An optimization model for market basket analysis with allocation considerations: A Genetic Algorithm based approach

Majeed Heydari, Amir Yousefli,

***Abstract*—** Nowadays market basket analysis is one of the interested research areas of the data mining that has received more attention by researchers. But, most of the related researches focused on the traditional and heuristic algorithms with limited factors that are not the only influential factors of the basket market analysis. In this paper to efficient modeling and analysis of the market basket data, the optimization model is proposed with considering allocation parameter as one of the important and effectual factors of the selling rate. Genetic algorithm is applied to solve the formulated non-linear binary programming problem and a numerical example is used to illustrate the presented model. The provided results reveal that the obtained solutions are seems to be more realistic and applicable.

***Keywords*—** Market basket analysis, Association rule, Non-linear programming, Genetic algorithm

I. Introduction

Nowadays, data mining is widely used in several aspects of science such as manufacturing, marketing, CRM, retail trade etc. Data mining or Knowledge discovery is a process for data analyzing to extract information from large databases. Artificial intelligence, neural network, statistical techniques, pattern recognition, clustering and classification approaches are areas included in the data mining. With increasing data mining popularity, most researchers conducted to apply data mining techniques to extract information from data sets.

Some authors used mathematical programming approaches to discover knowledge. Mangasarian [1] applied optimization models to classification and Vinod [2] and Rao [3] presented some optimal model for clustering. Padmanabhan and Tuzhilin, [4] used optimization models for improving eCRM problems. Bradley et al. [5] formulated the basic categories of data mining methods as optimization problems. Olafsson et al. [6] surveyed the intersection of operation research and data mining and they illustrated the range of interactions between them. On the other side, some authors focused on association rule mining as one of the most important techniques of data mining to catch out association rules which fulfill the predetermined minimum support and confidence from a given database [7]. Hegland [8] reviewed the most famous algorithms for producing association rules.

Market basket analysis is a well-known problem that numerous researchers have paid special attention to so far. Data mining techniques are frequently used for handling this problem. Tang et al. [9] proposed an approach to performing market basket analysis in a multi-store and multi-period environment. In another attempt Chen et al. [10] claimed that the most of models presented for dealing with market basket problem could not discover important purchasing patterns when multiple stores exist. So they developed a method to overcome this weakness. Yun et al. [11] clustered data of market basket using novel measurement that they named category-based adherence. Cavique [13] converted market basket problem into a maximum-weighted clique problem for discovering large item-set patterns. Russell and Urban [13] presented an optimization model for shelf-management problem in which products are grouped as families and the location of each family is determined on the shelves. They considered shelves location effect on sales but did not attend the cross selling effect and also purchase data never been used. Nierop et al. [14] proposed a method for dealing with shelf-management problem which consist of two parts. In the first phase, statistical model was provided to measure the impact of shelf layout on sales. In the second part simulated annealing (SA) is used to maximize expected total profit. Similar to Russell and Urban [13], they did not mined association rules from customers' purchasing data to maximize cross selling effect.

In the most recent research, considering a market basket analysis problem, Saraf and Patil [15] proposed a Bottom-up" hierarchical clustering approach for clustering a retail items. To do this, they applied the concept of 'distance' between the entities or, groups of entities to achieve the purpose of Market-Basket Analysis. For more information about the application of optimization models in data mining and application of evolutionary algorithms in association rule mining, please refer to [17,18].

Although there may be considerable number of research on the association rules mining techniques and optimization methods in a separate research however, to the best of our knowledge, mathematical optimization methods alongside

Majeed Heydari, Department of Mechanical and Industrial Engineering, University of Zanjan, Iran. (Corresponding author) PO Box: 45371-38791, Email: mheydari@znu.ac.ir

Amir Yousefli, Department of Industrial Management, Imam Khomeini International University, Qazvin, Iran , PO Box: 34148 – 96818 Email: yousefli@soc.ikiu.ac.ir

association rule mining as well as applying meta-heuristic methods has not been properly incorporated to formulate the market basket analysis problem. In this paper we present a non-linear zero-one optimization model for mining association rules and allocating products to shelves. It is worth to note that , the proposed mathematical model and applying proper meta-heuristic is not addressed in the previous research and we believe that the proposed model provide a comprehensive framework to more realistic formulation of the real world problems.

The organization of this paper is as follows: in the next section, problem description and formulation is presented. In the third section, a Genetic Algorithm is used to solve proposed model. An illustrative example is provided to clarify the proposed model in the fourth section. Finally, conclusions are remarked.

II. Problem Description and Formulation

Consider market data logs that include the items purchased by the customers. The manager of supermarket is interested in maximize the interestingness of the product placement on shelves. That is, interestingness value related to mined association rules and the location of shelves. The rationale of interestingness maximization with location considerations is based on this fact that, association rule mining helps to maximize cross selling effect however it is clear that the location of shelves have undeniable impact on the selling rate. For example the products that are placed into near the entrance or exit doors have more chance to be purchased. So, it can be said that, the preference function of the supermarket manager depends on the following parameters: selling benefit, support and confidence of each pair of products and the selling possibility of each shelf for each product. These parameters are integrated in the following preference function:

Preference function:

$$\sum_{i=1}^{m-1}\left[\sum_{l=i+1}^{m}\left[C_{il}+C_{li}\sum_{k=1}^{P}[b_i v_{ik}+b_l v_{lk}x_{ik}x_{lk}]\right]\right] \quad (1)$$

Where m is number of products is, P *is the* number of shelves, C_{il} is the confidence of the rule (product *i*⟶ product *l*), b_i also is selling benefit of the ith product, v_{ik} is selling possibility degree of the product *i* when placed into the kth shelf and x_{ik} is the binary decision variable that takes 1 when product *i* is allocated to shelf *k*, otherwise x_{ik}will be 0.

As it is expected, there are some restrictions that limit the preference function value. At first, the capacity limitation of each shelf must be considered as the following constraint.

$$\sum_{i=1}^{m} x_{ik} \le U_k \quad ;k=1,2,\dots,P \quad (2)$$

Where U_k is the capacity of the kth shelf. The second constraint is the association constraint: Support of the rule (product *i*⟶ product *l*) must greater than the minimum threshold determined by the decision maker.

$$x_{ik}x_{lk}(S_{il}-S_{min}) \ge 0 \quad ;\forall i,l \in \{1,2,\dots,m\} \quad (3)$$

Where, S_{il} is the support of rule (product *i*⟶ product *l*) and S_{min} is the minimum support.

The third constraint says each product can be allocated to just one shelf according to the following equation.

$$\sum_{k=1}^{P} x_{ik} = 1 \quad ; i=1,2,\dots,m \quad (4)$$

According to this fact that the objective and constraints are non-linear functions in which decision variables are binary; we deal with a rough feasible space that increases the probability of trapping in the local optimum. Moreover, it could be proved that the developed model belongs to class of computationally hard problems which is called Np-hard problems. So in the next section a GAs based solution approach is developed to solve the proposed mathematical model.

III. GA Based Solution Approach

Genetic Algorithm belongs to a category of meta-heuristics methods known as stochastic search ones that uses randomized choice of operators in its search strategy. In this section, we implement GAs to develop a solution approach for presented model in the previous section. The general mechanism of the GAs is depicted in Fig.1.

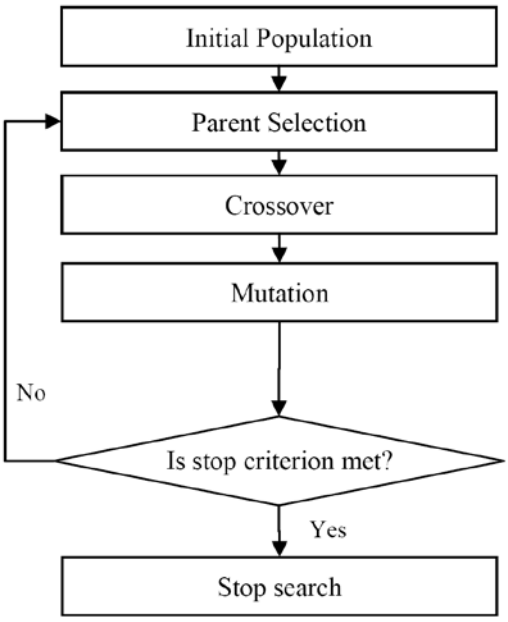

Fig.1. Genetic Algorithm mechanism

It has been proved that, the characteristics of the GA such as crossover, mutation and penalty function as well as the

selection mechanisms has a major impact on the quality of the provided solutions. For implementing the GA, we need to determine following essential concepts:

A. Chromosome representation

A chromosome shows the structure of the solution. The considered chromosome for each solution is depicted in Fig.2 in which the value of each gene is binary that takes 1 when product *i* is allocated to shelf *k*, otherwise 0.

Fig.2. Chromosome structure

B. Crossover and Mutation

In the Genetic algorithm, crossover and mutation operators are used to discover the unknown regions of the feasible space. Primarily, the crossover operator is considered as an exploitation mechanism while mutation is used for exploration of the feasible space. As shown in Fig.3, the Position Based Crossover is used. A set of positions from first parent is selected at random. These values are copied on the same situations of offspring and remained ones are fulfilled by the same genes of the second parent. Other offspring is generated in the same manner.

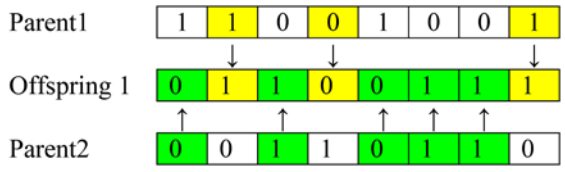

Fig.3. Crossover operator

Mutation operator is organized in a way that one position is selected at random and its value is flip flopped. Its mechanism is depicted in Fig. 4.

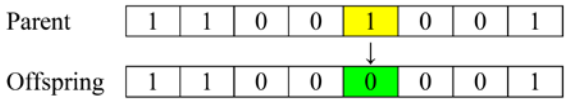

Fig.4. Mutation operator

C. Selection strategies

Here we use two selection strategies, roulette wheel for selecting parents to generate offspring and elitism to select next generation from the offspring and parents as survivor selection mechanism. Interested readers are referred to [19] for more information about roulette wheel and elitism mechanisms.

D. Feasibility checking and fitness evaluation

In the optimization problems with discrete solution space, generating feasible solutions is a major concern. There are two approaches to deal with this problem. The first is searching the feasible space via generating feasible solutions that is a time-consuming way and also may be not lead to effective exploitation and exploration. Other approach could be use of the penalty function. In this way all produced offspring either feasible or infeasible are accepted and the penalty value is assigned to infeasible solutions based on their infeasibility degree. The infeasibility degree is computed based on the proportion of the violated constraints. So the fitness value of the each offspring is included objective function and the penalty values.

In the next section, an illustrative numerical example is presented to clarify the developed model and proposed solution approach.

IV. Numerical Example

Here, an example of market basket data is simulated to describe the proposed model and GAs based solution approach. To do this, ten goods are considered that must be allocated into three shelves. Based on the shelves position each shelf has a different impact on the selling possibility of the allocated goods. These selling possibilities presented by an expert. These values are presented in the Table 1.

Table 1. The selling possibility of each goods

Goods / Shelf	1	2	3	4	5	6	7	8	9	10
1	0.8	0.6	0.1	0.9	0.1	0.1	0.5	0.1	0.2	0.1
2	0.5	0.2	0.5	0.4	0.5	0.3	0.5	0.1	0.3	0.8
3	0.9	0.5	0.4	0.1	0.9	0.7	0.5	0.7	0.8	0.1

The other characteristic that has major impact on the allocating products is the selling benefit. So, it is logical that for maximizing the expected benefit of the selling, the products with the higher benefits must be allocated to shelf so those have higher selling possibilities. Table 2 shows the values of the products' benefit.

Table 2. The benefit of each product ($/unit)

Goods	1	2	3	4	5	6	7	8	9	10
Benefit	40	15	70	20	15	25	10	10	22	5

For simulated data, the values of confidence and support are obtained as Tables 3 and 4 respectively.

Table 3. The confidence values for simulated data

Goods	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	0.67	0.42	0.44	0.36	0.28	0.28	0.47	0.31	0.33
2	0.51	1	0.38	0.38	0.38	0.28	0.23	0.47	0.3	0.3
3	0.33	0.39	1	0.48	0.43	0.26	0.3	0.46	0.37	0.3
4	0.36	0.4	0.49	1	0.36	0.31	0.29	0.53	0.4	0.27
5	0.36	0.5	0.56	0.44	1	0.22	0.25	0.44	0.39	0.36
6	0.28	0.36	0.33	0.39	0.22	1	0.56	0.53	0.28	0.28
7	0.32	0.35	0.45	0.42	0.29	0.65	1	0.39	0.23	0.32
8	0.35	0.45	0.43	0.49	0.33	0.39	0.24	1	0.47	0.33
9	0.35	0.45	0.55	0.58	0.45	0.32	0.23	0.74	1	0.29
10	0.4	0.47	0.47	0.4	0.43	0.33	0.33	0.53	0.3	1

Table 4. The support values for simulated data

Goods	1	2	3	4	5	6	7	8	9	10
1	0.36	0.24	0.15	0.16	0.13	0.1	0.1	0.17	0.11	0.12
2	0	0.47	0.18	0.18	0.18	0.13	0.11	0.22	0.14	0.14
3	0	0	0.46	0.22	0.2	0.12	0.14	0.21	0.17	0.14
4	0	0	0	0.45	0.16	0.14	0.13	0.24	0.18	0.12
5	0	0	0	0	0.36	0.08	0.09	0.16	0.14	0.13
6	0	0	0	0	0	0.36	0.2	0.19	0.1	0.1
7	0	0	0	0	0	0	0.31	0.12	0.07	0.1
8	0	0	0	0	0	0	0	0.49	0.23	0.16
9	0	0	0	0	0	0	0	0	0.31	0.09
10	0	0	0	0	0	0	0	0	0	0.3

The capacity of each shelf (U_k) are presented in Table 5.

Table 5. The capacity of shelf k (U_k)

Shelf	1	2	3
Capacity	2	4	4

Based on the above information, the market manager decides to maximize his preference function. The problem is formulated using developed model and to solve it described GAs is used with the following parameters. The rate of mutation and crossover operators are 0.6 and 0.4 respectively. Pop-size is considered 500 and one percent of pool including parents and offspring are selected as elites to pass the next generation directly. The GAs is run with maximum 100 generations as a termination criterion. Genetic algorithm is run with 1000 cycles and the best solution is reported in table 6 as optimal allocation.

Table 6. The optimum solution

Goods / Shelf	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	1	0	0	0	1
2	0	0	1	0	1	0	0	1	0	0
3	1	1	0	1	0	0	1	0	0	0

For the above solution, the objective value is obtained as 306.9.

V. Conclusion

In this paper a new model was developed to formulate the market basket analysis with allocation considerations. Beside the support and confidence as two important factors of market basket analysis, the products' location was regarded as another parameter which affects the selling rate. The formulated non-linear zero-one programming was solved using genetic algorithm based solution approach. Because of the realistic assumptions, the obtained solutions seem are more suitable for real world problems. For the future research, interested readers are advised to extend the proposed model to the uncertain environments for efficient modeling of the real world problems.

References

[1] O.L. Mangasarian, "Linear and nonlinear separation of patterns by linear programming". *Operations Research* vol. 13, 1965, pp. 444–452

[2] H.D. Vinod, "Integer programming and the theory of grouping" *Journal of the American Statistical association*, vol. 64, 1964, pp. 506–519.

[3] M.R. Rao, "Cluster analysis and mathematical programming". *Journal of the American Statistical Association* vol. 66, 1971, pp. 622–626.

[4] B. Padmanabhan, A. Tuzhilin, "On the use of optimization for data mining: Theoretical interactions and eCRM opportunities", *Management Science*, vol.49- 10, 2003, pp.1327–1343.

[5] P.S. Bradley, U.M. Fayyad, O.L. Mangasarian, "Mathematical programming for data mining: Formulations and challenges". *INFORMS Journal on Computing*, vol. 11, 1999, pp. 217–238.

[6] S. Olafsson, X. Li, S. Wu, "Operations research and data mining", *European Journal of Operational Research*, vol.187, 2008, pp. 1429–1448.

[7] S. Kotsiantis, D. Kanellopoulos, "Association Rules Mining: A Recent Overview GESTS", *International Transactions on Computer Science and Engineering*, vol. 32:1, 2006, pp.71-82.

[8] M. Hegland, "Algorithms for Association Rules", *Lecture Notes in Computer Science*, vol. 2600, 2003, pp. 226 – 234.

[9] K. Tang, Y. Chen, H. Hu, "Context-based market basket analysis in a multiple-store environment", *Decision Support Systems*, vol. 45, 2008, pp. 150–163.

[10] Y. Chen, K. Tang, R. Shen, Y. Hu, "Market basket analysis in a multiple store environment", *Decision Support Systems*, vol. 40, 2005 , pp. 339– 354.

[11] C. Yun, K. Chuang, M. Chen, "Adherence clustering: an efficient method for mining market-basket clusters", *Information Systems*, vol. 31, 2006 , pp.170–186.

[12] L. Cavique, "A scalable algorithm for the market basket analysis", *Journal of Retailing and Consumer Services*, vol. 14, 2007, pp. 400–407.

[13] R.A. Russell, T.L. Urban, "The location and allocation of products and product families on retail shelves", *Annuals of Operation Research*, vol. 179:1, 2010, pp. 131-147.

[14] E. Nierop, D. Fok, P. Franses, "Interaction between shelf layout and marketing effectiveness and its impact on optimizing shelf arrangements", *Marketing Science*, vol. 27:6, 2008 , pp.1065 – 1082.

[15] R. Saraf and S. Patil, "Market-Basket Analysis using Agglomerative Hierarchical approach for clustering a retail items", *International Journal of Computer Science and Network Security*, vol. 16:3, 2016, pp. 47-56.

[16] V. Badhe, P. Richharia, "Survey on Association Rule Mining for Finding Frequent Item Pattern", *International Journal of Scientific Research in Science, Engineering and Technolog*, vol. 2:2, 2016, pp. 1349-1355

[17] N. Tomar, A.K. Manjhvar, "A Survey on Data Mining Optimization Techniques", *International Journal of Science Technology & Engineering*, vol. 2:6, 2015, pp. 130-133.

[18] S. Shrivastava, V. Rajput, "Evolutionary Algorithm based Association Rule Mining: A Brief Survey", *International Journal of Inovation in Engineering research and Management*, vol. 2:1, 2015, pp. 1-7.

[19] M. Mitchell, "An introduction to genetic algorithm", Prentice Hall of India, 2002.

Dr. Majeed Heydari is an Assistant Professor in Department of Mechanical and Industrial Engineering at University of Zanjan, Iran. He received his BS as a top undergraduate student in Applied Mathematics from University of Guilan. He also received his MS and PhD (with honors) in Industrial Engineering from Iran University of Science and Technology. His research interests include Risk and Reliability Engineering, Warranty Modeling and Analysis, Statistical Quality Control and Decision Making.

Dr. Amir Yousefli is an Assistant Professor in Department of Industrial Management at Imam Khomeini International University of Qazvin, Iran. He received his BS in Industrial Engineering from University of Kurdistan. He also received his MS and PhD (with honors) in Industrial Engineering from Iran University of Science and Technology. His research interests include Operation Research and more specifically Uncertain Programming and Management, Fuzzy modeling and analysis as well as Decision making.