International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Association Rule Mining for Profit Patterns Using Genetic Algorithm

Sameer Kumar Vishnoi[1], Vivek Badhe[2]
[1]M.Tech. Scholar Gyan Ganga College of Technology Jabalpur, India
[2]Department of CSE, Gyan Ganga College of Technology, Jabalpur, India

***Abstract*-- Association Rule Mining for profit patterns focuses the important issues related with business and commercial aspects. To generate profit patterns we propose a genetic algorithm based approach. In our proposed approach we have added the power of genetic algorithm to the conventional association rule mining.. The literature shows that Genetic Algorithm improves the accuracy and efficiency of association rule mining, as genetic algorithms are capable to deal with the problems related with the global search, optimization and non-linearity. Even though a lot of research has been carried out in association rule mining using genetic algorithms, none of them dealt with the profit as a factor of interestingness. In our approach we are applying genetic algorithms with association rule mining to generate profit patterns which assure the business goals.**

***Keywords*-- Association Rule Mining, Genetic Algorithm**

I. INTRODUCTION

Data Mining [1] Techniques is being highly used for extracting the hidden predictive information from large databases. It is a powerful new technology with great potential to analyze important information from the large databases. Discovery of hidden pattern is an important database-mining problem. The area of data mining or knowledge discovery has recently attracted much attention from computer scientists. While being an important tool for many practitioners, data mining is also an interesting research area that raises several challenging problem.

II. DATA MINING METHODOLOGIES

The two fundamental goals of data mining: Prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, and description focuses on finding patterns describing data and the subsequent presentation for user interpretation. The relative emphasis of the both prediction and description differ with respect to the underlying application and the technique. There are several data mining techniques such as classification, clustering, outlier analysis and association rule mining fulfilling these objectives.

III. ASSOCIATION RULE MINING

Association Rule Mining [2] techniques can be used to discover unknown or hidden correlation between items found in the database of transactions. An association rule is a rule, which implies certain association relationships among a set of objects such as occurs together or one implies to other in a database. Association rules identify relationships among sets of items in a transaction database. Ever since its introduction in (Agrawal, Imielinski and Swami 1993), Association Rule discovery has been an active research area. Association Rule Mining finds interesting association or correlations among a large set of data items.

The discovery of association rules for a given dataset D is typically done in two steps: discovery of frequent itemsets and the generation of association rules. The first step is to find each set of items, called as itemsets, such that the co-occurrence rate of these items is above the minimum support, and these itemsets are called as large itemsets or frequent itemsets. In other words, find all sets of items (or itemsets) that have transaction support above minimum support. The second step is to find association rules from the frequent itemsets that are generated in the first step. The second step is rather straightforward. Once all the large itemsets are found, generating association rules by the user defined minimum confidence.

IV. PROFIT PATTERN MINING

Profit Pattern Mining [5] is a new direction of Association Rule Mining it aims to discover those patterns which provides maximum profit. As the major obstacle in the Association Rule Mining application is the gap between the statistical based patterns extraction and valued based decision making, Profit Pattern mining reduces this gap.In Profit Pattern Mining a set of past transaction and pre selected target item is given and a model is constructed for recommending target items and promotion strategies to new customers, with the goal of maximizing the net profit.

V. GENETIC ALGORITHM

Genetic Algorithm (GA) [3] was developed by Holland in 1970. It incorporates Darwinian evolutionary theory with sexual reproduction.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. 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.

0 START : Create random population of n chromosomes
1 FITNESS : Evaluate fitness f(x) of each chromosome in the population

2 NEW POPULATION

0 SELECTION : Based on f(x)
1 RECOMBINATION : Cross-over chromosomes
2 MUTATION : Mutate Chromosomes
3 ACCEPTION : Reject or accept new one
3 REPLACE : Replace old with new population: the new generation
4 TEST : Test problem criterion
5 LOOP : Continue step 1 – 4 until criterion is satisfied

Figure-1

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.

VI. Data Mining Using Genetic Algorithms

The application of the genetic algorithm in the context of data mining is generally for the task of hypothesis testing, refinement and optimization. Data mining can be thought of as a search problem. The problem is to search a large space for interesting information (rules). The absolute size of the search spaces involved in data mining requires that algorithm be explored that can determine interesting rules by examining subsets of this data.

The main motivation for using GAs in the discovery of high-level rules by optimization is that they perform a global search and cope better with attribute interaction.

GA requires no prior knowledge about the search space and discontinuities preset on the search space have little effect on overall search process. Genetic Algorithms are robust and they approach uniformly to large number of different classes of problems. If the solution for given problems exists, the Genetic Algorithms with proper coding, operators and fitness function will find it. This is an obvious advantage over methods such as regression models that can only be used in specific cases. Such generality is desirable in Data Mining where the search space is complex noise. Most important feature of Genetic Algorithms is that they are easily parallelizable and have been used for Association, classification as well as other optimization problems. In Data Mining, they may be used to evaluate the fitness of other algorithms.

VII. Related Work

Ke Wang, Senqiang Zhou, and Jiawei Han in 2002 presented a profit mining [5] approach to reduce the gap between the statistic-based pattern extraction and the value-based decision making. They took a set of past transactions and pre-selected target items, and intended to build a model for recommending target items and promotion strategies to new customers, with the goal of maximizing the net profit. They identified several issues in profit mining and proposed solutions. They evaluate the effectiveness of this approach using data sets of a wide range of characteristics. The key to profit mining is to recommend “right” items and “right” prices. If the price is too high, the customer will go away without generating any profit; if the price is too low or if the item is not profitable, the profit will not be maximized. The approach is to exploit data mining to extract the patterns for right items and right prices. The key issues in this context are Profit based patterns, shopping on unavailability, explosive search space, optimality of recommendations, and interpretability of recommendation.

Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad et. al.2004 [6] proposed 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.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba et. Al 2008 [7] presented a Pareto based multi objective evolutionary algorithm rule mining method based on genetic algorithms. They used confidence, comprehensibility, interestingness, surprise as objectives of the association rule mining problem. Specific mechanisms for mutations and crossover operators together with elitism have been designed to extract interesting rules from a transaction database. Empirical results of experiments carried out indicate high predictive accuracy of the rules generated.

Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K et. al. 2009 [8] is to find all the possible optimized rules from given data set using genetic algorithm. The rule generated by association rule mining algorithms like priori, partition, pincer-search, incremental, border algorithm etc, does not consider negation occurrence of the attribute in them and also these rules have only one attribute in the consequent part. By using Genetic Algorithm (GAs) the system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The major advantage of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach.

Xiaowei Yan, Chengqi Zhang, Shichao Zhang et al. 2009 [9] designed a genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. In their approach, they developed an elaborate encoding method and the relative confidence is used as the fitness function. With genetic algorithm, a global search can be performed and system automation is implemented, because the model does not require the user-specified threshold of minimum support.

Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar in 2010 [10] proposed method for frequent itemset mining using Genetic algorithm. The major advantage of using GA in the discovery of frequent itemsets is that they perform global search and its time complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach.

Sandhu, P.S.; Dhaliwal, D.S.; Panda, S.N.; Bisht, A. in 2010 [11] proposed an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach makes use of the traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent.

Subsequently, the set of association rules mined are subjected to weight age (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined Utility Weighted Score (UW-Score) is computed. Ultimately, they determined a subset of valuable association rules based on the UW-Score computed. The experimental results demonstrate the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development.

Indira K and Kanmani S in 2012 [12] presented method for performance analysis of Genetic algorithm for Mining Association rules. This paper analyzes the performance of GA in Mining ARs effectively based on the variations and modification in GA parameters. The recent works in the past seven years for mining association rules using genetic algorithm is considered for the analysis. Genetic algorithm has proved to generate more accurate results when compared to other formal methods available.

VIII. Proposed Work

In general, the association rules are generated in two steps. First, frequent itemsets are found through user defined minimum support and secondly, rules are generated using the frequent itemsets found in first step using the user defined value called confidence.

In our proposed approach we have followed the conventional association rule mining algorithm to generate rules from the database. We then optimized those rules using Genetic Algorithm implemented in MATLAB Version 7.6.0.324 (R2008a). For optimization of rules the fitness function is designed as

$$\text{Fitness} = \frac{\text{comp} * w1 + \text{intr} * w2}{w1 + w2}$$

Where w1 is ratio of percentage profit and w2 is ratio of margin of profit and both are user defined factors. The value of w1 and w2 are calculated as

$$w1 = \frac{\text{Percentage profit of B}}{\text{Percentage profit of A}}$$

$$w2 = \frac{\text{Margin profit on B}}{\text{Margin profit on A}}$$

And comp & intr are defined as

$$\text{Comp} = \frac{TP}{TP + FP}$$

$$\text{Intr} = \frac{TP}{TP + FN}$$

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Completeness (comp): Those rules are considered as complete rules where the item having lower percentage of profit implies the item having higher percentage of profit.

Interestingness (intr): Those rules are considered as rules of interest where the item having lower margin of profit implies the items having higher margin of profit.

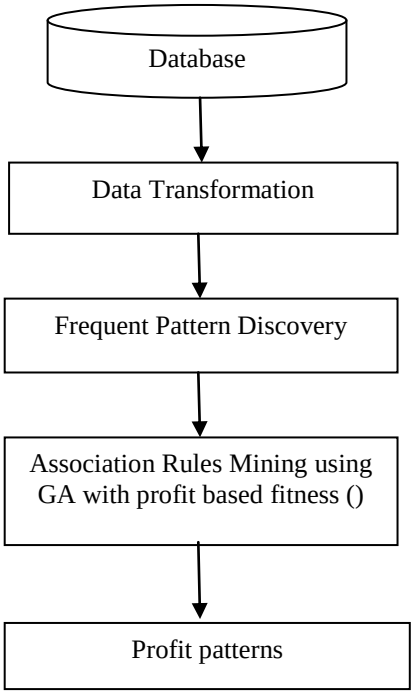

Figure-2: Proposed Methodology

IX. Result

Initially the FMCG goods data base (in MS Access) is taken and converted into Flat file: text tab delimited format, then applying Apriori algorithm on the processed data and generating the rules. Now using Genetic Algorithm tool in MATLAB Version 7.6.0.324 (R2008a) above rules are optimized to produce the desired profit oriented rules. The figure below shows the value of best individual and best fit value of the rules.

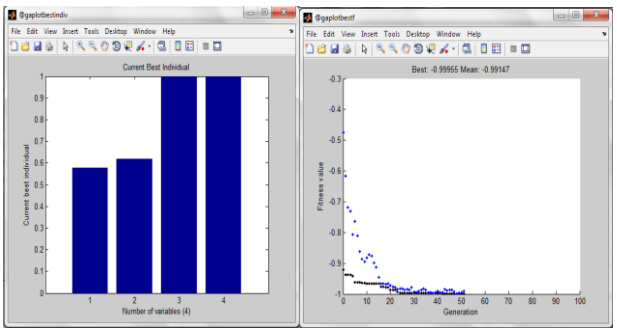

Figure-3 Best individual & Best Fitness

X. Conclusion

Mining profit pattern mixes the statistic based pattern extraction with value-based decision making to achieve the business goals. Using Genetic Algorithm to optimized rules not only improves the mining process but also provide the accuracy and efficiency to association rule mining. Although a many researches has been carried out in association rule mining but still it requires more attention for defining the notion of profit which would help in improving business strategies.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and techniques", Morgan Kaufmann Publishers, Elsevier India, 2001.

[2] R Agrawal, T.Imielinski, and A.Swami, 1993. "Mining association rules between sets of items in large databases", in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216.

[3] Melanie Mitchell, An Introduction to Genetic Algorithms, PHI, 1996

[4] A. Tiwari, R.K. Gupta and D.P. Agrawal "A survey on Frequent Pattern Mining : Current Status and Challenging issues" Information Technology Journal 9(7) 1278-1293, 2010.

[5] Ke Wang, Senqiang Zhou, and Jiawei Han, Profit Mining: From Patterns to Actions, C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 70–87, 2002.Springer-VerlagBerlin.

[6] Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms"IEEE 2004

[7] Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba, "Extraction of Interesting Association Rules Using Genetic Algorithms", Advances in Systems Modelling and ICT Applications, pp. 101-110. G

[8] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., "Optimized association rule mining using genetic algorithm", Advances in Information Mining, ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04.

[9] Xiaowei Yan, Chengqi Zhang, Shichao Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support", Expert Systems with Applications 36 (2009) 3066–3076

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

[11] Sandhu, P.S.; Dhaliwal, D.S.; Panda, S.N.; Bisht, A., "An Improvement in Apriori Algorithm Using Profit and Quantity" ICCNT Year: 2010, IEEE conference publication.

[12] Indira K and Kanmani S "Performance Analysis of Genetic Algorithm for Mining Association Rules", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814.