Clustering of Customers Based on Shopping Behavior and Employing Genetic Algorithms

Elham Photoohi Bafghi
Department of Computer Engineering
Bafgh branch, Islamic Azad University
Bafgh, Iran

***Abstract*—Clustering of customers is a vital case in marketing and customer relationship management. In traditional marketing, a market seller is categorized based on general characteristics like clients' statistical information and their lifestyle features. However, this method seems unable to cope with today's challenges. In this paper, we present a method for the classification of customers based on variables such as shopping cases and financial information related to the customers' interactions. One measure of similarity was defined as clustering and clustering quality function was further defined. Genetic algorithms been used to ensure the accuracy of clustering.**

Keywords-classification; customers; shopping; genetic algorithm

I. INTRODUCTION

Traditional methods of mass marketing are no longer responsive to customers' needs diversity. Such diversity must be managed by clustering methods which puts the customers with the same need and similar shopping behavior in the same clusters [1]. Using the proper clustering, companies can deliver the customers' goods, services and specific resources via close relationship. Customer clustering is one of the components in modern and successful marketing leading to improved customer relationship management (CRM) is [2]. Clustering is an important issue in selecting the appropriate variables. Clustering variables have two local variables and parameters are based on the product [3]. General variables are included of statistical information (age, gender, etc.), life-style and includes variables based on the buying patterns of customers. Different works have been taken in the field of customer clustering based on variables [4, 5]. In a study described in [6], the SOM (Self-Organizing Map) algorithm and K-Means were used for clustering of customers based on public information. In this method, first SOM helped to determine the number of clusters and then clustering was performed using the K-Means. Clustering based on general variables is more understandable, but the assumption that the customers have the same statistical information (age, etc.) as well as a similar life style and the same shopping habits, to some extent are doubtful. Today, customers can purchase from different parts of the planet, and this makes clustering partly more difficult. On the other hand some information about the local variables may not be provided by the customer. Although information may be available they may vary over time. For example, income, marital status, occupation and the similar cases makes the clustering through the use of public variables more doubtful [7]. In this paper we propose a method of clustering of customers based on information about the goods. Genetic algorithms were also used in order to enhance the quality of these clusters which also have been used to determine the cluster centers.

II. CLUSTERING METHOD BASED ON CUSTOMER PURCHASES

In this section, methods of clustering of customers have been provided by the help of customers' purchase behavior. This method involves pre-processing core, similarity criteria based on purchase, clustering algorithm and the quality function of the cluster.

A. Data pre-processing

The purpose of this selection process is to transform and integrate data from one or more datasets into one data set. Suppose 'I' as an entire collection of items and T^0 a transaction. T^0 transaction includes fields such as client ID, time of transaction, items purchased and financial information. If id_i is the customer ID number of ci customer, and $itemset_i=\{i_{ia}|i_{ia} \in I\}$ are goods purchased by ci, as well as financial information $moneyset_i=\{m_{ia}|a=1,\ldots,||itemset_i||\}$, a record of $t^c_i=(id_i,itemset_i,moneyset_i)$ describes ci customer's behavior based product and which will be displayed.

B. Criteria of similarity based on purchase

After preparing the data, it is necessary to specify the similarity criteria of users or to each other and goods. In this respect, there are two famous criteria of matching coefficient and Jaccard's coefficient [8, 9]. But these two criteria are not most appropriate because:

- Existence of a small number of goods compared to the total of goods in a user set items
- Inequality of goods (considered equal in the above two methods)
- Imbalance in the importance of customers for the organization

So in this article we have introduced a new similarity measure that has overcome some of these shortcomings. Similarity measures presented here considered goods purchased together and the profitability of customers. The criteria of goods purchased together were inspired by the concept of Support provided in [10]. Suppose Supp ({ii, ij)} of transactions involving {ii, ij} is the total transactions:

Suppose Supp({i_i,i_j)} as the ratio of transactions involving {i_i,i_j} to the total transactions:

$$Supp(\{i_i, i_j\}) = \frac{\left\| t^0 \in T^0 \middle| t^0 contains \{i_i, i_j\} \right\|}{\left\| T^0 \right\|} \quad (1)$$

In the above equation i_i,$i_j \in I$. It should be noted that if the items are rarely purchased together, support value has a low amount [11] and to solve the problem we use intimacy criteria as follows:

$$Int(\{i_i, i_j\}) = \frac{Sup(\left|\{i_i, i_j\}\right|)}{Supp(i_i) + Supp(i_j) - Supp(\left|\{i_i, i_j\}\right|)} \quad (2)$$

In the above equation numerical value of Int was between 0 and 1. This solves the problem of backup.
Finally, the proposed formula to measure the similarity of two customers is presented in the following form:

$$sim(c_i.c_j) = \frac{\sum_{a=1}^{s}\sum_{b=1}^{t}\left[m_{ia} \times m_{jb} \times Int\left(\left\{i_{ib}, i_{jb}\right\}\right)\right]}{\sum_{a=1}^{s}\sum_{b=1}^{t}\left[m_a^i \times m_{jb}\right]} \quad (3)$$

C. Clustering algorithm based on purchase

The proposed purchase algorithm of the customers is based on (3). In the algorithm first customers are classified into K clusters. The value of K can be determined by the help of marketers or the method proposed in this paper (see section 4.2). After determining K, initial cluster centers are chosen from T^c records. These centers may also be selected randomly or based on heuristic method (GA). Suppose G={c^n|n=1,…,K} is a total cluster centers where c^n is the center of n^{th} cluster, so that $c^n \in T^c$. So set(T^c-G)={c_i|i=1,…,||T^c-G||} includes customers not selected as the center of the cluster. Then for all customers who are not cluster center, their similarity to all cluster centers are measured and added to the clusters with the maximum similarity to the center. The problem can be seen in the equation below:

$$Max_{1snk}\left\{sim\left(c_i.c^n\right)\right\} where c_i \in \left(T^C - G\right) \quad (4)$$

After each customer was placed in clusters, cluster centers are calculated again. Let c_i and c_j are two customers in the G^n cluster, and the center of Gn cluster is equal to c_n. In this case priority of ci customer is calculated based on following relationship:

$$Pio(c_i) = \frac{\sum_{c_j \in G^m, j \neq i} Sim(c_i, c_j)}{\sum_{c^m \in G, m \neq n} Sim(c_i, c^m)} \quad (5)$$

Where c^m is the center of G^m cluster and $\sum_{c_j \in G^m, j \neq i} Sim(c_i, c_j)$ represents the sum of similarities between the c_i customer and other customers in cluster G^n. While $\sum_{c^m \in G, m \neq n} Sim(c_i, c^m)$ represents the sum of similarities between the client c_i and the other cluster centers except G^n. For all customers existed in a G^n cluster, customer is selected as the center of the cluster with the highest priority value. This concept is presented in the following equation: $c^n \equiv \arg Max_{c_{i \in G^n}}\left\{pio(c_i)\right\}$ (6)

After specifying the new centers, algorithm has done a repetition pattern. The algorithm repeats this process until no clustering center change. The overall process of algorithm is shown in Figure 1.

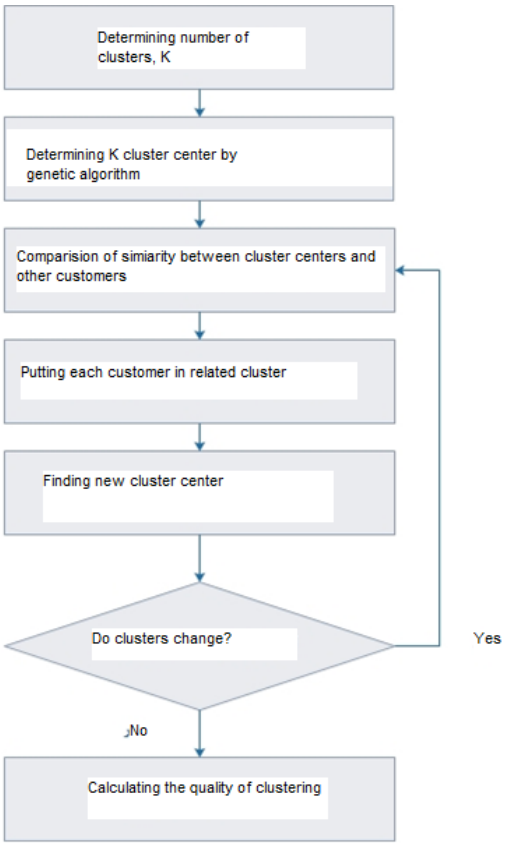

Fig. 1. Flowchart of Clustering using Genetic Algorithms

D. Clustering quality function

The purpose of clustering is to maximize the similarity between the centers of the clusters with the cluster's customers as the similarities between the centers of the clusters are

minimal. So the quality of clustering based on the following equation can be expressed as:

$$\rho(K)=\frac{1}{k}\sum_{n=1}^{k}\left(Min_{1\le m\le k,m\ne n}\left\{\frac{\eta_n+\eta_m}{\eta_{nm}}\right\}\right) \quad (7)$$

$$\eta_n=\frac{1}{\left\|G^n\right\|}\sum_{c_i\in G^n}Sim(c_i,c^n) \quad (8)$$

$$\eta_m=\frac{1}{\left\|G^m\right\|}\sum_{c_j\in G^m}Sim(c_j,c^m) \quad (9)$$

$$\sigma_{nm}=Sim(c^n,c^m) \quad (10)$$

Equation (8) considers η_n as the average similarity between the center of the c^n cluster and existing customers in the cluster G^n. Equation (9) shows that η_m determines the relationship between the average similarity between the center of the cluster c^m and existing customers in the cluster G^m. Equation 10 determines the similarity between c^n and c^m.

Using the clustering quality defined in (7), we can define the appropriate value for K in accordance with the following formula:

$$\hat{k}\equiv\arg Max_{ss'ks't}\{\rho(k)\} \quad (11)$$

In the above equation K value is determined between the lower bound and upper bound of t

III. DETERMINING THE CENTER OF CLUSTERS USING GENETIC ALGORITHM

Genetic Algorithms [12] can be used to select primary centers of clusters. A genetic algorithm is an abstract computational models of biological evolution that is used in optimization problems, in which the mutation operation, crossover and production of new population are used.

A. Encoding the chromosome

Each chromosome in the genetic algorithm represents a solution to the problem that is investigated. In this studied case, each chromosome represents a set of K centers of initial cluster. If f_i is a chromosome, $f_i=[y_1,\ldots,y_j,\ldots,y_k]$ where y_j is the j^{th} gene and K is the total number of genes (In Figure 2).

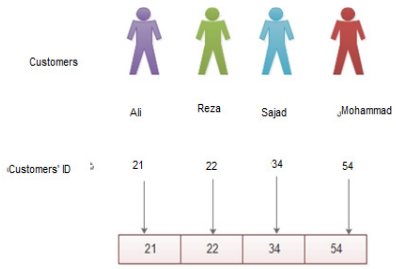

Fig. 2. Coding of chromosomes

B. set values to the initial population

Let P^e represents the population in the e^{th} replication, and 0<=e<=E, where E is the maximum number of times to repeat the genetic algorithm. The number of chromosomes is constant in all occurrences. So $P^e=\{f_i|i=1,\ldots,L\}$ where f_i is the i^{th} chromosome and L is the total number of chromosomes in the population. L number is an even number specified by the user.

C. The fitness value of a chromosome

Fitness value of each chromosome means to determine the suitability of chromosomes in order to survive it. The fitness of each cluster is determined based on the equation 7. So the formula for calculating the fitness of each f_i chromosome is calculated as follows:

$$Fitness(f_i)=\rho(K)\ ,\ f_i=[y_1,\ldots,y_j,\ldots,y_K] \quad (12)$$

In (12) chromosomes existed in the population P^e into two categories: good (P^e_{bad}) and bad (P^e_{good}). Good chromosomes are the sets of chromosomes with high fitness value and bad chromosomes are the sets of chromosomes with low fitness value as their numbers are equal$||P^e_{good}||=||P^e_{bad}||=I/2$.

D. Production of the new population

The purpose of generating new populations is to remove the chromosomes with low fitness value and copy the chromosomes with high fitness value in the new population. So P^e_{bad} is removed and P^e_{good} chromosomes take their place. The higher the fitness value of a chromosome, the higher the probability of selection. The formula for calculating the probability of selection of chromosomes as shown below:

$$\Pr ob(f_i)/\sum_{f_i\in pg^a_{good}}fitness(f_i) \quad (13)$$

The production process of new population has ensured that the new generations are created of the parents with high fitness value. This process is shown in the following figure.

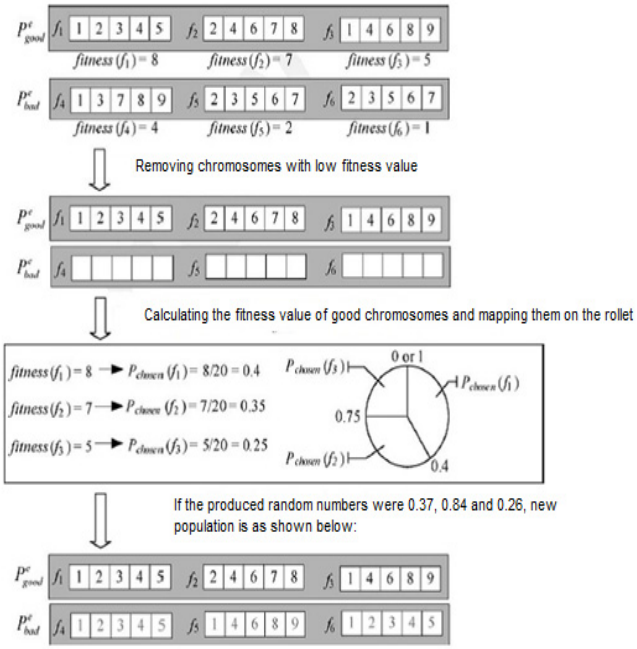

Fig. 3. The process of creating a new population

E. Mutate and crossover

After generating a new population, mutation and crossover operations are applied over the population. First, a f_i chromosome and from P^e_{good} and fj from P^e_{bad} set are elected. If the genes of f_i and f_j are not equal, the crossover operation is applied on them. How to crossover them is shown in Figure 4. If the two genes are equal, then the gene mutation operation is performed on them. For this purpose, created chromosome is placed instead of one of the two chromosomes. This operation is shown in Figure 5.

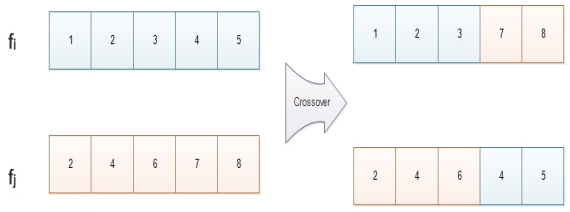

Fig. 4. Crossover operation

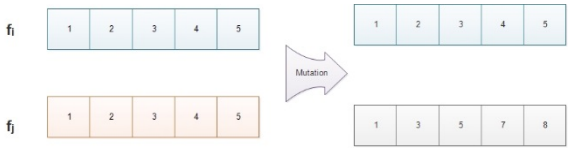

Fig. 5. Mutuation operation

IV. RESULTS

A. Assessment of the proposed method

In order to evaluate the proposed method, the information collection of shopping store with 9729 customer transaction information relating to 4223 customers and 1560 item were applied. The purpose of the proposed algorithm is that customers in each cluster have the highest similarity with each other. So items purchased by the customer in a cluster are similar. Although these items are distinct, customers' buying patterns are definable by the help of the frequency of each item. This concept is known by the Support, as shown in Equation 1. To assess this concept, we compared the K-means clustering methods [13] with the proposed method and achieved to the following results. In the results value of the K was set at 30 and calculated the mean support in each cluster of five maximum support besides standard deviation for all items. Results of K-means algorithm and the proposed method are shown in Tables I and II, respectively. The ANOVA test was used to express the differences of obtained results. The test results are shown in Table III. It can be seen that the mean value of support for 5 supported items were significantly higher than the results of K-means. Therefore, it can be argued that in the proposed approach, customers in each cluster have more similar buying pattern than the K-means algorithm. Also the standard deviation of proposed method is higher than the K-means which also confirms the previous term.

TABLE I. PRODUCT PURCHASE PATTERN WITH THE HELP OF K-MEANS METHOD

Customer's ID	Five maximum support values of all items						Mean of support for all items	Standard deviation of support for all items
	1	2	3	4	5	mean		
01	0.0236	0.0189	0.0189	0.0189	0.0142	0.0189	0.0063	0.0025
02	0.0208	0.0156	0.0156	0.0156	0.0256	0.0172	0.0063	0.0027
03	0.0138	0.0138	0.0138	0.0138	0.0138	0.0138	0.0047	0.0025
04	0.0175	0.0175	0.0175	0.0175	0.0175	0.0175	0.0056	0.0028
05	0.0259	0.0185	0.0185	0.0148	0.0148	0.0185	0.0054	0.0025
06	0.0131	0.0114	0.0114	0.0098	0.0098	0.0111	0.0035	0.0018
07	0.0294	0.0221	0.0221	0.0221	0.0221	0.0236	0.0084	0.0033
08	0.0107	0.0107	0.0107	0.0107	0.0092	0.0104	0.0035	0.0017
...	...	...	...	...	...	...	...	...
28	0.0209	0.0167	0.0167	0.0167	0.0167	0.0175	0.0054	0.0026
29	0.0244	0.0244	0.0244	0.0244	0.0244	0.0244	0.0094	0.0035
30	0.0158	0.0158	0.0158	0.0158	0.0126	0.0152	0.0044	0.0023

TABLE II. PRODUCT PURCHASE PATTERN WITH THE HELP OF PROPOSED METHOD

Customer's ID	Five maximum support values of all items						Mean of support for all items	Standard deviation of support for all items
	1	2	3	4	5	mean		
01	0.0934	0.0588	0.484	0.450	0.0381	0.0568	0.0052	0.0058
02	0.1019	0.0906	0.0340	0.0302	0.0302	0.0574	0.0060	0.0070
03	0.0799	0.0523	0.0303	0.0275	0.0275	0.0453	0.0050	0.0051
04	0.0403	0.0361	0.0361	0.0297	0.0297	0.0344	0.0040	0.0045
05	0.0985	0.0606	0.0492	0.0455	0.0379	0.0583	0.0061	0.0028
06	0.0909	0.0455	0.0455	0.0420	0.0350	0.0518	0.0052	0.0064
07	0.0466	0.0443	0.0373	0.0350	0.0326	0.0392	0.0042	0.0055
08	0.0463	0.0379	0.0379	0.0385	0.0295	0.0275	0.0039	0.0070
...	...	...	...	...	...	...	...	...
28	0.0591	0.0540	0.0411	0.0411	0.0360	0.0463	0.0047	0.0039
29	0.1412	0.0734	0.0621	0.0565	0.0452	0.0759	0.0080	0.0043
30	0.0606	0.0404	0.0379	0.0303	0.0303	0.0400	0.0048	0.0041

TABLE III. EVALUATION USING ANOVA TEST

=5%α	P-Value	F-value	Mean square	DF	Sum of Squares		
P<α	0.000	245.882	1.938*10-2	1	1.938*10-2	Variance among groups	Mean of 5 maximum supports
			7.883*10-5	58	4.572*10-3	Variance inside the groups	
				59	2.395*10-2	Total variance	
P>α	0.788	0.073	1.402*10-7	1	1.402*10-7	Variance among groups	Mean of support for all items
			1.918*10-6	58	1.113*10-4	Variance inside the groups	
				59	1.114*10-4	Total variance	
P<α	0.000	103.242	1.176*10-6	1	1.176*10-4	Variance among groups	Standard deviation of support for all items
			1.139*10-6	58	6.607*10-5	Variance inside the groups	
				59	1.837*10-4	Total variance	

The proposed method has a good convergence rate. As you can see in the following figure, when the replication frequency raised the total similarities between customers, cluster centers was also increased and number of change of cluster centers were also reduced. It was shown after the 9 replications, cluster centers did not change and the algorithm was finished.

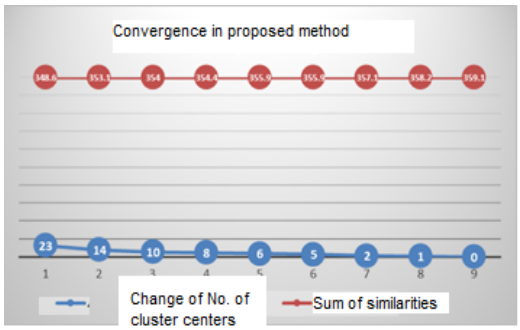

Fig. 6. Convergence in proposed method

B. Evaluating the performance of Genetic Algorithm

As mentioned, the initial cluster centers can be determined randomly or using genetic algorithm. Some experiments were taken in order to show the usefulness of genetic algorithms in which the number of clusters was changed from 30 to 80. In the following figure the effectiveness of using genetic algorithms was shown compared to the initial cluster centers randomly.

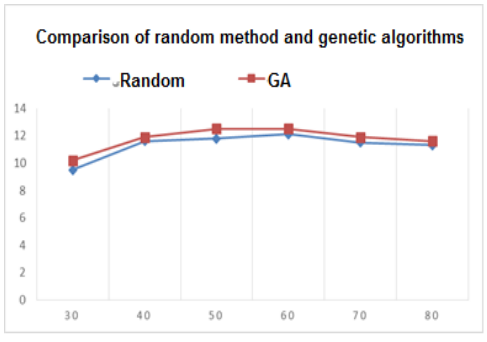

Fig. 7. Comparison of genetic algorithm and random method

V. CONCLUSION

Mass marketing no longer meets the diverse needs of customers. Customer clustering techniques have to be employed so that customers with the same buying patterns are put in a similar cluster. In this paper we presented a method of clustering customers employing genetic algorithms. Results are presented and further discussed. As shown, an improved performance is achieved through the discussed technique.

REFERENCES

[1] S. Dibb, L. Simkin, The market segmentation workbook: target marketing for marketing managers, Routledge, London, 1996

[2] A. Berson, S. Smith, K. Thearling, Building data mining applications for CRM, New York: McGraw-Hill, 2000

[3] S. Wedel, W. Kamakura, Market segmentation: Conceptual and methodological foundations, Boston: Kluwer, 1997

[4] T. P. Beane, D. M. Ennis, "Market segmentation: a review", European Journal of Marketing, Vol. 21, No. 5, pp. 20–42, 1987

[5] K. Hammond, A. S. C. Ehrenberg, G. J. Goodhardt, "Market segmentation for competitive brands", European Journal of Marketing, Vol. 30, No. 12, pp. 39–49, 1996

[6] R. J. Kuo, L. M. Ho, C. M. Hu, "Integration of self-organizing feature map and K-means algorithm for market segmentation", Computers and Operations Research, Vol. 29, No. 11, pp. 1475–14931, 2002

[7] R. G. Drozdenko, P. D. Drake, Optimal database marketing: Strategy, development, and data mining, London, Sage, 2002

[8] C. D. Manning, H. Schutze, Foundations of statistical natural language processing, Cambridge, MA, MIT Press, 1999

[9] H. C. Romesburg, Clustering analysis for researchers, Belmont Lifetime Learning Publications, 1984

[10] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", 20th International Conference on Very Large Databases, pp. 487–499, 1994

[11] H. Mannila, Database methods for data mining, 4th International Conference on Knowledge Discovery and Data Mining, New York, 1998

[12] J. H. Holland, Adaptation in natural and artificial systems, Ann Arbor, MI: The University of Michigan Press, 1975

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations", 5th Conference on Mathematical Statistics and Probability, Vol. 1, pp. 281–297, 1967