A MODEL FOR PROFIT PATTERN MINING BASED ON GENETIC ALGORITHM

Vivek Badhe[1], R.S. Thakur[2], G.S. Thakur[3]

[1]*Research Scholar, Maulana Azad National Institute of Technology, Bhopal, India*
[2]*Associate Professor, Maulana Azad National Institute of Technology, Bhopal, India*
[3]*Assistant Professor, Maulana Azad National Institute of Technology, Bhopal, India*

Abstract

Mining profit oriented patterns is a novel technique of association rule mining in data mining, which basically focuses on important issues related with business. As it is well known that every business aims to generate the profit and find the ways to improve the same. In earlier days association rule mining was used for market basket analysis and targeted only some of the business and commercial aspects. Afterwards the researchers started to aim the most prominent element of any business i.e. Profit, and determined the innovative way to generate the association rules based on profit. Profit oriented patterns mining approach combines the statistic based pattern mining with value-based decision making to generate those patterns with the maximum profit and some ways to generate recommenders for future strategy. To achieve the desired goal the traditional association rule mining alone is not effectual, so we combine the strength of genetic algorithm with association rule mining to enhance its capability. The study shows that Genetic Algorithm improves the effectiveness and efficiency of association rule mining outcome, since genetic algorithms are competent to handle the problems related with the uncertainty, multi-dimensional, non-differential, non-continuous, and non-parametrical, non-linearity constraint and multi-objective optimization problems. In this paper we apply the concept of profit pattern mining with genetic algorithm to generate profit oriented pattern which help out in future business expansion and fulfill the business objective.

Keywords: Data Mining, Association Rule Mining, Profit Pattern Mining, Genetic Algorithm

--***--

1. INTRODUCTION

Data mining [1] refers to the discovery of new information in terms of patterns or rules from vast amount of data. It the technique basically used for extracting the hidden predictive information from large database repository. It is a modern and powerful methodology with immense potential to analyze significant information from the huge databases. Discovery of unseen pattern is an essential database-mining task. While being a vital tool for several practitioners, data mining is also an attractive research area that raises many challenging problem [7].

Data Mining is the methodology to finding significant, new correlation pattern and trends by sifting through huge amount of data stored in repository, using patterns discovery technique as well as statistical and mathematical techniques. The two primary aims of data mining [2] are 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.

However data mining sometimes viewed as s a multi-objective task the Genetic Algorithm [3] with rule mining

achieve the target. Since data mining may also be viewed as the process of turning the data into information, the information into action, and action into value or profit, and profit pattern mining is the way to achieve this by using profit as measure of interest.

## 2. PRELIMINARIES

2.1 Association Rule Mining

Extraction of Association rules is one of the most important techniques which were given by R. Agrawal et. al. in 1993 [4]. It provides the information like "if-then" statements. These rules are invoked from the dataset. It generates from calculation of the support and confidence of each rule that can show the frequency of occurrence of a given rule. Association Analysis is the method of discovering unseen pattern or correlation condition that occurs frequently together in a given dataset but not visible due to large volume of data. Association Rule mining techniques looks for interesting associations and correlations among data set. An association rule is a rule, which entails probabilistic relationship, with the form $X \Rightarrow Y$ between sets of database attributes, where X and

Y are sets of items, and $X \cap Y = \phi$. Given the set of transactions T, we are interested in generating all rules that satisfy certain constraints. These constrains are *support* and *confidence*. The *support* of the rule is the fraction of the transactions in T that satisfy the union of items in X and Y.

The probability, measured as the fraction of the transactions containing X also containing Y, is called the *confidence* of the rule. Confidence is a measure of the rule's strength, support corresponds to statistical importance. With the help of these constraints, rules are computed from the data and, association rules are calculated with help of probability. Mining frequent itemsets [7] is a fundamental and vital task in many data mining applications such as the discovery of association rules, strong rules, correlations, multi-dimensional patterns, and many other important discovery tasks. The most important and principal algorithm that was given to generate association rules was *apriori* [5].

2.2 Profit Pattern Mining

The goal of Profit Pattern Mining (PPM) [8] is to develop a model which generates profitable rules, recommender rules that recommends target items and promotion strategy for future customer. It is a novel way of association Rule Mining which aims to find out 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.
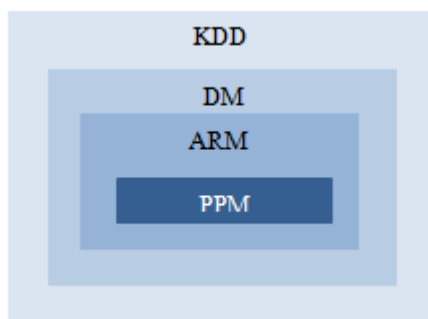

Fig 1: Hierarchy of Profit Pattern Mining

The Hierarchy of PPM shown in figure 1, In Profit Pattern Mining a set of preceding transaction and preferred target item is given and build a model for recommending target items and marketing strategies to new customers, with the aim of maximizing the business profit

2.3 Genetic Algorithm

The Theory Genetic Algorithm (GA) was given by John Holland in 1970. It incorporates Darwinian evolutionary theory survival of the fittest. Genetic algorithm [6] is a type of searching algorithm. It looks for an entire solution space for an optimal result to a problem. The main feature of the genetic algorithm is how the searching is made. The algorithm creates a "population" of possible solutions to the problem and lets them "evolve" over multiple generations to find better and better solutions. The algorithm operates through a simple cycle Population creation of strings, Evaluation of each string, Best string selection, Genetic manipulation to create a new population of strings. Figure below shows the interconnection of these four stages. Each cycle (show in figure 2) produces a new generation of possible solutions (individuals) for a given problem.
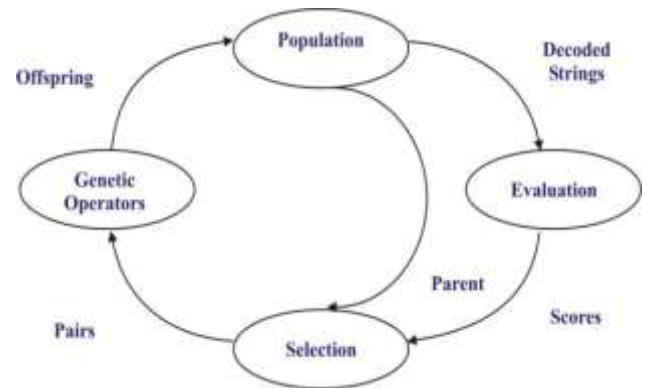
Fig 2: Genetic Algorithm Cycle

The manipulation process enables the genetic operators to produce a new population of individual the offspring, by manipulation the genetic information processed by the pairs chosen to reproduce. The information stored in strings (chromosomes) that describes the individuals. Genetic operators are used. The offspring generated by this process take the place of the older population and the cycle is repeated until a desired level of fitness is attained, or a determined number of cycles are reached.

The algorithm begins by creating a random initial population. The algorithm then produces a sequence of next population, or generations. At each step, the algorithm uses the individuals in the current generation to create the next generation. For creating the subsequently generation, the algorithm executes the following iterative steps: Scores each member of the present population by computing its fitness value. Scale the raw fitness scores to convert them into a more usable range of values. Selects parents based on their fitness. Produce children from the parents. Children are produced either by making random changes to a single parent – mutation – or by combing the vector entries of a pair of parents – Crossover. Replace the current population with the children to form the next generation. The algorithm stops when one of the five stopping criteria ie generations, time limit, fitness limit, stall generations or stall time limit is met and finally the ultimate beauty of Genetic Algorithm [16] is adaptability itself.

3. ASSOCIATION RULE MINING WITH GENETIC ALGORITHM

The utility of the genetic algorithm in the perspective of data mining is that Genetic Algorithms are robust, amendable and they approach uniformly to large number of different classes of data mining problems [3]. 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 traditional methods and models that can only be used in specific cases. Such generality is desirable in Data Mining where the search space is complex noise.

Genetic Algorithm is used with Data mining in one of the three different ways: GA in Pre-Mining: Although Genetic Algorithm is rarely applied in pre-mining ie. Before mining

process, still for some non linear constraints the GA is used in a pure traditional way. GA in In-Mining: Genetic Algorithm is applied during mining process in two ways: Traditional Genetic Algorithm: During the mining process Genetic Algorithm is applied between the mining methods ie. Association rule mining, clustering or outlier analysis etc. to optimize the results. The traditional GA is rarely used in-mining process. Modified Genetic Algorithm: During the mining process Genetic Algorithm itself is modified ie. The steps of Genetic Algorithm are altered accordingly and this is commonly used approach in in-mining process. GA in Post mining: In post mining, first the traditional mining is done and then Genetic Algorithm is applied to optimize the results. In general Genetic Algorithm is applied in a traditional way but sometimes the encoding method is modified as per the need. The use of Genetic Algorithm in Post mining is most commonly done.

Association Rule Mining uses primarily Apriori method for generating rules by using statistical measures, but there is no surety whether these rules formed are best at hand. A proposition was given by Manish Saggar et al. [9] suggesting that the induction of Genetic Algorithm will facilitate the optimization of rules being generated by Apriori while also taking into consideration the negative incidence. Now, optimizing the rules is one thing but generating optimized rules is another. The two metrics syntactic and transactional superiority proposed in 2008 [10] serves as the plummet in ascertaining the existence of interesting rules form a resultant set of rules. These two metrics are responsible for maintaining the quality of the optimized interesting rules generated in spite of the dominance of the regular rules of Apriori. Subsequently, the algorithm takes a bit longer to generate optimized rules and allows a better representation of results to the users.

The Apriori algorithm and is extensions like Partition, Pincer-Search, Incremental etc, uses properties that allow only those objects (attributes) to be candidates that acquire a certain threshold. Hence, the objects that do not fit the criteria are eliminated from the algorithm, and there is no sure way to keep these objects with negative occurrences. To find all the possible rules being generated form a given dataset GA is applied [11]. If Genetic Algorithm (GA) is used, a predictive analysis could be done for the generated rules that contain negative objects. Also these rules have more than one object in consequent part as compared to the regular ones. The predictive rule discovery using GA proves to be advantageous as it performs a global search and is based on the greedy approach thus have less complexity as compared to other algorithms. The traditional association rule algorithms use support and confidence as threshold value to generate interesting rules. The genetic algorithm-based strategy designed by Xiaowei Yan et al. [12] identifies association rules without considering the actual support threshold. Their approach employs an elaborated encoding method where the relative confidence threshold is used as the fitness function for the algorithm. As the proposed model did not require minimum support threshold, the GA performs a global search and a system automation in enforced.

The GA algorithm has less complexity as compared to other algorithms because it follows the greedy approach. Soumadip Ghosh [13] using this property mines frequent itemsets using GA. The frequent itemsets discovered by GA is done by performing a global search on the dataset. In 2012, K Indira and S Kanmani [15] proposed a technique which was able to analyze the performance of Genetic Algorithm when used for Mining ARs. The works considered for analysis were over a period of seven years for mining ARs using GA. The performance analysis of GA was done effectively by making a small yet beneficial amount of modifications to GA operators and monitor how the parameter functions in some particular interim.

4. RELATED WORK

In 2002 Ke Wang, Senqiang Zhou, and Jiawei Han presented a concept of profit mining [8], this approach to reduce the gap between the statistic-based pattern mining and the value-based decision making. They obtained a set of past transactions and pre-selected target items, and intended to construct a model for recommending target items and promotion strategies to new customers, with the goal of maximizing the net profit. They identified some issues in profit mining and proposed solutions. They evaluate the effectiveness of this approach using data sets of a wide variety of characteristics. The key to profit mining is to suggest "right" items and "right" cost. 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 discover the patterns for right items and right cost. The major issues in this context are Profit oriented patterns, shopping on unavailability, huge search space, optimality and interpretability of recommender.

Another efficient approach based on weight factor and utility for proficient mining of significant association rules proposed by Sandhu, P.S. et. al. in 2010 [14]. Initially, the proposed approach makes use of the conventional Apriori algorithm to generate a set of association rules from a repository. 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 show the effectiveness of the proposed approach in generating high utility association rules that can be profitably applied for business growth

5. PROPOSED WORK

The proposed framework and design shown in figure 3 for mining the profit patterns using Genetic Algorithm covers the following tasks and they are Data Preprocessing, Implementing ARM Algorithm on processed data and optimizing the Rules using GA. Initially we have perform the

conventional association rule mining algorithm implemented in C++ language to generate rules from the preprocessed database than we optimized those generated rules by Genetic Algorithm implemented in GA Solver of optimization toolbox called optimtool of MATLAB®.
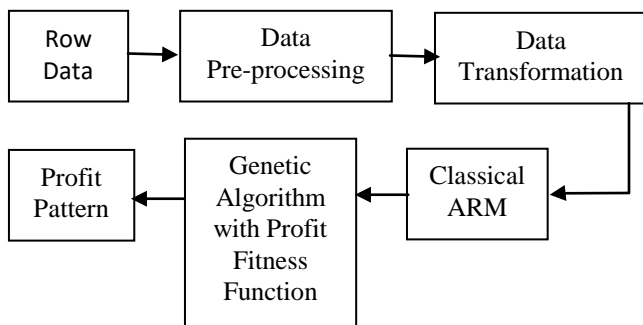

Fig 3: Block Diagram of Methodology

To design the fitness faction for the genetic algorithm we use measure called profit and define the notion of profit because the profit in any business is the key element and the notion of profit may vary depending upon the type of business but in general the notion of profit could be categorized as under:

Value profit: Value profit is simply the difference of selling price and the cost price of any product. It is also called the margin of profit.

Percentage of profit: It is the percentage of margin of profit with respect to the cost price of any product.

Quantitative profit: It is the profit based on the number of items sold, and sometimes it is known as weighted factor.

For each rule

$$Item1 \rightarrow Item2$$

The fitness function designed as

$$Fitness\ (FF) = \frac{C * w1 + I * w2}{w1 + w2}$$

Where w1 and w2 are the user defined weighted factors, and C and I namely Completeness and Interestingness. The value of w1 and w2 are calculated as:

$$w1 = \frac{Percentage\ profit\ of\ Item1 * Quantity\ of\ Item1}{Percentage\ profit\ of\ Item2 * Quantity\ of\ Item2}$$

$$w2 = \frac{Value\ profit\ on\ Item1 * Quantity\ of\ Item1}{Value\ profit\ on\ Item2 * Quantity\ of\ Item2}$$

Completeness and Interestingness are defined & calculated as:

Completeness (C): Those rules are considered as complete rules where:

$$Item1 \rightarrow Item2$$

Item1 is having lower percentage of profit & Item2 having higher percentage of profit.

Interestingness (I): Those rules are considered as rules of interest where

$$Item1 \rightarrow Item2$$

Item1 is having lower value profit & Item2 having higher value profit.

$$C = \frac{TP}{TP + FP} \qquad I = \frac{TP}{TP + FP}$$

Where TP, FP, FN are defined as

(For sample rule Item1 $\rightarrow$ Item2)

True Positive (TP):
No. of rules satisfying both Item1 & Item2

False Positive (FP) :
No. of rules not satisfying Item1 but satisfying Item2.

False Negative (FN):
No. of rules satisfying Item1 but not satisfying Item2.

Pseudo code of proposed methodology
1. Start
2. Preprocess the Row Dataset |RD|
3. Transform the |RD| to Relevant Transaction Dataset |D|
4. Load the Sample Transactions |S| from Dataset |D|
5. Apply Apriori Algorithm to |S| for Rule Generation with defined parameter Support and Confidence.
6. Store the output of Apriori to rule set |R|
7. Apply the GA Cycle on |R|
   i.   Selection - Tournament
   ii. Crossover – Single Point
   iii. Mutation - uniform
   iv. Check fitness – Defined Fitness Function FF
   v.   Check termination Condition Stall 100
8. Store the outcome of GA as final result to |F|, which contains the optimized (profitable) rules.
9. Mapped the |F| Rule with desire format
10. Stop

6. IMPLEMENTATION AND RESULT

The dataset that we gathered from departmental store contains retail data of Fast Moving Consumer Goods (FMCG) during sales of Second-Third quarter of year 2003. The dataset contains 9 attributes and 16293 records. The records hold repeated yet different transactions for a number

of products. The 9 attributes of the dataset are BillNo, ProductCode, ProductName, Packaging, ProductCategory, Quantity, PurchasedPrice, SellingPrice, BillDate. All 16293 records have a number of items purchased on a distinct bill.

RawData: The above dataset is then pre-processed in order to be more relevant for the mining task. This is done by first finding the number of distinct BillNo, ProductCode, ProductName and arranging all column-wise in ascending order. On processing the dataset the findings are listed in Table 1.

Table 1: Dataset Description

S.No.	Category	Count
1	Product List	16293
2	BillNo	1703
3	ProdName	1743
4	ProdCodes	2249

Firstly the database is taken which is shown below and converted into Flat file text tab delimited format as show in figure 4.

Fig 4: Snapshot of Row Data

Applying Apriori algorithm on the processed data and generating the following rules shown in figure 5.

Fig 5: Snapshot of Apriori Generated Rule

Now using Genetic Algorithm toolbox shown in figure 6 in MATLAB Version 7.6.0.324 (R2008a) above rules are optimized to produce the desired profit oriented rules, table2 and table3 shows the generic and profit rules respectively.
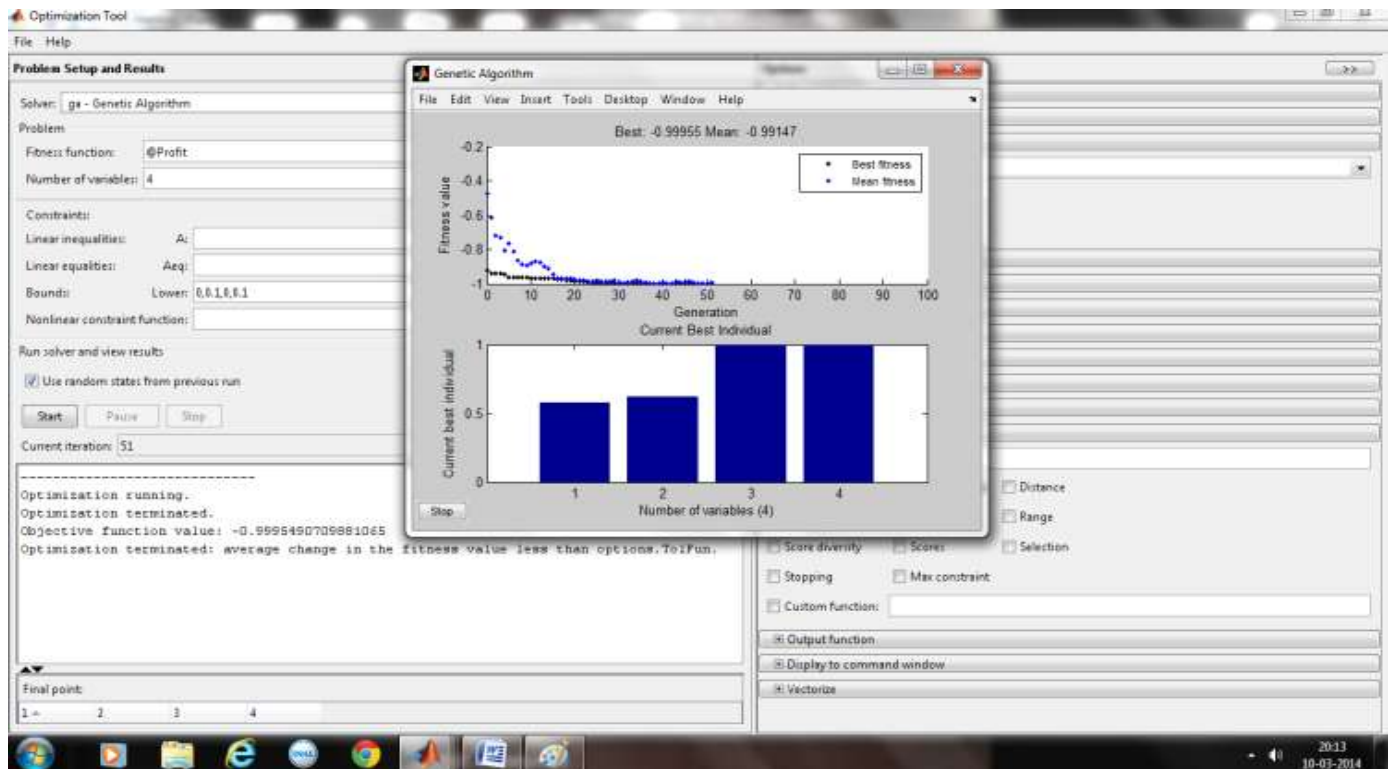

Fig 6: Snapshot of MATLAB Toolbox

Table 2: Generic Rule with Fitness value

Rules						C	I	W1	W2	Fitness
AAG111	<--	AAG005				1.00	0.50	0.13	0.11	0.8
AAG292	**<--**	**AAG235**				**1.00**	**0.60**	**0.23**	**0.07**	**0.9**

Table 3: Profit Oriented Rule with Fitness value

Rules					C	I	W1	W2	Fitness
AAG111	**<--**	**AAG005**			**1.00**	**0.50**	**0.13**	**0.11**	**0.8**
AAG220	<--	BAG049			1.00	0.25	0.21	0.52	0.5
AAG105	<--	AAG005			1.00	0.50	0.31	0.34	0.7
AAG295	<--	AAG292			0.67	0.40	0.06	0.06	0.5
AAG295	<--	AAG235			0.50	0.20	0.13	0.04	0.4
AAG296	<--	AAG292			0.67	0.50	0.13	0.13	0.6
AAG296	<--	AAG235			0.33	0.20	0.30	0.09	0.3
BAG186	<--	BAG049			0.13	0.25	0.10	0.05	0.2
AAG154	<--	BAG049			0.14	0.25	0.06	0.15	0.2
AAG198	<--	BAG049			0.25	0.25	0.09	0.09	0.3
BAG049	<--	AAG198			0.50	0.13	0.11	0.11	0.3
BAG186	<--	AAG010			0.10	1.00	0.10	0.08	0.5
AAG235	<--	AAG292			1.00	0.20	0.04	0.14	0.4
AAG292	**<--**	**AAG235**			**1.00**	**0.60**	**0.23**	**0.07**	**0.9**
AAG154	<--	BAG186			0.29	0.33	0.06	0.30	0.3
BAG186	<--	AAG154			0.50	0.71	0.17	0.03	0.5
AAG198	<--	BAG186			0.50	0.33	0.09	0.19	0.4
BAG186	<--	AAG198			0.29	0.29	0.11	0.05	0.3

AAG198	<--	AAG154			0.14	0.25	0.16	0.06	0.2
AAG154	<--	AAG198			0.57	0.50	0.06	0.16	0.5
AAG295	<--	AAG292	AAG235		0.33	0.50	0.41	0.03	0.3
AAG296	<--	AAG292	AAG235		0.33	0.50	0.91	0.05	0.3
BAG049	<--	AAG154	AAG198		0.50	0.50	0.66	0.04	0.5
AAG010	<--	BAG186	AAG154		0.67	0.67	0.63	0.03	0.7
AAG010	<--	BAG186	AAG198		0.67	0.67	0.51	0.04	0.7
AAG198	<--	BAG186	AAG154		0.25	0.33	0.58	0.05	0.3
AAG154	<--	BAG186	AAG198		0.20	0.33	0.30	0.10	0.2
BAG186	<--	AAG154	AAG198		0.17	0.33	0.67	0.02	0.2
AAG010	<--	BAG186	AAG154	AAG198	0.33	1.00	0.40	0.03	0.4

7. CONCLUSION

Association Rule Mining for profit pattern combine the statistic based pattern extraction with value-based decision making to achieve the commercial goals. In our proposed approach we have propose a model in which we apply classical association rule mining followed by genetic Algorithm techniques. Genetic Algorithm not only improves the mining process but also provide the optimized rules. Although a several 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 strategy and better decision making

ACKNOWLEDGMENT

REFERENCES

Books

[1] J. Han and M. Kamber, "Data Mining: Concepts and techniques", Morgan Kaufmann Publishers, Elsevier India, 2001.
[2] A. K. Pujari, Data Mining Techniques, University Press 2001.

Journals

[3] Satchidananda Dehuri, Ashish Ghosh, R Mall, Parallel Multi-objective Genetic Algorithm for Classification Rule Mining, IETE Journal of Research,vol 53,No .5, PP475-483
[4] 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.
[5] R.Agrawal and R.Shrikanth."Fast Algorithm for Mining Association Rules." In Proceeding of VLBD Conference, Santigo, Chile-1994,pp 487-494
[6] Melanie Mitchell, An Introduction to Genetic Algorithms, PHI, 1996

[7] 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.
[8] 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.
[9] Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms"IEEE 2004
[10] Peter P. Wakabi-Waiswa , Dr. Venansius Baryamureeba, "Extraction of Interesting Association Rules Using Genetic Algorithms", Advances in Systems Modelling and ICT Applications, pp. 101-110. G
[11] 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.
[12] 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
[13] 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
[14] 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.
[15] Indira K, Kanmani S, Performance Analysis of Genetic Algorithm for Mining Association Rules, International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814.
[16] Mehmed Çelebi, A new approach for the genetic algorithm, Journal of Statistical Computation and Simulation, Taylor & Francis group, volume 79, issue 3 March 2009,, PP 275-279