A Model for Clustering and Optimizing Portfolio: Tehran Stock Exchange using data mining algorithms

Siyamak Goudarzi *
Faculty of Engineering, Qom University
Tehran, IRAN
goudarzi.siamak@gmail.com

Ali Teymornejad
Faculty of Engineering, Qom University
Tehran, IRAN
ali.teymurnejad@gmail.com

Mohammad Javad Jafari
Faculty of Engineering, Qom University
Tehran, IRAN
mjafari.360@gmail.com

***Abstract*— Management of investment basket and selecting assets is one of the problems of decision making in financial area. In the competitive business environment, in order to confront complex competitions in the market, financial institutes try to consider the best policy of investment basket that in turn leads to an increase in the output for the investors. The goal of this study is to develop a portfolio by considering the behavior of investors in risk taking in a realistic method. This research aims at supporting investors, experts and intermediate managers in establishing optimized portfolio of stocks. The proposed model has used the data of 66 stockholders who were enlisted in Stock Exchange Market by using the five indexes of risk, output, skewness, liquidity and current ratio and clustered different companies by using the neuro- networks SOM algorithm The results show that the function of model to general index, the industry index and the index of 50 more active companies are better in Tehran Stock Exchange.**

I. INTRODUCTION

One of the decision problems in the financial domain is portfolio management and asset selection. The issues of optimizing portfolios attracted the attention of researchers in 1952, the modern theory of portfolio which was introduced by Markowitz for the first time and was the most essential and important success in the approach of the development and understanding financial markets and financial decision making established an organized paradigms towards establishing a portfolio with the highest output rate expectable in a certain level of risk (the characteristics of all existing portfolio in an efficient complex). According to Markowitz, for a certain level of an output, a person can minimize the portfolio variance through minimizing investment risk, or, in a certain level of risk that could be tolerable for the investor, the person can consider maximum output that would increase the output rate as expected by the portfolio [1].

Markowitz proposed that investors should consider both risk and output in associated way and select the amount of capital allocation in various investment opportunities based on the interaction between the two [2].

Since the time that Markowitz published MPT model, the model created many changes and improvements in the people's view to investment and portfolio and changed into an efficient tool for optimizing portfolio [3].

The primary assumption of MPT is market output, nevertheless, Grossman and Stieglitz (1980) stated that obtaining information on markets is costly and it is practically impossible to acquire full information in connection with the initial stocks; therefore, price could not fully reflect market information and efficiency; thus, identifying low value stocks is important for investors.

Another criticism on MPT is the calculation charge created by second rank functions and covariance matrix tools. This charge causes challenging problems in the actual life plans due to high number of stocks. It is for this reason that investors use other models for MPT models.

1. In first stage, the suitable stock is selected for the basket structure.

2. In the second stage, the amount of capital for the investment in each one of the selected stocks has been specified.

However, Markowitz's portfolio theory only yields a solution for capital allocation. There are hundreds types of capital in very good to very bad qualities are offered in the capital market and the capital faces a mass of information which make it very difficult to select. Markowitz's model is solvable by using mathematical programming modes; nevertheless,

when the limitations of the actual world such large number of capitals, limitations of weight values of stocks and as such are also added to it, its search space becomes very large and discontinued that make it practically impossible to use the mathematical models and it is in this point that using innovative algorithms such as Genetic Algorithm, Neural Networks, Ant Colony Algorithm and etc. [4].

The behavior of stock in market is non-linear like many of the natural phenomena. The linear models are unable to properly detect the non-linear behavior and are able to identify only the linear part of the good behavior; therefore, needing non-linear patterns and models has significant effect in prediction of stocks and taking suitable decisions in identifying the behavior of stock.

Following Markowitz's theory, various portfolio models have been introduced on portfolios that considered both the output and risk such as average- variance, mean- half variance model, etc. and it will be helpful if references are also provided for those models.

Therefore, with respect to uncertainty that governs stock exchange market and considering different trends and preferences of investors, it seems necessary to select a suitable set of stock exchange that one could overcome different uncertainties and preferences of individuals.

In most researches on selecting the portfolio, the first two momentums of distribution and return are used. In fact, in those researches, it was tried to select a portfolio that had the least risk and highest output. Many researchers believe that higher momentums such as skewness could be used only if either the return distribution is symmetric (for example normal) or the return distribution is not effective in selecting the investor. One of the solutions in optimized formation of investment basket is to select the portfolio from stocks of companies that do not have behavioral similarity in financial terms and somehow variable. By using this method, the existing risk reduces significantly and one could high output for the investment which is made. Since in ordinary methods, no other criteria than risk and output is considered and the companies are not ranked and clustered prior to investment basket formation, and also, no effect of skewness and liquidity in their calculations for forming stocks basket, the method which is used in present research could be an effective contribution in selecting the portfolio and gaining profit for investors.

This research tries to optimize the existing investment methods with less risk and higher outputs by using data mining techniques for clustering and classification of the existing companies in stock exchange. The goal of this research is to provide a model to create baskets with low risk and high output by classifying companies with similar characteristics in one group and by considering financial and personality characteristics of individuals.

II. METHODOLOGY

A. 1. Data collection and preparing data for performing the research

The model proposed in this research provides a flexible and realistic support to the investors, experts and intermediate managers in their decision making in the assessment and making portfolio. The stages of the proposed model of the research are shown in figure 1.

1.1. Deletion of unacceptable stocks and normalizing data

In this stage, we delete stocks that do not want investors. The data of the companies enlisted in Stock Exchange of Tehran that had no transaction in one month were deleted. In addition, some companies lacked historical data and therefore, the data of those companies were deleted as well. In addition, companies with low transaction in the month or those with no activities in a some part of time intervals concerned by the research were deleted. Furthermore, stocks with negative price to income were deleted. In sum, among all companies enlisted in Stock Exchange of Tehran, with respect to conditions mentioned above, 66 companies were selected for the research. In this research, the concerned data is the data of 2 years prior to investment date.

To perform more detailed processing, we normalized data before starting the calculation. This was done by using variance functions (1).

$$X=(x-\text{mean}(x))/\text{STD}(x) \quad (1)$$

2. Data clustering by using data mining algorithm

Clustering as defined by Mirkin (1996) is ''a mathematical technique designed for revealing classification structures in the data collected in the real world phenomena". Clustering is among data-mining classification algorithm. The clustering algorithm places information with close and similar characteristics in separate parts, called cluster. In another word, clustering is the very simple classification that we do in our daily works. In this research, we have used K-means and SOM to cluster the data of Stock Exchange market and ultimately, a combined method is done and finally, by comparing the results, optimized methods are used for this task. There are papers showing comparison of different clustering methods [5][6][7] and also adapting different clustering methods for a particular problem. In case based reasoning (CBR) [8][9] the problem of cluster indexing the case base to build a hybrid CBR has adapted many clustering methods.

In this paper we consider the K-means and self-organizing maps for clustering stock data. We will use validity indexes in each case to find the optimal number of clusters.

2.1. K-means clustering

In this stage, by using K-means algorithm to cluster data to K two to two cluster becomes incompatible. For the purpose of data clustering, the output, risk, and skewness indexes, the current and liquidity ratios are used. To perform this, the 24-month historical data for 66 companies of the present companies that is enlisted in Stock Exchange Market. The K-means algorithm divides the set of data into k sub-complex (cluster) as all elements of each sub-complex would have the

closest distance to the center of that sub-complex. The criteria that must be minimized in K-means include:

$$E_{K-means} = \frac{1}{C}\sum_{k=1}^{C}\sum_{x\in Q_k}\|x - C_k\|^2 \qquad (2)$$

In the above relations, C is the number of clusters; ،is the kth cluster and is the center of cluster. There are different indexes in determining the optimized number of clusters. Due to the importance of the number of clusters in using this algorithm, to specify the number of suitable clusters, the Davis-Bouldin criteria (Davis- Bouldin, 1979) and sum of squares errors (SSE) is used and the number of optimized clusters was assessed accordingly. The Davis- Bouldin in fact calculates the ratio of intra-cluster dispersion to the inter-cluster distances from following relations:

$$I_{DB} = \frac{1}{C}\sum_{k=1}^{C}\max_{l\neq k}\left\{\frac{S_c(Q_k)+S_c(Q_l)}{d_{ce}(Q_k,Q_l)}\right\} \qquad (٤)$$

$$S_c(Q_k) = \frac{\sum_i \|x_i - c_k\|^2}{|Q_k|} \qquad (٥)$$

Table 1. Davis- Bouldin Parameter rate and sum of errors squares (SSE) in K-means

Number of clusters	Davis- Bouldin	SSE
2	2.370704	7782.921
3	2.28979	7249.487
4	1.91615	6741.876
5	٨1.889458	6524.555
6	2.205832	6239.272
7	1.63858	5841.7
8	1.972306	5869.368
9	2.115385	5901.755
10	2.607626	6087.181

The results obtained show that the most s number of clusters is 7; as in Davis Bouldin and sum of squares of errors (SSE) have lower values in this state. Following figure shows the Davis-Baldwin and SSE fluctuation.

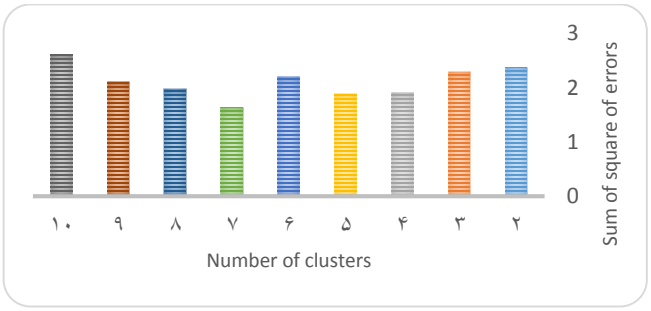

Diagram 1: Davis- Bouldin Parameter diagram for clusters in K-means

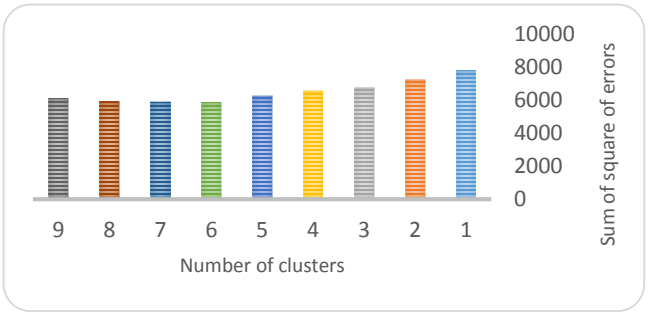

Diagram 2: Diagram of the parameter of sum of errors squares for different clusters in K-means

2.2. SOM Clustering

Neural networks are made of a series of non-linear knots which are in relations with each other in form of a network. The weights of this relationship help neural networks with their changes to see suitable trainings for solving a given problem. The SOM clustering algorithm is a type of neural network which was invented by Kohnen in 1987. Sum of square of errors

3. A combination of SOM and K-means clustering

In this method, according to the method introduced by Ku and Hoa, first, the optimized number of clusters and centers of each cluster is determined in accordance with SOM Clustering algorithm. After specifying the optimized number of clusters and clusters centers, they are used as input in K-Means algorithm and the clustering is performed by using those inputs. One of the reasons that we use in this method is that the hierarchical clustering methods such as SOM have a better performance in determining the number of clusters, the primary points for clustering process and clusters centers; and on the other hand, the non-hierarchical methods such as K-means show better performance when being used to determine the members of each cluster by using the obtained information. The results of executing SOM and K-means combined algorithm in MATLAB software for network 5 x 5 and 1000 frequencies of education are shown in the below figures.

The following figures show the distance between neighboring neurons from each other. As the amount of distances is bigger, it is shown with a darker color and as this distance, paler colors

are used for showing them. As it could be observed in following picture, the neurons with closer distance could be considered as a cluster. With respect to table 4-5 and subjects expressed above, we could use the SOM and K-means combined methods with 7 clusters in clustering the concerned data. The clusters labels for each neuron are shown in figure 1 and 2.

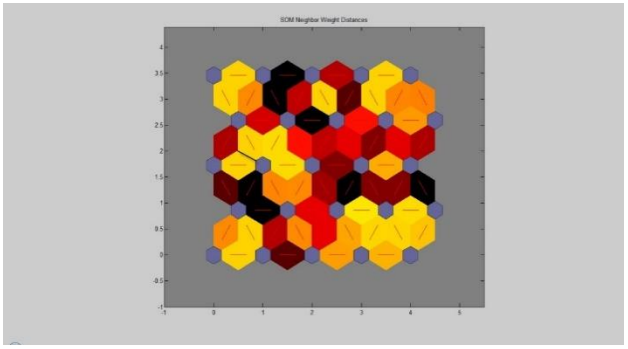

Figure 1. The distance of neurons from each other

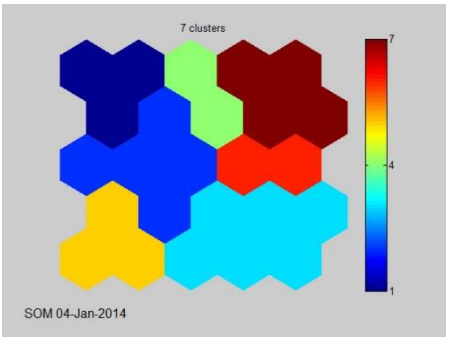

Figure 2. Clusters obtained

Ultimately, Davis- Bouldin indexes and SSE indexes were calculated in clustering results. As the following table shows, when using combined method, the most optimized results are when the number of cluster is considered as seven (7).

Table 3- Davis- Bouldin Parameter rate and sum of errors squares (SSE) in combined model

Number of clusters	Davis- Bouldin	SSE index
2	0.898988	292.6045
3	1.043345	205.0454
4	0.768323	124.123
5	0.811777	101.8414
6	0.816296	82.42638
7	0.667118	51.53343
8	0.697289	56.194.2
9	0.727447	69.478.2
10	0.710846	72.32169

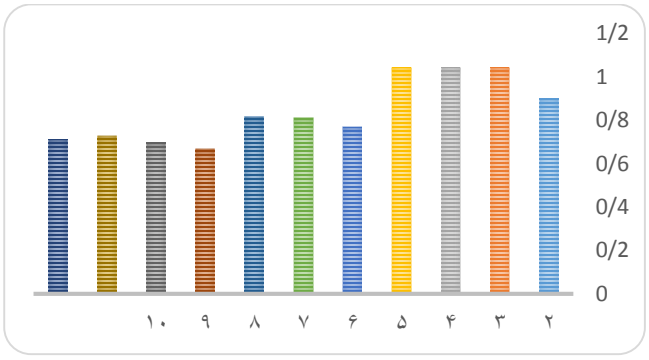

Diagram 3: The diagram of Davis- Bouldin parameter for different clusters in combined methods

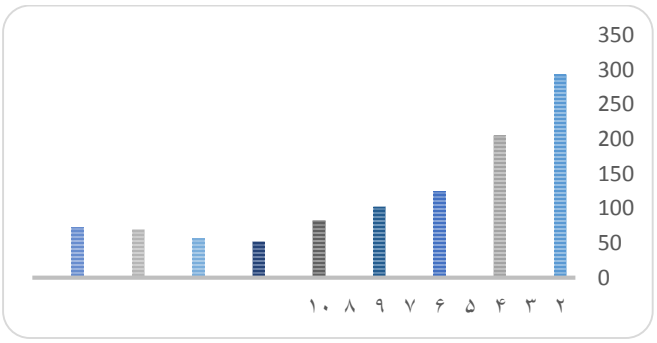

Diagram 4. The diagram of SSE parameter for different clusters in combined methods

4. Comparison between clustering methods

As it could be seen in the three methods of clustering, the best mode of the number of clusters is the 7 clusters, showing that the companies are grouped in suitable number. To benefit from the most detailed results of clustering in continuation of modeling, we compared the Davis- Bouldin and SSE indexed values, we compared the three methods in the most optimized states. As it could be seen in the table below, the third method, which is a combination of SOM and K-means has less SSE and Davis-Bouldin value and the main focus of this research is on this method and its resulted output.

Table4. Values of Davis- Bouldin Parameters and optimized SSE in the three methods subject of study

Methods used	Amount of Davis-Bouldin Index	Amount of SSE index
K-means	1.892833	3687.601
SOM	0.697289	64.95587
Combined methods of SOM and K-means	0.653226	56.19402

5. Discussion and conclusion

In present research, unlike ordinary models for the formation of portfolios, first superior and profitable companies are identified by using data clustering techniques and data ranking and in this path, it also considers the risk criteria and output of effective criteria, skewness, liquidity and current ratio for its calculations; and subsequently, it could form a basket consisting superior companies that are selected by considering several criteria in line with risk reduction and increase in investors output.

In brief, the innovations of the present research are outlined as follows:

1. Clustering the present companies in stock exchange market of Tehran

2. Using neural networks for clustering the companies

3. To use genetic algorithm ultra-explorative technique for selecting and optimization of portfolio

4. To use a three-stage process to perform companies clustering in first stage, clusters ranking in the second and optimize the portfolio among the superior stocks in the third stage.

5. Combining skewness and liquidity criteria in Markowitz for optimizing and completing the Markowitz model

The results show that the basket formed in present model provides a better result than total index, industry index and the index of the 50 more active companies and is a suitable guideline for the investors and immediate managers for the investment and consultation in investments.

Suggestions for the future researches

In this research, a model is presented for forming the portfolios that was compared with the baskets formed based on actual output in one month and the market, general and the 50 more active companies, and SOM, K-means algorithms and genetic algorithm were used for executing the model; thus, for studying the results of this research and finding a suitable algorithm for this model it is suggested:

1. The model presented with historical data to be executed in longer time intervals for acquiring better results.

2. To use other algorithms such as c-means for clustering and compare the results that were obtained with the results obtained from SOM and K-means algorithm.

3. Use other algorithms such as ants' algorithms for basket formation and compare the results with genetic algorithm.

4. Compare the research model with the variance mean model and other models.

5. Use other methods such as AHP and other methods for ranking.

This template has been tailored for output on US letter-sized paper.

Refrences

[1]. Lin, Chi-Ming, and Mitsuo Gen. "An effective decision-based genetic algorithm approach to multiobjective portfolio optimization problem." Applied Mathematical Sciences 1, no. 5 (2007): 201-210.

[2] Fabozzi, Frank J., Petter N. Kolm, Dessislava Pachamanova, and Sergio M. Focardi. Robust portfolio optimization and management. John Wiley & Sons, 2007.

[3] Lai, Kin Keung, Lean Yu, Shouyang Wang, and Chengxiong Zhou. "A double-stage genetic optimization algorithm for portfolio selection." In Neural Information Processing, pp. 928-937. Springer Berlin Heidelberg, 2006.

[4] Aranha, Claus, and Hitoshi Iba. "The memetic tree-based genetic algorithm and its application to portfolio optimization." Memetic Computing 1, no. 2 (2009): 139-151.

[5] Budayan, Cenk, Irem Dikmen, and M. Talat Birgonul. "Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping." Expert Systems with Applications 36.9 (2009): 11772-11781.

[6] Delibasis, Konstantinos K., et al. "MR functional cardiac imaging: Segmentation, measurement and WWW based visualisation of 4D data." Future Generation Computer Systems 15.2 (1999): 185-193.

[7] Mingoti, Sueli A., and Joab O. Lima. "Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms." European Journal of Operational Research 174.3 (2006): 1742-1759.

[8] Chang, Pei-Chann, and Chien-Yuan Lai. "A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting." Expert Systems with Applications 29.1 (2005): 183-192.

[9] Jo, Hongkyu, and Ingoo Han. "Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction." Expert Systems with applications 11.4 (1996): 415-422.