4D International Journal of Management and Science
ISSN No: 2250-0669
@4D Crossconnect.com, Inc, 2012
www.4dinternationaljournal.com
Vol. 3, Issue2, July 2013

ANALYSIS OF STOCKS USING CLUSTERING TECHNIQUE

Ms. Vandana Gupta[1] and Dr.Vikrant Agarwal[2]

ABSTRACT

The rapid worldwide increase in the data available leads to the difficulty for analyzing those data. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Thus, a proper data mining approach is required to organize those data for better understanding. Clustering is one of the standard approaches in the field of data mining. The main approach is to organize a dataset into a set of clusters, which consists of "similar" data items, as calculated by some distance function. Clustering is mainly applied in document categorization, scientific data analysis, and customer/market segmentation. Data clustering is considered as a key data mining technique for knowledge discovery. There are various clustering techniques available in the literature; but all the existing algorithms misclassify the data when large data are involved. This drags the researchers to provide a better clustering technique to deal with large data. However the present study is based on secondary data. This paper provides various available data clustering techniques with its merits. It will also depict the usage of data clustering in the field of Stock Analysis.

Keywords: Data Mining, Clustering, PE Ratio, cluster analysis, EPS (Earnings Per Share), MPS (Market Price per Share)

INTRODUCTION:

"Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns"

The growth and development in sensing and storage technology and drastic development in the applications such as internet search, digital imaging, and video surveillance have generated many

[1] ***Assistant professor, Department of Management, Kasturi Ram College of Higher Education Narela, Delhi. Affiliated by Guru Gobind Singh Indraprastha University Delhi.***

[2] ***Associate Professor, Department of Management, Kasturi Ram College of Higher Education Narela, Delhi. Affiliated by Guru Gobind Singh Indraprastha University Delhi.***

high-volume, high-dimensional data sets. It has been calculated that the digital universe consumed around 281 Exabyte's (One Exabyte is ~1018 bytes or 1,000,000 terabytes) in 2007, and it is estimated to be ten times higher by 2011. As the majority of the data are stored digitally in electronic media, they offer high prospective for the development of automatic data analysis, classification, and retrieval approaches. In addition to the growth in the amount of data, the variety of available data (text, image, and video) has also increased. Inexpensive digital and video cameras have made available huge records of images and videos. The usage RFID tags has increased mainly due to their low cost and small size has resulted in the use of millions of sensors that transmit data regularly. Several terabytes of new data has been generated due to E-mails, blogs, transaction data, and billions of Web pages. Majority of these data streams are unstructured, which makes them very difficult to analyze.

Methodology of the Paper:

The study is based on secondary data. Secondary data had been and writings of various authors in the stream of industry, academician, and research. The Journals and books have been referred were described in the bibliography

Data Warehouse:

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

Data Clustering:

Data mining, the extraction of hidden descriptive or predictive information from large databases, is a powerful new technology with great potential to help companies and data analysts focus on the most important information in their data repositories. Class identification, i.e. the grouping of the objects of a database into meaningful subclasses such that similarity of objects in the same group is maximized and similarity of objects in different groups is minimized, is called clustering. Representing data by fewer clusters necessarily loses certain fine details but achieves simplification. Thus the objects are clustered or grouped based on the principle of "maximizing the interclass similarity and minimizing the interclass similarity". Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. There exist many clustering algorithm from partition based, model based, non parametric density estimation based methods, graph theoretical based, to empirical and hybrid approaches. They all are underlying some concept about data organization and cluster characteristics to find interesting patterns or clusters in the given dataset. The partition-based algorithms are more suitable for clustering large datasets as they can be easily implemented and are most efficient one in terms of the execution time.

The most commonly used algorithm is K-Means clustering which is partitioning based algorithm. As this is the simplest algorithm but there are some limitations with this algorithm like the number of clusters (K) needs to be determined beforehand, means the algorithm is sensitive to an initial seed selection, it cannot be used for clustering problems whose results cannot fit in main memory,

which is the case when data set has very high dimensionality or desired number of clusters is too big. So it is proposed to develop an efficient algorithm which tries to remove some or all the above given limitations.

Cluster Analysis:
Cluster analysis is a collection of statistical methods, which identifies groups of samples that behave similarly or show similar characteristics. The key idea is to identify classifications of the objects that would be useful for the aims of the analysis. Cluster analysis is an important human activity. Early in childhood, one learns how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. The data stream model has recently attracted attention for its applicability to numerous types of data, including telephone records, Web documents, and large dataset. Clustering is an unsupervised classification as it has no predefined classes and few of its typical application include: As a stand-alone tool to get insight into data distribution. As a pre-processing step for other algorithms .A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method can be measured by its ability to discover some or all of the hidden patterns. For analysis of such data, the ability to process the data in a single pass, or a small Number of passes, while using little memory, is crucial. In this paper, K-Means Clustering algorithm is dynamically implemented with the cosine distance similarity which also helps in achieving global optimality.

Cluster
A cluster is a set of objects in which each object is closer to every other object. The types of clusters includes; well-separated clusters, prototype-based clusters, graph-based clusters, density-based clusters, and shared-property (conceptual clusters). The important characteristics of cluster include; data distribution, shape, different size, different density, poorly separation, relationships among clusters, and subspace

Clustering
Clustering is a class or group of objects that share common characteristics and play an important role in how people analyze and describe the world. It is dividing the objects into groups (clustering) and assigning particular objects to these groups (classification). Clustering aims to find useful groups of objects, where usefulness is defined by the goals of the data analysis. An entire collection of clusters is commonly referred to as clustering. There are three types of clustering namely; hierarchical versus partitional, exclusive versus overlapping versus fuzzy and complete versus partial. A partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each object is exactly in one subset. Partitional algorithms typically determine all clusters at once. The partitional clustering can be obtained by taking any member of that sequence.

Cluster Analysis
It groups data objects based only on information found in the data that describes the objects and their relationships. It is also a class or group of objects that share common characteristics and play an important role in how people analyze and describe. The goal is that the objects within a group be similar to one another and different from the objects in the other groups. The greater the

similarity within a group and greater the difference between groups, the better or more distinct is the clustering. Cluster analysis is sometimes referred to as unsupervised classification. When the term classification is used without any qualification within data mining, it typically refers to supervised classification.

K-means clustering algorithm

K-means method is widely used due to rapid processing ability of large data. K-means clustering proceeds in the following order. Firstly, K number of observations is randomly selected among all N number of observations according to the number of clusters. They become centers of initial clusters. Secondly, for each of remaining N–K observations, find the nearest cluster in terms of the Euclidean distance with respect to xi ¼ ðxi1; xi2;…; xip;…; xiPÞ After each observation is assigned the nearest cluster, re compute the center of the cluster. Lastly, after the allocation of all observation, calculate the Euclidean distance between each observation and cluster's center point and confirm whether it is allocated to the nearest cluster or not.

Price-Earnings Ratio - P/E Ratio Mean?

A valuation ratio of a company's current share price compared to its per-share earnings.

Calculated as:

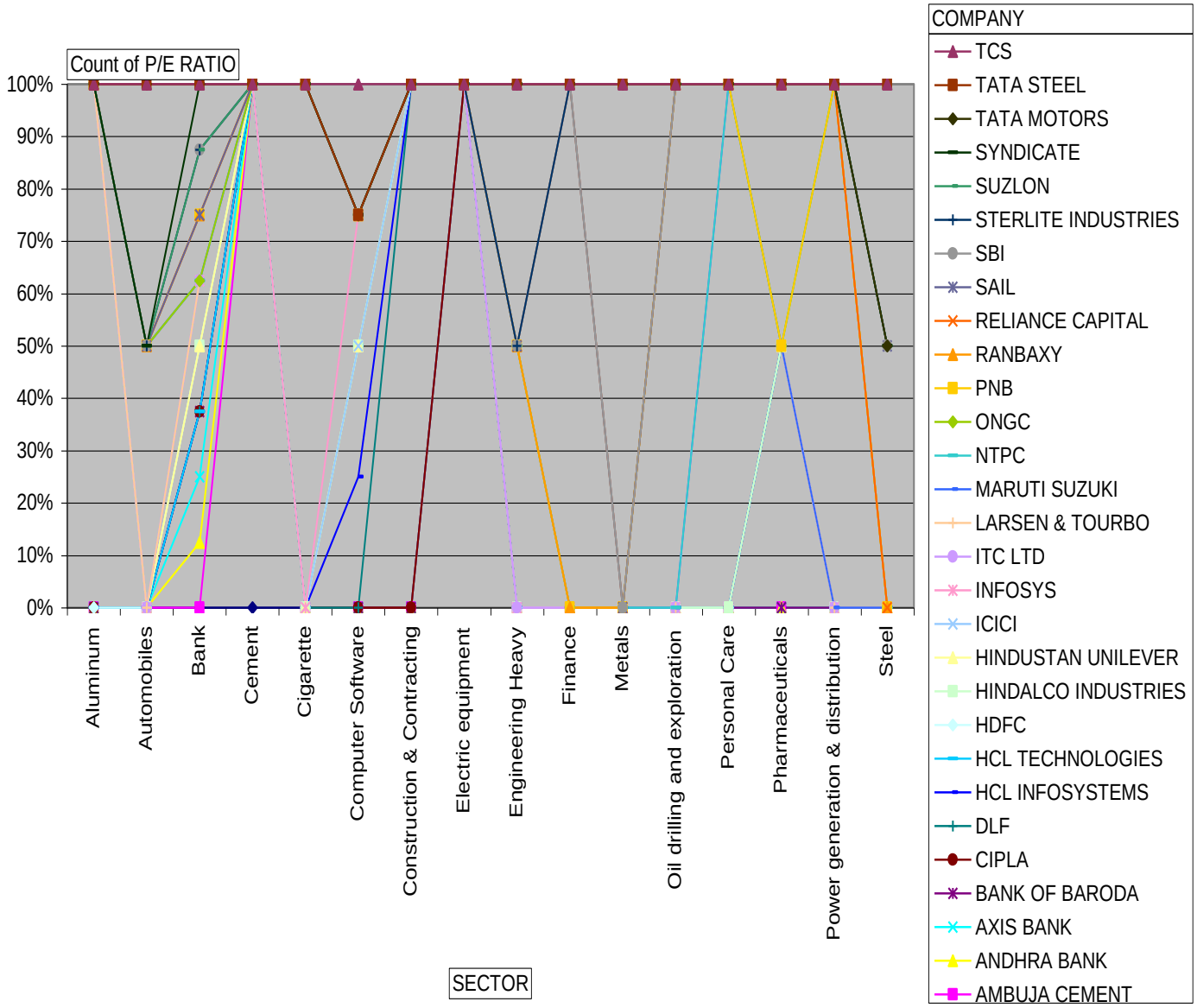
$$\text{Price-Earnings Ratio} = \frac{\text{Market Value per Share}}{\text{Earnings per Share (EPS)}}$$

Also sometimes known as "price multiple" or "earnings multiple".

Price/Earning ratio gives you fair idea of how a company's share price compares to its earnings. If the price of the share is too much lower than the earning of the company, the stock is undervalued and it has the potential to rise in the near future. On the other hand, if the price is way too much higher than the actual earning of the company and then the stock is said to overvalued and the price can fall at any point. The most commonly used guide to the relationship between stock prices and earnings is the P/E ratio. P/E ratio is volatile and may fluctuate considerably. In general, a high P/E suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E. However, the P/E ratio doesn't tell us the whole story by itself. It's usually more useful to compare the P/E ratios of one company to other companies in the same industry, to the market in general or against the company's own historical P/E. It would not be useful for investors using the P/E ratio as a basis for their investment to compare the P/E of a technology company (high P/E) to a utility company (low P/E) as each industry has much different growth prospects.

P/E RATIO (Analysis is done for the 1st week of May 2012)

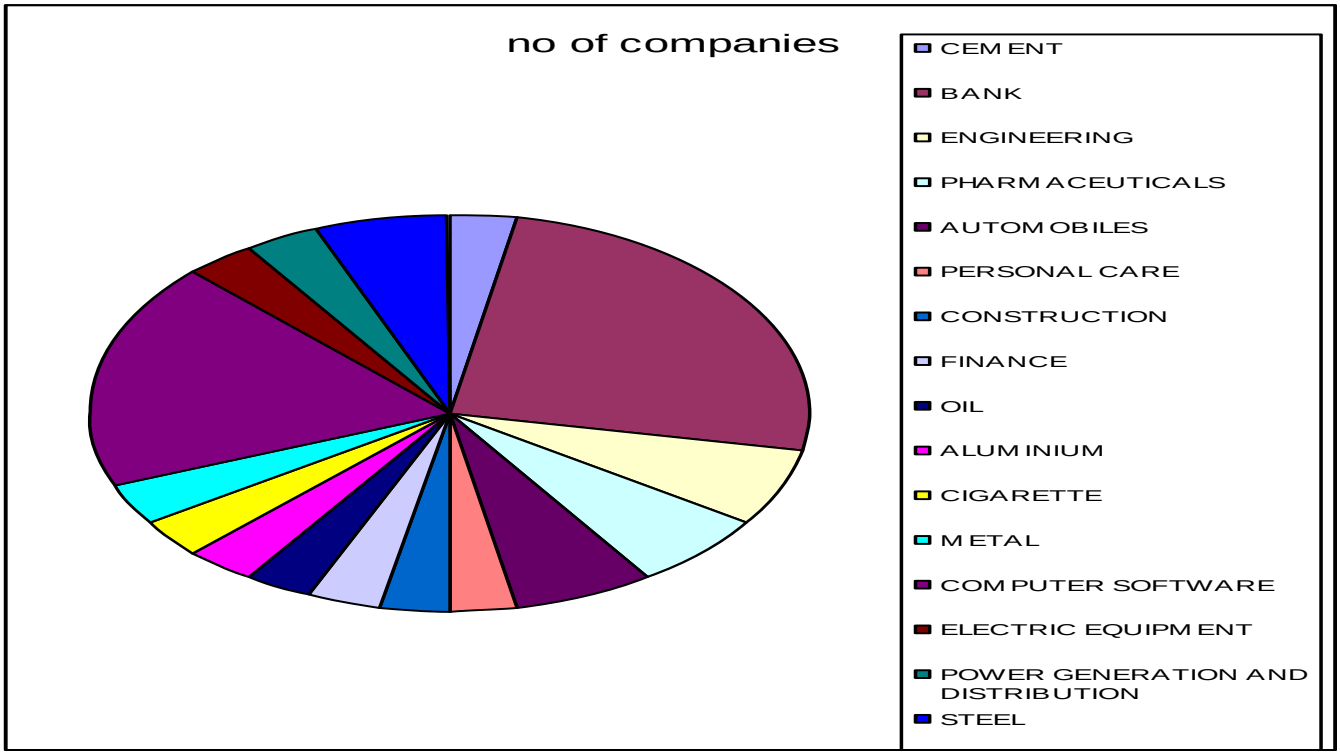
COMPANY	SECTOR	P/E RATIO=MV/EPS
ABB	Electric equipment	160.78
AXIS BANK	Bank	15.38
AMBUJA CEMENT	Cement	15.96
ANDHRA BANK	Bank	5.94
BANK OF BARODA	Bank	8.16
CIPLA	Pharmaceuticals	26.85
DLF	Construction & Contracting	31.24
HINDUSTAN UNILEVER	Personal Care	30.93
HINDALCO INDUSTRIES	Aluminum	15.82
HCL INFOSYSTEMS	Computer Software	8.10
HCL TECHNOLOGIES	Computer Software	32.61
HDFC	Bank	30.32
ICICI	Bank	23.70
INFOSYS	Computer Software	23.52
ITC LTD	Cigarette	31.19
LARSEN & TOURBO	Engineering Heavy	27.92
MARUTI SUZUKI	Automobiles	14.83
NTPC	Power generation & distribution	17.13
ONGC	Oil drilling and exploration	12.57
PNB	Bank	8.15
RANBAXY	Pharmaceuticals	102.20
RELIANCE CAPITAL	Finance	63.67
SAIL	Steel	11.04
SUZLON	Engineering Heavy	-
STERLITE INDUSTRIES	Metals	38.64
SYNDICATE	Bank	6.50
SBI	Bank	18.99
TATA MOTORS	Automobiles	36.47
TATA STEEL	Steel	8.01
TCS	Computer Software	29.68


Count of P/E RATIO
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Aluminum
Automobiles
Bank
Cement
Cigarette
Computer Software
Construction & Contracting
Electric equipment
Engineering Heavy
Finance
Metals
Oil drilling and exploration
Personal Care
Pharmaceuticals
Power generation & distribution
Steel
SECTOR
COMPANY
TCS
TATA STEEL
TATA MOTORS
SYNDICATE
SUZLON
STERLITE INDUSTRIES
SBI
SAIL
RELIANCE CAPITAL
RANBAXY
PNB
ONGC
NTPC
MARUTI SUZUKI
LARSEN & TOURBO
ITC LTD
INFOSYS
ICICI
HINDUSTAN UNILEVER
HINDALCO INDUSTRIES
HDFC
HCL TECHNOLOGIES
HCL INFOSYSTEMS
DLF
CIPLA
BANK OF BARODA
AXIS BANK
ANDHRA BANK
AMBUJA CEMENT

CLUSTERING OF COMPANIES ON THE BASIS OF P/E RATIO

SECTOR	P/E RATIO <1	P/E RATIO <10	P/E RATIO 10-20	P/E RATIO>20	P/E RATIO>30
CEMENT			AMBUJA CEMENT		
BANK		ANDHRA BANK, BANK OF BARODA, PNB,SYNDICATE	AXIS BANK, ,SBI	ICICI	HDFC
ENGINEERING	SUZLON			LARSEN & TOURBO	
PHARMACEUTICALS				CIPLA	RANBAXY
AUTOMOBILES			MARUTI SUZUKI		TATA MOTORS
PERSONAL CARE					HINDUSTAN UNILEVER
CONSTRUCTION					DLF
FINANCE					RELIANCE CAPITAL
OIL			ONGC		
ALUMINIUM			HINDALCO INDUSTRIES		
CIGARETTE					ITC LTD
METAL					STERLITE INDUSTRIES
COMPUTER SOFTWARE		HCL INFOSYSTEMS		TCS, INFOSYS	HCL TECHNOLOGIES
ELECTRIC EQUIPMENT					ABB
POWER GENERATION AND DISTRIBUTION			NTPC		
STEEL		TATA STEEL	SAIL,		

NUMBER OF COMPANIES WHICH ARE TAKEN FOR THE ANALYSIS FROM VARIOUS SECTORS.



FINANCIAL DATA ANALYSIS BY DATA MINING

Data mining of financial data has proven to be very effective and very Two natural problems arise from industrial categories: classification and association. We wish to develop methods that can answer the following two questions:

1. Can we determine a stock's industrial category given a historical record of the stock's prices?
2. How are the movements in stock prices across the various industries associated?

We describe association rule mining to answer the second question.

Time-Series Clustering

Any attempt at clustering the stocks is based on the following crucial assumption:

A time-series clustering will be valid if and only if the price fluctuations of stocks within a group are correlated, but price fluctuations of stocks in different groups are uncorrelated or not as strongly correlated.

This statement can be interpreted two ways. The forward version of this statement says that if we obtain a good clustering with respect to our distance measure, then stocks tend to move as a group. This implication in and of itself would be useful to the financial world. Moreover, clustering statistics can quantize to what extent stocks move as a group. External clustering statistics such as entropy, purity, and cohesion tell us how closely stocks within an industry resemble each other. Internal statistics such as separation and the silhouette coefficient can tell us to what degree the industries' behaviors are separate from each other. The reverse version of the assumption says that if stocks tend to move as a group, then we should be able to obtain a sensible clustering. Without this assumption, any clustering would be meaningless even if the clustering was perfect with respect to the chosen distance measure. Also, we need to see differentiation between the groups or else our clustering will fail. This also points out a possible application of financial clustering. Since our clustering is based solely on the historical price record, the clustering will determine which group the stock most behaves like, which is not necessarily the group the NYSE categorized it as. The stock market is based on perception, not reality. Perception has become even more important in recent years with the advent of on-line trading, which released a flood of anxious and often ill-informed day traders.

Also, by examining class statistics such as purity and entropy, we can say to what degree a stock follows each group. Clustering could also be very helpful in analyzing the time series of a portfolio including several stocks. The behavior of an investment portfolio is not necessarily determined by the stock that makes up the largest monetary share of the investment. Clustering a portfolio with stock data could identify which industrial groups have the greatest influence on the portfolio.

CONCLUSION

The P/E tells us what the market is willing to pay for the company's earnings. If a stock has a P/E of 15, that means the market is willing to pay 15 times its earnings for the stock. Companies with good growth potential will have a higher P/E because investors are willing to pay a premium for future profits.

High-risk companies will typically have low P/Es, which means the market is not willing to pay a high price for risk. High P/E suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E. It's usually more useful to compare the P/E ratios of one company to other companies in the same industry, to the market in general.

According to the above clustering and analysis it is observed that high P/E ratio is observed in pharmaceuticals, personal care, bank, automobiles and computer software and within banking industry we can conclude that HDFC is ruling with the highest P.E Ratio however a very low P.E Ratio is observed in engineering sector with the presence of one company i.e salon with a very low P/E Ratio i.e. less than one.

References

1. Traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications, 36(9), 11772–11781.

2. Chang, P.-C., & Lai, C.-Y. (2005). A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting. Expert Systems with Applications, 29, 183–192.

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

4. Chiu, C.-Y., Chen, Y.-F., Kuo, I. T., & He, C. K. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. Expert Systems with Applications, 36, 4558–4565.

5. Delibasis, K. K., Mouravliansky, N., Matsopoulos, G. K., Nikita, K. S., & Marsh, A. (1999). MR functional cardiac imaging: Segmentation, measurement and WWW based visualisation of 4D data. Future Generation Computer Systems, 15(2), 185–193.

6. Fernandez, L. (2005). A diversified portfolio: Joint management of non-renewable and renewable resources offshore. Resource and Energy Economics, 27, 65–82.

7. Isa, D., Kallimani, V. P., & Lee, L. H. (2009). Using the self organizing map for clustering of text documents. Expert Systems with Applications, 36, 9584–9591.

8. Jo, H., & Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. Expert Systems with Application, 11(4), 415–422.

9. Kasturi, J., Acharya, R., & Ramanathan, M. (2003). An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics, 19, 449–458.

10. Kim, K.-j., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. Expert Systems with Applications, 34, 1200–1209.

11. Krazanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum of squares clustering. Biometrics, 44, 23–34.

12. Kuo, R. J., Wang, H. S., Hu, T.-L., & Chou, S. H. (2005). Application of ant K-means on clustering analysis. Computers and Mathematics with Applications, 50, 1709–1724.

13. Markowitz, H. (1952). Portfolio selection. Journal of Finance, 7, 77–91.

Michaud, P. (1997). Clustering techniques. Future Generation Computer System, 13(2), 135–147.

14. K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174, 1742–1759.

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

16. Oh, K. J., Kim, T. Y., & Min, S. (2005). Using genetic algorithm to support portfolio optimization for index fund management. Expert Systems with Applications, 28, 371–379.

17. Östermark, R. (1996). A fuzzy control model (FCM) for dynamic portfolio management. Fuzzy Sets and Systems, 78, 243–254.

18. Ozkan, I., Türks_en, I. B., & Canpolat, N. (2008). A currency crisis and its perception with fuzzy C-means. Information Sciences, 178, 1923–1934.

19. Prieto, M. S., & Allen, A. R. (2009). Using self-organising maps in the detection and recognition of road signs. Image and Vision Computing, 27, 673–683.

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

21. Shu, G., Zeng, B., Chen, Y. P., & Smith, O. H. (2003). Performance assessment of kernel density clustering for gene expression profile data. Comparative and Functional Genomics, 4(3), 287–299.