ISSN:0975-9646

Problem Solving Strategy for Decision Making in Market Basket Analysis

Sneha Sawlani[1], Piyush Vyas[2]

[1]*Research Scholer BCA M.K.H.S. Gujarati Girls College DAVV,Indore, India*
[2]*IIST, Indore, India*

***Abstract*- These days many big organizations are striving for lower consumption rate of goods especially in retail market. Artificial intelligence is the key to resolve it. Artificial intelligence is known for solving real time problems. In this research paper we concentrate over decision making problems in retail market to improve productivity. Present day scenario clearly indicates about the competition among the many retail marts or stores. Our approach clearly helps in increasing productivity with the help of data mining and evolutionary algorithms like genetic algorithm and particle swarm optimization. In this paper we represent the pseudo algorithm of our approach with the sample data set collected by a short survey of retail store in locals.**

***Keywords*— Artificial intelligence, PSO, Genetic algorithm, Association Rule mining.**

I. INTRODUCTION

Basically data mining includes discovering meaningful new correlations, patterns and trends by searching through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining refers to extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information from large database such as: relational database, data warehouses, XML repository, etc. Knowledge discovery from databases is also known as data mining. It is treated as synonym for another popular term, knowledge Discovery in Database (KDD)[1].

A. Association rule mining

Association rule mining is one of the important techniques of data mining. Its purpose is to extract interesting correlations, frequently occurring patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. For Example in a retail store daily consumption of milk , bread and butter is 40 %, 25% and 25% respectively. Suppose, from the daily consumption store data we found that, 25 % of customer who brought bread also brought milk and 25 % of customer who brought butter also brought milk. This conclusion is treating as an association rule. Due to these rules, in future lot of business policies would make sense in increasing productivity. Basically association rule mining is used for market basket analysis. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc [1].

1) Basic concepts

As mentioned before association rule mining depends upon some basic terminologies like support count and confidence.

2) Support

Suppose we have nine transactions and five items. If an item set contain two or more items in a particular transaction then support count of the item is fraction of no. of that item that comes in all transaction and total no. of transactions. Suppose support of an item is 0.1%, it means only 0.1 percent of the transaction contains purchasing of this item. The retailer will not pay much attention to such kind of items that are not bought so frequently because a high support is desired for more interesting association rules.

$$Supp(A \Rightarrow B)=supp(A \cup B)=P(A \cup B) \quad \text{Eqn. (1)}$$

In a single line we can say that "The support of an item-set is defined as the proportion of transactions in the data set which contain the item-set".

3) Positive association rule

In this work we also did research over positive association rule mining. I first define positive association rule? Some strong pattern such as if a customer buys milk he/she is likely to buy bread or if when 80% of the time customer buys milk and bread he/she is likely to buy butter in 20% of the time. This rule is known as positive rule, from the example we conclude that milk and bread combination are really helpful to increasing productivity and butter is also a great option to make pair with bread for increase productivity. Support calculation and confidence calculation for positive rule is showed earlier.

4) Negative association rule

In association rule mining we evaluate negative association rule with the help of absence of item with regular item set. Suppose an item set contains A and B, and other item set contains B and C, so here in first item set, item C is absent which shows the negation of C. Similarly in second item set, item A is absent, so here it shows the negation of A. With this strategy we find out negative rules practically. Calculation of support and confidence for negative rules

The support is given by the following formulas,

$$supp(\neg A)=1-supp(A) \quad \text{Eqn. (3)}$$

$$Supp (\neg A=>B) =supp (B)-supp (A \cup B) \quad \text{Eqn. (4)}$$

$$Supp (A=>\neg B) =supp (A)-supp (A \cup B) \quad \text{Eqn. (5)}$$

$$Supp (\neg A=>\neg B) =1-supp (A)-supp (B) +supp (A \cup B) \quad \text{Eqn. (6)}$$

The confidence is given by the following formulas:

$$Conf (\neg A=>B) =supp (B)-supp (A \cup B)/1-supp (A)$$

Eqn. (7)

Conf (A=>¬B) =supp (A)-supp (A∪B)/supp (A)

Eqn. (8)

Conf (¬A=>¬B) =1-supp (A)-supp(B)+supp(A∪B)/1-supp(A)

Eqn. (9)[14].

5) Transactional database

Transactional database refers to the collection of transaction records. Mostly transactional database are available in super markets or retail stores. Due to enlargement of computerized selling strategies, transactional database are taking important place. Basically transactional database are used for association rule mining. Association rule mining applies to it in finding out relationship between item set of products or goods.

II. Releted Work

Generally at inception of data mining in 1993 first R. Agrawal, T. Imielinski and A. Swami researched about association rules mining between sets of items in large databases. They stated about an efficient algorithm that generates all significant association rules between items in the database. Rakesh Agrawal Tomasz Imielinski_ Arun Swami: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993 .[2]. After a while in year 1999 Georges R. Harik, Fernando G. Lobo, and David E. Goldberg gave an idea about The Compact Genetic Algorithm in which they concentrated toward the new modified Genetic to optimize[8]. In year 2006, Chris Cornelis, Peng Yan, Xing Zhang, Guoqing Chen worked over Mining Positive and Negative Association Rules from Large Databases[13].Till the time many researchers did work on positive rule mining and optimization. After a while, Ling Zhou, Stephen Yau did work on Association Rule and Quantitative Association Rule Mining among Infrequent Items [14]. In year 2012,]Asst. Prof. Nirupama Tiwari, Anubha Sharma1 did A Survey of Association Rule Mining Using Genetic Algorithm, National Conference on Security[10]. These all works for association rule mining and optimization motivates us to apply genetic algorithm for market basket.

III. Genetic Algorithm

GA is based on Darwinian evolutionary theory with sexual reproduction. GA has been successfully applied in many search, optimization, and machine learning problems. Fitness measure of every string indicating its fitness for the problem and a function which evaluated fitness is known as fitness function. 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 [3] [8].

A. Basic terminology of GA

1) Chromosome

A chromosome is also known as genome in biological manner. It is a set of proposed solution parameters for problems solved by GA. The chromosome is not only represented by simple string but also represented by wide variety of other data structures. After applying GA operators we got new strong genomes or chromosomes.

2) Gene

It is a part of chromosome and contains a part of solution. For example if ABCDE is a chromosome then A, B, C, D and E are its genes.

3) Fitness

Fitness is key factor of "Survival of fitness" theory and an important idea for evaluation methods. Fitness depends upon function which evaluates a value after the decision about whether a chromosome is best fit or worst fit according to value takes place. That function may be maximize or minimize according to problems' requirement [9] [10].

IV. Paritcle Swarm Optimization

PSO imitates behaviours of bird flocking. For example: a group of birds are randomly searching for food in a region. There is only one piece of food in the area being searched. No bird knows where the food is. But they know how far the food is. So what's the best strategy to find the food? The effective one is to follow the bird which is nearest to the food. Cultured from this scenario we can use PSO to solve the optimization problems. In PSO, each single solution is a "bird" in the search space. Call it "particle". All of particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. The particles fly through the problem space by following the currently best possible particles. PSO is initialized with a group of random solutions and then searched for optimal one by updating generations. In all iterations each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called p_{best}. Another "best" value that is tracked by the particle swarm optimizer is the best value obtained so far by any particle in the population. This best value is a global best and called g_{best}. When a particle takes part in the population as its topological neighbours, the best value is a local best and is called l_{best}.[12]

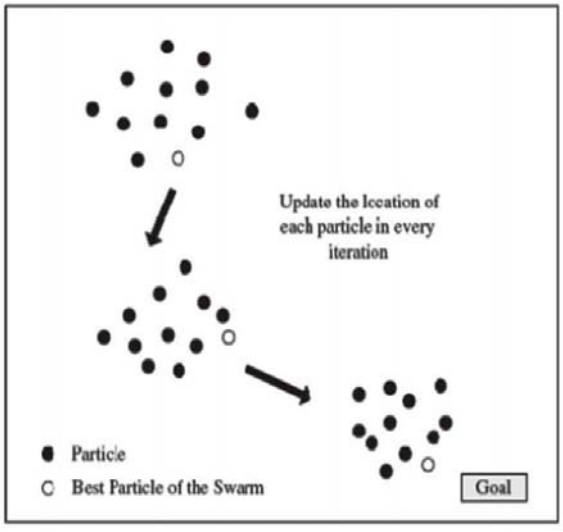

Fig. 1.Flocking of particles

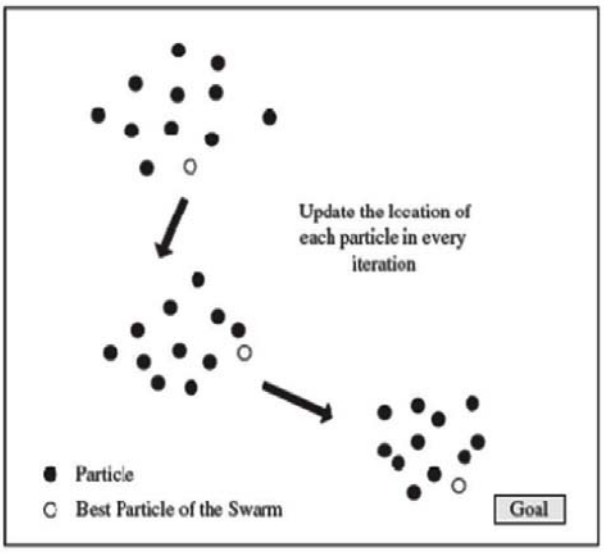

Fig. 1.Flocking of particles

After finding the two best values, the particle updates its velocity and positions with following equation

$v[] = v[] + c1 * rand() * (p_{best}[] - present[]) + c2 * rand() * (g_{best}[] - present[])$ Eqn. (10)

$present[] = present[] + v[]$ Eqn. (11)

v[] is the particle velocity, present[] is the current particle (solution). $p_{best}[]$ and $g_{best}[]$ are defined as stated before. rand () is a random number between (0,1). c1, c2 are learning factors. Usually c1 = c2 = 2.

From the procedure, we can find out that PSO has many common things with GA. Both algorithms take a group of a randomly generated population as input, both have fitness values to evaluate the population. Both update the population and search for the optimum solution with random techniques. Both systems do not guarantee success. But PSO does not have genetic operators like crossover and mutation. Particles update themselves with the internal velocity. They also have memory, which is important to the algorithm. Compared with genetic algorithms (GAs), the information sharing mechanism in PSO is significantly different. In GAs, chromosomes share information with each other. So the whole population moves like a one group towards an optimal area. In PSO, only g_{Best} (or l_{Best}) gives out the information to others. It is one way of information sharing mechanisms. The evolution only looks for the best solution.

V. PROPOSED APPROACH

Genetic algorithm and particle swarm optimization both are the evolutionary algorithms. Many real time problems are resolved by them. So we also take these to resolve market basket problem. We took a sample data from fruit mart as,

TABLE I
SAMPLE DATA SET OF MARKET BASKET

T.id.	ITEM SET			
1	mango	oranges	apple	milk
2	tea	sugar	milk	jam
3	tea	oranges	apple	milk
4	sugar	tea	milk	apple
5	mango	sugar	bread	jam
6	mango	sugar	milk	bread
7	tea	sugar	jam	bread
8	mango	tea	bread	milk
9	oranges	tea	apple	milk

1. Apply sample transitional data set of fruit mart.
   Above data set indicates item set with their transitional id.
2. Apply Apriori algorithm for getting frequent item set from transactional data set.
   Aprioro algorithm works over sample data set and find out frequent data set (how frequently an item occure in transitional data set).
3. Apply genetic and PSO algorithm over frequent item set to get association rules.
   After getting frequent data set Genetic algorithm and PSO both find the association rule between frequently occurring data set. We apply both at prior level to compare both optimizations. Both algorithms randomly select the set and GA uses the flippant mutation for new generation of solution. We provide sample threshold fitness value to rules is 0.02.
4. Separate the items from the rules for further decision making to increase the productivity of infrequent items through frequent items. Generated rules are very efficient because they formed by lot of filtering. Now we compare both the sample data set and the generated rules data set which help us or any analyst to make a different decision for improving productivity of those items which are rest in a rack from very long time.

A. Results by our basic approach

After applying GA, with Min Support = 0.2, Min Confidence = 0.2, Min fitness = 0.02 fitness value evaluated by product of support and confidence values.

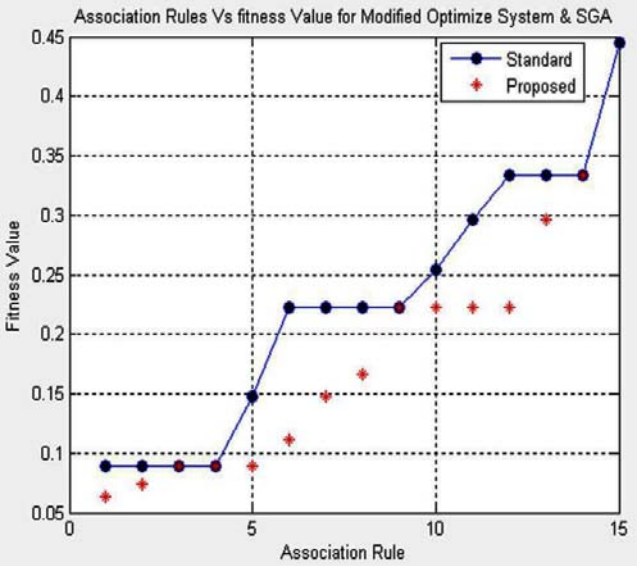

Fig.2.Association rule fitness values after GA

After applying PSO, with Min Support = 0.2, Min Confidence = 0.2, Min fitness = 0.02 fitness value evaluated by product of support and confidence values.

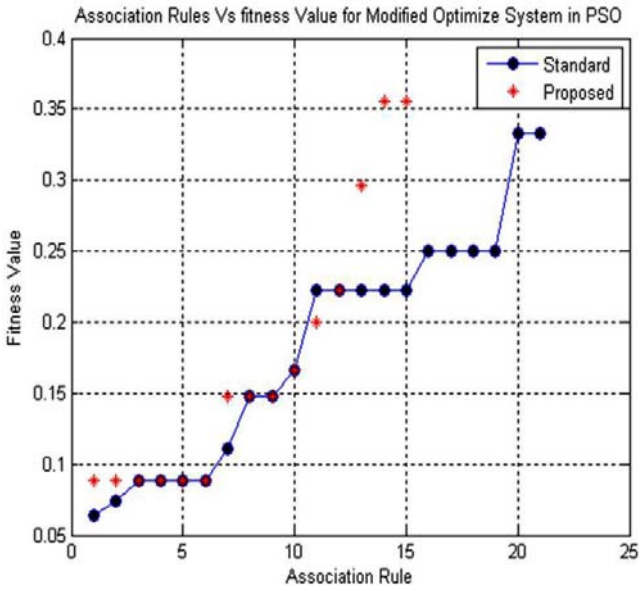

Fig.3.Association rule fitness values after PSO

VI. Conclusions

In this presented work we applied both genetic and particle swarm optimization algorithm over market basket sample data. Now we conclude that if data set is large enough and we want to get rules in short time, PSO is a good solution but for efficiency or for optimum solution genetic algorithm is best solution. Now, in future work we concentrate over genetic algorithm with some modifications in it and going to take such huge data sets like, Banking mortgage approval, loan underwriting, fraud analysis and detection, Finance analysis and forecasting of business performance, stock and bond analysis and Medicine Epidemiological studies.

References

[1] Qiankun Zhao: Association Rule Mining: A Survey, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[2] Rakesh Agrawal Tomasz Imielinski_ Arun Swami: Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993.

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

[4] Georges R. Harik, Fernando G. Lobo, and David E. Goldberg: The Compact Genetic Algorithm, IEEE Transactions On Evolutionary Computation, Vol. 3, No. 4, November 1999, pp. 287-297.

[5] Chris Cornelis, Peng Yan, Xing Zhang, Guoqing Chen: Mining Positive and Negative Association Rules from Large Databases, 2006 IEEE.

[6] Ling Zhou, Stephen Yau: Association Rule and Quantitative Association Rule Mining among Infrequent Items, 2007 ACM.

[7] Honglei Zhu, Zhigang Xu: An Effective Algorithm for Mining Positive and Negative Association Rules, 2008 International Conference on Computer Science and Software Engineering, 2008 IEEE.pp. 455-458.

[8] Sufal Das & Banani Saha: Data Quality Mining using Genetic Algorithm, International Journal of Computer Science and Security, (IJCSS) Volume (3) , Issue (2), pp. 105-112.

[9] Sanjay S , Pradeep S , Manikanta V , Kumara S.S , Harsha P: Genetic Algorithm Based Approach For The Selection Of Projects In Public R&D Institutions, Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 4 Aug -Sep 2011, pp. 523-532.

[10] Asst. Prof. Nirupama Tiwari, Anubha Sharma1: A Survey of Association Rule Mining Using Genetic Algorithm,National Conference on Security Issues in Network Technologies (NCSI-2012).

[11] S.N. Sivnandam, S.N. Deepa: Principles of Soft computing, II-edition, Wiley India pvt. Ltd, New Delhi, 2011, pp 385-464.

[12] Kennedy J and Eberhart R, Swarm intelligence ,Morgan Kaufmann Publishers,Inc San Francisco.