Figure 6: Representation of Algorithm.

1. Start
2. Load a sample of records from the database that fits into the memorization
3. Apply Apriori algorithm to find the frequent itemsets with the minimum support. Suppose S is set of the frequent item set generated by Apriori algorithm.
4. Apply NRGA (Negative rules generating Algorithm) for generation of all rules.
5. Set Q=Ø where Q is the output set, which contains the all association rule.
6. Set the Input termination condition of genetic algorithm.
7. Represent each frequent item set of S as binary encoding, string using the combination of representation specified in method above.
8. Select the two members (string) from the frequent item set using Roulette Wheel sampling method.
9. Apply GA operators, crossover and mutation on the selected members (string) to generate the association rules.
10. Find the fitness function for x=>y each rule.
11. If generated rule is better then previous rule then
12. Set Q = Q U {x =>y}
13. If the desired number of generations is not completed, then go to Step 3.
14. Stop.

VI. CONCLUSION

For resent scenario in market base analysis, Negative rules play very important role in decision making. In this paper we deal with an association rule mining problem for finding Negative and optimized association rules. The frequent item sets are generated using the Apriori association rule mining algorithm. NRGA use modified CRC to generate all negative association rules. After all rule generation, GA are apply to optimize generate rule. The results reported in this project are very promising since the discovered rules are of optimized rules.

REFERENCES

[1] Alex A. Freitas, "Understanding the crucial differences between classification and discovery of association rules - a position paper" ACM SIGKDD Explorations, 2(1):65-69, 2000.

[2] Agrawal R., Imielinksi T. and Swami A. "Database mining: a performance perspective", IEEETransactions on Knowledge and Data Engineering 5 (6), (1993), pp: 914–925.

[3] A. Savasere, E. Omiecinski, and S. Navathe, "Mining for strong negative associations in a large database of customer transactions," In Proc. of ICDE, 1998, pp. 494-502.

[4] Colombetti M. and Dorigo M... "Training Agents to Perform Sequential Behavior"., Italian National Research Council, 1993, pp: 93- 023.

[5] Das Sufal and Saha Banani" Data Quality Mining using Genetic Algorithm" International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (2)

[6] Manish Saggar and Agarwal Ashish Kumar "Optimization of Association Rule Mining using Improved Genetic Algorithms" 2004 IEEE Computer Society Press.

[7] Olafsson Sigurdur, Li Xiaonan, and Wu Shuning.Operations research and data mining,in: European Journal of Operational Research 187 (2008) pp:1429–1448.

[8] Wook J. and Woo S.. New Encoding/Converting Methods of Binary GA/Real-Coded GA. IEICE Trans, 2005 Vol.E88-A, No.6, 1545- 1564.

[9] W. Teng, M. Hsieh, and M. Chen, "On the mining of substitution rules for statistically dependent items," In Proc. of ICDM, 2002, pp. 442-449.

[10] X. Dong, S. Wang, H. Song, and Y. Lu, "Study on Negative Association Rules," Transactions of Beijing Institute of Technology, Vol. 24, No. 11, 2004, pp. 978-981.

[11] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," ACM Transactions on Information Systems, Vol. 22, No. 3, 2004, pp. 381–405.