Application of Principal Component Method and k-me ans clustering algorithm for Khartoum stock Market

Abdalla Suliman Mhmoud [1], Sharaf Obaid Ali [2]

[1.] Department of Mathematics, College of Arts and Sciences, Taif University, kingdom of Saudi Arabia
Department of Statistics, College of Economics and Political Sciences ,Omdurman Islamic University, Sudan
[2.] Department of Mathematics, College of Sciences, Shaqra University, kingdom of Saudi Arabia
College of Computer Science, Alzaeim alazhari University, Sudan
abdallsuli@hotmail.com

Abstract: This paper proposes a method in which the high dimensional data is reduced through Principal Component Analysis and then bisecting k-means clustering is performed on the reduced data. K-means technique of data mining was discussed and applied to maintain suitable way for selecting cluster from the huge datasets in order assist investors to have a proper guidance for making portfolio. The study shows that low price, high price, and close price are major variables that determine the trend of the share price in Khartoum market exchange. K-means results clustered the financial sectors in three homogenous clusters, most of objects in clusters consist of objects from the same sector.

[Abdalla Suliman Mhmoud, Sharaf Obaid Ali. **Application of Principal Component Method and k-means clustering algorithm for Khartoum stock Market.** *Nat Sci* 2013;11(7):41-44]. (ISSN: 1545-0740). http://www.sciencepub.net/nature. 8

Keywords: cluster, k-means cluster, stock exchange, classify, portfolio

1. Introduction:

Financial market is a complex, non stationary, noisy, chaotic, nonlinear and dynamic system but it does not follow random walk process, (Lo & Mackinlay, 1988; Deng, 2006). Therefore, predictions of stock market price and its direction are quite difficult. In response to such difficulty, clustering k-means techniques have been introduced and applied for financial prediction. Most of the studies have focused on the accurate forecasting of the value of stock price. Stock of exchange is a form of exchange which provides services for stock brokers and traders to trade stocks, bounds, and other securities. Stock exchange also provide facilities for issue and redemption of securities and other financial instruments, and capital events including the payment of income dividends.

Generally, they believes that there is an opinion about stock market like high risk and high returns. Even though we have a huge number of potential investors, only very few of them are invested in the stock market. The main reason is the inability of risk taking skill of investors. Though get low returns, they want to save their money. One important reason for this problem is that, they not have a proper guidance for making their portfolio. (3)

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information aboutthe market behavior for investment decisions. Nowadays, instead of a single method, traders need to use various forecasting techniques to gain multiple signals and more information about the future of the market. Existing clustering algorithm face difficulty in handling multidimensional data The inherent scarcity of the points makes multidimensional data a challenge for data analysis. Several attempts were made by researchers for improving the performance of the k-means clustering algorithm. Typically, the dimensionality reduction is accomplished by applying techniques from linear algebra or statistics such as Principal Component Analysis(6).

High dimensional data is phenomenon in real-world data mining applications. Developing effective clustering methods for high dimensional dataset is a challenging problem due to the curse of dimensionality. Usually k-means clustering algorithm is used but it results in time consuming, computationally expensive and the quality of the resulting clusters depends on the selection of initial centroid and the dimension of the data. The accuracy of the resultant value perhaps not up to the level of expectation when the dimension of the dataset is high because we cannot say that the dataset chosen are free from noisy and flawless. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be pre-processed by an efficient dimensionality reduction method. This paper proposes a method in which the high dimensional data is reduced through Principal Component Analysis and then bisecting k-means clustering is performed on the reduced data where there is no initialization of the centroids.(5)

2- Literature review:

April Kerby & James Lawrence (2003) introduced questions such as; is there a method to predict the stock market? What factors determine if a company's stock value will rise or fall in a given year? Using the multivariate statistical methods of principal component analysis and discriminate analysis, they aimed to determine an accurate method for classifying a company's stock as a good or a poor investment choice. Additionally, they will explore the possibilities for reducing the dimensionality of a complex financial and economic dataset while maintaining the ability to account for a high percentage of the overall variation in the data.

Dr. G. Manoj Someswar, B. Satheesh, G. Vivekanand (2012) apply a pair wise clustering approach to the analysis of the Dow Jones index companies, in order to identify similar temporal behavior of the traded stock prices. The objective of this attention is to understand the underlying dynamics which rules the company's stock prices. In particular, it would be useful to find, inside a given stock market index, groups of companies sharing a similar temporal behavior.

S.R. Nanda, B. Mahanty, M.K. Tiwari (2010) presented in their paper a data mining approach for classification of stocks into clusters. After classification, the stocks could be selected from these groups for building a portfolio. It meets the criterion of minimizing the risk by diversification of a portfolio. The clustering approach categorizes stocks on certain investment criteria. They have used stock returns at different times along with their valuation ratios from the stocks of Bombay Stock Exchange for the fiscal year 2007–2008. Results of their analysis show that K-means cluster analysis builds the most compact clusters as compared to SOM and Fuzzy C-means for stock classification data. Then, they selected stocks from the clusters to build a portfolio, minimizing portfolio risk and compared the returns with that of the benchmark index.

Anthony J.T. Lee, et al (2010) assumed an effective clustering method, Hierarchical agglomerative and Recursive K-means clustering (HRK) to predict the short-term stock price movements after the release of financial reports. The proposed method consists of three phases. First, they converted each financial report into a feature vector and use the hierarchical agglomerative clustering method to divide the converted feature vectors into clusters. Second, for each cluster, they recursively applied the K-means clustering method to partition each cluster into sub-clusters so that most feature vectors in each sub cluster belong to the same class. Then, for each sub-cluster, they chose its centroid as the representative feature vector. Finally, they employed the representative feature vectors to predict the stock price movements.

Harinath Kopeti (2010) showed that, the share price moving of each individual company forms a time series graph with different patterns. To pick up the right companies to buy, the investors need to identify and understand all the patterns of the share price moves and choose the share prices with the promising pattern. The objective of this project is to classify the stock prices into different groups based on their price moving patterns using clustering algorithms (Simple K-means clustering algorithm, Two Step and Hierarchical clustering algorithm).

Ehsan Hajizadeh*, Hamed Davari Ardakani and Jamal Shahrabi (2010) showed that one of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. Potential significant benefits of solving these problems motivated extensive research for years. Recent research suggests PCA data reduction can assist in market or data segmentation. Principal component analysis (PCA) provides a way to reduce large datasets to essential variables. It turns out to be more efficient, in many cases, to conduct segmentation analysis on this reduced dataset, rather than the total dataset.

Tajunisha1 and Saravanan (2011) proposed a method to make the algorithm more effective and efficient by using PCA and modified k-means. In this paper, the researchers have used Principal Component Analysis as a first phase to find the initial centroid for k-means and for dimension reduction and k-means method is modified by using heuristics approach to reduce the number of distance calculation to assign the data-point to cluster. By comparing the results of original and new approach, it was found that the results obtained are more effective, easy to understand and above all, the time given to process the data was substantially reduced.

Sakthi and Thanamani (2011) proposed an effective determination of initial centroids in k-means clustering using kernel principal component method. The initial centroids in k-means is generating randomly, the accuracy and time complexity affected is highly affected because of highly dimensionality of data. Hence, the principal technique for data reduction is used in their proposed algorithm.

3-Methodology:

3-1 Principal component method**:**

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal

transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (**WIKIPEDIA).** Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data. K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables(4). Combining PCA and k-means, helps improve visualization of data. The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set (11).

3-2 Clustering:

Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups (4).

3-3 The basic K-means Algorithm

The K-means algorithm, probably the first one of the clustering algorithms proposed, is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, and then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Although K-means has the great advantage of being easy to implement, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success.

If instead of the Euclidean distance the 1-norm distance is used:

$$d(x,y)=\sum_{i=1}^{n}\|y_i-x_i\| \qquad \text{.....(1)}$$

a variation of K-means is obtained and it is called K-median. The authors claim that this variation is less sensitive to outliers than traditional K-means due to the characteristics of the 1-norm. The algorithm is as follows:

1. Select k objects as initial centers;
2. Assign each data object to the closest center;
3. Recalculate the centers of each cluster;
4. Repeat steps 2 and 3 until centers do not change;

4- Experiment results:

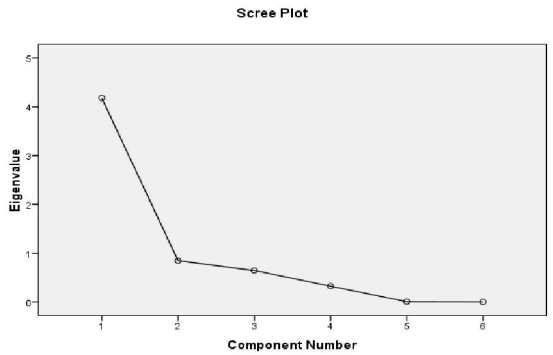
In this paper, the researchers gathered the financial data from Khartoum stock market database. We focused on the listed financial institutions in the market. There are four sectors in the market which are bank sector consist of 7commercial banks, companies sector (9 companies), services sector(9), and government bonds sector (16) issued. Monthly financial report was collected released from (Jan.2.2011 to Dec.31.2011). Besides , the daily predictors variables that determine the indices of the market, which are, number of shares, exchange size, open price, low price, high price, close price and the number of contracts executed by the financial institution. In this paper two different data mining techniques were applied, PCA method was applied in order to reduce dimensionality of data see table(1), and then apply K-means algorithm technique in order to construct a sample of homogenous clusters based on the results of PCA. We find that the major variables as PCA indicate that determined the share price of a sector in Khartoum market are open price, high price, and close price. By application of k-means algorithm see table (2) construct three clusters each of which consist of financial institution similar in their behavior in the market , for example investment in company 3 not similar as invest in company 1 because each of them belong to different cluster, so we differentiate these sectors according to PCA results.

Table (1) PCA for the four sectors

	Component for sector 1	Component for sector 2	Component for sector 3	Component for sector 4
close	.992	.968	.980	.937
low	.987	.960	.965	.916
high	.979	.954	.964	.906
open	.958	.905	.745	.612
exchange	.685	-.615-	-.134-	.368

Table (2) Cluster member ship for each of the financial institution

sector	Cluster(1)	Cluster(2)	Cluster(3)	Number of object in a sector
banks	0	1,2,3,4,5,6,7	0	7
companies	3,5	2,4,6,7,9	1,8	9
services	3,5,6,7,8	1,2,4	9	9
Governmental bonds	1,6,7,8,12,13,14	4,5,11	2,3,9,10,15,16	16
total	14	18	9	41



5-Conclusion

In this study, daily data was collected from Khartoum stock market, for a period of one year. This data covers four financial sectors, each of which consists of different financial institution listed in the market. Data is composed of several variables the market index. This work describes the modern finance finding efficient way to summarize and visualize the stock market data to give individuals useful information about the behavior for investment market decision and the data mining classification for stock markets . In this paper, Principal Component Analysis was applied in the first phase to reduce data dimensionality and to determine the most important variables affecting on the share price changes. We find that the major variables as PCA indicate that determined the share price of a sector in Khartoum market are open price, high price, and close price. K-means algorithm was applied to divide the market for different groups (clusters), each of them consist of homogenous objects. And these clusters give the investor guide to decide in which cluster to invest. Appling this algorithm leads to a natural partition of the data, as companies belonging to the same industrial branch are often grouped together. This algorithm is applied in order to identify similar temporal behavior of the traded stock prices. The identification of clusters of companies of a given stock market index can be exploited in the portfolio optimization strategies.

Reference:

[1] Anthony J.T. Lee, Ming-Chih, Rung-Tai Kao, Lin, Kuo-Tay Chen. An effective clustering approach to stock market prediction. Department of Information Management, National Taiwan University, 2010.

[2] April Kerby James Lawrence, A Multivariate Statistical Analysis of Stock Trends, Alma College Miami University, Alma, MI Oxford, OH. 2003.

[3] B. Uma Devi D. Sundar Dr.P. Alli , A Study on Stock Market Analysis for Stock Selection – Naïve Investors' Perspective using Data Mining Technique , International Journal of Computer Applications– Volume 34– No.3, November 2011.

[4] D.Napoleon and S.Pavalakodi, . A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set. International Journal of Computer Applications– .Volume 13– No.7, January 2011.

[5] Dr.G.Manoj Someswar, B. Satheesh, G.Vivekanand. Finance Mining – Analysis Of Stock Market Exchange For Foreign Using Classification Techniques. International Journal of Engineering Research and Applications (IJERA) ISSN: www.ijera.com Vol. 2, Issue 4, June-July 2012,

[6] Ehsan Hajizadeh, Hamed Davari Ardakani and Jamal Shahrabi. Application of data mining techniques in stock markets, Journal of Economics and International Finance Vol. 2(7), ISSN 2006-9812 ©2010 Academic Journals.

[7] Harinath Kopeti, Pattern classification of stock moving, A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences , School of Computer Science 7 HARINATH KOPETI 7562567, 2010.

[8] M.Sakthi and Dr. Antony Selvadoss Thanamani.An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (3),2011, 959.

[9] S.R. Nanda, B. Mahanty, M.K. Tiwari, (2010), Clustering Indian stock market data for portfolio management,Indian Institute of Technology, Expert Systems with Applications 37 (2010)

[10] Tajunisha1 and Saravanan. An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means.Coimbatore, Tamilnadu, International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011.

5/5/2013