Clinical Cancer Investigation Journal-ISSN: 2278-0513

Improving customer relationship management through data mining based on the GMDH algorithm

Abstract

Effective customer relationship management has become a major challenge in business competition. Stores need information about who their customers are, what their expectations and needs are, and how their needs should be met. One of the problems of grocery stores in previous years was the lack of sufficient information of managers about the preferences and shopping carts of customers and, thus, the lack of goods needed by customers in the warehouse and improper layout of stores to provide customers with the goods they need. The present study aims to analyze the data of store customers using neural networks and obtain sufficient information from the customer's shopping cart in the grocery store. The results showed that good clustering enables companies to increase their relationship with customers and increase their sales. Also, the layout of each store can be defined according to preferences for the goods needed by customers who visit the store more so that increased comfort leads to increasing their purchases. Prominent goods in each cluster, business units (goods groups) with a share of goods at least twice as high as the mean share in the sample can be distinguished.

Keywords: *Customer Relationship Management, Data Mining, GMDH Algorithm, Data analysis, Methods.*

First author(Responsible) : **Mohammad Davoodi***

Master of Computer Science, majoring in Software, Roshd Danesh University, Semnan, Iran.

Mohamad.dawoodi1373@gmail.com

Second Author: **Hamed Davoodi**

M.Sc. in Software Orientation from Roshd Danesh University, Semnan, Iran

Introduction

Environmental dynamics and increasing competition have led stores to become competitive to operate, gain customer satisfaction, and gain new customers. The basis of all store activities, especially marketing activities, is finding customers, retaining profitable customers, designing and presenting the desired values for customers, and creating value for them. Given what was stated, customer relationship management in stores is a kind of strategy [1]. Customer satisfaction is becoming the main goal of organizations and managers are well aware that achieving the goals of the company depends on the satisfaction of customers. This competitive advantage is achieved when an effective relationship can be established with its customers. This goal is achieved when the behavior of potential and actual consumers can be analyzed from different dimensions so that organizations can use a customer-centric attitude in society [2]. Organizations that seek customer satisfaction and loyalty need to interact more with them. Information technology and the use of various databases for the acquisition, storage, and dissemination of information and knowledge of customers, provide the conditions for organizations and companies to be aware of the needs and interests of their customers more quickly [3].

Data mining can predict the profitability of potential customers who can become actual customers, how long they will remain loyal customers and how they may leave us. Some customers are constantly shifting towards companies to take advantage of the competition created between them. In this case, companies can focus on more profitable customers. Thus, through data mining, the value of customers can be determined, their future behavior can be predicted and informed decisions can be made in this regard [4]. Chang et al. (2014) believe that customer relationship development results in customer loyalty. Customer relationship management leads to improved efficiency and effectiveness in customer service, integration of customer communication channels, an increase in new business opportunities, improvement of competitive capabilities and performance, customer classification (profitable and non-profitable), etc. [5].

According to Battle [6], one of the major methods of calculating customer lifetime value that has been used in many studies is called RFM. RFM calculation is an effective way to evaluate the customer's lifetime value and includes elements of exchange novelty and exchange volume, number of exchange iterations over a period of time, and volume of exchange. In this method, different weights should be assigned to RFM variables according to the different types of industries. Rinatz & Kumar [7] used another method called SOW as the basis for calculating customer value. It is the ratio of the sales of a particular product by the organization to the total customer purchase of the same product in the entire market and over a period of time. In other words, this method considers the extent to which customer needs are met in the organization as a measure.

Abdu, Pointon, and Al-Masri [8] examined the customer data of one of the Egyptian banks from which they had borrowed. The algorithms used in this paper were probabilistic and multilayer neural networks and logit regression. In the mentioned study, the performance of probabilistic and multilayer neural networks was compared with conventional methods such as discriminant analysis, logical regression, and analysis based on the minimum deviation from the average.

The results showed that the neural network has the best performance with 96% accuracy. Mohammadi and Sheikh [9] predicted customer behavior based on the rough set theory. They categorized new customers by examining the behavior of current customers. In other words, in this study, the Rough rule was used to segment new customers. The researchers categorized them into similar categories based on the response that customers gave to two questionnaires based on logical rules (if ... then). Then, they adapted the new customer behavior to this model and predicted the closest category to the customer.

Sometimes access to information in a system is difficult or even impossible, so the use of "univariate time series" models can be an effective factor in achieving a solution to this limitation. Time series prediction based on past perception has been done mostly in the form of using the ARIMA model and its sub-branches (AR, MA, ARMA, and SARIMA) with the initial assumption of series linearity and specific distribution when solving the problem and for non-static data. The main problem with this system is the assumption that the series model is linear, so the researchers turned to a combination of artificial intelligence methods such as neural networks and optimization algorithms. The main goal of this dissertation is to gain appropriate and correct knowledge of customer information and improve the relationship between managers and customers. In this study, customer information is first received and this information is analyzed through one of the neural network models.

After analysis, by using the knowledge gained, we can well identify customers' shopping carts and categorize the customers into several different categories. The gained knowledge can be used to buy more and less acceptable goods for customers and can also be used in warehousing and better store layout to make it easier for customers to find the items needed in their shopping cart. This dissertation also tried to use data mining and extract knowledge from customer information to help meet the needs of customers and ultimately increase their satisfaction with a store. Customer satisfaction leads to store loyalty, which ultimately leads to more profits for stores and improves the relationship between store managers and customers.

Suggested method

This plan aims to differentiate customers based on their purchasing methods. To achieve this goal, a default for purchases based on the set of products that are most commonly purchased is set, and then shopping carts are grouped using segmented cluster analysis and products. Customers can be separated in two ways. In the first method, customers are classified based on the number of purchases in 8 weeks and the mean money spent on each purchase, based on the Ballet model (1995) entitled "Frequency and monetary value". Another method is based on customer interest, which is the grouping factor for the purchased product category. It means that the clustering algorithm will be executed using the similarity of the purchased products and their classification in predefined portfolios.

GMDH neural network

The Group Method of Data Handling (GMDH) was introduced in 1968 by Ivakhnenko as a linear modeling and regression method that works with iterative, incremental algorithms and the use of natural selection patterns such as evolutionary algorithms. This structure is also known as a polynomial neural network and its main basis is the quadratic polynomial model and the least-squares error algorithm.

Network design

The process of creating a system consists of two steps: creating a diverse set of basic models and combining the output produced by these models using the appropriate criteria. Based on the basic model, there are two types of criteria. The first is the criterion that can be represented by the most probabilistic method. In this method, a single learning algorithm is applied to different subsets of training data. Two of the best-known models are convenience sampling by placement and re-weighting incorrectly classified training data. The second method of creating a group system is to integrate the outputs produced by the base models. The idea behind the development of group systems is to take advantage of all the unique features of the constituent models to capture the various patterns that exist in the database. Using successful generations of partial equations, the GMDH algorithm provides an ideal format of the model that is considered quadratic polynomials with two inputs. The transfer function of each equation is obtained using the following formula:

$$Y = a + bU + cV + d(UV) + e(U^2) + f(V^2) \quad (1)$$

This analysis is a systematic process for overcoming statistical shortcomings and weaknesses and neural networks. Modeling of the GMDH algorithm has two principles.

(1) Decomposition of complex systems with m input variables into several component systems.

(2) Integrating two-component systems with ideal accuracy.

However, for modeling partial systems, some methods used are:

1- SNE method: In this method, similar to the inverse of the real matrix A, A * is the unique matrix, which is calculated according to the following equation:

$$A^* = (A^T A)^{-1} A^T \quad (2)$$

2- Single value decomposition method (SVD): In this method, the matrix A with M rows and N columns, which has more than

or equal to the number of columns, can be written as a matrix multiplied by U in W and the output of a square matrix written in that matrix U is an orthogonal column matrix with dimensions M * N and W is a diagonal matrix with positive or zero values on its diameter. Hence, the inverse matrix of A will be:

$$A^{-1} = V.\left[diag(1/W_j)\right].\ U^T \quad (3)$$

However, the design of this network has three main methods:
Increased selection pressure: In this method, the number of layers, as well as the number of neurons in each layer, is determined completely automatically and optimally. Therefore, the model algorithm operates in such a way that for each layer, we determine the base error o$\overline{r^2}$, which is the mean of the errors $\overline{r_j^2}$ of the definite number V of the best neurons in terms of error. Then, we compare the base error $\overline{r^2}$ with the error $\overline{r_j^2}$ of each neuron. If $\overline{r_j^2} \leq \overline{r^2}$, the corresponding neuron is selected as the winning neuron and remains in the main network structure, otherwise, it is removed from the structure as a dead neuron.
2-Predefined structure: In this method, the main parameters of the network structure, which includes the number of layers and the number of neurons, are determined directly and completely optionally and without any constraints. The desired design is such that repeatedly selecting the parameters and forming different structures provide the basis for proper identification and optimal modeling. In principle, the performance of this method is somewhat similar to the trial and error, through which the ideal structure is identified.
Evolutionary design: In the third method, a genetic algorithm is used for converging neural networks. In the evolutionary design method, the constraint due to the placement of the error as a criterion for determining the structure of the network is removed and all neurons are given an equal chance to participate in the formation of the neural network. In fact, there are no constraints on creating a network and it is a random and purposeful process to find the most optimal or ideal structure. The only fit criterion for selection can be two parameters of the number of neurons in the whole network and also the rate of network output error compared to the value tested. It means that we ultimately seek to minimize network output error for the least number of neurons used.
By reviewing the first method, we observe that this method is based on errors in each layer. In creating the network structure and deciding to select neurons in each layer, the same layer is formed at the moment of formation of the same layer. In this method, the alpha parameter directly affects the selection pressure and changes the network structure by changing the alpha, but the alpha sensitivity works to a certain extent and this sensitivity varies depending on the type of input data. In other words, alpha has the same structure in an interval, and after leaving this interval, the network structure changes, but these changes are very large. For example, $\alpha_1$creates a 3-layer structure. α_2 creates a six-layer structure. In this regard, 4- and 5-layer networks are not considered at all. In the third method, there is no monitoring of the selection of neurons based on the error obtained from modeling, and the important issue is the modeling of each network. This method allows all neurons to participate in the modeling. The result of a neuron that was not accepted by the first and second methods might be better in connection with another neuron, which this chance in the third method is given to all neurons.

Training algorithm

To calculate the estimated output, a PD equation is formed for each input pair. Model parameters are obtained from the minimum error of training data. In addition, we select the best model to form the first layer. Finally, we create new PDs from the median variables (Z_ms) located in the new iteration. Now, a pair of new input variables is received and operations are performed on them until a stop criterion is obtained. Once the final layer is created, the node with the best performance is considered as the output. The remaining nodes of that layer are removed. This operation is performed in the previous layers up to the first layer.

Execution

1-First, the input variables are normalized and divided into input and output variables
2-Data should be divided into two categories: testing and training.
3- Creating PNN structure: This structure will be selected based on the number of inputs and the order of PD of each layer.
4-Determining the number of input variables and the order of polynomials making up PD: The number of partial equations of each layer varies depending on the number of input variables selected from the nodes of the previous layer. As a result, the number of nodes is N! / (N-r)! r! that r is the number of input variables selected.
5- Estimation of PD constants calculated using the Equation (4):

$$\begin{aligned} C_i &= (X_i^T X_i)^{-1} X_i^T Y \\ X_i &= [X_{!i}, X_{2i}, \dots X_{ki} \dots X_{n_{tr}i}]^T \\ X_{ki} &= [1 X_{ki1} X_{ki2} X_{ki1} X_{ki2} X_{ki1}^2 X_{ki2}^2] \\ C_i &= [C_0 C_1 C_2 C_3 C_4 C_{n'}] \end{aligned} \quad (4)$$

Where i is the number of nodes, k is the number of data, ntr is the number of training data, n is the number of selected inputs, m is the maximum degree, and n' is the estimated number of constants.
6-Selecting the PD with the best performance: After evaluating each PD by training and testing data, we compare all the items

and select the PDs that have the best performance. Usually, a predetermined number of Ws will be selected from these PDs.
7-Checking the stop condition: the algorithm will stop,
A. If the error of the new step is greater than the error of the previous step.
B- If the number of layers reaches the specified number in the algorithm.
Selecting new variables for the next layer and repeating steps 4 to 7 until the stop condition is reached: The inputs of each layer will always be the outputs of the previous layer.
The MATLAB software algorithm is used to run the programs on 2GHz processors.

Results

To perform the test, a five-layer network with different input conditions was tested in 4 different input modes. In each case, a set of inputs (product groups) were given to the network and due to the number of iterations in the customer's shopping cart, it was given to the network to predict the purchase procedure or classification of this group of goods.
Experiment parameters are checked in table 1 and you can see the minimum error and in table 3 Measures for four experiments it is shown that the experiment is shown in 4 stages. You can see the results of the variance tests shown in Table 4 Analysis of variance test and Table 5 General rating of the algorithm according to five measures is shown, which you can see below these tables.

Table 1: Experiment parameters

Layer	Number of neurons	Least error
First experiment		
1	10	0.03
2	20	0.023
3	20	0.021
4	20	0.02
5	1	0.019
Second experiment		
1	10	0.123
2	30	0.122
3	30	0.11
4	30	0.1
5	1	0.093
Third experiment		
1	10	21.12
2	10	19.82
3	10	19.2
4	10	19.12
5	1	18.98
Fourth experiment		
1	5	15
2	5	14
3	5	12.82
4	5	12.38
5	1	12.62

Table 2- an error rate of experiments

	STD	MEAN	RMSE	MSE
First experiment				
Training	0.019	-0.00013	0.019	0.00039
Testing	0.092	-0.0095	0.093	0.0087

Total data	0.053	-0.003	0.054	0.0029
Second experiment				
Training	0.098	-0.00127	0.098	0.0096
Testing	0.094	0.0213	0.094	0.0088
Total data	0.098	0.00095	0.0975	0.0095
Third experiment				
Training	20.8	-0.32	20.8	431.5
Testing	22.02	0.6	22	483.5
Total data	21	-0.23	20.9	436.8
Fourth experiment				
Training	16.07	0.66	16.02	256.8
Testing	15.8	1.88	15.4	235.6
Total data	16	0.77	16	254.7

Results of accuracy estimation

After comparing the results of different GMDH network test cases, the best result was extracted. The results of Table 2 show that the artificial neural network model based on a genetic algorithm has a good prediction criterion for grouping goods.

Table 2: Error rates for the two GMDH models

Model	MSE	RMSE
GMDH	0.0017	0.04

Table 3: Measures for four experiments

Experiment	Measures	Problem number				
		1	2	3	4	5
		Number of goods				
		5	10	35	80	315
1	distance MID variety NOS time	2.10E+06 2.26E-03 3.13E+07 6 15.13	4.59E+06 7.68E-04 3.14E+07 9 31.81	3.19E+06 2.77E-02 4.14E+08 3 61.88	5.69E+07 4.13E-02 8.44E+09 7 114.74	2.71E+06 1.01E+00 5.49E+09 10 198.71
2	distance MID variety NOS time	6.84E+06 6.55E-04 6.18E+07 22 28.71	1.22E+07 1.98E-03 4.19E+08 21 52.72	3.22E+07 8.13E-04 8.92E+09 22 121.17	1.32E+08 9.98E-02 3.24E+10 24 179.61	1.21E+07 4.44E-03 4.81E+09 19 301.61

3	distance MID Variety NOS time	4.40E+06 3.13E-04 4.39E+07 31 31.71	9.19E+06 6.22E-01 4.93E+07 27 61.51	7.34E+06 2.81E-02 4.93E+08 25 119.63	7.79E+07 5.67E-01 4.24E+10 33 181.61	1.39E+06 6.79E-02 2.44E+08 20 229.81
4	distance MID Variety NOS time	9.12E+06 6.62E-02 9.82E+07 21 26.71	3.89E+07 2.83E-02 4.93E+07 20 44.71	7.91E+07 6.90E-03 4.03E+09 23 89.61	3.83E+08 9.07E-01 8.22E+10 22 167.91	3.42E+07 2.73E-02 4.68E+09 17 283.81

To examine the results of different GMDH network test cases more accurately, five effective factors (distance to the solution, variety of goods, execution time, NOS or number of solutions, and MID or mean ideal distance) were used to analyze the results. The results of the following table show that the artificial neural network model based on a genetic algorithm has a good prediction measure for classifying the goods.

Table 4- Analysis of variance test

Name of measures	-value P	Test results
Variety	0.9	The hypothesis is not rejected
MID	0.35	The hypothesis is not rejected
distance	0.3	The hypothesis is not rejected
NOS	0	The hypothesis is rejected
Time	0.57	The hypothesis is not rejected

Table 5 - General rating of the algorithm according to five measures

	Experiment			
measure	First	Second	Third	Fourth
Variety	4	2	3	1
MID	2	1	3	4
Distance	1	3	2	4
NOS	4	2	1	3
time CPU	1	3	4	2

Execution cost analysis

Figure (1) shows the cost of executing the proposed algorithm. This figure shows that the proposed algorithm is in good condition in terms of the number of iterations to achieve the ideal solution and requires less than 50 iterations of the algorithm.

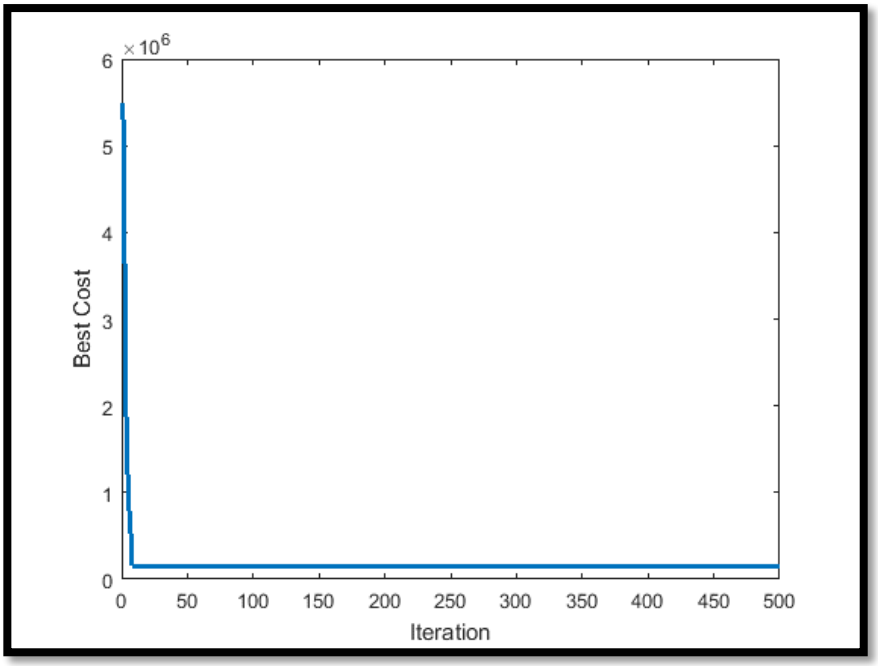


Figure 1- Cost of the proposed algorithm in 500 iterations

Analysis of clustering results

Finally, a time period was analyzed from the database of a large shopping center. A database containing a two-month purchase history that was integrated and its noise has been removed was used. Out of 5265 databases of customers, only those who have purchased at least 5 different goods in the two months were used. We randomly selected 1000 customers from the sample to perform clustering, and to facilitate the work, out of 36581 purchases, we examined only the goods that were provided by at least 1000 customers (315 goods).

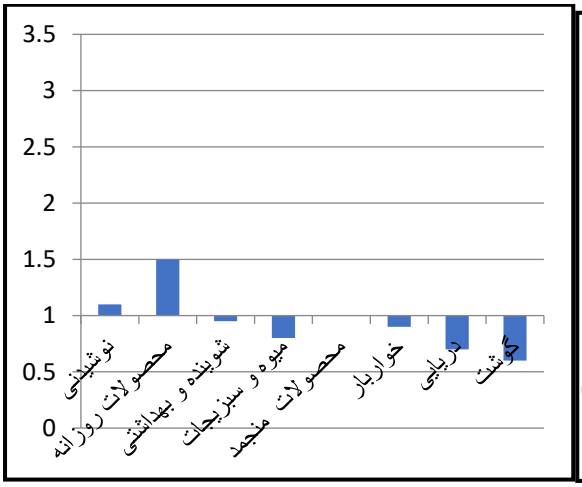
Table 6- Number of goods in each cluster

Category	Number of goods	Percentage of goods
Category 1	36	11.43
Category 2	35	11.12
Category 3	93	29.52
Category 4	28	8.88
Category 5	75	23.82
Category 6	48	15.23
Total	315	100

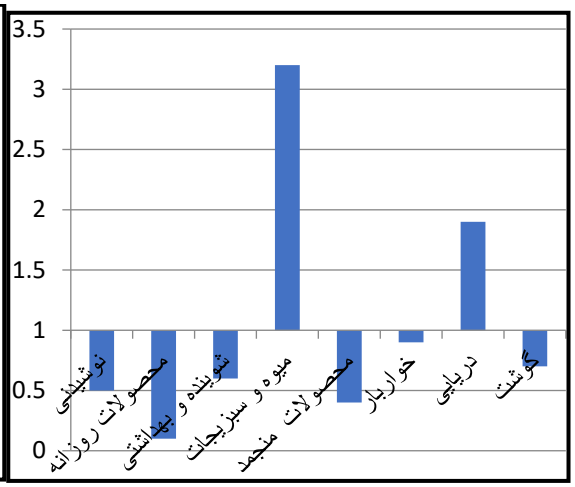
The distribution of 315 goods used for cluster analysis is shown in Table (6). Each good is assigned to one of 6 designated clusters. To discover the purchasing patterns, the ratio of the share of goods in each cluster belonging to each business unit and their mean share in that business unit in the studied sample was calculated. The results are shown in Figure 2.

نوشیدنی: drinks
محصولات روزانه: daily products
شوینده و بهداشتی: detergents and health products
میوه و سبزیجات fruits and vegetables
محصولات منجمد frozen products
خواربار foodstuff

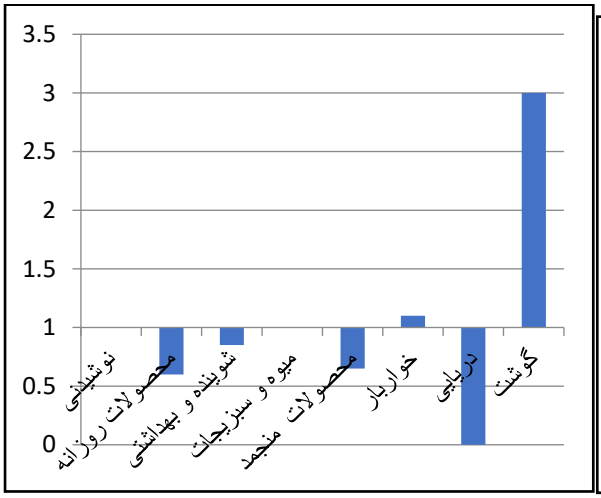
دریایی seafood
گوشت meat



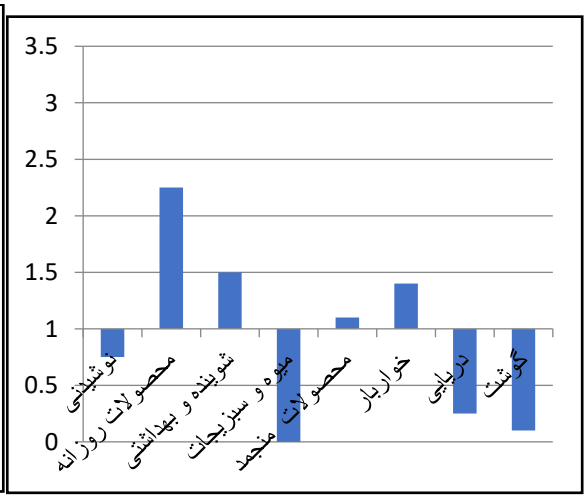
(A)



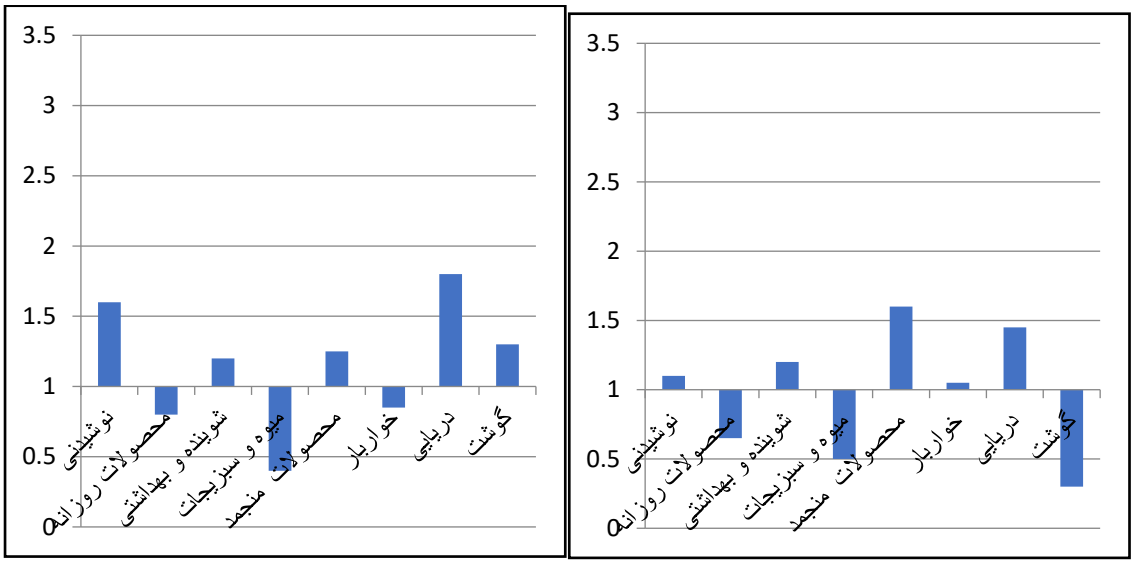
(B)



(C)



(D)

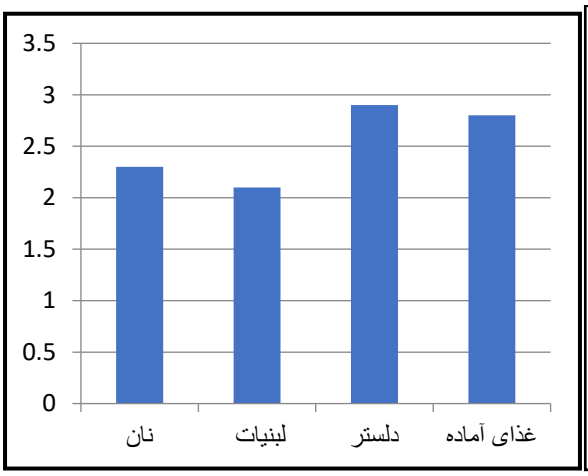


(E) (F)

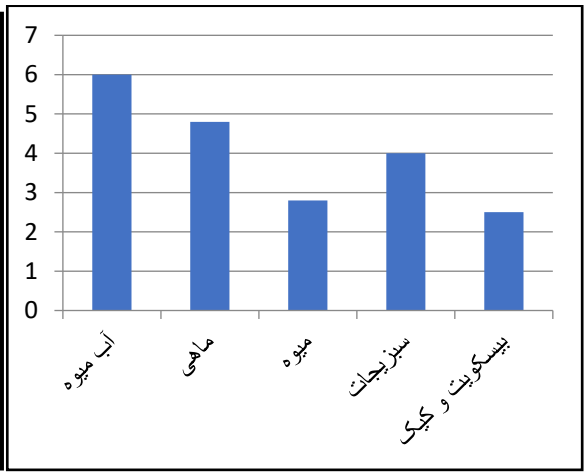
Figure 2- Ratio of goods share and mean share in cluster 1, b) ratio of goods share and mean share in cluster 2, c) Ratio of goods share and mean share in cluster 3, d) Ratio of goods share and mean share in cluster 4, Ratio of goods share and mean share in cluster 5, c) Commodity Ratio of goods share and mean share in cluster 6

To determine the important goods, we first separated the business units (goods groups) in each cluster with the share of goods at least twice as high as the mean share in the sample

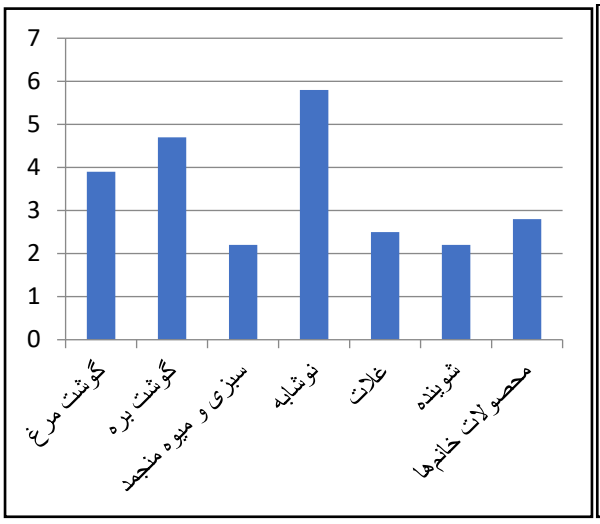
ابمیوه Fruit juice
ماهی Fish
میوه fruit
سبزیجات vegetables
بیسکویت و کیک biscuit and cake
غذای اماده prepared food
دلستر non-alcoholic beer
لبنیات dairy
نان bread
عسل و مربا honey and jam
تخم مرغ: egg
کنسروها canned foods
ادویه spices
سوسیس sausage
غذای منجمد frozen meat
گوشت گوساله veal
محصولات خانم ها female products
شوینده Detergent
غلات cereals
نوشابه soft drinks
سبزی و میوه منجمد frozen fruit and vegetables
گوشت بره lamb
گوشت مرغ: chicken meat
روغن oil
غذای بچه baby food
گوشت منجد frozen meat
کباب barbecue
جوجه کباب chicken barbecue
محصولات اقایان male products
محصولات بهداشتی health products



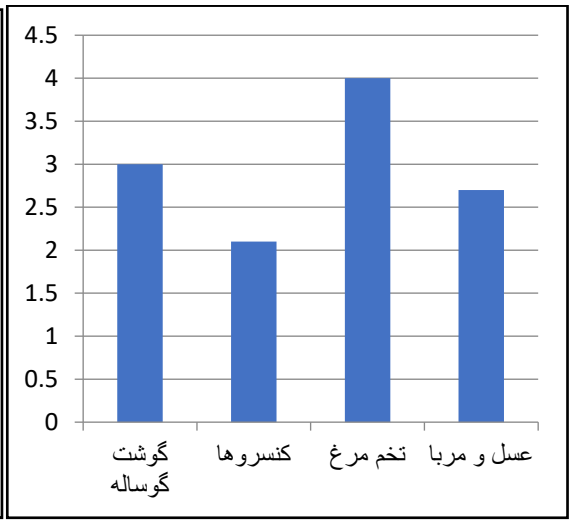
A



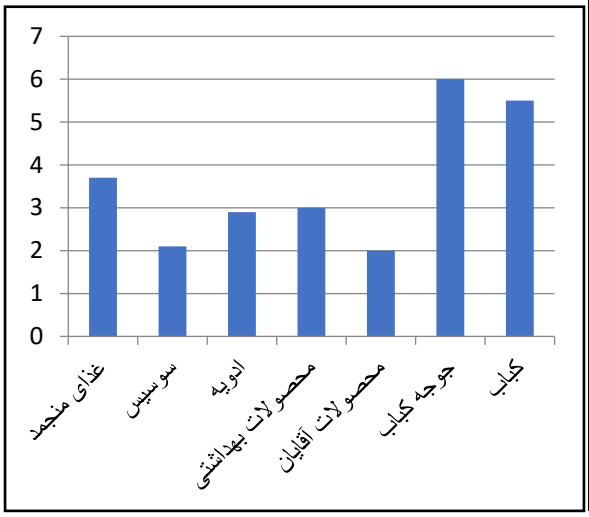
B



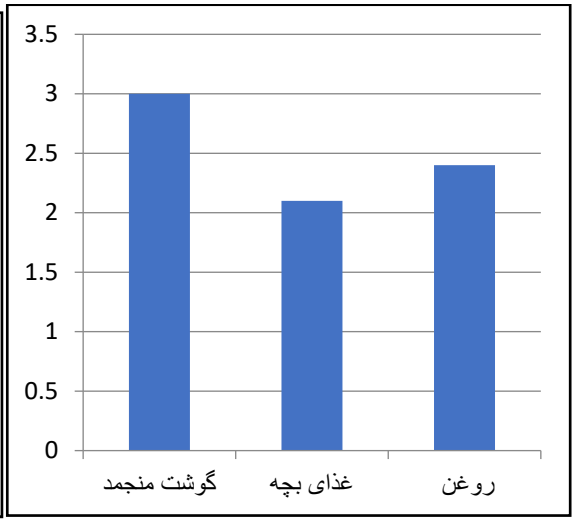
C



D



E



F

Figure 3- Share of goods in the main category of cluster 1, b) Share of goods in the main category of cluster 2, c) Share of goods in the main category of cluster 3, d) Share of goods in the main category of cluster 4, e) Share of goods in the main category of cluster 5, c) Share of goods in the main category of cluster 6

Finally, Figure 3 shows the most important goods for each cluster, each of which provides a stylistic representation of the purchase. The way of classification of customers in clusters is presented in Table (7)

Table 7 - Distribution of customers by clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
Percentage of customers	24.2	32.1	14.2	18.5	3.8	7.2	100

Good clustering enables companies to enhance their relationship with customers and increase their sales. Accordingly, the layout of each store can be defined according to preferences for the goods needed by customers who visit the store more often, so that increasing comfort will lead to an increase in their purchases.

Conclusion

The present study aimed to enhance customer relationship management through data mining based on the GMDH algorithm. This study provided a way to categorize customers based on purchased goods and based clustering techniques, to achieve layout and inventory per the purchase demand. Recognizing customer differences can be the key to successful marketing, as it can lead to more effective customer satisfaction. This methodology will include inference of customers according to each cluster of purchased goods by analyzing the type of goods in each cluster. The study is also very effective in setting the range of stores goods and products and the layout of stores, which can help enhance the relationship between companies and customers.

The results revealed that good clustering enables companies to increase their relationship with customers and increase their sales. Also, the layout of each store can be defined according to preference and for the goods needed by customers who visit the store more often, so increasing comfort will lead to increasing their purchases. Important goods in each cluster, business units (goods groups) with a share of goods at least twice as high as the mean share in the sample can be distinguished. The model can be developed by considering the layout of the framework in the warehouse. Other quasi-revelatory methods with good convergence and a variety of solutions can also be considered in multi-purpose optimization in inventory control articles.

Acknowledgments: I appreciate and thank all my professors, friends and family who have supported me throughout my studies, and I dedicate this article to my family who have always supported me in life.

Conflict of Interests: Non

Ethical Considerations: Non

Financial Disclosure: Non

Funding/Support: Non

References

1- Belaghi Inalo, R, Investigating the Impact of Customer Relationship Management Dimensions on Customer Satisfaction,

2-Mohammadi, M, Sohrabi, T, The Impact of Electronic Customer Relationship Management on Customer Satisfaction, Quarterly Journal of Intelligent Business Management Studies, Volume 6, Issue 22, Winter 2017

3-Fazli, S, Rashidi Astaneh, m, The Role of Factors Affecting the Success of Customer Relationship Management Strategy in Car Dealers in Gilan Province, Business Management, Volume 6, Issue, Spring 2014,

4- Vali Mohammadi, S, Shokrizadeh Esfahani, A, and Shafiei, Application of Data Mining in Customer Relationship Management, First National Conference on Industrial and Systems Engineering, December 2012,

5- Bonyadi, A, Naeini, S. Ghodsi, A, Kheibari, N, The Impact of Customer Relationship Management on Organizational Performance, Business Management Perspective, Issue 27, Fall 2016,

6-Zarrin Moghadam, P, Qarekhani, M, Designing a life insurance customer segmentation model based on LRFM model using artificial intelligence,

7- Qarekhani, M. Abolghasemi, M, Data Mining Applications in the Insurance Industry, News of the Insurance World,

8- Ghasemi, S, Improving the evaluation of computer simulation using data mining techniques,

9- Khalili Nejad, M, Minaei Bighdeli, B, Data Mining in Medicine,