Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1248 – 1255

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

CLUSTERING AND REGRESSION TECHNIQUES FOR STOCK PREDICTION

Bini B.S[a*],Tessy Mathew[b]

[a]Department of computer science and technology ,Marbaselios College of Engineering,Trivandrum,Kerala.,India.
[b]Department of computer science and technology ,Marbaselios College of Engineering,Trivandrum,Kerala.,India.

Abstract

Stock market prices keep on varying day by day. It is very difficult to foresee the future value of the market by the sellers and buyers. In this paper, an analysis system which helps the people to identify the more profitable companies using data mining approaches is proposed. The clustering and regression are the two techniques of data mining used here, Validation index is used for analysing the performance of different clustering methods such as partitioning technique, hierarchical technique, model based technique and density based technique.

Among the different clustering techniques experimented, partitioning technique and model based technique give high performance i.e. K-means and EM clustering algorithm respectively. For prediction of future stock price multiple regression technique is used which helps the buyers and sellers to choose their companies from stock.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

Keywords: K-means; EM; Hierarchical; Density based; Model based; Multiple regression.

1. Introduction

A collection of buyers and sellers of stock is the stock market, were stocks are released by the companies for elevating the capitals and are bought by the investors in order to get a portion of the company. Stock markets are always aggressive. It is very difficult to predict the future stock price of the companies since it keeps fluctuating every day. The booming prediction of a stock's future price could give important profit. In data mining a number of algorithms were designed to overcome this incertitude.

* Corresponding author. Tel.: +91-7736643083.
E-mail address: binibabu9590@gmail.com

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.104

Data mining is classified into predictive and descriptive. In the present work the techniques of data mining used are clustering and regression, were clustering is a descriptive and regression is a predictive method. The predictive method makes prediction about values of data and descriptive method identifies the relationship between the data. Data Mining mainly aims at extraction of previously unknown and substantially helpful information from the available dataset. The main categories of data mining technique include clustering, classification and regression.

In this present work clustering and regression are considered. Clustering is the process of creation of clusters of similar objects which is an unsupervised technique since it does not make use of class labels. The partitioning technique, hierarchical technique, model based technique and density based technique are different clustering techniques which are discussed here. Regression is a supervised technique, which has a predefined target. Multiple regression in regression technique is used for this work. The major difference between classification and regression is that classification depends on variables which are unordered whereas regression depends on variables which are ordered.

The work focuses mainly to find out the top companies in the market using different clustering techniques and to predict the future stock price for that companies using regression technique.

2. Literature Review

Data mining algorithms are classified into Clustering, Regression and Classification. The present work focuses on clustering and regression techniques. Review of different clustering techniques and regression technique used for the work is presented in this chapter.

Hailong Chen and Chunli Liu [1] made a comparative study on different clustering algorithms namely CLARANS of Partitioning Methods, CURE and BIRCH of Hierarchical Methods, DBSCAN and OPTICS of Density Based Methods, STING and CLIQUE of Grid-Based Methods, COBWEB of Model-Based Methods. The performance of all the above algorithms were compared. The performance is compared on the basis of six aspects namely scalability, the shape of the cluster, sensitivity to the "noise", sensitivity to the data input sequence, high dimension and algorithm efficiency. Different algorithms showed their best performance in their different aspects. BIRCH and STING showed better performance when the aspects efficiency, sensitivity to noise and sensitivity to data input sequence is considered. CLARANS, CURE, DBSCAN, OPTICS, STING and CLIQUE showed less sensitivity to noise. CURE, DBSCAN, OPTICS, STING and COBWEB has the freedom to form arbitrary shaped clusters. When considering the sensitivity to data input sequence BIRCH, DBSCAN, OPTICS, STING and CLIQUE shows the least sensitivity. Both DBSCAN and OPTICS algorithm have the same time complexity. CURE algorithm forms spherical clusters when trying to overcome the defects by inheriting BIRCH algorithm and combining hierarchical algorithms and partitioning algorithm. They concluded that K-means are comparatively well to do and understand, also bare to implement for large databases the algorithm is relatively scalable and efficient. Hierarchical clustering algorithm is costly in terms of computation and storage requirements [1].

Joel Joseph and Indratmo made a tool intended to assist users in identifying groups of stocks that have similar price movement patterns over a period of time. Here they cluster the stock market data and then visualizes it using an unguided clustering algorithm known as the Self-Organizing Map algorithm. Visualization was done using Yahoo Finance line charts. The dataset is taken from yahoo finance charts precisely S&P 100 stock data from Jan1, 2012 to Jul 30, 2012. The application of SOM is to map data represented in high dimensions to two –dimensional plane. The algorithm initially assigns weights on spatially organized nodes in a random fashion and then iteratively modifies the weights. Euclidean Distance is used to determine the neighbourhood of a data point .The major problem in using SOM algorithm is its scalability[2].

S.R. Nanda, B. Mahanty and M.K. Tiwari made an analysis on stock market data. They initially the stock is classified to clusters and then from that a portfolio is built. The stock data used for clustering is the Bombay Stock Exchange data for the fiscal year 2007 to 2008. The portfolio is built using Markowitz model. The clustering is done using three different techniques namely K-means, Fuzzy C-means and SOM. Among the three K-Means showed better performance. They suggest how to integrate different clustering techniques into portfolio management. The hybridization of different techniques would further increase the portfolio efficiency [3].

Shalini S. Singh and N. C. Chauhan compared the partitioning techniques k-means and k-medoid. The performance was compared on the basis of eight aspects namely Complexity, Efficiency, Implementation, Sensitivity to Outliers, Necessity of Convex shape, Advanced specification of number of clusters “k”, Effect of initial partitioning on result and run time and for which type of clusters or data sets are they optimized for K-means showed better results in all aspects but sensitivity to outliers.So it was concluded that K-means outrun k-Medoid.[4].

M.Suresh Babu, Dr. N. Geethanjali and Prof B. Satyanarayana compares three major clustering algorithms K- Means, Hierarchical clustering algorithm and reverse K means and compare the performance of these three major clustering algorithms on the building ability for correct class wise cluster of algorithm They also proposes an efficient clustering method Hierarchical and Recursive K-means clustering (HRK) to predict the short-term stock price movements after the release of financial reports. Initially, The financial report is converted to feature vectors. Then the hierarchical agglomerative clustering method is employed to perform clustering on the converted feature vectors. Next, K-Means clustering method is applied in order to partition each

cluster into further sub- clusters still belonging to the same parent class. Then, for each sub cluster, its centroid is chosen as the representative feature vector. Finally, these representative feature vectors are employed to predict the stock price movements. The proposed clustering and classification framework for stock time series, benefits from the multi- resolution capability to analyze the nonlinear similarities between stock price time-series at different time horizon. It is observed that the proposed method outperforms SVM in terms of accuracy and average profits. This can be attributed to three aspects. First,in financial reports both qualitative and quantitative features are considered. Second, the advantages of two clustering methods are combined thus coming up with an effective clustering method. Third, as an optimal number of splits are calculated in HAC thus localizing the clusters generated and as a result improvise the performance [5].

Han Lock Siew and Md Jan Nordin made a study on different regression techniques, also they Predict the Trend of the Stock Price . The data set were taken from various companies in Bursa Malaysia.The stock price movement of various companies in Bursa Malaysia is considered .Here various regression techniques is compared using the tool weka. Five weka classifiers namely additive regression, linear regression, regression by discretization, simple linear regression and SMO regression were employed for comparing the performance.Among the five classifiers SMO out performed all [10].

3. Clustering Techniques

The process for grouping the similar objects together were different clusters will be dissimilar to each other is clustering. The similarities and dissimilarities are based on the attribute values and frequently involve distance measures. There are different techniques used for clustering some are Partitioning based technique, Hierarchical technique, Density based technique, Model based technique, Grid based technique. Among these techniques Partitioning technique, Hierarchical technique, Density based technique and Model based technique were employed in this work.

3.1. Partitioning techniques

Partitioning method creates k partitions or clusters, from a given dataset having „x‟ data objects. The algorithm starts its working by assigning different objects to different dataset randomly and then at each iteration it reallocates the data objects to another partition. Each partition is represented as a centroid were it is an average of all data objects in a partition. Here every partition must contain leastways a single data object, and each data objects must contain exactly one cluster. Different clustering techniques are k-means, k-medoids, “Partitioning Around Medoids” (PAM) “Clustering Large Applications based upon RANdomized Search” (CLARANS) and “Clustering LARge Applications” (CLARA) etc [1]. In this paper, the main focus is given to K-means.

K-Means is a widely known unsupervised techniques. It assigns a given set of data objects into „k‟ clusters, where k shows the number of clusters and it should be mentioned in advance. Each cluster should have a centroid. Based on the distance of each data point to centroid. Data point is assigned to cluster having closest centroid. Then at each iteration the data points are reassigned to different clusters by recalculating the distance. This process continues until there is no further change in centroid location [4].

Algorithm [4]:
Input: The number of clusters created: k The number of objects assigned: x
Output: Based on the given similarity function „k‟ clusters are obtained.
Steps :
i)select „k‟ data entity randomly and assign it as the first cluster centroids;
ii)Continue,
a.With respect to similarity function set the remaining data points to each cluster.
b.Revise the centroid for each cluster by taking the cluster mean
iii) Until no further change occurs.

3.2.Density Based techniques

Density-based techniques are for the purpose of unwrapping clusters of arbitrary shape. The main idea is that, for every information within the given category, during a given vary of space should contain a minimum of a particular range of points. This methodology can be used to filter the "noise" outlier data. In this paper,“Density Based Spatial Clustering of Applications with Noise” (DBSCAN) is reviewed.

The basic difference of density-based technique with that of partitioning technique is that it is not based on distance, but on density[1].

Algorithm[9]:
i. Randomly select a point q
ii.Fetch all points which are density-reachable from q based on “MinPts‟ and “ε‟.
iii.Form a cluster when q is a core point.

iv.If q is a boundary point, then none of the points are density-reachable from q, thus the algorithm visits the next point in the database.
v.Repeat the procedure until every data point is processed.

3.3.Hierarchical Techniques

Hierarchical clustering technique creates cluster by data hierarchy and forms a tree based on cluster nodes. In accordance with different directions of decomposition in different hierarchies, this kind of clustering algorithm can be divided into agglomerative and divisive method. If it is bottom-up hierarchical decomposition, which is called agglomerative hierarchical clustering; this strategy first treat each object as a cluster, and then merge the cluster adjacent to each other into a big cluster until all objects belong to same cluster, or a stopping condition is met. Major part of the hierarchical clustering methods belongs to this class, which is just different on the inter-clustersimilarity definition, and in this paper also this algorithm is reviewed; while top-down hierarchical decomposition is called divisive hierarchical clustering. In contrast to agglomerative hierarchical clustering, all objects are in one cluster, and then gradually divide into smaller and smaller clusters until each object belong to its own cluster, or reach a termination conditions.

Hierarchical clustering technique includes BIRCH algorithm, CURE algorithm and so on. The defect of hierarchical methods is that once merge or split is completed, it cannot be undone and can't exchange between clustering objects. The strict rules are useful, thus we don't have to worry about the different choices of different combinations, and computational cost will be smaller. However, one of main problems of the technology is that it can't correct the wrong decision. If in one step without well choosing to merge or split decision, which may lead to low quality of clustering results. Moreover, since the decision to merge or split need to check and estimate amounts of objects or cluster, this clustering method has poor scalability [1].

3.4. Model based Technique

Model based technique is probability based model. Maximum Likelihood Estimation (MLE) is used for finding the parameter inside the probability model. The model based technique attempts to optimize the fit between the given data and some mathematical models. It finds out characteristic description for each group, where each group represents a class.

The EM (Expectation Maximization) algorithm finds the maximum likelihood parameters of a model where the equations can‟t be resolved straightly. Generally these models demand latent variables along with the known data observations and unknown parameters. i.e., sometimes there will be missing values among the data.

4. Regression Techniques For Stock Prediction

Regression is used for predicting an outcome based on a given input. The simplest regression technique is linear regression and advanced regression technique is multiple regression. If a single descriptive variable is used then it is known as simple linear regression and if more than one descriptive variable is used then the technique is multiple regression.

4.1.Linear Regression

Linear Regression is statical technique used to predict the relationship between the dependent and an independent variable. Generalities represented as V=Y+WX, where Vis the dependent variable, X is the independent variable ,Y is a constant and W is the slope of regression line.

4.2.Multiple regression

Multiple regression is a technique for modeling the association among the scalar dependent variable “V‟ and one or more descriptive variables indicated by “U‟. It predicts the future value of variable with respect to other variables.

$V = w0+w1\ y1+ ... +wnyn+ \epsilon$

where ,V implies the dependent variable,w0 to wn implies the co-efficients, y1 to yn implies the independent variables, and ϵ implies the random error. In this work Multiple regression technique is used for predicting the future stock price

5. Validation indexes for Clustering Algorithms

Index measure helps to seek out the accuracy of result obtained. Once clustered, it determines the quantity of tuples that are properly labeled and clustered. Number of the validity indexes mentioned in this paper are C- Index, Jaccard Index, Rand Index and Silhouette-Index [6].

5.1.C- index:

The C-index is specified as follows: C = S-Smin / Smax – Smin

Where, "S" represents the sum of distances of all values belongs to an equalent cluster. Total number of value is denoted as "q". Smin denotes the total amount of q smallest distances of the values of the considered attribute, and Smax is the total amount of q greatest distance of the values. Smaller value of C denotes better clustering.

5.2.Jaccard index

In this index a collection of class labels "c" and outcome of cluster is "k" which is set by count of set of points allotted to the identical cluster in each division.

J (c, k) = p / (p + q + r)

Where, "p" implies the count of set of points with identical label in "c" and assigned to identical cluster in "k", "q" implies the count of pairs with the identical label, but in dissimilar clusters and "r" implies the count of set in the identical cluster, but with dissimilar class labels. Index ranges between zero and one, where one shows that "c" and "k" are identical.

5.3.Rand index

Rand index is similar to Jaccard index. Only dissimilarity is the count of set with a different label in "c" that were assigned to a dissimilar cluster in "k", which is denoted with "s". It is calculated as follows:

J (c, k) = (p +s)/(p + q + r + s).

The index ranges between zero and one. If the value for index is one, which implies 100% accurate.

5.4.Silhouette

Here, the tuples are grouped into k clusters. For each tuple J, let p(j) indicates the mean dissimilarity of j with other tuples in the identical cluster. After that finds the mean dissimilarity of J with data of another cluster. Repeat this for all the cluster until J is not an element. The least mean difference of J is represented as q(j) Then,

s(j) = q(j) - p(j) / (max.(p(j),q(j))

If s(j) is near to one, then the data is clustered properly clustered.

6. Data Description

Required data is collected from National Stock Exchange (NSE) which includes WIPRO, TCS, ROLTA, POLARIES, PERSISTENT, NIITTECH, NAUKRI, MINDTREE, INFY, and HCLTECH. The period under consideration, for dataset of each company was taken us six months. The selected attributes are open, close, high, low, previous close and average price for clustering and classification. Fig. 1. shows the graphical representation of dataset for all companies in the month of January 2015.

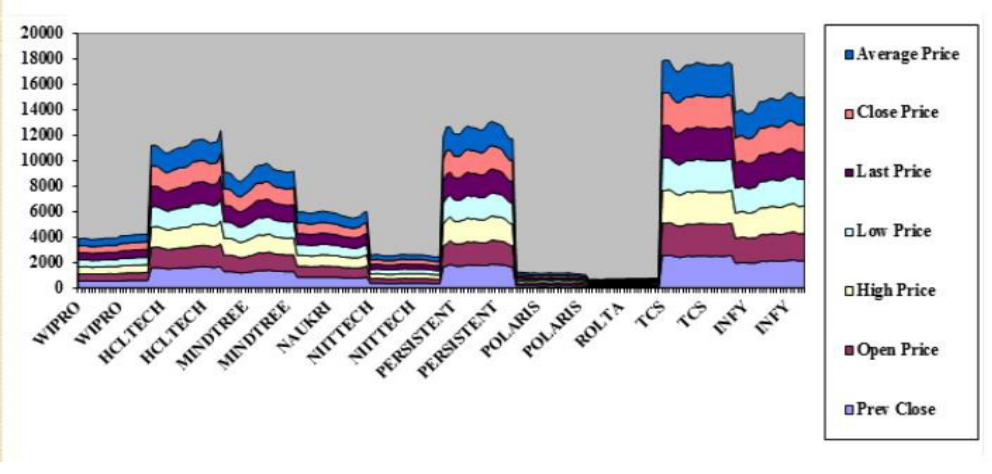

Fig.1. Stock details for the month of January 2015

7. Experimental Result

This section briefs the result obtained by implementing the clustering and regression techniques considered in this work.In this model the dataset is collected from the National Stock Exchange (NSE) then it undergoes different Clustering techniques such as the partitioning technique were K-means algorithm is used, in hierarchical technique agglomerative algorithm is used, in model based technique EM algorithm is used and in density based method DBSCAN is used.The algorithms showing better performance were find out by using validation index.

The validation index used are C-Index, Jaccard Index, Rand Index and Silhouette-Index. While analyzing the results obtained by the indexes the K-means and EM algorithm shows better performance than agglomerative and DBSCAN.The companies from these algorithm under goes the regression method called multiple regression for predicting the future stock.

The partitioning, hierarchical and density based techniques were performed on stock¨s dataset during the period of january 1st 2015 to june 30th 2015. The given dataset is clustered into three ie; K=3 and gets the output as shown in the Table.1.

Table 1.Clustering results of four clustering techniques (K=3;N=1232)

Cluster Label	K-means	EM	Agglomerative	DBSCAN
Cluster 1	363	369	862	370
Cluster 2	503	528	246	492
Cluster 3	365	334	123	369

The main objective is to identify a clustering algorithm that will generate best companies from the above techniques which helps the investors to choose the companies for investment.

For getting the best algorithm validation Index is calculated.The validation index used are C-Index, Rand Index, Jaccard Index and Silhouette Index. Fig.2. shows the graphical representation of validation Index of different clustering techniques.

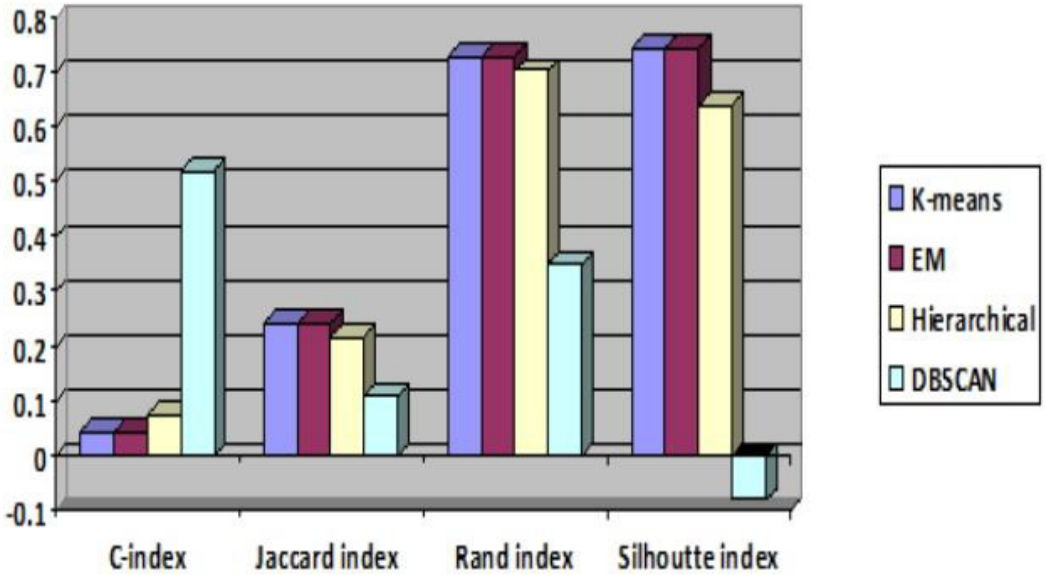

Fig.2. Validation Index for Clustering Methods

From the graph it is clear that K-means and EM algorithm are better as compared with hierarchical and DBSCAN since C-index of K-means and EM is having value near to zero whereas jacard ,Rand and silhouette index is having value near to one.

Companies from the best clustering technique under goes the regression technique for predicting the future stock.The technique used is multiple regression. The fig.3. shows the predicted values of TCS for the month of January using the multiple regression technique, this data set contains 21 days stock price and its close price is taken for fitting the price which is one of the attribute which means that the close price is taken as the dependent variable and rest of the attribute as the independent variables. The fig.3 gives the result for the prediction i.e. it results the future price of February which is approximately equal to the stock price of February. This will help the buyers and sellers to choose the correct company for their investment.

	fit	lwr	upr
1	2538.601	2484.866	2592.336
2	2567.754	2514.280	2621.227
3	2533.174	2479.633	2586.716
4	2440.331	2388.195	2492.467
5	2411.601	2361.118	2462.083
6	2421.809	2371.525	2472.094
7	2480.512	2429.403	2531.620
8	2484.126	2432.161	2536.092
9	2483.787	2431.781	2535.793
10	2505.974	2453.653	2558.295
11	2545.378	2492.513	2598.242
12	2484.026	2431.525	2536.527
13	2508.713	2456.197	2561.230
14	2486.602	2434.546	2538.657
15	2486.327	2434.021	2538.634
16	2495.515	2443.135	2547.895
17	2502.837	2450.657	2555.018
18	2488.573	2436.312	2540.834
19	2521.211	2469.042	2573.381
20	2526.357	2473.695	2579.020
21	2477.643	2424.451	2530.835

Fig 3. Predicted values of TCS for the month of January

8. Conclusion and future work

In the present work clustering is performed on stock data obtained from NSE, which produces the name of the best companies as output. Then comparison between partitioning based, hierarchical, model based and density based techniques are performed with the help of validation index such as c-index, Jaccard index, rand index and silhouette index. K-means algorithm in partitioning based technique and EM algorithm in model based technique shows better performance than hierarchical and density based technique. Then the clustered result is given to multiple regression which is one of the regression technique for getting the future stock price.

As future work an online stock prediction system can be created using Partitioning based or model based technique along with multiple regression technique.

References

[1]Hailong Chen ,Chunli Liu Research and Application of Cluster Analysis Algorithm,2013 2nd International Conference on Measurement, pp.575-579.

[2]Joel Joseph and Indratmo, Visualizing Stock Market Data with Self-Organizing Map,2013 Proceedings of the Twenty-SixthInternational Florida Artificial Intelligence Research Society Conference, pp.488-491.

[3]S.R. Nanda, B. Mahanty, M.K. Tiwari, Clustering Indian stock market data for portfolio management, 2010, pp. 8793–8798.

[4]Shalini S Singh, N C Chauhan, K-means v/s K-medoids: A Comparative Study, National Conference on Recent Trends in Engineering Technology,13-14 May 2011.

[5]M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana, Clustering Approach to Stock Market Prediction, Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291 (2012).

[6]Parul Agarwal, M. Afshar Alam, Ranjit Biswas, Issues, Challenges and Tools of Clustering Algorithms, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 ISSN (Online): 1694-0814.

[7]Saeid Fallahpour 1, Mohammad Hendijani Zadeh & Eisa Norouzian Lakvan, Use of Clustering Approach For Portfolio Management, Inter- national SAMANM Journal of Finance and Accounting ISSN 2308-2356 January 2014, Vol. 2, No. 1

[8]Anil K.Jain,Data clustering:50years beyond k-means, 2009, pp.651-666.

[9]M.Parimala, Daphne Lopez, N.C. Senthilkumar,A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases, International Journal of Advanced Science and Technology Vol. 31, June, 2011.

[10]Han Lock Siew, Md Jan Nordin,Regression Techniques for the Prediction of Stock Price Trend, Statistics in Science ,Business and Engineering (ICSSBE) 2012 International Conference, 12th sept 2012.