Optimization of Association Rule Mining through Genetic Algorithm

RUPALI HALDULAKAR

School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, Madhya Pradesh India

Prof. JITENDRA AGRAWAL

School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya
Bhopal, Madhya Pradesh India

Abstract:-

Strong rule generation is an important area of data mining. In this paper we design a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that we use the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules .In this direction for the optimization of the rule set we design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set.

Keywords: Data Mining, Association Rule, genetic algorithm

1.0 Introduction

Concepts of mining associations are given as follows [5]. Let $I = \{I1 , I2 , ..., \text{Im} \}$ be a set of items and $D = \{\text{t1,t2,…………tn}\}$ be a set of transactions, where *ti* is a set of items such that $ti \subset I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$ and $X \cap Y =¢$. The rule $X \Rightarrow Y$ holds in the set D with *support* and *confidence*, where *support* is the percentage of transactions in D that contain both X and Y and *confidence* is the percentage of transactions in D containing X that also contain Y. An association rule must satisfy a user-set minimum support (*minsup*) and minimum confidence (*minconf*). The rule $X \Rightarrow Y$ is called a strong association rule if *support* ≥ *minsup* and *confidence* ≥ *minconf*. Usually association analysis is not given decision attributes so that we can find association and dependence between attributes to the best of our abilities. But the aimless analysis may take much time and space. Decision attributes determined can reduce the amount of candidate sets and searching space, and then improve the efficiency of algorithms to some extent [12].In addition, users are not interested in all association rules, but they are just concerned about the associations among condition attributes and decision attributes. So, in this paper we just mine association rules with decision attributes, in that, we consider attributes which users are concerned about as decision attributes and other attributes as condition attributes. If mining association rules from continuous attributes data, the continuous attributes have to be discretized first. The essence of discretization is to use the selected cut-points to divide the values of the continuous attributes into intervals. The methods of dividing determine the quality of association rules [2].While the genetic algorithm has good optimization. In this paper, genetic algorithm is used to search the cut-points of the continuous attributes. This paper start with as in section 2 describes genetic algorithm and genetic operators. Section 3 we define the rule generation by apriori algorithm. In section 4 we explain the optimization through GA. Section 5 represent the flow chart of proposed work. The section 6 defines Experiment and result and the last one section 7 define conclusion and future trends.

2.0 Genetic algorithm

A genetic algorithm ([7]) is a type of searching algorithm. It searches a solution space for an optimal solution to a problem. The algorithm creates a “population” of possible solutions to the problem and lets them “evolve” over multiple generations to find better and better solution. Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. Cycle of the Algorithm: The algorithm operates through a simple cycle

- Creation of a population of strings.
- Evolution of each string.
- Selection of the best string.
- Genetic manipulation to create a new population of strings.

Operators of Genetic Algorithm

The Genetic operators determine the search capability and convergence of the algorithm. Genetic operators hold the selection crossover and mutation on the population and generate the new population.

Selection

Chromosomes are selected from the population to be parents to crossover. The problem is how to select these chromosomes. According to Darwin's evolution theory the best ones should survive and create new offspring. There are many methods how to select the best chromosomes, for example roulette wheel selection, Boltzmann selection, tournament selection, rank selection, steady state selection and some others.

Encoding of a Chromosome

The most used way of encoding is a binary string. The chromosome then could look like this:

Chromosome 1	1101100100110110
Chromosome 2	1101111000011110

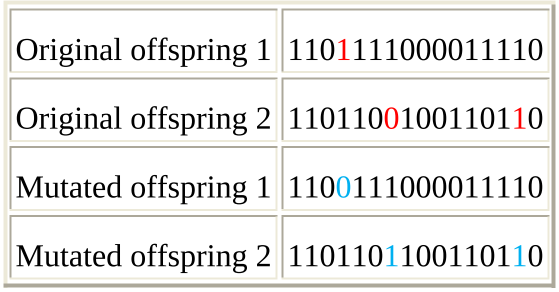
Crossover

After we have decided what encoding we will use, we can make a step to crossover .Crossover selects genes from parent chromosomes and creates a new offspring. There is different way like single point crossover other is multipoint crossover.

Chromosome 1	11011 | 00100110110
Chromosome 2	11011 | 11000011110
Offspring 1	11011 | 11000011110
Offspring 2	11011 | 00100110110

Mutation

After a crossover is performed, mutation takes place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1. Mutation can then be following:



Original offspring 1	1101111000011110
Original offspring 2	1101100100110110
Mutated offspring 1	1100111000011110
Mutated offspring 2	1101101100110100

Outline of Basic Genetic Algorithm

1. **[Start]** Generate random population of n chromosomes (suitable solutions for the problem)

2. **[Fitness]** Evaluate the fitness f(x) of each chromosome x in the population

3. **[New population]** Create a new population by repeating following steps until the new population is complete

Selection: Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)

Crossover: With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.

Mutation: With a mutation probability mutate new offspring at each locus (position in chromosome).

Accepting: Place new offspring in a new population

4. **[Replace]** Use new generated population for a further run of algorithm

5. [**Test]** If the end condition is satisfied, stop, and return the best solution in current population

6. **[Loop]** Go to step 2.

A typical flowchart of a genetic algorithm is shown in Figure. One iteration of the algorithm is referred to as a generation

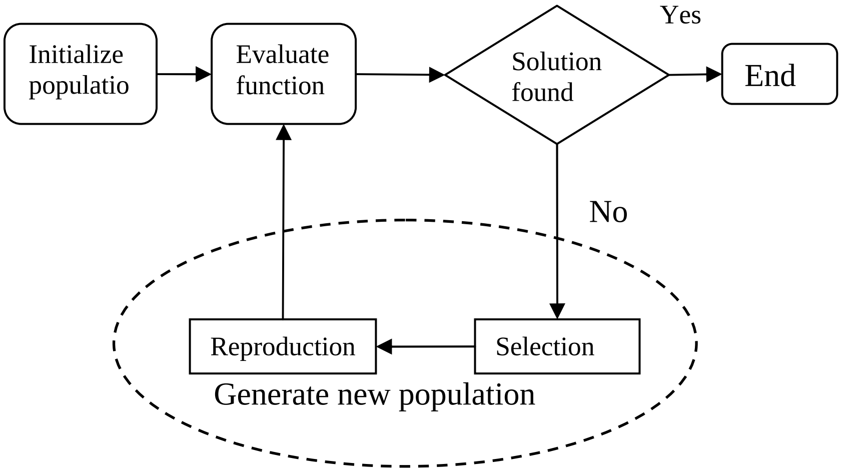


Fig.1 flow chart of genetic algorithm

3.0 Rule generation by A priori algorithm

The basic concept of Mining association are given as follows

Let

$I = \{I_1, I_2 \ldots . I_n\}$ be a set of item

$D= \{T_1, T_2 \ldots T_n\}$ be a set of transaction

Where ti is a set of transaction $t_i \in$ I, An association rule is transaction of the form X=>Y where X, YCI and X^Y=Ø .The rule X=>Y holds in the set D with Support and Confidence.

Input: a transaction database *D*, the minimum threshold of support *minSupNum*

Output: the set of frequent item sets *L*

```
1) for all transaction $t \in D$ do
2) Generate $TV$;
3) $L_1$ = (frequent 1-itemsets);
4) $C_2 = L_1 \infty L_1$;
5) $L_2 = \{c \in C_2 \mid sup(c) \geq MinSupNum\}$; // $sup(c)$ is the result of the formula (4)
6) for (k=3; $L_k$-1$\neq$ ø; $k$++ ) do begin
7) for ( $j=k$; $j \leq m$; $j$++ ) do
8) Generate $CIV_{ij}^{k-1}$
       $C_k$ = candidate_gen ($L$k-1);
9)     $L_k = \{c \in C_k \mid sup(c) \geq MinSupNum\}$;
10) end
11) Return $L = \in L_k$;
```

candidate_gen (frequent item sets L_k-1)

```
1) for all ($k$-1)-item set $l \in L_k$-1 do
2) for all $ij \in L_k$-1 do
3) // $S$ is the result of the formula (2) if for every $r(1 \leq r \leq k)$ such that $S[r] \geq k$-1 then
4) add $l \in \{ij\}$ to $C_k$;
```

4.0 Rule Optimization by Genetic algorithm

a.)Data Encoding

Here we have used Abolan Dataset basically these dataset used for the classification. In this Database 8 attributes are exited. It is provided by the MCI. Genetic Algorithm directly not work on the raw data then whole data we have encoded in the form of Binary representation technique (0 and 1).

b.) Fitness function

The most important part of Genetic Algorithm is a design of Fitness Function:

$$f(x) = \frac{\text{Support}(x)}{\text{minsupport}} \begin{cases} p(\text{support}(x)>\text{minsupport}) \\ q(\text{support}(x)<\text{minsupport}) \end{cases}$$

Support is the Support of New rule generated through genetic operation. Normal case the value of q(Support(x)<minsupport)is rejected for the better performance of genetic algorithm. We have used class-learned classifier for the prediction for rejected those value near to the Maximum value.

The value of q class is divided into two parts C_1 and $C_{2.}$

$q = \{C_1, C2\}$

C_1 = {those value or Data minsupport less than 0.5}

C_2 = {those value or Data minSupport greater than 0.5}

Now

$$f(q) = \frac{\text{Support }(C_2)}{\text{minsupport}} = \begin{cases} \alpha= (\text{Support }(C_2))>C_1 \\ \beta= (\text{Support}(C_2))=C_1 \end{cases}$$

c.)Selection Strategy

The selection strategy based on the basis of individual fitness and concentration pi is the probably of selection of individual whose fitness value is greater than one and f(α) is a those value whose fitness is less than one but near to the value of 1.

Now

$$Pi = \frac{f(x_i)}{\sum_1^M f(x_j)} \quad e^{-\alpha f(\alpha)}$$

Where α is an adjustment factor.

d.) Genetic Operation

The Genetic operators determine the search capability and convergence of the algorithm. Genetic operators hold the selection crossover and mutation on the population and generate the new population.

Select operation: In this algorithm it restores each chromosome in the population to the corresponding rule, and then calculate selection probability pi for each rule based on above formula.

Crossover operation: In which multi point crossover are used. It classifies the domain of each attribute into a group and classifies the cut point of each continuous attributes into one group .And the crossover carried out between the corresponding groups of two individuals by a certain rate.

Mutation Operation: Any bit in the chromosomes is mutated by a certain rate, that is, changing “0”to”1”,”1”to”0”.

e.)Terminating condition

The algorithm finally extracts the rules that meet the confidence threshold given by users, so the final output is not one optimal rule, but rather a set of rules that meet the threshold. **f.) Rules extraction**

The frequent rules are generated according to the fitness function and genetic operators. In order to mine the strong association rules finally, these rules must be extracted again. Extraction criteria are: output the rule which meets the minimum confidence given by users, otherwise abandon it.

5.0 Flow chart of proposed work

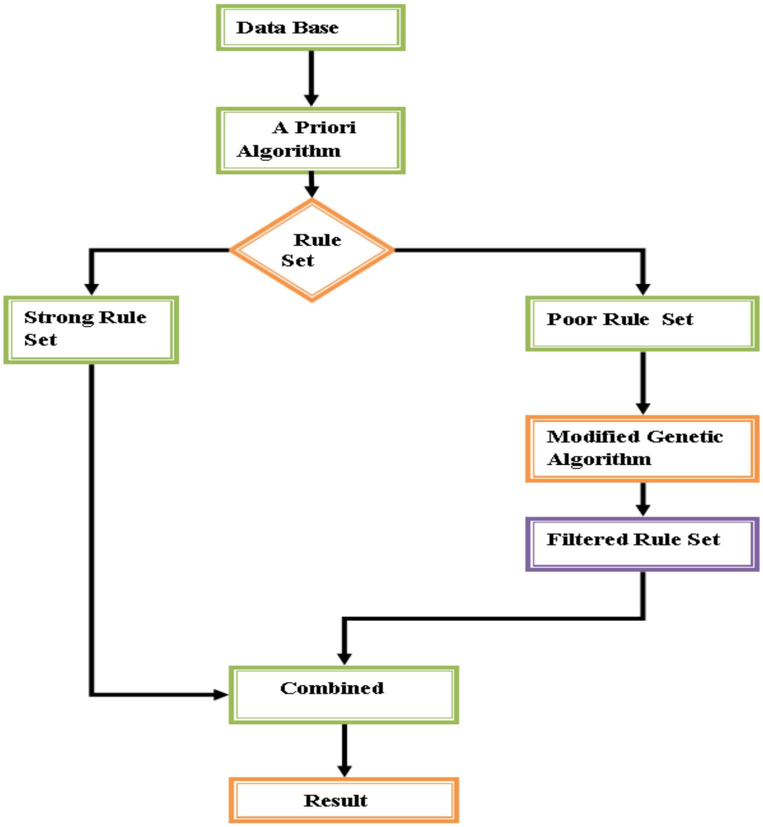


Fig.2 Flow chart of proposed work

6.0 Presents the Experiments and Results

The experiment uses Abalone dataset obtained from UCI machine learning repository. The data set has 4177 samples. It is composed of a discrete attribute and 8 continuous attributes. In this paper, we only mined such association rules $X \Rightarrow Y$ that Y was age. The setting of parameter: the size of evolutionary population N=100, crossover rate=0.006, mutation rate=0.001. The experiment was executed on Pentium IV CPU 2.58GHZ machine and software was MATLAB (2007). The below table represent the rule generation with genetic and without genetic algorithm for particular support and confidence

TABLE 1.Represent the result of rule generation

Attribute	A	B	C	D	E	F
Min support	2	4	6	8	9	7
Min confidence	0.1	0.2	0.3	0.1	0.2	0.5
Apriori algorithm	121	137	50	98	84	28
Genetic algorithm	94	115	42	75	70	22

The below graph represent the result analysis.

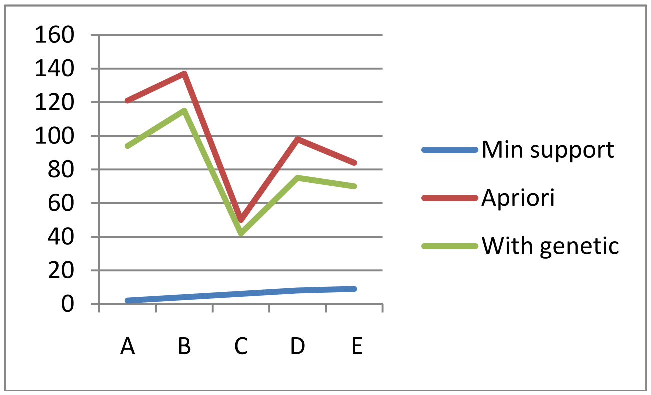


Figure.3 support/confidence vs. rule generation graph

7.0 conclusion & Future trend

In this direction we optimize association rule mining using new fitness function. In which fitness function divide into two classes' c1 and c2 one class for discrete rule and another class for continuous rule. Through this direction we get a better result.

To make genetic algorithm more effective and efficient it can be incorporated with other techniques so it can provide a best result.

REFERENCES

[1]. C.Y.Jia and X.J.Ni, "Association rule mining: A survey,"Computer Science, vol. 30, ,pp.145-149,2003.

[2]. F.Q.Shi, S.Q.Sun, and J.Xu, "Association rule mining of Dansei knowledge based on rough set," Computer Integrated Manufacturing Systems, vol.14,pp.407-411,2008.

[3]. F.Y.Li, L.P.Zhao, and H.Y.Wang, "Improved mining method for association rules based on genetic algorithm," Computer Engineering and Applications, vol. 44, pp.155-158,2008.

[4]. H.S.Nguyen and A.Skowron, "Quantization of real value attributes: rough set and boolean reasoning approach," Proc of 2th Joint Annual Conf. on Information Science, IEEE Press,pp.34-37,1995.

[5]. J. Han and M. Kamber, Data Mining:Concepts and Techniques, San Francisco: Morgan Kaufmann Publishers,2001.

[6]. J.G.Zheng and X.Y.Wang, "DNA-Immune-Genetic algorithm based on information entropy," Computer Simulation,vol.23,pp.163- 165,2006.

[7]. Kalyanmoy Deb, "Introduction to Genetic Algorithms", Kanpur Genetic Laboratory (Kangal), Depart of Mechanical Engineering, IIIT Kanpur 2005.

[8]. W.S.Yao, L.Shang, and Z.Q.Chen, “A quantization of real-value attributes based on evolution algorithm,” Computer Applications and Software, vol. 22, pp.37-39, 2005.

[9]. W.Ning and M.Q.Zhao, “More improved greedy algorithm for discrimination of decision table,” Computer Engineering and Applications, vol.43, pp.173-178. , 2007

[10]. X.P.Wang and L.M.Cao,Genetic Algorithm-Theory, Application and Software Realization, Xi’an: Xi’an Jiaotong University Press, 2001.

[11]. Y.Zhu,H.Zhang and L.D.Kong,“Research and application of multidimensional association rules minng based on artificial immune system,” Computer Science, vol. 36pp.239-242,2009.

[12]. Z.Tong, K.Luo, “Mining of association rules with decision attributes based on rough set,” Computer Engineering and Applications, vol 42,pp.166-169,2006.