DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2015-0001

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

A COMPARISON OF *K*-MEANS AND FUZZY *C*-MEANS CLUSTERING METHODS FOR A SAMPLE OF GULF COOPERATION COUNCIL STOCK MARKETS

Salam Al-Augby, Ph.D. Student at University of Szczecin

University of Kufa
Research and Information Qualifying Centre
Kufa, P.O. Box (21), Najaf Governorate, Iraq
e-mail: salam.alaugby@gmail.com

Sebastian Majewski, Ph.D.
Agnieszka Majewska, Ph.D.

University of Szczecin
Faculty of Economics and Management, Institute of Finance
Department of Insurance and Capital Markets
Mickiewicza 64, 71-101 Szczecin, Poland
e-mail: masaj@wneiz.pl; e-mail: magnes@wneiz.pl

Kesra Nermend, Ph.D. Eng.

University of Szczecin
Faculty of Economics and Management, Institute of IT in Management,
Department of Computer Methods in Experimental Economics,
Mickiewicza 64, 71-101 Szczecin, Poland
e-mail: kesra@wneiz.pl

Received 3 February 2014, Accepted 24 October 2014

Abstract

The main goal of this article is to compare data-mining clustering methods (*k*-means and fuzzy ***c***-means) based on a sample of banking and energy companies on the Gulf Cooperation Council (GCC) stock markets. We examined these companies for a pattern that reflected the effect of news on the bank sector's stocks throughout October, November, and December 2012. Correlation coefficients and ***t***-statistics for the good news indicator (GNI) and the bad news indicator (BNI) and financial factors, such as PER, PBV, DY and rate of return, were used as diagnostic variables for the clustering methods.

Keywords: news, k-means, GCC, stock market, fuzzy c-means.

JEL classification: A12, A13, C02, C63, G11.

Introduction

Data mining (DM) analyzes (often large) observational data sets to find unsuspected relationships and summarizes the data in novel ways that are understandable and useful for the data owner[1]. The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business and education. Data mining attempts to formulate, analyze, and implement basic induction processes that help extract meaningful information and knowledge from unstructured data[2]. In today's knowledge-driven economy, DM is an essential tool in pursuing enhanced productivity, decreased uncertainty, delighted customers, mitigated risk, maximized returns, refined processes, and optimally allocated resources[3]. DM can be used in financial application such as businesses[4], banking and marketing[5] to gain significant advantages in today's competitive global marketplace[6].

Clustering is an important data-mining technique for extracting useful information from various high-dimensional datasets[7]. Clustering is a process of grouping a set of objects into clusters so that the objects are quite similar in the same cluster but very dissimilar compared to objects in other clusters. Various types of clustering methods have been proposed and developed[8] or can be defined as a mathematical technique designed for revealing classification structures in the data collected in real-world phenomena[9]. Clustering methods organize a data set into clusters so that data points in one cluster are similar and data points in other clusters are dissimilar[10].

The k-means algorithm is an efficient and a well-known algorithm in clustering large data sets[11]. Ruspini[12] and Bezdek[13] adduced the fuzzy versions of the -means algorithm, where each pattern is allowed to have membership functions in all clusters rather than a distinct membership in exactly one cluster. However, working only on numeric data limits, the use of these means algorithms in such areas as data mining where large categorical data sets are frequently encountered[14].

There are two main types in applying k-means algorithms in cluster analysis − "hard" non-fuzzy[15] or fuzzy. In the first type the number of clusters k must be determined in advance as an input to these algorithms. In a real data set, k is usually unknown. In practice, different k values are tried, and cluster validation techniques are used to measure the clustering results and determine the best value of k[16].

The k-means algorithm is a classic technique, and many descriptions and variations are available[17]. In addition, it is popular because it is conceptually simple and is computationally fast as well as memory efficient. Nonetheless, various limitations in the *k*-means algorithm make extraction difficult[18].

K-mean clusters observations into k groups, where k is provided as an input parameter[19]. *K*-means clustering starts with a single cluster with its center as the mean of the data. This cluster is split into two, and the means of the new clusters are trained iteratively. These clusters again split, and the process continues until the specified number of clusters is obtained[20].

In clustering algorithms, points are grouped by some notion of “closeness” or “similarity.” In *k*-means, the default measure of closeness is the Euclidean distance[21]. The idea of fuzzy clustering was first introduced as an alternative to the traditional cluster analysis by applying membership values to points between clusters Ruspini[22].

The fuzzy *c*-means clustering approach is also known as fuzzy *k*-means[23]. It is analogous to traditional cluster analysis[24]. Fuzzy *c*-means developed by Bezdek in 1981 adapted the fuzzy set theory which assigns a data object (observation) to more than one cluster.

The essential difference between fuzzy *c*-means clustering and standard *k*-means clustering is the partitioning of objects into each group. Rather than the hard partitioning of standard *k*-means clustering, where objects belong to only a single cluster, fuzzy *c*-means clustering considers each object a member of every cluster, with a variable degree of “membership”[25].

The similarity between objects is defined by a distance measure, which plays an important role in obtaining correct clusters. For simple datasets where the data are multidimensional, the Euclidean distance measure[26] can be used. However, there are several types of distance measure that can be used for obtaining clusters of the same data, for example the Manhattan distance can be used for Euclidean data[27].

The squared Euclidean distance is another distance measure, mathematically speaking; it uses the same equation as the Euclidean distance metric but does not take the square root. As a result, clustering with the squared Euclidean distance metric is faster than clustering with the regular Euclidean distance[28]. Applying some other distance measure than the most commonly applied Euclidean distance has been reported in several articles and it has obtained better clustering accuracy[29].

The Gulf Cooperation Council (GCC) markets are the most advanced in economic reforms in the Middle East and proceeded solidly toward regional integration during the early 2000s[30]. The members of the GCC, Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates, represent very promising emerging markets[31], and their financial sectors remain dominated by banks[32]. Development of physical infrastructure has been assisted in a large part by labor cost competitive advantages and facilitated, at the technical and management level, by imported Western expertise[33]. GCC banks, in turn, are often partly state-owned, reflecting the continuing large role of GCC governments in local economic development[34]. The GCC markets differ from those of developed countries and from other emerging markets in that GCC markets are largely segmented from the global equity markets and are overly sensitive to regional political events[35].

A stock market is a public market for trading company stock and derivatives at an agreed price; these are securities listed on a stock exchange as well as those traded privately. It is an organized set-up with a regulatory body, and the members that trade in shares are registered with the stock market and regulatory body[36]. A financial market is a complex, non-stationary, noisy, and chaotic system, but it does not follow a random walk process[37]. Financial markets contain many uncertainties, and interact with various economic, political, and social factors. Since change in the stock market is more disordered, the system is hard to define as merely a linear or nonlinear system[38]; therefore, predictions of stock market price and its direction are quite difficult. Some recent studies show that media has a systemic impact on financial markets, and it can effectively attract the investors' attention but also affect stock prices[39]. Stock movements might be random but may be correlated with some (also randomly occurring) economic or political news[40]. Recently, various studies have analyzed the link between news coverage and stock prices[41]. Many experiments have shown an influence of information on the future valuation of stocks. For example, Andreassen[42] presented fictitious news and stock quotes (positive and negative) for a selected group of investors[43]. Being able to grasp the information to make the right decision is a very important issue for short-term investors. However, "good news" and "bad news" contain a large amount of repetitive keywords, thus decreasing the accuracy of clustering[44].

1. Data Mining

Data mining and knowledge discovery is a family of computational methods that aim at collecting and analyzing data related to the function of a system of interest to gain a better understanding of the system[45]. Data mining analyzes massive observational data sets to find unsuspected relationships and summarizes the data in novel ways that are understandable and useful for the data owner[46] and refers to extracting or "mining" knowledge from large amounts of data[47]. Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning.

Data mining and knowledge discovery (data mining or KDD for short) has emerged to be one of the most vivacious areas in information technology in the last decade. Cao, Yu, Zhang, & Zhang[48], which includes pre-processing and post-processing tasks. Pre-processing includes data extraction, data cleaning, data fusion, data reduction, and feature construction, whereas post-processing steps include pattern and model interpretation, hypothesis confirmation, and generation, and so on. This knowledge discovery and data mining process tends to be highly iterative and interactive[49].

DM is involved in predictive and descriptive models that are applied in many different tasks[50].

The descriptive model includes the following tasks[51]:

- association rules,
- sequence discovery,
- summarization,
- clustering.

1.1. Clustering

Clustering is a process of grouping a set of physical or abstract objects into a set of classes, called clusters, according to some similarity function. Cluster is a collection of objects that are similar to one another within the cluster and dissimilar to objects in other clusters[52]. There are different types of clustering paradigms such as representative-based, hierarchical, density-based, graph-based, and spectral clustering depending on the data and desired cluster characteristics[53]. One of the mostly commonly used clustering algorithms is the *k*-means algorithm[54].

1.2. K-means algorithm

The *k*-means is one of the simplest unsupervised learning algorithms for clustering problems. The algorithm aims to form *k* clusters of n objects, resulting in intra-clusters[55]. The *k*-means algorithm is a simple, iterative, clustering algorithm that partitions a given dataset into a user-specified number of clusters, *k*. One of the main advantages of this algorithm is that it is simple to implement and run, relatively fast, easy to adapt, and common in practice[56]. The *k*-means is an efficient centroid-based algorithm that has been widely used in various key areas, such as micro-array datasets, high-dimensional data sets, etc. Two terms, a cluster and distance, should be defined:

A cluster is an ordered list of objects that have common characteristics. The objects belong to an interval [a, b], in our case [0, 1].

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed[57].

The *k*-means clustering technique begins with a description of the basic algorithm[58]. Choosing k initial centroids is the first step, where k is a user-specified parameter, namely, the number of desired clusters. The second step is to assign each point to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. The assignment steps are repeated and updated until there are no point change clusters, or equivalently, until the centroids remain the same[59].

An algorithm for partitioning (or clustering) n data points into k disjoint subsets (Si) containing data points minimizes the sum-of-squares criterion, such that

$$Jmin = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \tag{1}$$

where x_j is a vector representing the j_{th} data point and μ_i is the geometric centroid of the data points in S_i [60].

In clustering algorithms, points are grouped by some notion of "closeness" or "similarity"[61]. A more common measure in *K*-mean is Euclidean distance, which is computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle[62]; the distance between two points on the real line is the absolute value of their numerical difference. Thus, if x and y are two points on the real line, then the distance between them is computed as[63]:

$$\sqrt{(x-y)^2} = |x-y| \tag{2}$$

As in Cartesian coordinates, if $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ are two points in Euclidean n space, then the distance from p to q or from q to p is given by[64]:

$$d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots . + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \tag{3}$$

The issue of determining "the right number of clusters" in *k*-means has attracted considerable interest, especially in recent years. Cluster intermix affects the clustering results the most[65]. Researchers have suggested several other procedures for determining the number of clusters in a dataset. The first method, suggested by Calinski[66], chooses the number of clusters as the argument maximizing, where B (k) and W (k) are the between and within cluster sum of squares (SS) with *k* clusters, respectively. CH (k) has the form of an analysis of variance (ANOVA) *F*-statistic for testing the presence of distinct groups[67].

1.3. Fuzzy *c*-means

Fuzzy clustering is based on Zadeh's idea which was introduced in 1965[68].This idea refers to the similarity a point shares with each cluster with a function (termed the membership function) whose values (called memberships) are between zero and one[69]. Fuzzy approach in clustering analysis gives the opportunity to describe groups or clusters that can at best be imprecisely articulated. Data points can be partitioned into a specific number of overlapping natural groups, i.e., fuzzy clusters, with each data point in each cluster to some degree specified by a membership value[70].

The fuzzy *c*-means algorithm is based on minimizing the objective function shown below, for a given fuzzy partition of the data, n, and a set of *k* cluster centroids[71].

The fuzzy *k*-means algorithm classifies each vector to all clusters with different values of membership between 0 and 1. This membership value indicates the association of a vector to each *k* cluster. Notice that the fuzzy *c*-means algorithm does not classify fuzzy data, but crisp data into fuzzy clusters[72].

For a set of n individuals classified into *k* classes with conventional (Boolean) classification, the membership function equals $F = \mu_{ij} = 1$, where individual *i* belongs to class *j*, and $F = \mu_{ij} = 0$, when individual *i* does not belong to class *j*. Three conditions ensure that conventional sets are exclusive and jointly exhaustive:

$$\sum_{j=1}^{k} \mu_{ij\,=}\ 1, \qquad 1 \ \leq\ i\ \leq\ n \tag{4}$$

$$\sum_{i=1}^{n} \mu_{ij}\ > 0, \qquad 1 \ \leq\ i\ \leq\ k \tag{5}$$

$$\mu_{ij} \in \{0, 1\}, \quad 1 \leq I \leq n,\ 1 \leq j \leq k \tag{6}$$

The sum of membership of an individual across all classes is 1and it is indicated in equation (4). Equation (5) shows that the classes are not empty because one individual belongs to each class at least, so finally, Eq. (6) suggests that an individual belongs to a class or does not belong at all. This equation presents the difference between non-fuzzy and fuzzy classes. Fuzzy set theory allows Eq. (6) so that class memberships are let to be partial and can take on any value between and including 0 and 1 (Eq. 7)[73].

$$\mu_{i,j} \in [0,1], \quad 1 \le i \le n,\ 1 \le j \le k \tag{7}$$

Fuzzy *C*-Means can be computed using several algorithms[74].

By minimizing the objective function the optimal fuzzy classification will be achieved to satisfy the conditions in Eqs. (4), (5), and (6). The generalized objective function is given in Eq. (8).

$$J(F, z) = \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_{ij}{}^{\emptyset}\ d_{ij}{}^{2} \tag{8}$$

where *k* is the number of the clusters, *n* is the number of data points. The expression of $d_{i,j}{}^2$ shows the distance between the individual *i* and the class center *j*, $F = [u_{i};_{j}]$ is an n-by-k partition matrix, $u_i;_j$ represents the association degree of membership of the i_{th} object x_i to the j_{th} cluster z_j, $Z = [z_1, z_2, ..., z_k]^T$ is an k-by-m matrix containing the cluster centers. Euclidian distance measure is used. The parameters affecting the results are number of clusters (k) and degree of fuzziness (ϕ). Equation (8) assigns intermediate memberships and solved the problem of intergrades, which are data points between two classes.

2. Methodology and data

In this case study, press economic information, taken from Alarabia.net (2012), a semi-official press agency, and Reuters.com (2012), one of the world's largest international multimedia news agencies, is treated as a source of media noise that influences the value of stocks quoted on the GCC stock market (SMs) for the top market capitalization (mark cap) companies in the banking sector. The period chosen for data collection was October, November, and December 2012. All economic news was categorized into three types of information, neutral, positive, and negative, after which another group, called the most tragic news in the economic sense, was created by selecting this type of news from the negative information group. To determine the most tragic news, the researchers selected news that contained words such as "crisis", "depression", and "collapse" in its title. An indicator of media expansiveness was constructed based on these data[75]. This indicator is a quotient of the number of articles in a chosen information group and the number of all Alarabia.net and Reuters.com information pieces published on the same day (t).

$$BNI_t = \frac{NBN_t}{TNN_t} \cdot 100\% \tag{9}$$

where BNI_t is the bad news indicator, NBN_t is the number of negative headlines, and TNN_t is the total number of headlines. This indicator was calculated daily and provides information on the strength of the negative information obtained from press coverage. Analogous tragic news indicators and a good news indicator ($\mathrm{GNI_t}$) were also calculated.

The second type of data was the daily rates of return of the stock exchange ($\mathrm{R_t}$), β risk coefficient, mark cap, price to earnings (P/E) ratio, price to book value (P/BV) ratio (P/B), and dividend yield (DY) ratio published on KAMCO, Abu Dhabi Securities Exchange, Dubai Financial Market, Kuwait Stock Exchange, Qatar Exchange, Saudi Stock Exchange, Bahrain Bourse, and Muscat Securities corresponding to the same time period[76].

In the present study, we used correlation and regression analysis to identify the behavioral character of the dependency between the analyzed variables. Since this research was conducted over a short period, statistical verification of the significance of the correlations is important. The confidence level was set at (0.05). First, the correlation coefficients between the economic ratios of the chosen securities with the news indicators were calculated. The second step of this research applied the *k*-means clustering algorithm to the bank and energy sectors of the GCC stock markets (86 banks: 25 Islamic, 50 conventional, and 11 mixed; and 19 energy companies) by using the Euclidean distance measure method and then applied fuzzy *c*-means for the same companies. The third step applied the chi square test to the *k*-mean and fuzzy *c*-mean similarity of a number of clusters. The aim of clustering in this study is to get homogenous groups of stock market ratios that have the same reaction to the news.

3. Empirical results

We analyzed the correlation between *GNI*, *BNI*, and *NNI* and the changes in the rates of return of the stock exchange indexes R_t, mark cap, *P/E*, *P/B*, β, and *DY* for the same period (63 trading days) on the GCC stock markets on the banking and energy companies. We observed changes in the correlation coefficients that were the result of companies. We observed that the correlation coefficients were not statistically significant; therefore, we divided the period into sub-periods of 25, 26, 27, 28, 29, and 30 trading days to obtain significant correlation coefficients.

We used *t*-tests to analyze the statistical significance of the relationship between *GNI*, *BNI*, and R_t, mark cap, *P/E*, *P/B*, β, and *DY*. The correlation analysis and *t*-test were used to estimate the optimum period that could give the best reaction of the market movement to the (*BNI* and *GNI*) indicators.

Based on the Euclidean distance method, we determined the optimum period for the stock market ratios, as shown in Table 1.

Table 1. Optimum time period for stock market ratios

Ratio	Rt	Mark cap	P/E	P/B	DY
Best Period (days)	25	28	26	25	27

Source: original data.

Based on these values, we clustered the banking and energy companies using the *k*-means algorithm, with Euclidean distance measure methods. The number of clusters for each ratio to each distance measure method was chosen based on ANOVA (according to values of the SS, between separated clusters, this should be the maximum and minimum distances between the object within the same cluster, and the *p*-value); therefore, we chose, as an example, four clusters instead of three or five clusters for the P/E, using Euclidean distance, as shown in Table 2.

Table 2. Number of clusters of stock market ratios based on Euclidean distance measure

Ratio	Rt	Mark cap	P/E	P/B	DY
No. of Clusters	4	3	4	5	3

Source: authors' calculations.

Depending on the number of *k*-mean clusters, we applied the fuzzy *c*-mean to each stock ratio. Table 3 shows the similarity between the membership of the stocks for *k*-mean and fuzzy *c*-mean for the DY ratio.

Table 3. Similarity between memberships of stocks for *k*-mean and fuzzy *k*-mean for the DY ratio

	FKM			
KM	DY	Cluster 1	Cluster 2	Cluster 3
	Cluster 1	4	15	12
	Cluster 2	1	17	19
	Cluster 3	2	19	16

Source: authors' calculations.

The chi square test of dependency is used to prove that the distribution of stocks clustered by *k*-means and fuzzy *c*-means is completely independent. Thus, choosing the clustering methodology could give us different conclusions. Table 4 shows a summary of values for (chi square, alpha, *p*-value, and correlation coefficient) for stock exchange indexes R_t, mark cap, *P/E*, *P/B*, and *DY*.

Table 4. Summary of statistics (chi square, alpha, *p*-value, and correlation) for stock exchange ratios R_t, mark cap, *P/E*, *P/B*, and *DY*

Ratio	Chi square	Alfa	p-value	Correlation coefficient
R_t	35.3958	3.3251	0.9999	0.1124
Mar Cap	14.6500	0.7107	0.9945	0.0697
P/E	60.7865	3.3251	1.0000	0.1930
P/B	68.4248	7.9616	1.0000	0.1629
DY	3.5100	0.7107	0.5240	0.0167

Source: authors' calculations.

An Intel® Core (TM) i7-2670QM CPU at a 2.20 GHz Workstation with an 8 GB RAM computer was used to conduct the research experiments. The program STATISTICA version 10 was used to data mine the clustering methods.

Conclusions

The purpose of the current study was to compare two clustering methods (*k*-mean and fuzzy *k*-mean) in clustering banking and energy companies and to test the dependency relation of these clustering methods. The results showed the following:

1. The chi square test of dependency was used to prove that the distribution of stocks clustered by *k*-means and fuzzy *c*-means are completely independent. The chi square values were greater than the alpha values; the *p*-values were greater than the level of confidence (0.05), and the correlation was statically insignificant.
2. Choosing the fuzzy *c*-means method to identify groups of homogenous stocks in terms of reactions to news uses news indicators (GNI, BNI, NNI) effectively for diagnosing the market's emotional state.

We suggest using fuzzy *k*-means as a statically proven method of clustering for these variables because they are closely tied to investor behavior, and this method is more flexible than the standard *k*-mean.

Notes

[1] Larose (2005).
[2] Dunham (2002).
[3] Kudyba (2004).
[4] Kumar et al. (2003).
[5] Elmasri, Navathe (2011).
[6] Kumar et al. (2003).
[7] Kumar, Wasan (2010).
[8] Jain, Dubes (1988).
[9] Mirkin (1996).
[10] Nanda et al. (2010).
[11] Anderberg (1973).
[12] Ruspini (1969).
[13] Bezdek (1980).
[14] Huang, Ng (1999).
[15] Bezdek et al. (1984).
[16] Li et al. (2008).
[17] Witten, Eibe (2005).
[18] Singh et al. (2011).
[19] Ibidem.
[20] Nanda et al. (2010).
[21] Ghosh, Liu (2009).
[22] Ruspini (1969).
[23] Bezdek (1981).
[24] Gorsevski et al. (2003).
[25] Gasch, Eisen (2002).
[26] Vimal et al. (2008).
[27] Tan et al. (2006).
[28] Santosh, Nattee (2009).
[29] Koloseni et al. (2013).
[30] Simpson (2008).
[31] Hammoudeh, Choi (2006).
[32] Hertog (2012).
[33] Simpson (2008).
[34] Hertog (2012).
[35] Hammoudeh, Choi (2006).
[36] Setty et al. (2010).
[37] Lo, MacKinlay (1988).
[38] Luo et al. (2010).
[39] Mitchell, Mulherin (1994).
[40] Zielonka (2000).
[41] Carretta et al. (2011).
[42] Andreassen (1987).
[43] Majewski (2009).
[44] Majewski et al. (2012).

[45] Triantaphyllou (2010).

[46] Larose (2005).

[47] Han, Kamber (2006).

[48] Cao et al. (2009).

[49] Zaki, Meira (2013)

[50] Dunham (2002).

[51] Bose, Mahapatra (2001).

[52] Blaiewicz et al. (2003).

[53] Zaki, Meira (2013).

[54] Jain, Dubes (1988).

[55] Mathuriya, Bansal (2012).

[56] Ghosh, Liu (2009).

[57] Singh et al. (2011).

[58] Tan et al. (2006).

[59] Ibidem.

[60] Ramamurthy, Chandran (2011).

[61] Ghosh, Liu (2009).

[62] Madhulatha (2012).

[63] Nikam et al. (2011).

[64] Deza, Deza (2009).

[65] Chiang, Mirkin (2010).

[66] Calinski (1974).

[67] Sugar, James (2003).

[68] Zadeh (1965).

[69] Bezdek et al. (1984).

[70] Marghescu et al. (2010).

[71] Gasch, Eisen (2002)

[72] Vassilios et al. (1999).

[73] Gorsevski et al. (2003).

[74] Bezdek (1981).

[75] Majewski (2009).

[76] KAMCO (2012).

References

Alves, A., Camacho, R. & Oliveira, E. (2004). Inductive Logic Programming for Data Mining in Economics. *The 2nd International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management.* Pisa: University of Porto.

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* New York: Academic Press.

Andreassen, P.B. (1987). On the social psychology of the stock market.Aggreagat attributional effects and the regressivness of prediction. *Journal of Personality and Socioal Psychology*, 53 (3), 490–496.

Bezdek, J.C. (1980). A convergence theorem for the fuzzy ISODATA clustering Algorithms. *IEEE Trans. Pattern Anal. Machine Intell*, 2, 1–8.

Bezdek, J.C. (1981). *Pattern recognition with fuzzy objective function algorithms.* New York: Plenum Press.

Bezdek, J.C., Ehrlich, R. & Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10, 191–203.

Błażewicz, J., Kubiak, W., Morzy, T. & Rusinkiewicz, M. (2003). *Handbook on Data Management in Information Systems.* Springer-Verlag.

Bose, I. & Mahapatra, R.K. (2001). Business data mining – a machine learning perspective. *Information & Management*, 39, 211–225.

Business (10, 11, 12.2012), www.reuters.com/finance/economy.

Bussiness and Technology (10, 11, 12.2012). From AL ARABIA NEWS: http://english.alarabiya.net/index.

Calinski, R.H. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1–27.

Cao, L., Yu, P.S., Zhang, C. & Zhang, H. (2009). *Data Mining for Business Applications.* New York: Springer.

Carretta, A., Farina, V., Martelli, D., Fiordelisi, F. & Schwizer, P. (2011). The impact of corporate governance press news on stock market returns. *European financial management*, 17 (1), 100–119.

Chiang, M.M.-T. & Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification*, 27, 3–40.

Clustering (2012, June 8). From Computer Science 831: Knowledge Discovery in Databases: www2.cs.uregina.ca/~dbd/cs831/notes/clustering/clustering.html (7.03.2013).

Deza, E. & Deza, M.M. (2009). *Encyclopedia of Distances.* Berlin, Heidelberg: Springer-Verlag.

Dunham, M.H. (2002). *Data Mining: Introductory and Advanced Topics.* New York: Prentice Hall.

Elavarasi, S.A., Akilandeswari, J. & Sathiyabhama, B. (2011). A Survey on Partition Clustering Agorithms. *International Journal of Enterprise Computing and Business Systems*, 1, 1–14.

Elmasri, R. & Navathe, S.B. (2011). *Fundamentals of database systems.* Boston, MA: Addison-Wesley.

Fairfield, P.M. (1994). P/E, P/B and the Present Value of Future Dividends. *Financial Analysts' Journal*, 23–31.

Field, A. (2009). *Discovering Statistics Using SPSS.* New Delhi: Sage Publications.

Fridson, M.S. (2011). *Financial Statement Analysis. A Practitioner's Guide.* New Jersey: John Wiley & Sons.

Gasch, A.P., & Eisen, M.B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3, 1–22.

Ghosh, J. & Liu, A. (2009). K-Means. In: W. Xindong, V. Kumar, *The top ten algoritms in Data Mining* (pp. 21–36). Boca Raton, Florida: Taylor & Francis Group.

Gorsevski, P.V., Gessler, P.E. & Jankowski, P. (2003). Integrating a fuzzy k-means classification and a Bayesian approach for spatial prediction of landslide hazard. *Journal of Geographical System*, 223–251.

Hammoudeh, S. & Choi, K. (2006). Behavior of GCC stock markets and impacts of US oil and financial markets. *Research in International Business and Finance*, 20, 22–44.

Han, J. & Kamber, M. (2006). *Data Mining:Concepts and Techniques.* San Francisco: Morgan Kaufmann Publishers.

Hertog, S. (November 2012). *Financial markets in GCC countries: recent crises and structural weaknesses.* Norwegian Peacebuilding Resource Centre.

Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Research Issues on Data Mining and Knowledge Discovery. Cite Seer*, 1–8.

Huang, Z. & Ng, M.K. (1999). A Fuzzy K-Modes Algorithm for Clustering Categorical Data. *IEEE Transactions on Fuzzt Systems*, 7 (4), 446–452.

Investmens Policy. (2013). Calgary.

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice Hall.

KAMCO (10, 11, 12.2012). *Research Reports*, www.kamconline.com (01.2013).

Kudyba, S. (2004). *Managing Data Mining, Advice from Experts.* USA: IT Solutions Series, Idea Group.

Kumar, P. & Wasan, S.K. (2010). Comparative Analysis of k-mean Based Algorithms. *International Journal of Computer Science and Network Security*, 10 (4), 314–318.

Kumar, V., Joshi, M.V., Han, E.-H.S., Tan, P.-N. & Steinbach, M. (2003). High performance data mining. *High Performance Computing for Computational Science – VECPAR 2002*, 111–125.

Larose, D.T. (2005). *Discovering Knowledge in Data (An Introduction to Data Mining).* Hoboken, NJ: John Wiley & Sons.

Levinson, M. (2006). *Guide to Financial Markets* (pp. 145–146). London: The Economist (Profile Books).

Li, M.J., Ng, M.K., Cheung, Y.-M, & Huang, J.Z. (2008). Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters. *IEEE Transactions on Knowledge and Data Engineering*, 20 (11), 1519–1534.

Lo, A.W., & MacKinlay, A.C. (1988). Stock Market Prices Do not Follow Random Walks: Evidence from a Simple Specification Test. *The Review of Financial Studies*, 41–66.

Luo, F., Wu, J. & Yan, K. (2010). A Novel Nonlinear Combination Model Based on Support Vector Machine for Stock Market Prediction. *8th World Congress on Intelligent Control and Automation* (p. 1). Jinan, China: IEEE.

Madhulatha, T.S. (2012). An Overview On Clustering Methods. *IOSR Journal of Engineering*, 2 (4), 719–725.

Majewski, S. (2009). The media and the prices creation in Poland. *International Journal of Management Cases*, 11 (1), 70–77.

Majewski, S., Nermend, K. & Al-augby, S. (2012). Media and Price Creation in Abu Dhabi Security Exchange. *Sientific Papers of the Polish Information Processing Society Sientific Council,University of Szczecin*, 81–93.

Marghescu, D., Sarlin, P. & Liu, S. (2010). Early-Warning Analysis for Currency Crises in Emerging Markets: A Revisit With Fuzzy Clustering. *Intellegent Systems in Accounting, Finance and Management*, 17, 143–165.

Mathuriya, N. & Bansal, A. (2012). Comparison of K-means and means and Back propagation Data Mining Algorithms. *International Journal of Computer Technology and Electronics Engineering*, 151–155.

McBratney, A.B. & De Gruijter, J.J. (1992). A Continuum Approach to Soil Classification by Modified Fuzzy K-means with Extragrades. *Journal of Soil Science*, 43, 159–175.

Mhmoud, A.S. & Ali, S.O. (2013). Application of Principal Component Method and k-me ans clustering algorithm for Khartoum stock Market. *Nature and Science*, 108–112.

Mirkin, B.G. (1996). *Mathematical classification and clustering.* Dordrecht: Kluwer Academic Publishing.

Mitchell, M.L. & Mulherin, J.H. (1994). The impact of public information on the stock market. *The Journal of Finance*, 49 (3), 923–950.

Mooi, E. & Sarstedt, M. (2011). *A Concise Guide to Market Research The Process, Data, and Methods Using IBM SPSS Statistics.* Berlin: Springer-Verlag.

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

Nikam, V., Kadam, V.J. & Meshram, B.B. (2011). Image Compression Using Partitioning Around Medoids Clustering Algorithm. *International Journal of Computer Science Issues*, 8, 6 (1), 399–401.

Ramamurthy, B. & Chandran, K.R. (2011). CBMIR: Shape-BasedImage Retrieval Using Canny Edge Detection and K-Means Clustering Algorithms for Medical Images. *International Journal of Engineering Science and Technology*, 3, 1870–1877.

Ruspini, E.R. (1969). A new approach to clustering. *Inform. Control*, 19, 22–32.

Santosh, K.C. & Nattee, C. (2009). A Comperhensive Survey on On-line Handwriting Recgnition Technology and Its Real Application to The Nepalese NaturalL Handwriting. *Kathmandu University Journal of Science, Engineering and Technology*, 5 (1), 31–55.

Setty, D.V., Rangaswamy, T.M. & Subramanya, K.N. (2010). A Review on Data Mining Applications to the Performance of Stock Marketing. *International Journal of Computer Applications*, 1 (3), 24–34.

Shiller, R.J. (2001). *Irrational Exuberance.* New York: Brodway Books, p. 95.

Shrestha, D. (2009). Text Mining with Lucene and Hadoop: Document Clustering With Feature Extraction. *Research Degree Thesis*. Wakhok University.

Simpson, J. (2008). Financial Integration In The GCC Stock Markets: Evidence From The Early 2000s Development Phase. *Journal of Economic Cooperation*, 1–28.

Singh, K., Malik, D. & Sharma, N. (2011). Evolving limitations in K-means algorithm in data mining and their removal. *International Journal of Computational Engineering & Management*, 12, 105–109.

StatSoft (2013). *StatSoft Electronic Statistics Textbook*. From Introduction to ANOVA/MANOVA: www.thefullwiki.org/Analysis_of_variance.

Sugar, C.A. & James, G M. (2003). Finding the number of clusters in a data set :An information theoretic approach. *Journal of the American Statistical Association*, 98 (463), 750–763.

Tan, P.-N., Steinbach, M. & Kumar, V. (2006). *Introduction to Data Mining.* Pearson Addison Wesley.

Thompson, B. (2002). "Statistical," "Practical," and "Clinical": How Many Kinds of Significance Do Counselors Need to Consider? *Journal of Counseling & Development*, 80, 64–71.

Triantaphyllou, E. (2010). *Data Mining and Knowledge Discovery Via Logic-Based Methods.* New York: Springer.

Vassilios, C., Adrian, G.B. & Ioannis, P. (1999). Multimodal Decision-Level Fusion for Person Authentication. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 674–680.

Vimal, A., Valluri, S.R. & Karlapalem, K. (2008). *International Conference on Management of Data COMAD 2008.* Mumbai: Computer Society of India.

Wei, Y. (2005, May). *Approximation To K-means Clustering*. Hamilton, Ontario, Canada: McMaster University.

Witten, I.H. & Eibe, F. (2005). *Data Mining Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann Publishers is an imprint of Elsevier.

Xu, R. & II, D.W. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neura Networks*,16 (3), 645–678.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8 (3), 338–353.

Zaki, M.J. & Jr., W.M. (2013). *Data Mining and Analysis:Fundamental Concepts and Algorithms.* Draft copy: Cambridge University Press.

Zielonka, P. (2000). *Biased Judgement on What Moves Stock Prices.* Warsaw: Institute of Philosophy and Sociology Polish Academy of Sciences.