

An Analysis of Performance of Multidimensional Stock Exchange Data using k-means Clustering

Susanta Kumar Sahoo¹

Department of CSE&A
IGIT Sarang, Odisha, India
susantasahoo79@gmail.com

Sasmita Mishra²

Department of CSE&A
IGIT Sarang, Odisha, India
sasmita.mishra@gmail.com

Dillip Kumar Swain³

Department of CSEA,
IGIT Sarang, Odisha, India
dillipswain41@gmail.com

Abstract-Multidimensional databases are used mostly for Online Analytical Processing (OLAP) and data warehousing. These can be used to show multiple dimensions of data to users. A multidimensional database is created from multiple relational databases and these relational databases allow users to access data in the form of different queries. These multidimensional databases also allow users to ask analytical questions related to business or market trends. The multidimensional database uses multidimensional online analytical processing to access its data. The stock market is one of the most unpredictable place in this world which directly or indirectly attracts many eyes due to its sensitivity and direct market involvement. There are several analytical systems to analyze this type of huge amount of data. Clustering of these types of data is more popular to analyze these data to carry out various researches and to produce meaningful result. The data in multidimensional databases is stored in a data cube format. This means that data can be seen and understood from many dimensions and many perspectives. A K-means clustering technique is one of the popular techniques for the clustering of these data to calculate the essentiality and the diversity of various clusters. K-means clustering is the simplest and popular unsupervised machine learning algorithms in which the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, with keeping the centroids as small as possible. The stock exchange data are different from each other by their uniqueness, so each data should be assign to the particular cluster so as to reach an optimized result. The stock exchange data analysis provides the price and demand of a particular product as well as the market trend of a company. It also provides the shares value as well as historical trends of various company and products. In this paper we have applied k-means clustering and linear regression basically for the market rate analysis and demand of Gold in India according to the price. This will provide information to merchants and share holders about the price and demand of Gold from time to time for various states in India to invest in Gold and purchase shares.

Key Words-OLAP, k-means clustering, multidimensional data set, Stock market, Centroids, Share holders, Linear regression.

I. INTRODUCTION

Share market is the unpredictable place for researchers as well as for the common people which attracts high interest in the world. Though there is no significant methods exist to predict the share price but the analysis of previous market trends and previous history can provide an overall result for further use and investment in the same sector. Mainly people use various ways such as statistical analysis, fundamental analysis and machine learning to predict the share price of share market but none of these methods provides consistently 100% optimized and accurate results. So development of a prediction tool is the challenging tasks so as to predict an effective prediction accuracy method. Economical indexes measure the performance of segments of the stock market and normally used for the performance of stock portfolio.

K-means clustering basically belongs to the field of signal processing which is used for cluster analysis in data mining concept and uses the method of vector quantization. K-means clustering divides n observations into k clusters in which each observation belongs to one of the cluster nearest to the mean. K-Means minimizes the cluster variances (squared Euclidean distances), but not regular Euclidean distances. K-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. In pattern recognition, the k-nearest neighbor algorithm is a non-parametric method used for classification and regression. In these cases the inputs consists of the k-closest training examples in the future space and the output depends upon the classification or regression for which k-NN is used. The centroid (Arithmetic mean) position is the mean position upon which the n observations are to be calculated. For example if we consider a triangle its centroid is the mean position of all the points in all of the coordinate directions.

This Paper is organized as follows. The Section II provides a brief literature review about Multidimensional Stock exchange data and its

Copyright © 2019 Authors

necessity with applications. In section-III we have discussed about k-means clustering and some applications of k-means clustering. In section-IV we have provided the application of k-means for stock exchange data basically the analysis of gold price as well as its demand from time to time for various states of India. The section-V provides in details the algorithms and simulation results. Finally the paper concludes in section-VI with conclusion and future work.

II. MULTIDIMENSIONAL STOCK EXCHANGE DATA

A multidimensional database is a database that is optimized for a data warehouse as well as online analytical processing. These type of databases are frequently created using inputs from existing relational database. In a multidimensional data base several mining techniques applies to carry out several meaningful and optimized result that can be applied from day by day though a multidimensional data base frequently accessed by various SQL query. In a multidimensional database different people having their own queries that are extracted for different purposes. Basically in a multidimensional stock exchange data base seller, buyer, traders, share holders have their own query and own intension to access the database.

A stock market, equity market or share market is basically the aggregation of buyers and sellers of stocks which represents ownership claim on business which may include securities listed on a public stock exchange. The example of shares includes shares of private companies which are sold to investors through equity crowd funding platforms. In a stock exchange stock brokers and traders can buy and sell shares of stocks and bonds and other securities. Many Big companies have their stocks listed on a stock exchange which attract many investors to invest upon share markets. The exchange may also act as a guarantor of settlement. There may be some companies whose stocks are listed over multiple country stock exchanges. In India the National Stock Exchange of India Limited (NSE) is the Only stock exchange which provides fully automated screen based electronic trading system and easy trading system to buyers. National Stock Exchange is world's 11th largest exchange (as on April 2018) which has a total market capitalization of more than US\$2.27 trillion.

In [1] A.R. Condrobimo et al. Proposed Data mining techniques with cluster analysis using k-means algorithm for Indonesia Stock Exchange data. This article intend to apply data mining techniques with cluster analysis on stock data registered in LQ45 in Indonesia Stock Exchange. The cluster analysis in represent the characteristics of data volumes and stock values, while the results in this study are presented in the form of cluster members visually. According to A.R. Condrobimo et al. [1] this cluster analysis technique can be used for quick and efficient identifier for each member of LQ45 index cluster based on share value for each cluster and its volume. The result can be utilized by beginner-level investors who are interested in stock investments.

In the pre-processing stage it determine the attributes for cluster analysis, i.e. 1) the code of the stock attribute, 2) the attribute of the value of the stock and 3) the transaction volume attribute. The stock code is used as the identifier of the attribute. The volume attribute describes the number of shares traded and the attribute value of the stock represents the total value of the transaction. The second cluster analysis uses Euclidian distance measurement method, by applying equations and inequalities between measurement data objects. From the findings of both traits is that the stock in LQ 45 is the most desirable investor is a combination of stocks with low value transactions and low transaction volumes [1].

In [2] S. Chaigusin et al. used Neural Network to predict the Stock exchange of Thailand. Neural networks are considered as suitable techniques for many prediction requirements. In this forecasting of Stock Exchange of Thailand (SET) performed by using feedforward backpropagation neural networks. Many combinations of parameters were investigated to identify the right set of parameters for the neural network models in the forecasting of SET. Several global and local factors influencing stock market were considered before development of the model like Dow Jones index, Nikkei index, Hang Seng index, gold prices, Minimum Loan Rate (MLR), and the exchange rates of the Thai Baht and the US dollar etc. Two years data are being used to train and test the models. Three neural network models identified by this research are a three layer, a four layer and a five layer neural network. To train the neural network models, daily data from 2003 were used. For testing, daily data in 2004, previously unseen in the creation of the neural network models, were used. The study has presented the prediction of the SET index using multilayer feedforward backpropagation neural networks. The

three suitable neural network models for the SET index prediction were found to be 7-3-1, 7-7-3-1, 7-13-7-3-1 and their prediction performance, measured by MAPE, were 1.26594, 1.14719 and 1.14578 respectively. The movements of the SET index were sensitive to the Dow Jones index, Nikkei index, Hang Seng index, gold prices, MLR and exchange rates of the Thai Baht and the US dollar [2].

III. K-MEANS CLUSTERING FOR DATA ANALYSIS

Normalization of data is very essential for multidimensional data for the efficient and significant computation. In [6] K. Aparna et al. proposed an upgraded technique i.e. genetic algorithm based bisecting K-Means clustering for multidimensional data set. There are basically two predominant issues with the multidimensional data. The first problem is the "curse of dimensionality" i.e. the number of features is much higher than the number of samples, while the second problem is there can be numerous features having minimal impact on the classification result. These two significant issues somewhere influence the overall clustering and associated classification process. Since the proposed method intends to perform bisecting K-Means clustering for multidimensional data with a different feature selection approach here a T-test Analysis approach can be applied.

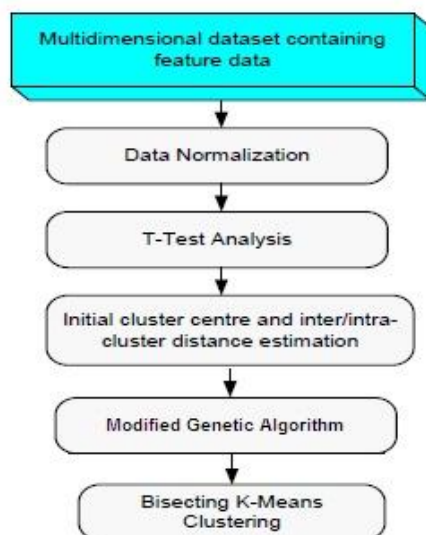


Fig. 1 genetic algorithm based bisecting K-Means clustering

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be

related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the probability of difference between two sets of data. To conduct a test with three or more variables, one must use an analysis of variance. In comparison to other conventional approaches and other existing systems, selecting the features using T-test intends as shown in Fig.1 significantly reduce the dimension of the data by finding a small set of important features which can facilitate better classification performance [6].

In [13] J. Wang et al. proposed an improved K-Means clustering algorithm using noise data Filter. The intension of this algorithm is to add density based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. The preprocessing of the data excluding these noise data before clustering the data set, the cluster cohesion of the clustering results is improved significantly and the impact of noise data on K-means algorithm is decreased effectively. First a pretreatment of data is performed to be clustered to remove the outliers using outlier detection method based on LOF, so that the outliers cannot participate in the calculation of the initial cluster centers, and excluded the interference of outliers in the search for the next cluster center point. In the next step fast global K-Means clustering algorithm is applied on the new data set which was generated. Fast global K-Means clustering algorithm is an improved global K-Means clustering algorithm [13]. The improved clustering algorithm is described as follows:

Input: The number of cluster K with data set M $\{x_1, x_2, \dots, x_m\}$ to be clustered.

Output: k clusters with the sum of dissimilarity between all objects and their nearest cluster centers is the smallest.

1) Traversal on data set M, calculate $lof(p)$,

$$lof(p) = \frac{\sum_{i=1}^k \frac{lrd(i)}{lrd(p)}}{k}$$

if $lof(p)$ is much larger than 1, remove the isolated point p, else leave p; finally get the new data set N;

2) Calculate the mean of data set N as the first clustercenter [13].

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i$$

3) Calculate the distancebetween the remaining points and the cluster center with the next cluster center [13].

$$b_n = \sum_{j=1}^n \max (d_{k-1}^j - |x_n - x_j|^2, 0)$$

Make the sample point X_n whose b_n is the largest as the next initial cluster center, calculate the distance between each object x_i and the center of each cluster, with assigning to the nearest cluster.

4) Calculate the mean of objects in each cluster as a newcluster center [13].

5) Repeat 3) 4) until the criterion function E converged, return $(m_1, m_2 \dots m_k)$.

6) Exit

IV. K-MEANS CLUSTERING FOR STOCK EXCHANGE DATA SET

Stock Exchange prediction is the most unpredictable and challenging task in these days. A gold standard is the monetary system in which the standard economic unit of account is based on a fixed quantity of gold. The gold standard was being used in the early 19th and 20th century. Gold Exchange Traded Funds (ETFs) are simple investment products that combine the flexibility of stock investment with the simplicity of gold investments. ETFs trade on the day to day market of the National Stock Exchange, like any other company stock, so that these can be bought and sold continuously with the market prices.

K-means clustering is used to cluster the data i.e. the numbers of shares sold in different clusters as well as the rate of sold according to the gold price. In India the Demand of gold is not unique in all part of the country. Basically the People of South and South-Eastern States prefer to buy more and more Ornaments that is the active investment i.e. investment upon the raw material that is purchasing of Gold Ornaments is high. Gold ETFs are passive investment instruments that are based on gold prices and invest in gold shares. Similarly people from North and North-Western states invest more upon Gold shares. K-means clustering is useful in Clustering all states of the country where the clusters

can be considered as North, South, East, West, North-East, South-East, South-west and North-west. Each state should be assigned to only one cluster nearest to it and the Gold purchase/Gold share purchase should be assigned respectively. Some of the Gold Share Trade fund upon India Stock Exchange includes AXISGOLD, HDFC Gold Exchange Traded Fund, ICICI Prudential Gold Exchange Traded Fund, Reliance Gold Exchange Traded Fund, SBI Gold Exchange Scheme etc.

Year Wise Gold Price in India(24K in Rupees)

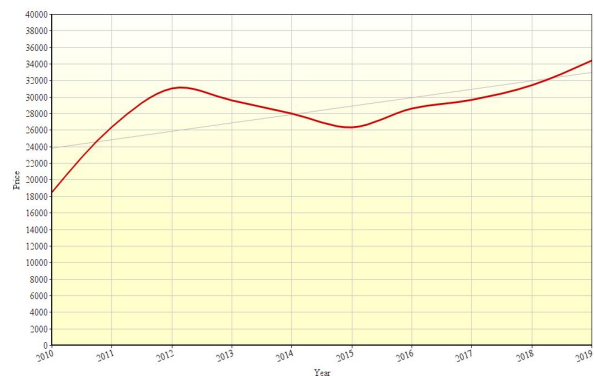


Fig. 2 Year wise Gold Price of 24K Gold

The price of gold is increasing day by day and a rapid growth basically observed in the last 10 years. Fig. 2 shows the 24K gold price and the rate of increment and decrement in its price in India in the last 10 years i.e. 2010 to 2019(Aug-2019).

Gold Investment and Demand Comparison

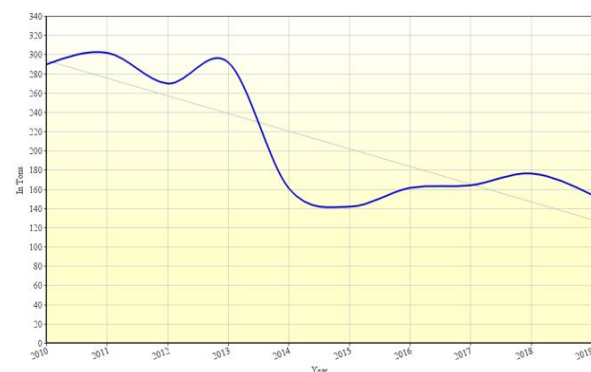


Fig. 3 Gold Investment and demand

Similarly the Fig. 3 represents in TONS, the demand of Gold in the last ten years i.e. the Investment upon Gold shares as well as the quantity of gold sold in the last 10 years. Though there is the significant growth in gold price in the last 10 years but the sale as well as investment upon Gold is

remarkable. Though there is a fall in investment observed from 2013 to 2015 but after that there is the growth in investment to till date.

V. ANALYTICAL RESULTS AND OUTPUTS

In statistics linear regression is the linear approach to establish the relationship between a dependent variable and one or more independent variable. For more than one independent variable the technique is known as multiple linear regressions. So the linear regression models can be used for analysis that is the analysis of multidimensional databases. Here we use a linear regression model to determine the trends of clusters and existence of different data element to the proper cluster. Extracting the trend from the slope of the model is straightforward. This model is also more reliable and robust than that of other complex models.

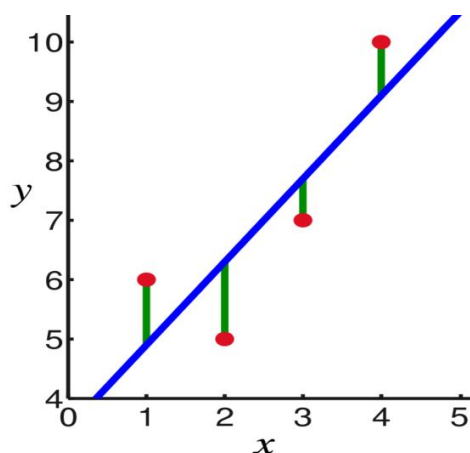


Fig. 4 Example of Linear regression

In linear the deviation from the underlying relation is always calculated to regenerate the optimal solutions. In Fig. 4 the 'x' is the independent variable and y is the dependent variable and observations (red dot) deviates (green line) from the underlying relationship (blue line) are to be calculated. So in our experiment the gold price should be taken as independent variable upon which the sale/purchase in gold share and sale/purchase in gold ornament i.e. the active and passive investment in gold are to be calculated.

The data source clustering is useful for multiple data mining sources which accept the concept of targeted learning. K-means grouping algorithm is used to optimize or improve the separation distance or similarity in the cluster. Thenew average for each cluster will be calculated and grouping of object into the respective should be

performed. Then re-grouping should be done for all objects using the new updated cluster center. The iteration process continues upon obtaining a stable grouping, which means that the group formed from the last iteration is equal to the group formed with the previous iteration [1].

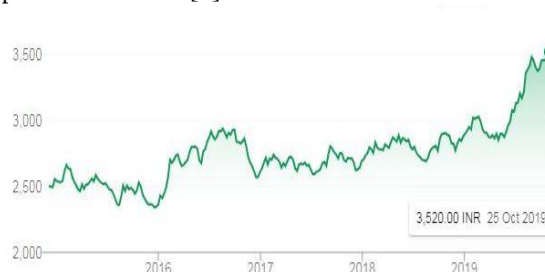


Fig. 5 HDFC gold ETF Share price

In India lot of investors wants to invest upon gold in terms of purchasing of Gold ornaments or purchasing of shares. Though the market value of Active and Passive investments on Gold varies from area to area and market price also varies from cluster to cluster though in this approach we have used eight clusters to carry out the research upon the Gold market value in India. Fig. 5 shows the gold exchange traded fund and gold share value of HDFC gold exchange traded fund in the last five years. The market value of share represents that on 25th October 2019 it touch the highest value i.e Rs.3520 which is highest in the last five years.

VI. CONCLUSION

K-means clustering as well as linear regression are two important methodologies to carry out analysis to generate fruitful and optimized results. Due to the rapid growth in multidimensional data it's always a challenging task for the data mining methodology to make proper analysis and to carry out proper and optimized result. Stock market and stock exchange is the most unpredictable system in which each moment of time there is the up-down in the sensex. There are very rare systems in this globe which provide more than 80% of efficiency. So designing an error free, robust and dependable model for the best prediction of the stock exchange data set is always a challenging task. In this paper we have provided the analysis of demand of gold in comparison to Price for various states in India for the last decades. This article provides the analysis of gold price according to the year. We have provided the analysis upon the gold price and the sale that various from time to time in India. The active and passive investments upon gold in India can provide a clean idea about the future trend and prospective which will be help full for the

Copyright © 2019 Authors

investor who wants to invest upon gold. For example we have provides the share value of HDFC gold traded fund and its value of share upon the last five years. The model can be used for the prediction of the demand according to price for the upcoming days too. This will be helpful for the merchants and researcher for investigation upon the market as well carryout research for the stock market respectively. In our upcoming research we want to focus upon the real time development of models for the prediction of stock trend and share value that will be helpful for the investors want to invest in stock market.

ACKNOWLEDGMENT

We are very much thankful to Prof. Sarojananda Mishra and staff members of IGIT Sarang, Odisha, India for providing resources for the data collection for the research. We are also thankful to the anonymous reviewers of this research article for their valuable suggestions and advice which improved the quality of the research paper.

REFERENCES

- [1] A. R. Condrobimo, B. S. Abbas, A. Trisetyarso, W. Suparta, C. H. Kang, "Data mining technique with cluster analysis use K-means algorithm for LQ45 index on Indonesia stock exchange", Proceedings of Int. Conf. on Information and Communications Technology, pp. 885-888, 2018.
- [2] S. Chaigusin, C. Chirathamjaree, J. Clayden, "The Use of Neural Networks in the Prediction of the Stock Exchange of Thailand (SET) Index", Proc. of Int. Conf. on Computational Intelligence for Modelling Control & Automation, pp. 670-673, 2008.
- [3] O. Lachheb, M. S. Gouider, L. Ben, "An Improved MapReduce Design of Kmeans with Iteration Reducing for Clustering Stock Exchange Very Large Datasets", Proc. of 11th Int. Conf. on Semantics, Knowledge and Grids, pp. 252-255, 2015.
- [4] C. Hendaheva, V. Pavlovic, "Analysis of Causality in Stock Market Data" 11th Int. Conf. on Machine Learning and Applications, pp. 288-293, 2012.
- [5] C. Fonseka, L. Liyanage, "A Data mining algorithm to analyze stock market data using lagged correlation" Proc. of 4th Int. Conf. on Information and Automation Sustainability, IEEE, pp. 163-166, 2008.
- [6] K. Aparna, M. K. Nair, "A Pragmatic Approach for Multidimensional Data Clustering", Proc. of 8th Int. Conf. on Computing Communication and Networking Technologies, 2017.
- [7] Z. Yixin, J. Zhang, "Stock data analysis based on BP neural network", Second International Conference on Communication Software and Networks, pp. 396-399, 2010.
- [8] N. Sirisha, K. V. D. Kiran, "Stock Exchange Analysis using Hadoop userExperience", Proc. of Int. Conf. on Intelligent Sustainable Systems, pp. 1141-1144, 2017.
- [9] K. H. Sadia, A. Sharma, A. Paul, S. Padhi, S. Sanyal, "Stock Market Prediction Using Machine Learning Algorithms", Int. Journal of Engineering and Advanced Technology, vol. 8(4), 2019.
- [10] Z. Peng, "Stocks Analysis and Prediction Using Big Data Analytics", Proc. of Int. Conf. on Intelligent Transportation, Big Data & Smart City, pp. 309-312, 2019.
- [11] C. T. Wang, Y. Y. Lin, "The Prediction System for Data Analysis of Stock Market by Using Genetic Algorithm" Proc. of 12th Int. Conf. on Fuzzy Systems and Knowledge Discovery, pp. 1721-1725, 2015.
- [12] S. Aghabozorgi, Y. W. Teh, "Stock market co-movement assessment using a three-phase clustering method," Expert Syst. Applications, vol. 41(4), pp. 1301-1314, 2014.
- [13] Y. Luo, J. Hu, X. Wei, D. Fang, H. Shao, "Stock trends prediction based on hypergraph modeling clustering algorithm," Proc. of IEEE Int. Conf. on Progress in Informatics and Computing, pp. 27-31, 2014.
- [14] X. Zhong, D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting," Neurocomputing, vol. 267, pp. 152-168, 2017.
- [15] S. Chaigusin, C. Chirathamjaree, J. Clayden, "Soft Computing in the Forecasting of the Stock Exchange of Thailand", Proc. of the fourth IEEE Int. Conference on Management of Innovation and Technology, pp. 1277-1281, 2008.
- [16] H. He, J. Chen, H. Jin, S. H. Chen, "Trading Strategies Based on K-means Clustering and Regression Models", Computational Intelligence in Economics and Finance, pp. 123-134, 2007.

[17] H. Jin, M. L. Wong, K.S. Leung, "Scalable model-based clustering for large databases based on data summarization" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(11), pp. 1710–1719, 2005.