Application of Genetic algorithms for the prioritization of Association Rules

M. Ramesh Kumar
Department of Computer Science (DDE),
Madurai Kamaraj University,
Madurai , Tamilnadu 625021, India

Dr. K. Iyakutti
School of Physics,
Madurai Kamaraj University,
Madurai, Tamilnadu 625021, India

ABSTRACT
A novel genetic algorithm based association rule mining algorithm is discussed in this paper. Prioritization of the rules has been discussed with the help of genetic algorithm. Fitness function is designed based on the two measures like all confidence and the collective strength of the rules, other than the classical support and the confidence of the rules generated. The algorithm is been tested for the four data sets like Adult, Chess, Wine, Zoo. The results are promising and lead to the future enhancements in this research area.

Keywords:
Apriori,Genetic algorithm, Prioritization

1. INTRODUCTION
Data mining techniques, extracting patterns from large databases, are the processes that focus on the automatic exploration and analysis of large quantities of raw data in order to discover meaningful patterns and rules [1]. Association rule mining forms an important research area in the field of Data mining. Association rules are used to identify relationships among a set of items in dataset. These relationships are not based on inherent properties of the data themselves. In existing data mining techniques, there exist some situations that make necessary the prioritization of rules for selecting and concentrating on more valuable rules due to the number of qualified rules and limited business resources [2]. The main association rule mining algorithm, Apriori, not only influenced the association rule mining community, but it affected other data mining fields as well. Apriori and all its variants like Partition, Pincer-Search, Incremental, Border algorithm etc. take too much computer time to compute all the frequent item sets.

The research area to be concentrated on is to prioritize association rules, by considering their importance by explicitly incorporating the conflicting criteria of consideration values and by the user's preference statements toward their trade-off conditions.

Genetic algorithm (GA) is the part of the evolutionary algorithm which has their base as the simulation of the human reproduction, used mostly in the optimization problems. Application of GA to the data mining techniques introduces the novel field of research, leads to the new arena in the dispensation of the knowledge discovery process. In this paper we introduce a GA based knowledge discovery process in the association rule mining to prioritize the rules generated to value add the data mining process.

Due to the prioritization of the rules in the association rule mining, the information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

The paper is organized as the section 2 deals with the related works, Section 3 gives the use of GA for association rule mining, Section 4 deals with the Methodology adopted and section 5 discusses the results obtained and section 6 gives the conclusion of the paper.

2. Related Works
In this section we discuss about the related works in Association rule mining and the genetic algorithm.

2.1 Association rule mining
Association rule mining has been widely used from traditional business applications such as cross-marketing, attached mailing, catalog design, loss-leader analysis, store layout, and customer segmentation [3]. Given a set of transactions where each of transactions is a set of items (item set), an association rule implies the form

X->Y, where X and Y are item sets; X and Y are called the body and the head, respectively. A rule can be evaluated by two measures, called confidence and support. A measure, support for the association rule X->Y is the percentage of transactions that contain both item set X and Y among all transactions. The confidence for the rule X->Y is the percentage of transactions that contain an item set Y among the transactions that contain an item set X. The support represents the usefulness of the discovered rules and the confidence represents the certainty of the rules.

Apriori algorithm is one of the most widely used and famous techniques for finding association rules [4]. Apriori operates in two phases. In the first phase, all item sets with minimum support (frequent item sets) are generated. This phase utilizes the downward closure property of support. In other words, if an item set of size k is a frequent item set, then all the item sets below (k-1) size must also be frequent item sets. Using this property, candidate item sets of size k are generated from the set of frequent item sets of size (k-1) by imposing the constraint that all subsets of size (k-1) of any candidate item set must be present in the set of frequent item sets of size (k-1). The second phase of the algorithm generates rules from the set of all frequent item sets.

The problem to be addressed in the Association rule mining is the rule quality and the rule quantity. If minimum support is set too high, the rules involving rare items that could be of interest to decision makers will not be found. Setting minimum support low, however, may cause combinatorial explosion [5]. Due to the limited resource available for the business, only a considerable small number of rules can be chosen for the execution[6].

2.2 Genetic algorithms

Genetic Algorithm (GA) [7] was developed by Holland in 1970. It incorporates Darwinian evolutionary theory with sexual reproduction. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution [8]. GA has been successfully applied in many search, optimization, and machine learning problems. The main motivation for using Genetic Algorithms in the discovery of high-level prediction rules is that they perform a global search [9]. GA works in an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem.

Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings. GA runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found. GA is appropriate for problems which require optimization, with respect to some computable criterion. The functions of genetic operators are as follows:

1) Selection: Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point.

3) Mutation: Alters the new solutions so as to add stochastic in the search for better solutions. This is the chance that a bit within a chromosome will be flipped (0 becomes 1, 1 becomes 0).

Essentially, Genetic algorithms are a method of "breeding" computer programs and solutions to optimization or search problems by means of simulated evolution. Processes loosely based on natural selection, crossover, and mutation are repeatedly applied to a population of binary strings which represent potential solutions. Over time, the number of above-average individual increases and highly-fit building blocks are combined from several fit individuals to find good solutions to the problem at hand.

Not only does GAs provide alternative methods to solving problem, it consistently outperforms other traditional methods in most of the problems link. Many of the real world problems involved finding optimal parameters, which might prove difficult for traditional methods but ideal for GAs.

This generational process is repeated until a termination condition has been reached. Common terminating conditions are[10] :

- A solution is found that satisfies minimum criteria
- Fixed number of generations reached
- Allocated budget (computation time/money) reached
- The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
- Manual inspection
- Combinations of the above

3. Use of Genetic algorithm for Association Rule Mining

In the existing works, we could cite a lot of works based on the genetic algorithm used for Association rule mining. The main motivation for using GA in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction that the greedy rule induction algorithms often used in data mining[11]. Due the high randomness and inbuilt parallelization, GA can be deployed to prioritize the rules needed for further manipulations. The rule mining process can use the searching capability of the GA for the undisclosed rules. Based on the above reasons, we tend to use GA for prioritization of the rules.

4. Proposed Methodology

We present a novel algorithm for the rule prioritizing, generated by the apriori algorithm through the application of genetic algorithm. The proposed methodology is as follows

Pseudo code

1.Start
2. Load a sample of records from the dataset
3. Apply apriori algorithm to find the frequent item sets with the minimum
Support. Let the frequent set be F
4. Set F_GA is the output set, which contains the association rule.
5. Input the termination condition of GA
6. Represent each frequent item set of F as binary string using the combination of
representation.
7. Select the two members from the frequent item set using Roulette Wheel
sampling method
8. Apply the crossover and mutation on the selected members to generate the
association rules.
9. Find the fitness function for each rule x->y and check the following condition.
10. if (fitness function > M)
11. set F_GA = F_GA U {x -> y}
12. If the desired number of generations is not completed, then go to Step 3.
13. Stop

The fitness function is designed based on the user's interesting measure and M is the threshold value of the interesting measure considered. In our approach the measures used other than the support and the confidence are

All confidence (AC) is the measure introduced by [12] , it has been used by Hahsler [13]

AC (F) = supp(F) / max(support(z element of F)) = P(F) / max(P(f element of F))

max(support(f element of F)) is the support of the item with the highest support in **F**. All-confidence means that all rules which can be generated from item set **F** have at least a confidence of **all-confidence(F)**. All-confidence possesses the downward-closed closure property.

Collective strength (CS) introduced by [14]

$$CS(F) = (1-v(F))/(1-E[v(F)]) * E[v(F)]/v(F)$$

where **v(F)** is the violation rate and **E[]** is the expected value for independent items. The violation rate is defined as the fraction of transactions which contain some of the items in an item set but not all. Collective strength gives 0 for perfectly

negative correlated items, infinity for perfectly positive correlated items, and 1 if the items co-occur as expected under independence. Problematic is that for items with medium to low probabilities the observations of the expected values of the violation rate is dominated by the proportion of transactions which do not contain any of the items in **F**. For such item sets collective strength produces values close to one, even if the item set appears several times more often than expected together.

Fitness function = W1 * AC(F) + W2 * CS(F) / (W1 +W2)

Where W1 and W2 are the weights assigned by the user, if the prioritization is based on the measure AC then W1 will have the importance and the M will be assigned to the minimum threshold that the user, prefer based on the AC. If the prioritization is based on the measure CS then W2 will have the importance and the M will be assigned to the minimum threshold that the user, prefer based on the CS.

5. RESULTS AND DISCUSSIONS

The sample data sets have been taken from the UCI data repository for the testing of our algorithm. The environmental measure we had for testing is the population size for performing GA is 200, the selection rate is 10% and the cross over rate is 6% and the mutation rate is fixed as 1%.

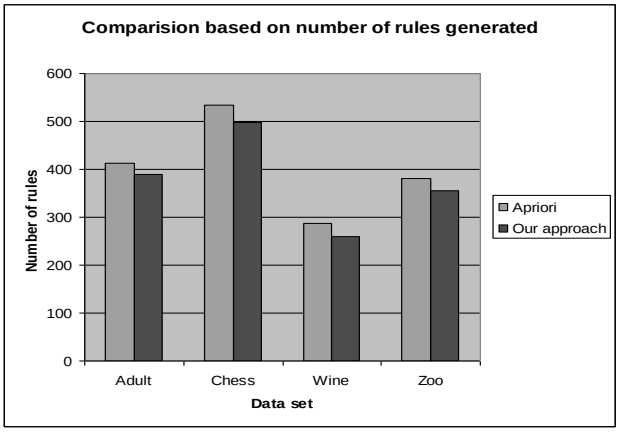

Fig1: Comparison of our approach with apriori when support = 30% and confidence = 50%

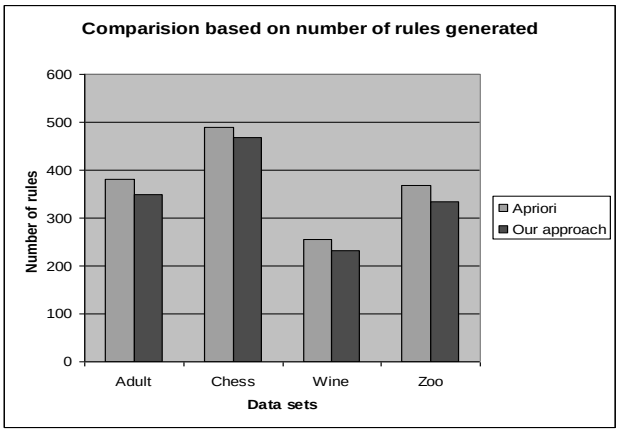

Fig2: Comparison of our approach with apriori when support = 50% and confidence = 50%

The number of rules produced by our approach is significantly less compared to the apriori approach. This is due to the application of the GA with the apriori algorithm.

6. Conclusion

We propose a novel Genetic algorithm based association rule mining algorithm for the prioritization of the rules. Our approach significantly reduces the number of rules generated in the four data sets we have used. The fitness function is designed in such a way that to prioritize the rules based on the user preference. The future work can be extended by the incorporation of the other interesting measures in the literature to our work, the business values can also be tested from the rules generated.

7. REFERENCES:

[1] Song, H. S., Kim, J. K., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, 21, 158–168.

[2] Tan, P. N., & Kumar, V. (2000). Interestingness measures for association patterns: A perspective. KDD 2000 Workshop on Post processing in Machine Learning and Data Mining, Boston, MA, August.

[3] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association between sets of items in massive database. International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).

[4] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the international conference on very large data bases (pp. 407–419).

[5] Duke Hyun Choi , Byeong Seok Ahn , Soung Hie Kim, Prioritization of association rules in data mining: Multiple criteria decision approach, Expert Systems with Applications: An International Journal, v.29 n.4, p.867-878, November, 2005.

[6] Choi et al., (2005).Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications. v29. 867-878.

[7] Pei M., Goodman E.D., Punch F. (2000) Feature Extraction using genetic algorithm, Case Center for Computer-Aided Engineering and Manufacturing W. Department of. Computer Science.

[8] Stuart J. Russell, Peter Norvig (2008) Artificial Intelligence: A Modern Approach.

[9] J.Arunadevi and V.Rajamani. ,Optimization of Spatial Association Rule Mining using Hybrid Evolutionary algorithm. International Journal of Computer Applications 1(1):86–89, February 2010.

[10] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar (2010) Mining Frequent Itemsets Using Genetic Algorithm, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, 133-143.

[11] Sufal Das, Banani Saha, "Data Quality Mining using Genetic Algorithm", International Journal of Computer Science and Security, ISSN: 1985-1553, 3(2): pp 105-112, 2009.

[12] Edward R. Omiecinski. Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering, 15(1):57-69, Jan/Feb 2003.

[13] M. Hahsler, A model-based frequency constraint for mining associations from transaction data, Data Mining and Knowledge Discovery 13 (2006), 137–166.

[14] C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In PODS 98, Symposium on Principles of Database Systems, pages 18-24, Seattle, WA, USA, 1998.