International Journal of Electrical, Electronics and Computer Engineering **1**(2): 25-27(2012)

ISSN No. (Online) : 2277-2626

Distance Weight Optimization of Association Rule Mining with Improved Genetic Algorithm

Nikhil Jain and Vishal Sharma

Department of Computer Science, JIT Khargone, (M.P.)

(Recieved 10 August, 2012 Accepted 11 September, 2012)

ABSTRACT : In this paper, the main area of concentration was to optimize the rules generated by Association Rule Mining (apriori method), using Genetic Algorithms. In general the rule generated by Association Rule Mining technique do not consider the negative occurrences of attributes in them, but by using Genetic Algorithms (GAs) over these rules the system can predict the rules which contains negative attributes. The main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. The improvements applied in GAs are definitely going to help the rule based systems used for classification as described in results and conclusions.

Keywords: Genetic Algorithms, Data Mining, Association Rule Mining.

I. INTRODUCTION

This large amount of data contains latent knowledge, which can be utilized to improve decision making process of an organization. This knowledge discovery can be done in various ways available today, like Decision Tree, Association Rule Mining, Bayesian Classifier and so on. The form of this latent knowledge also varies to a large extent from different kind of rules to prediction values. In this paper the authors have considered Association Rule Mining and tried to improve this technique by applying Genetic Algorithms on the rules generated by Association Rule Mining.

A brief introduction about Association Rule Mining and GA is given in the following sub-sections, followed by methodology in section 2, which will describe the basic implementation details of Association Rule Mining and GAs.

A. Association Rules

Introduced in [5], association rule mining has gained great deal of attention. Even today people use it for mining in KDD. In brief, an association rule is an expression *X* => *Y*, where X and Y are item sets.

The meaning of this kind of rule is : Given a database *D* containing say *N* tupples or transactions, where say *T* belongs to *D* is a transaction, then *X* => *Y* expresses that whenever a transaction *T* contains *X* than *T* probably also contains *Y*. This probability or confidence is defined as the percentage of transactions containing *Y* in addition to *X* with regard to overall number of transactions containing *X*. Thus the authors can represent can represent this probability as conditional probability $p(Y \in T | X \in T)$. The thrust behind introduction of these rules was there similarity with market-based data where rules like "A customer buys milk and Bread will also buy butter with a probability, say *x* %" is a famous example. Also, their direct applicability to business problems together with their inherent understandability, even for non-experts, made them a popular mining method. Further, it was also determined that their applications can be further extended from general dependency based rules to a wide range of business applications.

Mining Association rules is not full of advantages; it has some limitations too, first of all the algorithmic complexity. The number of rules grows exponentially with the number of items. But this complexity is tackled with some latest algorithms which can efficiently prune the search space. Secondly, the problem of finding rules from rules, i.e. picking interesting rules from set of rules. The work tackling the second problem mainly support the user when browsing the rule set, *e.g.* [4] and the development of further useful quality measures on the rules, *e.g.* [2; 6; 7].

Thirdly, the problem that is being discussed in this paper is that, association rules do not utter the rules in which the negation of attributes is there. Like, say there are three attributes in the database X_1, X_2, X_3, than rules like "If a customer takes X_1 and not X_2 than he will take X_3 with a confidence of say *c* %" will not be provided by normal association rule mining. In order to generate these kinds of rules and also to tackle the second problem discussed above, *i.e.* to evolve quality rules, this paper is using Genetic Algorithms.

B. Genetic Algorithms

As discussed in [1], in general the main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. This section of the paper discusses several aspects of GAs for rule discovery. The main areas of discussion include individual representation of rules, Genetic Operators involved and the choice of Fitness function.

Representation of rules plays a major role in GAs, broadly there are two approaches based on how rules are encoded in the population of individuals ("Chromosomes") as discussed in [1] Michigan and Pittsburgh; The pros and cons as discussed in [1] is as follows, Pittsburgh approach leads to syntactically-longer individuals, which tends to make fitness computation more computationally expensive. In addition, it may require some modifications to standard genetic operators to cope with relatively complex individuals. By contrast, in the Michigan approach the individuals are simpler and syntactically shorter. This tends to reduce the time taken to compute the fitness function and to simplify the design of genetic operators. However, this advantage comes with a cost. First of all, since the fitness function evaluates the quality of each rule separately, now it is not easy to compute the quality of the rule set as a whole *i.e.* taking rule interactions into account. In this paper Michigan's approach is opted *i.e.* each individual encodes single rule. The encoding can be done in a number of ways like, binary encoding or expression encoding etc. For example let's consider a rule " If a customer buys milk and bread then he will also buy butter", which can be simply written as If milk and bread then butter

Now, following Michigan's approach and binary encoding, for simplicity sake, this rule can be represented as 00 01 01 01 10 01 where, the bold di-digits are used as product id, like 00 for milk, 01 for bread and 10 for butter and the normal di-digits are 00 or 01 which shows absence or presence respectively. Now this rule is ready for further computations.

Second, area of concern is Genetic Operators. Mainly three operations are to be performed, selection, cross-over and mutation to robustly search the rule space for various options. Selection involves selecting two fit parents for evolving new children rules which are fit than the parents, and in this manner the average fitness of the rules can be increased. Cross-over and mutation provides the ways to evolve new rules. Third area of concern is fitness function. Since, the discovered rules should : (*a*) have a high predictive accuracy; (*b*) be comprehensible; and (*c*) be interesting, thus choice of this function is very important to get the desired results.

II. METHODOLOGY

In this paper the genetic algorithms are applied over the rules fetched from association rule mining. Now for demonstration its utility, the database is produced synthetically. This database contains the choice of electives by students during their 3rd year of course studies at Indian Institute of Information Technology, Allahabad, India. Students have to choose four subjects from eight based on their liking and area of interest. Now, the authors firstly implemented Association Rule mining (using a-priori technique) by the help of their toolkit [3]. And then the GAs are applied to evolve the rules which contains the negations in attributes and are of richer quality. In this section the paper discusses each step in detail.

A. Association Rule Mining (*a-priori*)

The Algorithm for its implementation is same as described in section 1.1. The rules came out of it looks like: IF QC and VLSI THEN IPR Where NFC, RIA, QC, VLSI, etc. are of the eight subjects out of which student has to choose four. The rules above shows that if a student takes NFC and RIA then the probability is high that he will choose BI too, similarly if he chooses QC and VLSI then the probability is high that he will take IPR. There is no boundation on the number of antecedents in the rules, but there is a constraint on the number of consequents, and i.e. number of consequents = 1 This boundation doesn't make any harm, because if in case the user wants to see the confidence value of a rule that contains the more than one consequents can do the same by taking two rules from our system and then by doing the intersection of it.

III. RESULTS

As described earlier, in this paper the implementation of GAs are applied to the rules obtained by applying association rule mining on synthetic database which is based on the selection procedure The database was made at random. The columns in the database are the subjects and their value is either one or zero depending on whether they are selected or not. Following figure shows a database glimpse.

IV. CONCLUSIONS AND FUTURE WORK

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] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery" Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba-PR. 80215-901. Brazil.

[2] C. Silverstein, S. Brin, R. Motwani and J.D. Ullan. Scalable techniques for mining causal structures. *In the Proc. of 1998 ACM SIGMOD Int'l Conf. on Management of Data*, Seattle, Washington, USA, June (2011).

[3] Manish Saggar, Ashish K. Agarwal, Abhishek Agarwal; Discovery- A Data Mining Tookit.

[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. *In Proc. of the 3rd Int. Conf. on Information and Knowledge Management, Gaithersburg, Maryland,* **29**. Nov-2. Dec. (1994).

[5] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In the Proc. of the ACM SIGMOD *Int. Conf. on Management of Data* (ACM SIGMOD '93), Washington, USA, May (2011).

[6] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalising association rules to correlations. In the Proc. of the ACM SIGMOD *Int. Conference on Management of Data (ACM SIGMOD* '97).

[7] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In the Proc. of the ACM SIGMOD *Int. Conf. on Management of Data,* (2012).