Mining Frequent Itemsets using Genetic Algorithm

Santosh Kumar Satapathy
Department of Computer Science & Engineering
Gandhi Engineering College, Bhubaneswar, Odisha,India

Santosh Kumar Moharana
Department of Computer Science & Engineering
Gandhi Engineering College, Bhubaneswar, Odisha,India

Suren Kumar Sahu
Department of Computer Science & Engineering
Gandhi Engineering College, Bhubaneswar, Odisha,India

Narendra Kumar Rout
Department of Computer Science & Engineering
Gandhi Engineering College, Bhubaneswar, Odisha,India

Abstract- In general frequent itemsets are generated from large data sets by applying association rule mining algorithms like Apriori, Partition, Pincer-Search, Incremental, Border algorithm etc., which take too much computer time to compute all the frequent itemsets. By using Genetic Algorithm (GA) we can improve the scenario. 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. The main aim of this paper is to find all the frequent itemsets from given data sets using genetic algorithm.

Keywords – Genetic Algorithm (GA), Association Rule, Frequent itemset, Support, Confidence,Data Mining.

I. INTRODUCTION

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions). Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. 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.

II. DATA MINING METHODOLOGIES

Frequent pattern mining is an important area of Data mining research. The frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well.

The process of discovering interesting and unexpected rules from large data sets is known as association rule mining. This refers to a very general model that allows relationships to be found between items of a database. An

association rule is an implication or if-then-rule which is supported by data. The association rules problem was first formulated in [3] and was called the market-basket problem. The initial problem was the following: given a set of items and a large collection of sales records, which consist in a transaction date and the items bought in the transaction, the task is to find relationships between the items contained in the different transactions. A typical association rule resulting from such a study could be "90 percent of all customers who buy bread and butter also buy milk" – which reveals a very important information. Therefore this analysis can provide new insights into customer behaviour and can lead to higher profits through better customer relations, customer retention and better product placements. The subsequent paper [4] is also considered as one of the most important contributions to the subject.

Mining of association rules is a field of data mining that has received a lot of attention in recent years. 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 itemsets. The papers [10, 11] contributed a lot in the field of Association Rule Mining (ARM). In this paper, an attempt has been made to compute frequent itemsets by applying genetic algorithm so that the computational complexity can be improved.

III.ASSOCIATION RULE MINING (ARM)

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.

Association Rule Mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [8]. The major aim of ARM is to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. ARM algorithms discover high-level prediction rules in the form: IF the conditions of the values of the predicting attributes are true, THEN predict values for some goal attributes.In general, the association rule is an expression of the form X=>Y, where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the support and confidence value.

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

A genetic Algorithm is an iterative procedure maintaining a population of structures that are candidate solutions to specific domain challenges. During each temporal increment (called a generation), the structures in the current population are rated for their effectiveness as domain solutions, and on the basis of these evaluations, a new population of candidate solutions is formed using specific genetic operators such as reproduction, crossover, and mutation. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem.

GAs are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space. The Genetic Algorithm [5] was developed by John Holland in 1970. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution [6].

GA has been successfully applied in many search, optimization, and machine learning problems. GA works in an iterative 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 [7].
. 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.

This generational process is repeated until a termination condition has been reached. Common terminating conditions are:
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
Simple generational genetic algorithm pseudo-code:
Choose the initial population of individuals
Evaluate the fitness of each individual in that population
Repeat on this generation until termination: (time limit, sufficient fitness achieved, etc.)
Select the best-fit individuals for reproduction
Breed new individuals through crossover and mutation operations to give birth to offspring
Evaluate the individual fitness of new individuals
Replace least-fit population with new individuals

VI.DATA MINING USING GENETIC ALGORITHMS

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.
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.

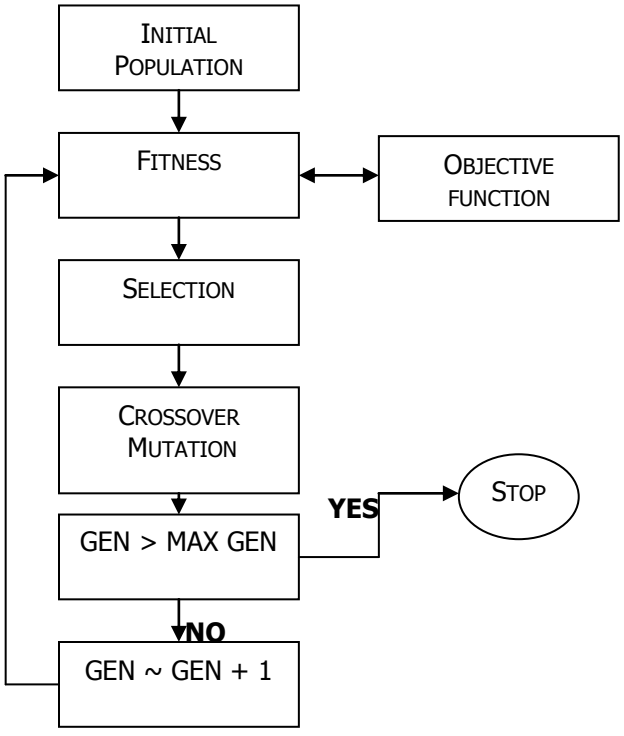
Crossover: Mate each string randomly using some crossover technique. For each mating, randomly select the crossover position(s).

Mutation: Mutation is performed randomly on a gene of a chromosome. Mutation is rare, but extremely important. As an example, perform a mutation on a gene with probability .005.If the population has g total genes (g = string length * population size) the probability of a mutation on any one gene is 0.005g, for example. This step is a no-op most of the time. Mutation insures that every region of the problem space can be reached.When a gene is mutated it is randomly selected and randomly replaced with another symbol from the alphabet.
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 individuals 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:
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
Simple generational genetic algorithm pseudo-code:
Choose the initial population of individuals
Evaluate the fitness of each individual in that population
Repeat on this generation until termination: (time limit, sufficient fitness achieved, etc.)
Select the best-fit individuals for reproduction
Breed new individuals through crossover and mutation operations to give birth to offspring
Evaluate the individual fitness of new individuals
Replace least-fit population with new individuals



VII.METHODOLGY

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

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

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

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

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

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.

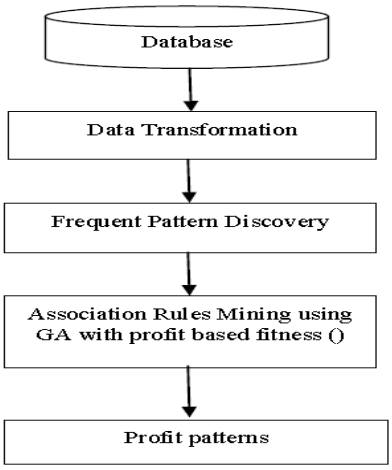
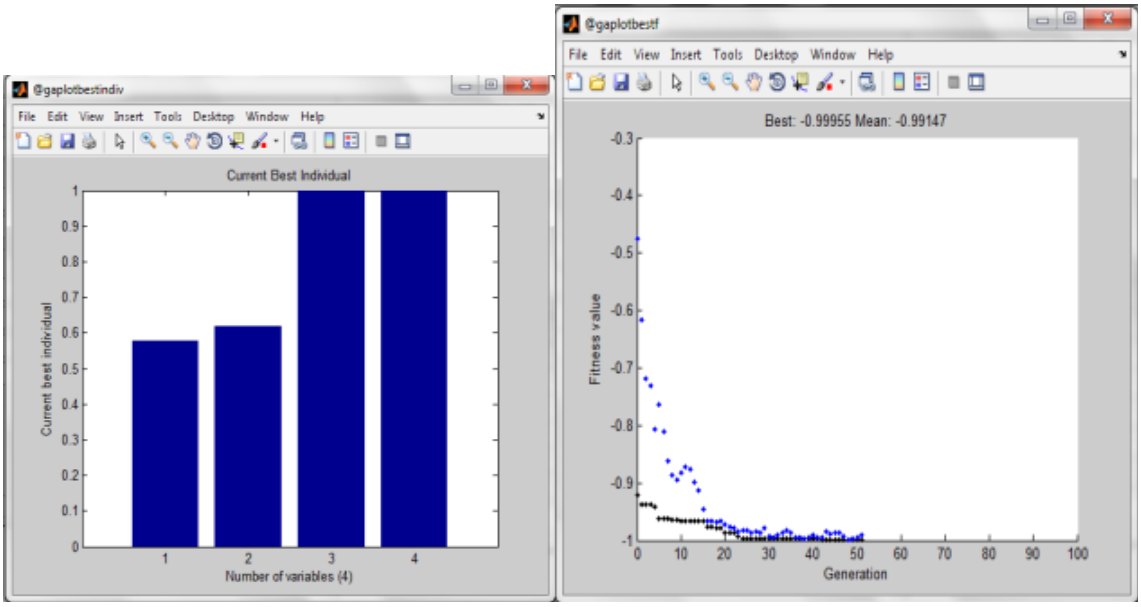


Figure-2: Proposed Methodology

VIII.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.



IX.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] L. I. Kuncheva, J.C. Bezbek, R.P.W Duin, ―Decision template for multiple classifier fusion: an experimental comparison‖, Pattern Recognition, Vol-34, pp.299-314, 2010.
[9] M. Re, G. Valentini, ―An ensemble based data fusion for gene function prediction, Multiple Classifier Systems‖, Springer, pp.448-457, 2009.
[10] H.R. Albert, R. Ko, R. Sabourin, A. S. Britto, L. Oliveira, ―Pair wise fusion matrix for combining classifiers‖, Pattern Recognition, Vol-40, pp. 2198-2210, 2007.
[11] J. Kennedy, R. Eberhart, ―Particle Swarm Optimization‖, Proc. of IEEE Int. Conf. on Neural Networks, pp.1942-1948, 1995.
[12] A.M. Sarhan, ―Cancer classification based on micro array gene expression data using DCT and ANN‖, Proc. Of Int. Conf. on General of Theoretical and Applied Information Technology, pp. 208-216, 2009.
[13] R. Kumar, M.S.B. Saithij, S. Vaddadi, S.V.K.K. Anoop, ―An intelligent functional link artificial neural network for channel equalization‖, Proc. of Int. Conf. on Signal Processing Robotics and Automation, pp. 240-245 2009.
[14] E.Peterson, ―Partitioning large –sample microarray –based gene expression profile using principal component analysis‖, Computer Methods Programming in Biomedicine, pp107-109 2003.
[15] W.Chen, S.Chen, C.Lin, ―A Speech Recognition Method Based on The Sequential Multi-layer Perceptrons‖, Neural Networks, Vol-9, pp.655-699, 1996