OPTIMIZATION IN ASSOCIATION RULE MINING USING DISTANCE WEIGHT VECTOR AND GENETIC ALGORITHM

Sanjay Tiwari#1, Mahainder Kumar Rao#2
#1HOD, Department of Information Technology, AIEE, Jaipur, India
#2M.Tech, Department of Computer Science and Engineering, AIEE, Jaipur, India
#1 sanjay76_tiwari@yahoo.com
#2 mahindersinghraoi@yahoo.com

Abstract

The increasing rate of data is a challenging task for mined useful association rule in data mining. The classical association rule mining generate rule with several problem such as pruning pass of transaction database, negative rule generation and supremacy of rule set. Time to time various researchers modified classical association rule mining with different approach. But in current scenario association rule mining suffered from supremacy rule generation. The problem of supremacy is solved by multi-objective association rule mining, but this process suffered continuity of rule generation. In this dissertation we proposed a new algorithm distance weight optimization of association rule mining. In this method we find the near distance of rule set using equalize distance formula and generate two class higher class and lower class .the validate of class check by distance weight vector. Basically distance weight vector maintain a threshold value of rule item set. In whole process we used genetic algorithm for optimization of rule set.

Keywords— Data Mining, KNN, GAPO-ARM, Genetic Algorithm, MORA.

Introduction

Data mining plays 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. Final paper submission is available from the conference website. Data mining is becoming more and more common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to decrease costs, improve research, and increase sales. In the public areas, data mining applications at first were used as a means to detect fraud and waste, but have grown to be used for purposes such as measuring and civilizing program performance. Though, some of the homeland security data mining applications represent an important expansion in the quantity and scope of data to be analysed. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a diversity of parameters to examine the data. They include association (patterns where one event is connected to another event, such as buying a pen and buying paper), sequence or path analysis (patterns where one event leads to some other event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape buying and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unidentified facts, such as geographic site and brand preferences), and forecasting (discovering patterns from which one can make sensible predictions regarding prospect activities, such as the prediction that people who join an athletic club may take work out classes)[2]. As an application, compared to other data analysis applications, such as structured queries (used in a lot of commercial databases) or statistical analysis software, data mining represents a difference of type somewhat than degree.

In principle, data mining is not exact to one type of media or data. Data mining should be appropriate to any kind of information repository. However, algorithms and approaches may different when applied to different kinds of data. Indeed, the challenges presented by different types of data vary significantly [1]

PROPOSED METHOD

We proposed a novel algorithm for optimization of association rule mining, the proposed algorithm resolve the problem of negative rule generation and also optimized the process of supremacy of rules[2]. Supremacy of association rule mining is a great challenge for large dataset. In the generation of supremacy of rules association existing algorithm or method generate a series of negative rules, which
generated rule affected a performance of association rule mining. In the process of rule generation various multi objective association rule mining algorithm are proposed but all these are not solve supremacy problem of association rule mining.

In this algorithm we used second odder quadratic equation and nearest neighbour classification technique for the selection of set of candidate of supremacy of key for generation of rules. In the generation of rule selection of support value of transaction data set is play a important role, for this role we used heuristic search algorithm for better searching of support value for generation of optimized association rule.

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[5,6], 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. For example, if the domain of the parameter x is [2,5] and the precision requirement is six places after the decimal point, then the domain [2,5] should be divided into 7,000,000 equal size ranges. The implies that the length of the string requires to be 23, for the reason that $4194304 = 2^{22} < 7000000 < 2^{23} = 8388608$ the decoding from a binary string $<b_22b_21…b_0>$ into a real number is straightforward and is completed in two steps.

(1). Convert the binary string $<b_22b_21…b_0>$ from the base 10 by

$$x' = \sum_{i=0}^{22} b_i 2^i$$

(2). Calculate the corresponding real number x by

$$x = -2.0 + x' \cdot \frac{7}{2^{23} - 1}$$

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

Genetic operators drive the evolutionary process of a population in GA, after the Darwinian principle of survival of the fittest and naturally occurring genetic operations. The most widely used genetic operators are reproduction, crossover and mutation. To perform genetic operators, one 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 [7]. 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=20). If

$$X_1 = (01001|101100001000101)$$
$$X_2 = (11010|011100000010000)$$

Then the two resulting offspring are

$$X_1' = (01001|011100000010000)$$
$$X_2' = (11010|101100001000101)$$

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 $p_m$. 

A. Initial Population

ISSN No: 2250-3536

Volume 4, Issue 1, Jan. 2014 80
Proposed Algorithm GAPO-ARM
(Genetic Algorithm Produce optimization of association rule mining)

The proposed algorithm is a combination of support weight value and near distance of superior candidate key. Support weight key is a vector value given by the transaction data set. The support value passes as a vector for finding a near distance between superior candidate key. When we finding a higher candidate key the nearest distance divides into two classes, one class take a higher order value and another class gain lower value for the rule generation process. The process of selection of class also reduces the passes of the data set dividing. After finding a class of lower and higher of giving 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 the steps of process of algorithm step by step and finally draw a flow chart of complete process.

In the process of novel algorithm for rule optimizations fist we discuss association rule mining, KNN and genetic algorithm and finally we proposed a hybrid method for optimization of association rule mining (GAPO-ARM).

Steps of algorithm (GAPO-ARM)

1. Select data set
2. Put value of support and confidence
3. Start scanning of transaction table
4. Count frequent items
5. Generate frequent item sets
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. Initialized a population set (t=1)
11. Compare the value of distance vector with population set
12. If value of support greater than vector value
13. Processed for encoded of data
14. Encoding format is binary
15. After encoding offspring are performed
16. Set the value of probability for mutation and the value of probability is 0.006.
17. Set of rule is generated.
18. Check supremacy of rule
19. If rule is not superior go to selection process.
20. Else optimized rule is generated.
21. Exit
The whole process of proposed algorithm is shown in the architecture of the proposed algorithm in Figure 1. In the architecture model ellipse represents input or output and rectangle represent the calculation process. In architecture model, it is shown that as an input datasets from UCI machine learning repository [10] are used, which consists of various attributes and number of instances. The input dataset is preprocessed to reduce negative rules and used to find the positive rules and insignificant data points from the dataset.

**IMPLEMENTATION AND EXPERIMENTAL RESULTS**

To investigate the effectiveness of the proposed method implemented on Mat lab 7.8.0.347(R2009a). In the research work, I have measured execution time of the rule generation procedure and no. of rule generation. To evaluate these performance parameters I have used five datasets from UCI machine learning repository [10] namely Wisconsin breast cancer (original), Iris, Glass identification, wine, and Chess dataset. Out of these five dataset, two are small dataset namely Iris and Glass identification dataset; and remaining four are large datasets namely Wisconsin breast cancer, White wine quality and Chess dataset. Here are the following table of all used data set. I have used five datasets from UCI machine learning repository [10] namely Wisconsin breast cancer (original), Iris, Glass identification, wine, and Chess dataset.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Data Set</th>
<th>Attributes</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cancer</td>
<td>10</td>
<td>699</td>
</tr>
<tr>
<td>2</td>
<td>Iris</td>
<td>04</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>Glass</td>
<td>10</td>
<td>214</td>
</tr>
<tr>
<td>4</td>
<td>Wine</td>
<td>13</td>
<td>4177</td>
</tr>
<tr>
<td>5</td>
<td>Chess</td>
<td>36</td>
<td>3196</td>
</tr>
</tbody>
</table>

Experimental result demonstrates that the proposed GAPO-ARM algorithm reduces the execution time of MORA algorithm. It also improves rule generation process and accuracy for all five dataset than MORA. The experimental results also demonstrate that the proposed GAPO-ARM method is best then MORA in association rule mining algorithm.

Figure 2: Shows the comparative value of execution time of MORA and GAPO-ARM algorithm for the processor extraction of rule with all data set. GAPO-ARM takes more time in comparison of sample MORA algorithm.
CONCLUSIONS AND FUTURE WORK

We proposed a novel method for optimization of association rule mining. Our propped algorithm is combination of distance function and genetic algorithm. We have observed that when we modify the distance weight new rules in large numbers are found. This implies that when weight is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm.

We theoretically proofed a relation between locally large and globally large patterns that is used for local pruning at each site to reduce the searched candidates. We derived a locally large threshold using a globally set minimum recall threshold. Local pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of exchanged messages. Suggestion for future work, in the process of distance weight calculation large number of rule set generated that rule set take huge amount of time in compression of MOARGM algorithm. In future we minimize the time complexity of our method.

VI. References


[8] By Chi-Ren Shyu1,2, Matt Klaric1,2, Grant Scott1,2, and Wannapa Kay Mahamaneerat1 Knowledge Discovery by Mining Association Rules and Temporal-Spatial Information from Large-Scale Geospatial Image Databases 0-7803-9510-7/06/$20.00 © IEEE 2006.


[22] By A.B.M.Rezbaul Islam, Tae-Sun Chung ” An Improved Frequent Pattern Tree Based Association Rule Mining Technique” 978-1-4244-9224-4/11/$26.00 ©IEEE 2011.


[29] By Xinqi Zheng1 Lu Zhao2 Association Rule Analysis of Spatial Data Mining Based on Matlab Knowledge Discovery and Data Mining 0-7695-3090-7/08 $25.00 © IEEE 2008.