Distance Weight Parity Optimization of Association Rule Mining With Genetic Algorithm

Nikhil Jain & Vishal Sharma
Department of Computer Science and Engineering
Jawaharlal Institute of Technology, Khargone(M.P.)

Abstract- Association rule mining is the method for finding interesting relations between variables in database. In this paper we focus on the problem of mining association rules. Association rule mining suffered by the generation of negative rules and superiority problem. To overcome this problem we use genetic algorithm which give us the optimized association rules. By using Apriori algorithm we prune the candidate itemsets that can't meet the minimum support count.

Keywords- Association rule mining; support; confidence

1 Introduction

Data mining is play very important role in current growing rate of internet data. It is also a vital field of research in the field of pattern extraction and gathering of information on given database. The task of data mining is to extract useful knowledge for human users from a database. Here as the application of evolutionary computation to data mining is not always easy due to its heavy computation load especially in the case of a large database [6].

2 Association Rule Mining

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 item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence[1]. Suppose one of the large item sets 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 item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item-sets, those item sets 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[3]. 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 "no redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

3 Apriori Algorithm

The work of the Apriori Algorithm is to find associations between different sets of data. Each set of data has a number of items and is called a transaction. The result of Apriori is sets of rules that tell us how often items are contained in sets of data. It is based on the breadth first search algorithm which is moving in upward direction. The first pass of the algorithm counts item occurrences to find the frequent itemsets .A subsequent pass k consists of two steps. First the frequent itemsets L_{k-1} found in the $(k-1)^{th}$ pass are used to generate candidate itemsets C_k. Next the database is scanned and the support of candidates in C_k is counted. The set of candidate itemsets is a pruning process to ensure that all the subsets of candidate sets are already known to be frequent itemsets. The pruning process and the candidate generation process is the part of the algorithm.

Candidate Generation is the process in which L_{k-1} the set of all frequent (k-1) itemsets, we want to generate superset of the set of all frequent k-itemsets.the motive behind the apriori candidate generation is that if an item set X has a minimum support, so do all subsets of X.

The Pruning steps eliminates the extension of (k-1)-itemsets which are not found to be frequent from being considered for counting support.

The Apriori algorithm uses these two functions at every iteration.

4 Genetic Algorithm

Genetic Algorithm (GA), first introduced by John Holland in the early seventies, is the powerful stochastic algorithm based on the principles of natural selection and natural genetics, which has been quite successfully, applied in machine learning and optimization problems. To solve a Problem, a GA maintains a population of individuals (also called strings or chromosomes) and probabilistically modifies the population by some genetic, operators such as selection, crossover and mutation, with the intent of seeking a near optimal solution to the problem. Coding to Strings in GA [4, 5], each individual in a population is usually coded as coded as a fixed-length binary string. The length of the string depends on the domain of the parameters and the required precision.

A. Initial Population

The initial process is quite simple. Create a population of individuals, where individual in a population is a binary string with a fixed-length, and every bit of the binary string is initialized randomly.

B. Evaluation

In each generation for which the GA is run, each individual in the population is evaluated against the unknown environment. The fitness values are associated with the values of objective function.

C. Genetic Operators

To perform genetic operators, must select individuals in the population to be operated on .The selection strategy is chiefly based on the fitness level of the individuals actually presented in the population. There are many different selection strategies based on fitness. The most popular is the fitness proportionate selection. After a new population is formed by selection process, some members of the new populations undergo transformations by means of genetic operators to form new solutions (a recombination step). Because of intuitive similarities, we only employ during the recombination phase of the GA three basic operators: reproduction, crossover and mutation, which are controlled by the parameter rp, cp and mp(reproduction probability, crossover probability and Mutation probability), respectively. Let us illustrate

these three genetic operators. As an individual is selected, reproduction operators only copy it from the current population into the new population (i.e., the new generation) without alternation. The crossover operator starts with two selected individuals and then the crossover point (an integer between 1 and L-1, where L is the length of strings) is selected randomly. Assuming the two parental individuals are x1 and x2, and the crossover point is 5 (L=15). If

X1= (11101|1001001110)

X2= (01110|0100111100)

Then the two resulting offspring are

X'1= (00010|0110110001)

X'2= (01110|1011000011)

The third genetic operator, mutation, introduces random changes in structures in the population, and it may occasionally have beneficial results: escaping from a local optimum. In our GA, mutation is just to negate every bit of the strings, i.e., changes a 1 to 0 and vice versa, with probability pm.

5 Proposed Methodology

The proposed algorithm is a combination of support weight value and near distance of superior candidate key and parity based selection of rule based on group value of rule. Support weight key is a vector value given by the transaction data set and plays a role of rule selection on the base of genetic parity order. The support value passes as a vector for finding a near distance between superior candidate key. After finding a superior candidate key the nearest distance divide into two classes, one class take a higher odder value and another class gain lower value for rule generation process. The process of selection of class also reduces the passes of data set. After finding a class of lower and higher of given support value compare the value of distance weight vector. Here distance weight vector work as a fitness function for selection process of genetic algorithm. Here we present steps of process of algorithm step by step

Steps of algorithm

1. Select data set
2. Put value of support and confidence
3. Start scanning of transaction table
4. Count frequent items
5. Generate frequent itemsets
6. Check the transaction set of data is null
7. Put the value of support as weight
8. Compute the distance with Euclidean distance formula
9. Generate distance vector value for selection process
10. Assign random parity of each group of selected vector
11. Initialized a population set (t=1)
12. Compare the value of distance vector with population set
13. Selected value of parity arrange by distance weight factor.
14. If value of support greater than vector value of parity odder.
15. Processed for encoded of data

16. Encoding format is binary
17. After encoding offspring are performed
18. Set the value of probability for mutation and the value of probability is 0.007.
19. Set of rule is generated.
20. Check superiority of rule in set of parity group
21. If rule is not superior go to selection process
22. Else optimized rule is generated.
23. Exit

6 CONCLUSION

We have dealt with association rule mining problem of finding optimized association rules. The frequent itemsets are found by Apriori algorithm and after that Genetic algorithm has been applied on frequent itemsets to generate the optimized rules.
in this paper the authors have tried to use the enormous robustness of GAs in mining the Association Rules. The results generated when the technique applied on the synthetic database, includes the desired rules, i.e. rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining. The authors believe that the toolkit can also handle other databases, after minor modifications. As for future work, the authors are currently working on the complexity reduction of Genetic Algorithms by using distributed computing.

References

[1] By Rakesh Agrawal Tomasz Imielinski Arun Swami Mining Association Rules between Sets of Items in Large Databases ACM SIGMOD Conference Washington DC, USA, May 1993.

[2] By Rakesh Agrawal Ramakrishnan Srikant Fast Algorithms for Mining Association Rules VLDB Conference Santiago, Chile, 1994.

[3] By Ramakrishnan Srikant Rakesh Agrawal Mining Generalized Association Rules VLDB Conference Zurich, Swizerland, 1995.

[4] By Pengfei Guo Xuezhi Wang Yingshi Han The Enhanced Genetic Algorithms for the Optimization Design 978-1-4244-6498-2/10/$26.00 © IEEE 2010.

[5] By Q. C. Meng, T.J. Feng I , *2.* Chen I , C.J. Zhou , J.H. Bo2 Genetic Algorithms Encoding Study and A Sufficient Convergence Condition of GAS 0-7803-5731-0/9)sk$l0.00 0 IEEE 1999.

[6] By Dieferson Luis Alves de Araujo' , Heitor S. Lopes', Alex A**.** Freitas2 A Parallel Genetic Algorithm for Rule Discovery in Large Databases 0-7803-5731-0/99$$10.00109 99 IEEE.