CLUSTERING STOCK MARKET COMPANIES VIA K- MEANS ALGORITHM

Mansoor Momeni
Full professor ,Tehran University
Maryam Mohseni*
Ph.D student ,Tehran University (Corresponding Author)
Mansour Soofi
Department of Industrial Management, Rasht Branch, Islamic Azad University, Rasht, Iran

Abstract

One of the main concepts in pattern recognition is clustering. This technique is used as important knowledge discovery tools in modern machine learning process. Clustering of high-performance companies is very important not only for investors, but also for the creditors, financial creditors, stockholders, etc. Hence, firms' clustering is considered as one of the important issues in Tehran Stock Exchange (TSE). To this end, we have used financial statement data of three industries in TSE for the year 2012. After selecting profit criteria (attributes) and prioritizing them using AHP, k means clustering algorithm is used to classify these companies. Also, to obtain the optimal number of clusters, different validity measures are presented. The identification of clusters of companies of TSE can be exploited to improve planning and get to more comprehensive decision making about companies.

***Keywords*:** Clustering, AHP, K- means algorithm, Validity.

Introduction

The purpose of any investor is to seek desirable investment opportunities for maximizing profit. Nowadays, investing in companies stocks needs financial knowledge, profitable stock selection and efficient use of capital. On the other hand, one of the important human activities is to classify complex phenomena by using their characteristics. Clustering or cluster analysis is the main method of classification *(Sharma, 1996)*. This technique has many applications in various sciences. Mirkin (1996) defined it as ''a mathematical technique designed for revealing classification structures in the data collected in the real world phenomena" *(Nanda et al. 2010)*. Clustering is used to divide a data set into classes using the principle of maximizing the intra class similarity and minimizing inter class similarity. It means that, clusters are formed so that objects which are similar are grouped together and objects that are very different fall into other clusters *(Babu et al. 2012)*. Cluster analysis is not the ultimate goal of research; rather it is

beginning for another works. In this paper, after determining criteria and prioritizing them by using Analytic Hieratical Process (AHP), we demonstrate well known clustering technique namely K-means as well as some validity indexes to obtain the optimal number of clusters.
The rest of this paper is organized as follows. Section 2 describes relevant literature review. Section 3 explains methodology of research. Section 4 shows main findings. Finally, in Section 5 conclusion is presented.

Literature review

In this section, concept of cluster, different methods of clustering and related works are briefly reviewed.

Clustering

Clustering is an unsupervised classification process *(Gan and Wu, 2007; Liu et al. 2010; Pakhir et al. 2004).* The aim of clustering is to find structure in data set. Cluster is collection of objects with similarity between them and dissimilarity to the objects in other clusters. Unlike classification, clustering doesn't rely on predefined classes *(Hajizadeh and Shahrabi, 2010*). Clustering has applications in several fields like math, multimedia, marketing (customers segmentation based on their similarities), meteorology, geology, medical, etc (*Gan and Wu, 2007).* First time, Tryon in 1939 used this term for grouping similar objects. There are two categories- crisp (or hard) and fuzzy (or soft) - for Clustering methods (*Gan and Wu, 2007 ; Halkidi and Batistakis, 2001*; *Pakhir et al. 2004).* In crisp clustering, data set is divided into distinct clusters. It means that data belongs to exactly one cluster. In fuzzy clustering, data belong to more than one cluster. It means that associated with each element is a set of membership levels. Hierarchical and Partitional algorithms are two categories of crisp clustering algorithms (*Gan and Wu, 2007).* Hierarchical clustering techniques divide the data set into smaller subsets and the result is a hierarchical structure of groups known as dendrogram. Partitional clustering techniques partition the data set into desired number of sets in a single step. They are based on minimizing an appropriate objective function.

Clustering techniques and stock market

There are some papers within literature that used various clustering methods in the field of financial markets and showed comparison of various clustering techniques. For example, *Doherty et al. (2005)* used TreeGNG, a hierarchical clustering algorithm, on a time series of share closing prices to identify groups of companies that cluster into clearly identifiable groups. *Rashidi and Analoui (2007*) proposed a modified k-means clustering algorithm to cluster stock market companies, based on similarity measure between time series. This algorithm is applied to the analysis of the Dow Jones (DJ) index companies, in order to identify similar temporal behavior of the traded stock prices. To cluster financial markets, Also, *Basalto et al. (2005)* applied chaotic map clustering algorithm to the analysis of the Dow Jones index companies, in order to identify similar temporal behavior of the traded stock prices.

Nanda et al. (2010) used stock returns at different times from the stocks of Bombay Stock Exchange for the fiscal year 2007–2008 in order to manage portfolio. Results of analyses indicated that *K*-means cluster analysis builds the most compact clusters as compared to SOM and Fuzzy *C*-means for stock classification data. Also, *K*-means clustering minimize portfolio risk. *Shin and sohn (2004*) used K-means, self-organizing map (SOM), and fuzzy K-means clustering techniques to segment stock market brokerage commission customers. Results of analysis showed that fuzzy K-means cluster analysis is the most robust approach for segmentation of customers.

According to the literature reviewed, we could see that there are very few studies and researches in clustering stock market companies. In this paper we consider the K-means technique for clustering stock market companies and weigh the selected criteria. Also, we will use validity indexes to find the optimal number of clusters.

Methodology

Determination industries and Selection criteria

In the literature of cluster analysis, two terms are used: objects (cases) and criteria (attributes).The purpose of this paper is to cluster companies (here called objects) - in three industries of TSE. These industries consist of automotive industry; cement and metal industries. Factors that affect the choice of the companies in each industry are (1) Dissemination of financial data, (2) Being active in TSE for the year of 2012.therefore, 87 companies are selected.

Also criteria are quantitative or qualitative variables that objects are categorized based on them. So the first step is to determine appropriate criteria for clustering. To this end, by interviewing with several of the capital market experts, five profitability ratios are selected. These ratios cause easier analysis. In Table 1, variables definition is showed.

Table 1. Variables definition

Variables	Definition
Return on assets(ROA)	Net profit / total assets
return on equity(ROE)	Net Income/ Shareholder Equity
Profit to sales ratio	net profit / net sales
Earnings Per Share(EPS)	Net Income - Dividends on preferred Stock/ Average Outstanding Shares
Operating profit margins	operating profit / sales

Prioritization of criteria (weighing)

In this step a list of the relative weights, importance, or value, of the criteria calculates. There are different methods for determining weights. In this paper, analytic hierarchal process (AHP) is used. In the following, a review of AHP as well as implementation of this technique is explained.

Analytic hierarchal process (AHP)

The analytic hierarchy process (AHP) was introduced by Thomas L. Saaty in the 1970s. It is a multi-criteria decision-making (MCDM) approach *(Saaty, 2012; Vaidya and Kumar, 2006*). Process of AHP consists of 5 stages:

1. ***Determining a Hierarchical Tree***: AHP uses a multi-level hierarchical structure that comprises a goal, criteria (and sub criteria) and options.
2. ***Finding priority of the criteria***: AHP uses a set of pairwise comparisons to calculate the relative weights of importance of the criteria.
3. ***Scoring of options based on each criterion***: in this stage like stage 2, pairwise comparison of options in terms of each criterion carry outs. Then, the ratings are normalized and averaged.
4. ***Obtaining Consistency Ratio (CR):*** The important stage is to obtain a CR to measure how consistent the judgments have been relative to large samples of purely random judgments. It is noteworthy that consistency ratio should calculate for each of pairwise comparisons. The CR should be ≤ 0.1. It means that, if the CR is much in excess of 0.1, the judgments are untrustworthy and the pairwise comparison is valueless and it must be repeated.
5. ***Calculating the final score*:** Finally, the option scores are combined with the criterion weights to make a final score for each option.

Sometimes, there are two or more decision makers (DMs). So, geometric mean method should be used to aggregate individual judgments.

According to this algorithm, in this paper, hierarchal tree contains of two levels:

- object (Prioritization of selected criteria for clustering companies of three industries in TSE) and;
- criteria (five selected criteria)

Then based on decision tree, a questionnaire prepared and five experts completed it. In this paper, AHP is implemented in the software Expert Choice (EC). Calculated consistency ratio by software is 0.06 that represents the relative consistency of decision makers' judgments.
Table 2 shows the results of using geometric mean to obtain group judgment. Also, in Table 3 output of software is represented.

Table 2.Combination matrix of DMs 'opinions

Prioritization of selected criteria for clustering companies of three industries in TSE	Return on assets(ROA)	Rate of return on equity(ROE)	Profit to sales ratio	Earnings Per Share (EPS)	Operating profit margins
Return on assets (ROA)	0.380	0.3309	0.459	0.3486	0.3060
Rate of return on equity(ROE)	0.2003	0.1760	0.2020	0.0958	0.2221
Profit to sales ratio	0.116	0.1225	0.1418	0.3304	0.1180
Earnings Per Share (EPS)	0.168	0.2855	0.0669	0.1563	0.2452
Operating profit margins	0.134	0.0855	0.1297	0.0687	0.1085

Table 3. Final weights

Final weight	criteria	Prioritization
0.364	Return on assets (ROA)	1
0.184	Earnings Per Share (EPS)	2
0.179	Rate of return on equity(ROE)	3
0.165	Profit to sales ratio	4
0.105	Operating profit margins	5

Determining appropriate method for companies clustering

The K-means is one of the most popular clustering algorithms *(Hajizadeh and Shahrabi, 2010; Halkidi and Batistakis, 2001; Kanungo and Netanyahu, 2002; Pakhira et al. 2004).* This clustering algorithm was first described by Macqueen (1967) *(Gan and Wu, 2007*). This algorithm is a partitioning clustering or non hierarchical method that splits dataset (objects) into k clusters (*Gan and Wu*, 2007 ; Jain, 2010; Wu, 2002). Stages of k-means clustering algorithm are described as follows:

- **Initial stage**: Partition the objects into k cluster randomly.
- **Repetition stage:**
- Calculate the center of each cluster as the mean of the data.
- Compute distance (like Squared Euclidean distance) from each object to each cluster.
- Compute objective function know as squared error function given by:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

["K": the number of cluster center, "n": the number of data in i^{th} cluster and" $\left\| x_i^{(j)} - c_j \right\|$ ": Euclidean distance of each object from its center (C_j)]

- **Improvement stage:**
- Assign each object to the cluster with the nearest center.
- **Stop stage:** This process continues until no object move clusters or the value of objective function doesn't reduce.

In this paper, due to plurality of data, SPSS software is used.

Validity

Cluster analysis is an unsupervised method and one of the biggest problems with Clustering is identifying the optimum number of clusters (*Gan and Wu, 2007; Liu et al.2010).* Therefore, determining the appropriate number of clusters for a data set is generally an important process. To obtain the optimal number of clusters, some kind of clustering results validation is necessary (*Liu et al.2010; Pakhira et al. 2004).* There are different indexes and functions to provide validity

measures. These indexes provide a clear picture on the optimal number of clusters. Generally, validity indexes are measures for assessing the results of clustering (*Gan and Wu, 2007*). In the following, some of validation indexes review briefly.

Davies–Bouldin index: The lower the value, the better the cluster structures *(Halkidi and Batistakis,2001; Maulik and Bandyopadhyay, 2002; Nanda et al. 2010; Pakhira et al. 2004;Vendramin et al, 2010).*
Modified Hubert's statistic (MH): The higher the value, the more compact clusters (*Bezdek, 2002*).
Silhouette index: Better quality of a clustering is indicated by a larger Silhouette value (Chen et al. 2002; Liu et al. 2010; *Vendramin et al, 2010).*
Dunn's index (DI): This index is proposed to use for the identification of ''compact and well-separated clusters". Large values indicate the presence of compact and well-separated clusters (*Halkidi and Batistakis,2001; Nanda et al. 2010; Pakhira et al. 2004*).
R -Squared index (RS): This index is to measure the dissimilarity of clusters. It measures the degree of homogeneity degree between groups. The values of RS range from 0 to 1. The value near 1 means that there is significant difference among the clusters (*Halkidi et al. 2002*).
SD Validity Index: this index defines based on average scattering and total separations. Lower SD index means better clustering (*Halkidi et al. 2002;Liu et al,2010*).
RMSSTD (root – mean – square standard deviation) index: it is the variance of the clusters and it measures the homogeneity of the clusters. The lower RMSSTD value means better clustering (*Liu et al,2010).*
In this paper, we give more importance to Davies–Bouldin, Silhouette, and Dunn's indexes.
The validity indexes calculated by Machaon CVE (one of the powerful software in calculating cluster validation and finding the optimal number of clusters).Table 4 shows validity indexes of K-means clustering.

Table 4. Validity indexes of K-means clustering

Metal industry		
k-means validity indexes	No. of clusters	
	2	3
Davies–Bouldin index	**0.969**	1.131
Dunn's index (DI)	**1.869**	1.052
Silhouette index	**0.64**	0.44

This calculation is done just for metal industry. The results of two other industries are the same. From Table 4 we can infer that number of clusters can be 2 for optimal clustering of the data set.

Main findings

The result of implementing k- means method by using SPSS software is presented in Table 5. The table contains the number of cases in each cluster and final cluster centers for three industries.

Table 5. Final cluster centers in each industry

final cluster centers in each industry						Attributes
Metal industry		**Automotive & parts industry**		**Cement industry**		
2	**1**	**2**	**1**	**2**	**1**	
0.120	-0.353	0.076	-0.182	0.306	0.106	**Return on assets (ROA)**
0.151	0.131	0.248	0.057	0.574	0.264	**Rate of return on equity(ROE)**
15.59	-82.14	10.34	-2.42	42.47	23.35	**Profit to sales ratio**
447.49	-1.993	550.39	-99.01	2717.01	590.75	**Earnings Per Share (EPS)**
0.1749	-0.514	0.093	0.02	0.338	0.268	**Operating profit margins**
23	6	16	12	7	23	**No. of cases**

Also, in Table 6 companies located in each cluster as well as the object's distance from the cluster center is displayed.

Table 6. List of companies of each cluster

Metal industry		**Automotive & parts industry**		**Cement industry**	
distance from the cluster center	companies	distance from the cluster center	companies	distance from the cluster center	companies
	Cluster1		**Cluster1**		**Cluster1**
90.714	Parsmetal	206.004	Pars Khodro	683.9	Chalk Iran
1270.14	Sadidpipe	249.6	Saipa Diesel	248.7	Orumieh cement
1728.13	Rolling and Pipe Mills Ahwaz	428.6	Tractor Forging	521.6	Elam cement
1638.12	Alumtekcorp	127.3	Electric khodro shargh	300.4	Bojnurd cement
1632.3	Aluminum Rods	131.5	Tractor Foundry	96.57	Tehran cement
1234.1	Aluminum Pars	129.3	Zamyad	47.46	Khash cement
1057.9	Feromolibden kerman	102.9	Saipa azin	146.8	Khazar cement
	Cluster2	719.7	Iran casting industries	95.48	Darab cement
694.08	Esfahan steel	52.2	Zar spring	163.6	Dashtestan cement
354.28	Sepahan group	167.3	Mehvarsazan	558.5	Dorud cement
292.71	Iran ferosilis	105.2	Mehrcampars	137.5	Saveh cement
213.7	Amirkabir steel	139.7	Indamin	432.1	Sepahan cement
74.07	Mobarake steel		**Cluster2**	321.1	Neyrize cement
425.8	Kavian steel	162.5	Iran khodro	46.42	Shahrood cement
52.17	Lule & Machine Iran(L.M.I)	255.55	Iran Khodro Diesel	335.04	Shargh cement
131.53	Production Steel & Navard	48.42	Saipa	382.8	Shomal cement
432.7	Navard sakhte foolad	103.9	Bahman group	672.4	Sofian cement
337.6	Iran aluminium	378.6	Motorsazan	196.9	Gharb cement
931.8	Madan Asia Zarin	197.3	Irka part	1064.6	Fars cement
81.44	National Iranian Lead Zinc Zanjan	666.9	Charkheshgar	567.04	Sabzever cement
35.42	mineral processing	398.02	Iran radiator	1013.7	Mazandaran cement

111.33	Iran alloy steel	259.2	Mashhad wheel	83.47	Hegmatan cement
126.49	Yazd alloy steel	2005.4	Sazepouyesh	150.7	Karon cement
1364.42	Khuzestan steel	188.5	Ravan fan avar		**Cluster2**
350.7	Khorasan steel	420.2	**Khavar Spring**	1042.8	Ardebil cement
291.5	Bahonar Copper	88.45	Iranlent	485.5	Esfahan cement
226.3	National lead &zinc	384.02	Vehicle axle manufacturing(VAMco)	797.6	Behbahan cement
438.9	National Iranian Copper Industries	287.3	Nasirmachine	3955.2	Ghaen cement
1405.3	Navard aluminium	216.32	Niroo Mohareke	730.05	Hormozgan cement
509.5	Calcimine			936.9	Kordestan cement
				1557.3	Kerman cement

Output of software only shows that each company located in what cluster (1 or 2). But given that all attributes (criteria) used in clustering are profitability ratios and these ratios analyze the level of success of company to make profit and provide it through income, sales and investment. So, the higher amounts of these ratios express proper performance of the company. Therefore, Based on outputs of software and attributes amounts, two clusters- appropriate and inappropriate performance identified. Since, all attributes are profitability ratios, the analysis will be easier. Identifying company's situation can help executive managers to make a better decision and improve their planning.

The reason of implementing cluster analysis for each separated industries is that effect of factors such as inflation, economic sanctions, etc. isn't the same for industries.

Conclusions

The purpose of this paper was to segment active companies of three industries- cement industry, metal and automotive and parts industries- in Tehran Stock Exchange (TSE), using financial ratios and cluster analysis. There are various financial ratios such as liquidity ratios, leverage ratios, profitability ratios and efficiency or activity ratios but, in this research, profitability ratios highlighted and used for clustering. So, five criteria selected. The result of criteria weighing by analytic hierarchal process showed that ROA is first priority between other ratios in this research and EPS, ROE, and Profit to sales ratio and Operating profit margins are in the next places in sequence.

K-means algorithm, one of the most popular methods, used for companies' clustering. Before implementing k-means, determining the appropriate number of clusters for a data set is an important process. So, different indexes introduced and three validity indexes (Davies–Bouldin index, Dunn's index and Silhouette index) calculated and optimal number of clusters determined.

The data set for companies listed in TSE collected from financial information databases. We considered data for the year 2012. The results of clustering listed in the main finding section. All companies located in two clusters. Given the economical and political conditions, including sanctions as well as the result of this research, most of companies located in cluster 1.

Since investors seek the best opportunities for investment, the results of clustering can help them to make better decision.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers and the editor for their insightful comments and suggestions.

References

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., and , Pascazio, S. 2005Clustering stock market companies via chaotic map synchronization, Physica A, 345 (1-2) ,196–206.

Bezdek, J.C. 2002.Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics.28 (3),301 – 315.

Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S.H., and Zhang, M.S.H. 2002.Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. Statistica Sinica, 12 ,241–262.

Doherty, K.A.J., Adams, R.G., Davey, N., and Pensuwon, W. 2005. Hierarchical Topological Clustering Learns Stock Market Sectors, ICSC Congress on Computational Intelligence Methods and Applications 1-6, Istanbul.

Gan, G., and Wu, C.Ma,J.2007. Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA 466 .

E.Hajizadeh, E., Davari., H., and Shahrabi, J. 2010. Appilication of data mining techniques in stock market. Journal of Economics and International Finance 2(7), 109-118.

M.Halkidi, M., Batistakis, Y., and Vazirgiannis, M. 2002. Cluster validity methods: part II, SIGMOD Rec, 31(3) , 19-27.

M.Halkidi, M., Batistakis,Y., and Vazirgiannis, M. 2001. Clustering algorithms and validity measures, Scientific and Statistical Database Management(SSDBM) Conference, Virginia, USA , 3-22.

Jain, A.K. 2010. Data clustering: 50 years beyond k -means, Pattern Recognition Letters 31 , 651 –666.

Kanungo, T., Netanyahu, N. S. and Wu, A. Y. 2002. An Efficient k -Means Clustering Algorithm: Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7).

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. 2010. Understanding of internal clustering validation measures, In ICDM , 911–916.

Maulik , U., and Bandyopadhyay,S.2002. Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach.Intell , 24(12) (2002) 1650–1654.

Nanda, S.R. Mahanty, B., and Tiwari, M.K. 2010. Clustering Indian stock market data for portfolio management. Expert Systems with Applications, 37 (12) , 8793–8798.

Pakhira, M. K., Bandyopadhyay, S., and Maulik, U. 2004. Validity index for crisp and fuzzy clusters, Pattern recognit. Lett,37(3), 487–501

Rashidi, P., and Analoui, M. 2007. Modified k-means algorithm for clustering stock market companies,1st Iran Data Mining Conference 201-21, Tehran: Amir Kabir University.

Saaty , T. L., and Vargas, L. G. 2012. Models, Methods, Concepts & Applications of the Analytic Hierarchy Process. Springer, 175.

Sharma, S. 1996. Applied multivariate techniques, John Wiley & Sons, Inc .

Shin, H.W., and Sohn, S.Y. 2004. Segmentation of stock trading customers according to potential value, Expert Systems with Applications, 27 (1) , 27–33.

Suresh Babu, M., Geethanjali, N., and Satyanarayana, B. 2012. Clustering Approach to Stock Market Prediction. Int. J. Advanced Networking and Applications. 3(4) , 1281-1291.

Vaidya, S., and Kumar, S. 2006. Analytic hierarchy process: An overview of applications, European Journal of Operational Research, 169(1) , 1-29.
Vendramin, L., Campello, R J. G. B. and Hruschka, E. R. 2010.Relative Clustering Validity Criteria: A Comparative Overview, Statistical Analysis and Data Mining 3(4) , 209-235.
Wu, A.Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7) , 881-892.