ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/108029, December 2016*

Stock Market Prediction using Hierarchical Agglomerative and K-Means Clustering Algorithm

T. Renugadevi, R. Ezhilarasie, M. Sujatha and A. Umamakeswari*

School of Computing, SASTRA University, Thirumalaisamudram, Thanjavur - 613401, Tamilnadu, India; aum@cse.sastra.edu, renugadevi@cse.sastra.edu, ezhil@cse.sastra.edu, sujatha@cse.sastra.edu

Abstract

Objectives: The stock market performance has more impact on national economy. The purpose of this work is to generate a portfolio to reduce the uncertainty of stock in short term basis. **Methods:** Hierarchical clustering is more efficient while non-determinism is of concern when compared with flat clustering. Hierarchical agglomerative Clustering is used, which results in more informative structure than flat clustering on unstructured data. Single-link clustering is taken into account as it does not pays more attention to outliers and amalgamation criterion is local than complete-link clustering and results in intuitive cluster structure. Dendrogram is used to represent the progressive formation of clusters in HAC. **Findings:** Flat clustering K-means algorithm is used to combine the clusters generated by Hierarchical agglomerative clustering (HAC). As the number of samples has been reduced, iterative use of k-means will choose better centroid. **Applications:** The final list of the recommended stocks is then showcased to the investor on short term basis. The baseline data is downloaded from National Stock Exchange (NSE).

Keywords: Dendrogram, Hierarchical Agglomerative Clustering (HAC), Single-Link Clustering

1. Introduction

With ever evolving technologies it becomes easy to recognize the stock market prices instantly. Despite the price of the selected stocks, all of the stock related news is also obtainable at the same instant. Yet the naive problem associated with the stock market prediction[1] of vague parameters is that influence of the market dynamics still survives. Regardless of the company's performance there exists certain factors such as political news, economic situation; financial reports influence the value of the share. Each day there is a battle sandwiched between bulls and bears. Investor constantly monitors the movement of stocks which has influence on future investments. Different scenarios like psychology and emotions drive our decisions other than unpredictable value of the stock. Each stock pulls the market in different directions causing frustration on the other side. The nature of the stock is volatile in nature. Hence it always requires risk taking factor before investment. A stock market has different sectors like automotive, information technology, power, services, banking, health care etc. The existing systems cater the prediction of a single stock exclusively using diversified methods.

Existing systems support prediction of single stock, by association rule mining. A rebalancing formula is used to remove the negatively performing stocks[2]. Financial report is converted to feature vectors. First HAC is applied to convert feature vectors to clusters and then k-means is applied recursively to form sub clusters. Centroid is chosen for each sub cluster as its representative feature vectors to determine the stock price[3]. Apriori is implemented to determine stock category association and K-means is implemented to determine the stock clusters, the parameter used in both the algorithm is stock indices[4]. Neural network models[5] for level estimation and classification and cross-validation technique are employed on the parameters like production price index, industrial production index, consumer price index on long term basis to forecast stock returns[6]. BSE-Sensex is

**Author for correspondence*

selected as a parameter for analysis. Genetic algorithm is used to optimize the parameter and Decision tree is used to improve accuracy. The hybrid approach is implemented using Support vector machine[7]. Survey of various techniques like decision tree, neural networks and ARM for various applications is discussed in[8]. Original data set is converted into smaller set of prototype vectors by cluster means and three methods like K-means, hierarchical clustering and self-organizing map are used[9]. A survey is done on the data mining techniques and how it is used in stock market, clustering is used to uncover hidden patterns and predict future trends and behaviors in financial markets[10]. As stock market is non linear, neural networks would more suited, and in this paper various factors like hidden layers, learning rate and momentum rate is considered to predict stock movements[11]. In this paper relationship between daily and intraday stock data is established using neural networks, in this exhaustive experiments are performed to compare ARMAX, LMS, BPN, Radial basis functional neural networks and the Recursive least square prediction, and finally ARMAX is recommended[12]. A linear relationship of non-model based co relation and linear regression based predictability and non linear neural network is found to exist between intraday as well as AHIPMI. NASDAQ datasets like open, close, high, low prices of intraday and AHIPMI are considered as parameters[13]. phase of stock price trends, and accuracy of results increases, so in this paper algorithms like pattern extraction algorithm, data stream correlation extraction algorithms has been employed on time series data[14]. In foreign exchange market movement is displayed by the best offer and best asking price, which in turn can be determined by neural networks and regression trees and it is found that it is more accurate than simple averaging and waning.

2. Clustering Analysis

2.1 Hierarchical Clustering

Connectivity based clustering results in clusters, with high within-cluster similarity and low inter-cluster similarity. There are many variants to define closest pair clusters. Single Link aims at grouping the instances based on most cosine-similar. Complete-link aims at grouping the instances based on least cosine similar. Centroid based aims at clustering data based on most cosine similar centroid. Average-link based on average cosine between pairs of elements. The main target is to single out distinguished features.

2.1.1 *Single-Link Agglomerative Clustering*

In this work single-link agglomerative clustering is used. The clusters are formed based on maximum similarity and progressive of cluster formation is visualized using dendrogram.

$$sim(c_i,c_i)=\max_{x\in c_i,y\in c_i} sim(x,y)$$

After merging *ci* and *cj*, the similarity of the resulting cluster to another cluster, *ck*, is

$$sim(c_i\cup c_j)=\max(sim(c_i,c_k)\ sim(c_j,c_k))$$

Procedure

- Initially consider each instance as disjoint clusters at level K (m), m=0.
- Calculate the dissimilarity matrix (NX N) to hold all distances d (i, j).
- Locate similar clusters based on minimum distance, $D(p,q)=\min_{i\in p,j\in q} d(i,j)$, D is the Euclidean distance between clusters p and q.
- Move to next sequence m=m+1. Merge clusters (p) and (q) into a single cluster to form the next level cluster k (m) = d [(p) (q)].
- Modify the similarity matrix D, by removing the row and column values corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster (p, q) and old cluster in the previous level (k) is denoted as d[(k),(p, q)]=min d[(k),(p)],d[(k),(q)].
- Repeat step 2 until single cluster formation.

2.2 K-Means

K-Means algorithm is a heuristic algorithm that converges to local optimum. K-means is more sensitive to outliers. Flat clustering used to create partitions independent of each other. For the given data set D the K-Means algorithm partitions the data set into K clusters. Each cluster has a cluster center called centroid. The outliers can affect the mean centroid a lot. K-median is a better alternate for data with outliers. K-means will result in better output for equal or density clusters and poor results for non-convex shape clusters.

Procedure

Objective of K-Means algorithm is to minimize the total distortion. K-means is more sensitive to cluster center initialization.

Consider the data set D= {s_1, s_2, s_3, s_4 ….. s_r}

Step 1: Let {k_1, k_2, k_3… k_n} be chosen as 'n' initial centroid.

Step 2: Calculate the distance for each instance {s_1, s_2, s_3, s_4 ….. s_r} in data set D to each centroid.

Step 3: Assign each Si {s_1, s_2, s_3, s_4 …s_r} in D to the closest centroid cluster.

$$C_n=\{i{:}n=\arg\min_n \left\| s_i - k_n \right\|^2\}$$

Where: C_n are the datasets closest to cluster k_n

Step 4: Update the centroid of each cluster based on the new instance in the cluster.

$$K_n=\frac{1}{|c_n|}\Sigma_{i\in c_n} S_i$$

where:

$|c_n|$= size of cluster C_i

K_n = is the centroid of the cluster 'n' after adding the new instance.

Step 5: Continue step 2 until the convergence criteria.

Convergence Criteria

- Insignificant re-assignment of instances to different clusters.
- No or insignificant change in centroid value of the clusters.
- Insignificant decrease in the sum of squared error

$$E=\Sigma_{i=1}^{n}\sum_{s\in c_i} d(s,m_i)^2$$

N=total number of clusters. C_i = ith cluster

M_i=Centroid of ith cluster

d (s, m_i)=distance between instance s and centroid of cluster i

3. Implementation and Results

In this work stock prediction using cluster analysis precisely employs two methods. Hierarchical agglomerative clustering is used as the hierarchical methods and the K-means algorithm as the partitioning method.

Hierarchical agglomerative clustering

There are two approaches in the hierarchical methods, agglomerative approach and divisive approach. In this work the former approach is used. As each share is considered as single object, objects are clustered bottom up. This means it starts with the single object and forms separate groups to form clusters. Dendrogram is represented in Figure 1.

From Table 1, the percentage of increase of different stocks for different days is calculated using Eq(1).

$$Input_{s_i}=[\Sigma_{d=0}^{n}(closeprice_d - Openprice_d)/openprice_d]*100/n \tag{1}$$

Where, s_i – stock, d – day (n business days in 2 weeks)

Dissimilarity matrix is calculated as shown in Table 2 to identify the similarity between two stocks which controls the formation of clusters.

$$\text{Dis-sim (i) (j) = d ((input } s_i\text{), input } s_j\text{))} \tag{2}$$

Where, Dis-sim is dissimilarity matrix

- Group the stocks with minimum similarity index to form new clusters.

$$S_i, S_j = \min(\text{Dis-sim } (S_i)(S_j)) \tag{3}$$

- Update dendrogram with S_i, S_j as follows.
- Replace zero to the values in the ith and jth column

$$Dis-sim(n)(i)^{no.ofstocks}{}_{n=0}=0 and Dis-sim(n)=0 \tag{4}$$

- Repeat until dissimilarity value is very large

In the above diagram Figure 3 – x axis depicts the sample stocks that are considered for analysis, y axis depicts the dissimilarity level at which stocks are clustered simultaneously a weightage factor for each stock is calculated which is used further rank the stocks as shown in Table 3.

$$Weight_s=\sum_{d=0}^{0}[(closeprice_d - openprice_d)/openprice_d]*w$$

K-Means

The input to the k-means algorithm would be the resultant stocks from the hierarchical agglomerative clustering and percentage of raise is calculated using Eq(5).

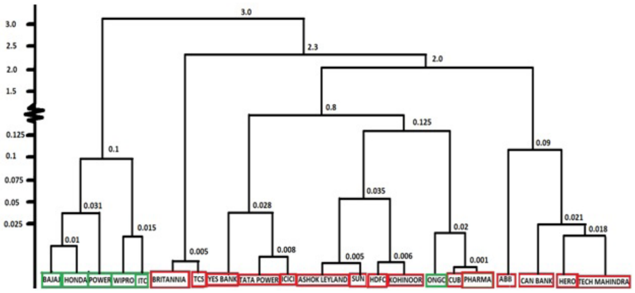


Figure 1. Dendrogram representation of HAC here.

Table 1. Sample Actual Data Sets downloaded from NS

YESBANK		ZEEL		ZENSARTECH		ZODIACLOTH	
Open Price	Close Price	Open Price	Close Price	Open Price	Close Price	Open Price	Close Price
380.00	379.55	290.10	287.80	388.55	382.95	240.00	227.00
365.00	379.20	287.00	290.25	395.00	387.50	222.00	233.40
362.45	362.00	288.50	285.90	390.70	393.30	214.90	220.65
360.60	360.35	280.90	287.95	393.00	390.70	201.90	215.30
365.90	362.05	280.55	278.35	391.00	390.95	198.00	196.55
367.25	365.40	283.50	280.15	405.90	390.85	200.95	199.75
359.50	363.70	274.30	282.35	387.85	402.20	196.10	196.00
355.80	355.85	279.30	276.90	376.00	387.30	197.00	198.00
368.10	361.35	285.00	279.20	386.70	374.25	197.50	196.00
370.00	364.75	279.20	287.75	378.05	386.70	197.05	197.50
WHIRLPOOL		WIPRO		WOCKPHARMA		WSTCSTPAPR	
Open Price	Close Price	Open Price	Close Price	Open Price	Close Price	Open Price	Close Price
230.65	227.10	557.35	548.50	436.70	441.05	52.35	51.75
235.80	230.10	571.00	562.80	435.60	438.55	51.65	52.70
230.75	233.65	567.90	569.65	442.00	439.95	51.50	52.05
226.95	229.60	558.65	568.20	439.00	440.35	52.95	52.00
217.70	224.35	547.00	552.35	449.90	443.65	51.05	51.90
205.35	216.95	540.55	545.10	459.85	448.35	52.65	51.05
205.90	205.35	550.00	540.45	453.00	459.40	52.30	51.80
205.90	205.40	564.00	547.60	455.70	449.80	52.50	51.85
204.85	203.75	562.00	563.60	446.35	457.35	51.35	52.30
205.05	203.55	559.90	566.75	441.30	445.25	52.55	52.30

Table 2. Sample dissimilarity matrix of sample 9 stocks

	3MINDIA	ABAN	ABB	ABGSHIP	ABIRLANUVO	ACC	ADANIENT	ADANIPORTS
3MINDIA	0	0.6114	1.187887	0.972478	0.384962	0.798992	0.89337	0.4051
AARTIIND	0.20381311	0.815214	1.3917	1.176291	0.588775	1.002805	1.097183	0.608913
ABAN	0.61140045	0	0.576486	0.361077	0.226439	0.187592	0.281969	0.2063
ABB	1.18788694	0.576486	0	0.215409	0.802925	0.388895	0.294517	0.782787
ABGSHIP	0.97247769	0.361077	0.215409	0	0.587516	0.173486	0.079108	0.567378
ABIRLANUVO	0.38496165	0.226439	0.802925	0.587516	0	0.41403	0.508408	0.020138
ACC	0.79899212	0.187592	0.388895	0.173486	0.41403	0	0.094378	0.393892
ADANIENT	0.89336976	0.281969	0.294517	0.079108	0.508408	0.094378	0	0.48827
ADANIPORTS	0.40510012	0.2063	0.782787	0.567378	0.020138	0.393892	0.48827	0

$$input_s = [(closeprice_m - openprice_m)/openprice_m]*100 \quad (5)$$

- Arbitrarily choose different values for k to decide the number of clusters needed. Initially stocks from datasets as the initial centers

$$m^k_{i=o} = rand(Input_s) \quad (6)$$

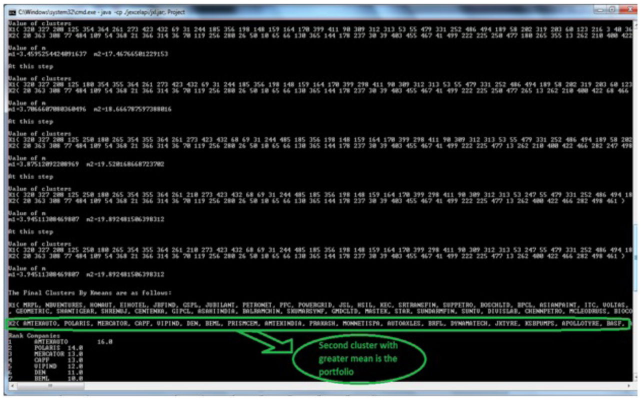
- Assign each stock to the cluster to which the stock is most similar based on the mean value of the stocks in the cluster.

$$cluster(x^{k1}_{i=0}, y) = \min(diff(m_o, input_s)\ diff(m_1, input_s)) \quad (7)$$

The Figure 2 shows the formation of clusters representation using the index representation of each stock. The

Table 3. Sample intermediate results of the Hierarchical Agglomerative Clustering (HAC)

Raising Stocks			
Rank	Weightage	Stocks	Stocks no.
1	16	AMTEKAUTO	20
2	14	POLARIS	363
3	13	MERCATOR	308
3	13	CAPF	77
4	12	VIPIND	484
5	11	DEN	109
6	10	BEML	54
6	10	PRISMCEM	368
6	10	AMTEKINDIA	21
7	9	PRAKASH	366
7	9	MONNETISPA	314
7	9	AUTOAXLES	36
8	8	BRFL	70
8	8	DYNAMATECH	119
8	8	JKTYRE	256
8	8	KSBPUMPS	280
8	8	APOLLOTYRE	26



C:\Windows\system32\cmd.exe - java -cp ./jexcelapi/jxl.jar; Project

```
Value of clusters
X1( 320 327 208 125 354 364 261 273 423 432 69 31 244 185 356 198 148 159 164 170 399 411 90 309 312 313 53 55 479 331 252 486 494 189 58 202 319 203 60 123 216 3 40 36
X2( 20 363 308 77 484 109 54 368 21 366 314 36 70 119 256 280 26 50 10 65 66 130 365 144 178 237 38 39 403 455 467 41 499 222 225 250 477 180 265 355 13 262 210 400 422

Value of m
m1=3.4595254424091637  m2=17.467665081229153

At this step

Value of clusters
X1( 320 327 208 125 180 354 355 364 261 273 423 432 69 31 244 185 356 198 148 159 164 170 399 298 411 90 309 312 313 53 55 479 331 252 486 494 189 58 202 319 203 60 123
X2( 20 363 308 77 484 109 54 368 21 366 314 36 70 119 256 280 26 50 10 65 66 130 365 144 178 237 38 39 403 455 467 41 499 222 225 250 477 265 13 262 210 400 422 68 466

Value of m
m1=3.7066607080360496  m2=18.666787597388016

At this step

Value of clusters
X1( 320 327 208 125 250 180 265 354 355 364 261 273 423 432 68 69 31 244 405 185 356 198 148 159 164 170 399 298 411 90 309 312 313 53 55 479 331 252 486 494 189 58 202
X2( 20 363 308 77 484 109 54 368 21 366 314 36 70 119 256 280 26 50 10 65 66 130 365 144 178 237 38 39 403 455 467 41 499 222 225 477 13 262 210 400 422 466 282 247 498

Value of m
m1=3.87512092208969  m2=19.520168668723702

At this step

Value of clusters
X1( 320 327 208 125 250 180 265 354 355 364 261 210 273 423 432 68 69 31 244 405 185 356 198 148 159 164 170 399 298 411 90 309 312 313 53 247 55 479 331 252 486 494 18
X2( 20 363 308 77 484 109 54 368 21 366 314 36 70 119 256 280 26 50 10 65 66 130 365 144 178 237 38 39 403 455 467 41 499 222 225 477 13 262 400 422 466 282 498 461 )

Value of m
m1=3.94511308469887  m2=19.892481506398312

At this step

Value of clusters
X1( 320 327 208 125 250 180 265 354 355 364 261 210 273 423 432 68 69 31 244 405 185 356 198 148 159 164 170 399 298 411 90 309 312 313 53 247 55 479 331 252 486 494 18
X2( 20 363 308 77 484 109 54 368 21 366 314 36 70 119 256 280 26 50 10 65 66 130 365 144 178 237 38 39 403 455 467 41 499 222 225 477 13 262 400 422 466 282 498 461 )

Value of m
m1=3.94511308469887  m2=19.892481506398312


The Final Clusters By Kmeans are as follows:

X1( MRPL, NBVENTURES, HONAUT, EIHOTEL, JBFIND, GSPL, JUBILANT, PETRONET, PFC, POWERGRID, JSL, HSIL, KEC, SRTRANSFIN, SUPPETRO, BOSCHLTD, BPCL, ASIANPAINT, ITC, VOLTAS,
, GEOMETRIC, SHANTIGEAR, SHREBUJ, CENTENKA, GIPCL, ASAHIINDIA, BALRAMCHIN, SKUMARSYNF, GMDCLTD, MASTEK, STAR, SUNDARMFIN, SUNTV, DIVISLAB, CHENNPETRO, MCLEODRUSS, BIOCO

X2( AMTEKAUTO, POLARIS, MERCATOR, CAPF, VIPIND, DEN, BEML, PRISMCEM, AMTEKINDIA, PRAKASH, MONNETISPA, AUTOAXLES, BRFL, DYNAMATECH, JKTYRE, KSBPUMPS, APOLLOTYRE, BASF,
Rank Companies
1       AMTEKAUTO       16.0
2       POLARIS  14.0
3       MERCATOR 13.0
4       CAPF     13.0
5       VIPIND   12.0
6       DEN      11.0
7       BEML     10.0
```

Second cluster with greater mean is the portfolio

Figure 2. Sample clusters formed in subsequent iterations.

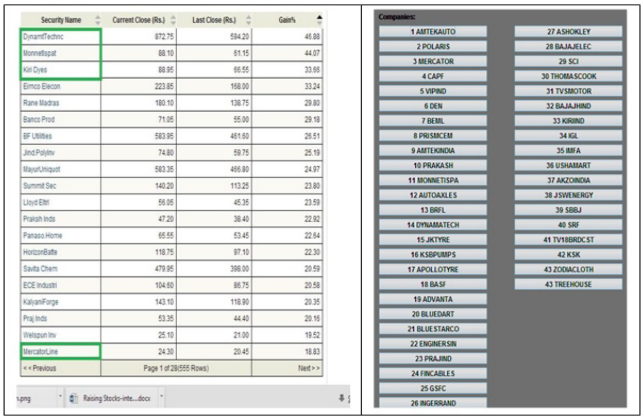
sample stock marked in Figure 2 is represented with the names of the stocks grouped together

- Now calculate the new centroid value for the newly formed cluster using Eq(8).

$$m_i = \sum_{j=0}^{k1} [cluster(i)(\,j)] \qquad (8)$$

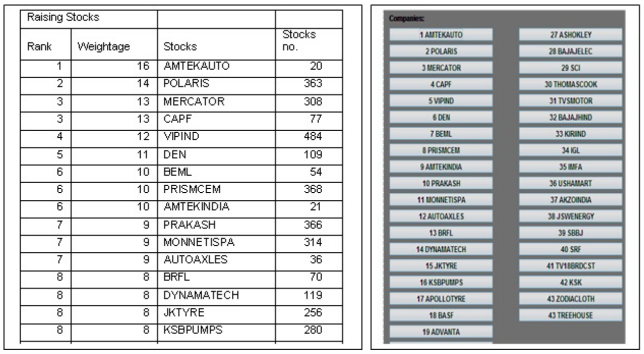
- Repeat from step 2 until no significant change in centroids for repeated iterations.

The Figure 3 depicts the results obtained by this work with the actual scenario at the end of the day. The advantage of k-means algorithm is it is more scalable and efficient. The algorithm is also sensitive to noise and outlier data. The final result obtained by portfolio generator is compared with the day's actual result for accuracy as shown in Figure 3. Comparison between results of HAC and k-means is shown in Figure 4. Figure 5 represents the movement of stock of AMTEKAUTO Company in the course of a day.



Security Name	Current Close (Rs.)	Last Close (Rs.)	Gain%
DynamtTechnc	872.75	594.20	46.88
Monnetispat	88.10	61.15	44.07
Kiri Dyes	88.95	66.55	33.66
Eimco Elecon	223.85	168.00	33.24
Rane Madras	180.10	138.75	29.80
Banco Prod	71.05	55.00	29.18
BF Utilities	583.95	461.60	26.51
Jind.Polyinv	74.80	59.75	25.19
MayurUniquot	583.35	466.80	24.97
Summit Sec	140.20	113.25	23.80
Lloyd Eltrl	56.05	45.35	23.59
Praksh Inds	47.20	38.40	22.92
Panaso.Home	65.55	53.45	22.64
HorizonBatte	118.75	97.10	22.30
Savita Chem	479.95	398.00	20.59
ECE Industri	104.60	86.75	20.58
KalyaniForge	143.10	118.90	20.35
Praj Inds	53.35	44.40	20.16
Welspun Inv	25.10	21.00	19.52
MercatorLine	24.30	20.45	18.83
<< Previous	Page 1 of 28(555 Rows)		Next >>

Figure 3. Comparison of the portfolio generator with actual result.



Raising Stocks			
Rank	Weightage	Stocks	Stocks no.
1	16	AMTEKAUTO	20
2	14	POLARIS	363
3	13	MERCATOR	308
3	13	CAPF	77
4	12	VIPIND	484
5	11	DEN	109
6	10	BEML	54
6	10	PRISMCEM	368
6	10	AMTEKINDIA	21
7	9	PRAKASH	366
7	9	MONNETISPA	314
7	9	AUTOAXLES	36
8	8	BRFL	70
8	8	DYNAMATECH	119
8	8	JKTYRE	256
8	8	KSBPUMPS	280

Figure 4. Comparison results of HAC and K-means.

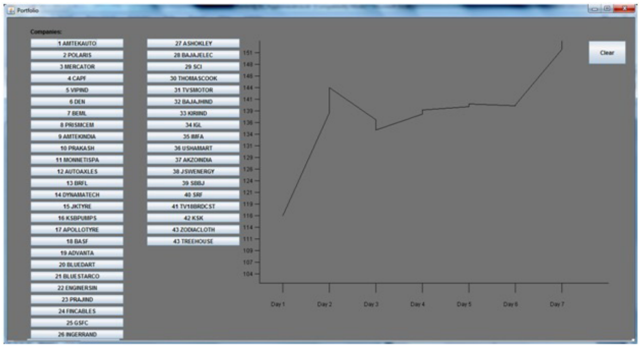


Figure 5. Final result with sample graph of AMTEKAUTO Company.

4. Conclusion

Since the stock market has huge amount of data sets, it is viewed as a complex domain of research analysis. This work depicts how clustering approaches in data mining such as hierarchical agglomerative clustering and the K-means algorithm is employed to predict the value of stocks to decide over future investments. The application is tested with the sample stocks collected from national stock exchange it could be extended to other stock exchanges also. The graph provided at the end proves us that the portfolio recommends the consistently performing stock through the specified period. Though we have restricted the data sets for a period of 22 business days, it could be extended in future to increase the reliability. As well extend the application for intraday trading exclusively.

5. References

1. Pauksto A, Raudys A. Intraday forex bid/ask spread patterns – analyzing and forecasting. IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr). 2013.
2. Nair RB, Mohandas VP, Sakthivel NR. A Genetic algorithm optimized decision tree- SVM based stock market trend prediction system. (IJCSE) International Journal on Computer Science and Engineering. 2010; 2(9):2981-88.
3. Cheu EY, Kwoh CK, Zhou Z. On the two-level hybrid clustering algorithm. International Conference on Artificial Intelligence in Science and Technology. 2004; p. 138-42.
4. Cheung WS, Ng HS, Lam KP. Intraday stock price analysis and prediction. Proceedings of the IEEE International Conference on Management of Innovation and Technology. 2000.
5. David E, Thawornwong S. The use of data mining and neural networks for forecasting Stock market returns. Expert Systems with Applications. 2005; 29(4):927–40.
6. Zhang D, Jiang Q, Li X. Application of Neural Networks in Financial Data Mining. World Academy of Science, Engineering and Technology. 2007; 1(1):225-28.
7. Yao J, Kong S. The application of stream data time series pattern reliance mining in stock market analysis. IEEE International Conference on Service Operations and Logistics, and Informatics. 2008.
8. Lam KP, Mok PY. Stock price prediction using intraday and AHIPMI data. Proceedings of the 9th International Conference on Neural Information Processing. 2002; 5:2167–71.
9. Voditel PP, Deshpande U. A stock market portfolio recommender system based on association rule mining. Applied Soft Computing. 2013; 13(2):1055-63.
10. Suresh Babu M, Geethanjali N, Satyanarayana B. Clustering approach to stock market prediction. Advanced Networking and Applications. 2012; 3(4):1281-91.
11. Liao SH, Ho HH, Lin HW. Mining stock category association and cluster on Taiwan stock market. Expert Systems with Applications. 2008; 35(1-2):19-29.
12. Hajizadeh E, Ardakani HD, Shahrabi J. Application of data mining techniques in stock markets: A survey. Journal of Economics and International Finance. 2010 Jul; 2(7):109-18.
13. Kambey S, Thakur RS, Jalori S. Applications of data mining technique in stock market (An analysis). International Journal of Computer & Communication Technology. 2012; 3(3):109-18.
14. Debashish D, Safa SA, Noraziah A. An Efficient Time Series Analysis for Pharmaceutical Sector Stock Prediction by Applying Hybridization of Data Mining and Neural Network Technique. Indian Journal of Science and Technology. 2016 Jun; 9(21):1-7.