Cluster Driven Candlestick Method for Stock Market Prediction

Yogita Patil
Computer Science and IT
Deogiri College
Aurangabad, India
patilmyogita@gmail.com

Manish Joshi
School of Computer Sciences
KBC North Maharashtra University
Jalgaon, India
joshmanish@gmail.com

***Abstract*—Trend prediction of the volatile stock market has been an interesting and challenging task for many researchers over many years. In this paper, we present how rough set-based BIRCH clustering can be used to develop stock data prediction model.**

The proposed model augments clustering with a popular technical analysis method called candlestick. BIRCH clustering algorithm is used to group stocks of varied sectors by taking into consideration the previous few days volatility. Further cluster analysis is carried out to predict stocks movement for next trading day. The proposed prediction model is different from existing models as it works on all NSE stocks from varied sector. Our model outperforms models that merely using clustering or candlestick techniques.

***Index Terms*—Candlesticks, Clustering, Rough BIRCH, Prediction, volatility**

I. Introduction

The financial market plays a vital role in promoting economic growth. The analysis of financial time series data such as stock exchange helps individual investor and financial institution to direct their funds to obtain profitable returns. We simply defined trading as an act of buying and selling company shares over the stock exchange to get the most effective returns. Making a smart trading decision without prior knowledge is like trying to play golf with a blindfold. To take smart trading decision every trader or individual requires a good understanding of market behavior, the performance of companies and knowledge about how and when to invest.

In order to get positive returns, every trader has to make the right choice of stock to trade. Apart from price and liquidity, one of most the important consideration in selecting stocks to trade is volatility. The term volatility refers to the upward and downward movement of the stock price. The stocks with higher volatility mean its price changes dramatically over a short period in either direction. Stock with small price volatility is unlikely to shoot up or fall on any given day. Moreover, many factors influence stock price, such as trading volume, news and financial report, political event, economic conditions and investor's expectation. Traditionally, high volatile stocks are considered riskier but on the other hand swing traders often seek out volatile stocks with the expectation of achieving higher returns.

To analyze stock market and behavior, many researchers have employed Machine learning, Regression, Artificial neural network, GA and Fuzzy logic based models. Generally, technical analysis tools, such as MACD (Moving Average Convergence Divergence), Exponential Moving Averages, RSI and Bollinger Bands, OHLC (Open, High, Low, Close) bar chart and candlestick patterns are used to investigate market trend and to develop a prediction model.

Generally, the regression model uses one month/six months/yearly stock data of specific stock or a few numbers of stocks for prediction of the stock price. Based on such prediction, investors take their trading decision. We reviewed many research papers and found prediction is done for a specific number of stocks or specific stocks of specific sectors. Apart from the existing prediction model, our model predicts the next day movement of like performing stocks of diverse sectors.

This paper presents cluster driven predictive model which will help traders or investors in trading decisions. Our focus was to predict the next day behavior of more than one stock. So we clustered stocks from varied sectors. We modeled the behavior and fluctuation of stocks and grouped them using Rough BIRCH clustering algorithm. In order to avoid the wrong prediction, we have explored to model the ambiguity while clustering stocks.

Rest of this paper is organized as follows. Section 2 reviews a few data mining techniques used in stock market prediction. Section 3 presents conventional candlestick basics. The problem definition and the data set used for the experiment is provided in section 4. The details of proposed framework for stock market prediction is given in section 5. The result analysis and observations are presented in section 6 and 7 respectively. Section 8 concludes the paper.

II. Related Work

Investing money in the unpredictable stock market is considered to be a very difficult task hence volatile stocks data must get investigated to get the insight of market trends and behavior. Many regression analysis tools are available to guide investors to take the right trading decision. Besides the regression technique, classification and clustering techniques are also used to find market trends and behavior.

Generally, to make trading decision stock traders combines

ISBN 978-1-7281-6202-7

both technical and fundamental analysis. In addition, data mining techniques like classification, clustering and regression are used to make a profitable decision. The purpose of clustering is to group data objects with similar characteristics in one group. Joel et al., [2] grouped stock data based on similar price movement pattern using self-organizing-map (SOM) clustering algorithm. [1] Proposed an unsupervised multi-scale data stream algorithm which detects trends for evolving time series based on a data driver data stream allowing for a simulation of on-the-fly trading decisions.

In order to manage the stock profile [4] studied selection and active trading of 138 stocks from many separate sectors and indices by the use of Partitioning Among Medoid (PAM) clustering algorithm. [5] Proposed a hybrid model to predict stock value movement using the opinion mining and clustering method to predict the National Stock Exchange (NSE). They have combined both the output from sentimental analysis and DENCLUE clustering to predict the stock market.

To provide a final prediction for each stock proposed hybrid model is analyzed based on the value of technical indicators. The research in [6] provides a list of the recommended stocks to the investor on a short-term basis. They have used hierarchical agglomerative clustering and the K-means algorithm to predict the value of stocks to decide over future investments. The author of [7] proposed an analysis system to help the investors to identify the more profitable companies using clustering and to predict the future price for that profitable company using a regression technique. The author has tested the performance of partitioning technique, hierarchical technique, model-based technique and density based technique using a number of the validity indexes like C-Index, Jaccard Index, Rand Index and Silhouette-Index. The clustered result is given as input to multiple regression techniques to predict the future stock price. The research by [8] proposed HRK (Hierarchical agglomerative and Recursive K-means clustering) predicts the short-term stock price movements after the release of financial reports. The proposed framework classifies stock time series based on similarity in the price trends. Proposed HRK is compared with Support Vector Machine (SVM). The proposed method outperforms SVM in terms of accuracy and average profits. Moreover, Candlestick analysis based prediction model is used to predict short term price fluctuation of stocks [13].

We proposed to use clustering as a pre-processing to candlestick approach for stock market prediction. The tails of candlestick approach are presented in next section.

III. Candlestick Approach

Candlestick graph has spread as a technical indicator to predict stock movement. Candlestick figures are associated with stock's behavior. Japanese Candlestick is a good technical indicator, which provides an educated guess about further stock movement. Many stock trading tools recognize candles and can predict the direction of the trend. Candlesticks chart are popular among traders of stocks and other financial markets because it provides market movement in detail. Conventional bar and line chart, however, need to be used in conjunction with other technical indicators to gain trading insight.

The candlestick is a plot with four attributes of stocks namely Open, High, Low and Close value of particular time period. Candlestick chat can be created for any time period like for monthly, weekly, hourly or even for one minute. Following figure 1 represents an image of a basic candlestick, which provides the information of stock opening price, closing price, highest price and lowest price and color for different time stamp.

1) The body of the candlestick is called the real body, which represents the price range between close and open prices.
2) The vertical lines above and bottom of the real body are called upper shadow and lower shadow, which represent the highest and the lowest price during the trading day.
3) The "Red" real body illustrates that the open price is higher than the close price, which shows the stock trend is decreasing, and when close price is higher than the open price, it shows the "Green" real body representing that the stock trend is increasing.

Candlesticks can accurately pick up the up on the changes in trend which occur in the financial markets. Candlesticks are able to give clues to price action and the mood of the market towards a certain stock or index. Traders often use candlesticks as a good visual aid to check a movement of particular price within a certain time period.

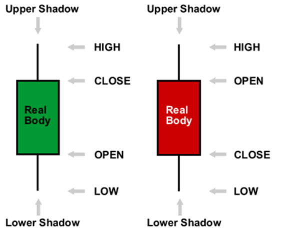


Fig. 1. Basics of Candlestick Figures

IV. Problem Definition

Predicting next day movement of the volatile stock market has been an interesting and challenging task for many researchers over many years. Many researchers have developed stock market prediction tools using various approaches such as regression, Artificial Neural Network, Support vector machines, genetic algorithms, machine learning algorithms and Fuzzy time series, etc. These tools predict next day movement of a few numbers of stocks or specific stock. We proposed a clustering driven stock data prediction model. Our objective is to predict next day movement of like performing stocks of varied sectors traded in the NSE (National Stock Exchange, India) using cluster driven candlestick method.

Let,

D = $\{d_1,d_2,d_2,....d_N\}$ is data set of N trading day,
where each d_i = $\{x_1,x_2,x_3,.....x_M\}$ is a trading day containing M stocks.
C=$\{c_1,\ c_2,....c_p\}$ is a set of clusters of stocks for every d_i (i=1 to N).

Problem:

1) Find like performing stock list for N trading days.
   i.e. SL = $\{x_i,\ x_j \in c_k\ \forall$ N days $\}$
   Where, i, j ranges from 1 to M and k range from 1 to p.
2) Predict $d_{N+1^{th}}$ day behavior of all like performing stocks.
   i. e. $\forall\ x_i \in$ SL predict N+1th day behavior.

V. PROPOSED FRAMEWORK FOR STOCK MARKET PREDICTION

Our objective is to check whether clustering can be helpful to get a moderate accuracy rate in prediction trend. Following figure 2 describes the proposed framework for stock prediction. First, clustering algorithm groups daily stock data based on fluctuation and behavior (bullish/bearish) parameters,broadly into three distinct cluster types namely positive, negative and mixed. Cluster analysis is carried out to find meaningful information out of it.

The main objective of cluster analysis task is to find out if clustered stock for a particular number of days does follow the same pattern? Based on our findings we have developed two predictive models which will help traders to take a smart trading decision for a group of stocks. Generally, conventional candlestick method and tools based on regression analysis are used for predicting stock movement. We aimed to design a predictive system augmented with a clustering solution to predict stock movement.

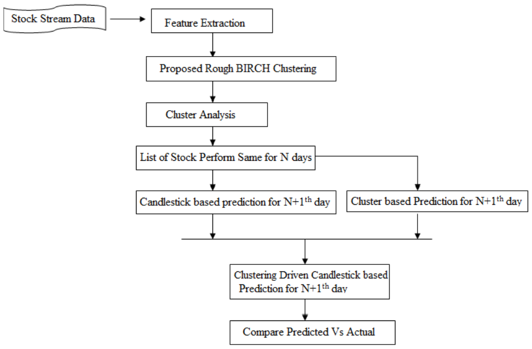

Stock Stream Data
Feature Extraction
Proposed Rough BIRCH Clustering
Cluster Analysis
List of Stock Perform Same for N days
Candlestick based prediction for N+1th day
Cluster based Prediction for N+1th day
Clustering Driven Candlestick based Prediction for N+1th day
Compare Predicted Vs Actual

Fig. 2. Block Diagram of Stock Market Prediction

A. Dataset

National Stock Exchange (NSE) Limited is leading stock exchange of India. It is established in 1994 and recognized as the largest stock exchange in India. In order to design the predictive trading strategy based on the clustering technique, we used day stock data from the NSE website. We have extracted and experimented on daily stock data from NSE for business days (Monday to Friday). Original stock data comes with many attributes but in order to find daily stock behavior and fluctuation we used only OHLC attributes of each working day.

The stock data which we collect from the National stock exchange website is in the form of raw data. So, we converted the data into a well-structured form using Z score normalization to re-scale transformed data.

B. Feature extraction

Stock data collected from the NSE website comprises of many attributes. Our objective is to predict the behavior of a group of stocks for next day. Original OHLC attributes of stock are transformed in such way that could significantly determine day behavior and fluctuation. Each attribute is derived from OHLC price in percentage.

C. Prediction Procedure

Clustering driven candlestick prediction model for financial stock data stream consist of following steps.

1) **Clustering**
   To increase the prediction accuracy rate of the stock market we have used Rough BIRCH clustering algorithm as a pre-processing technique.
2) **Cluster Analysis**
   We carried out cluster analysis task to obtain a list of stocks, which behaves same (fluctuate in a specific range) for N days. We have proposed a cluster driven candlestick based stock prediction model. We have developed a mathematical model to recognize a candlestick pattern. The mathematical model is used to predict next day movement of like performing stock before and after augmenting cluster statistics to OHLC data.

In order to develop cluster driven prediction model, we have modified the average N days OHLC value of the clustered stocks(like performing stocks belongs to SL) using cluster statistical information i. e. mean and standard deviation values. The behavior of each clustered stock is compared with the overall cluster behavior. If the behavior of each clustered stock does not match with overall cluster behavior, then we have used cluster statistical values to change N days average OHLC data of clustered stocks to modified O‘H‘L‘C‘ data. The supervised learning approach is used to design rules that will generate modified O‘H‘L‘C‘ data. We have explored different formulas to incorporate cluster information to OHLC data. Based on this learning by example approach it is found that the rules presented in Algorithm 1 are used to obtain cluster driven O‘H‘L‘C‘ data which can effectively predict next day movement of like performing stocks for considered time period.

The N+1th day prediction of clustered stocks is obtained through clustering, conventional candlestick method and clustering driven candlestick method using N days original and modified OHLC data.

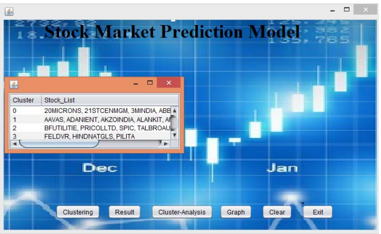
We have developed a mathematical model in JAVA to recognize candlestick figure and provide prediction based on predefined hypothesis. The rules for the candlestick figure identification were obtained from [11].

We have developed Stock Data Prediction Model using JAVA

Algorithm 1 Cluster Analysis:

```
Initialization:
$CC_p$ represents cluster Centroid of $c_p$ $^{th}$ cluster. p range from 1 to k
$\sigma_p$ represents standard deviation value of the p$^{th}$cluster.
$C_p$CLOSE represents average CLOSE value of $p^{th}$ cluster.
if $d_N$CLOSE $<$ $C_p$CLOSE then
    Incorporate CC$_p$ to N days OHLC data to obtain modified O'H'L'C' data.
else
    Incorporate CC$_p$ and $\sigma_p$ to N days OHLC data to obtain modified O'H'L'C' data.
end if
```

and executed on system with Intel Core 5 2.60GHZ, 4GB memory. Following figure 3 presents a screencap of Stock Data Prediction Model.



Cluster	Stock_List
0	20MICRONS, 21STCENMGM, 3MINDIA, ABB
1	AAVAS, ADANIENT, AKZOINDIA, ALANKIT, AI
2	BFUTILITIE, PRICOLLTD, SPIC, TALBROAU
3	FELDVR, HINDNATGLS, PILITA

Fig. 3. Stock Data Prediction Model

VI. Result Analysis

This section demonstrates the performance of predictive model designed to predict the next day movement of stocks. We have augmented clustering information to OHLC data to get modified OHLC values, which will further used in prediction of stock movement. Likewise, clustering information is used to design the clustering driven candlestick based prediction model. Morris found that candlestick charts are used for short term investment prediction purpose and the most efficient time for prediction is 10 days [10]. Hence, in order to validate our cluster analysis solution time window of 10 days is considered and accordingly, we have tested the actual behavior of like performing stocks for each month.

Proposed machine learning based technical analysis is suitable for the short term (one week to a month) stock behavior prediction. Generally, prediction of stock movement is based upon a series of successive candlesticks, as well as its shapes, their position relative to each other. In order to predict the future directional movement of a stock a series of candlestick figures for a particular time period is studied. Instead of relying only on a series of candlestick figures for a single stock, our proposed clustering approach for financial data stream gives a possible prediction for a group of stocks.

Cluster analysis result for the first quarter of the year that is for the month of January to March, and the last quarter of the year that is for the month October to December 2018 for N=10 days is presented in table I, II, III, IV, V and VI respectively. Performance of stock data prediction accuracy obtained through clustering, conventional candlestick method and clustering driven candlestick method using OHLC attributes is presented in table VII.

TABLE I
Stock list for the month of January for N=10 days

Clusters	Stock List				
1	BSE	IBULHSGFIN	PETRONET	RELDIVOPP	UTINEXT50
2	ALBK	AXISBANK	BANKBEES	COLPAL	GAIL
	GOLDIWIN	INDIGO	KOTAKNIFTY	LUPIN	MRF
	NIFTYBEES	PGHH	STAR	TATACOMM	
3	BOSCHLTD	SETFNN50			

TABLE II
Stock list for the month of February for N= 10 days

Clusters	Stock List				
1	ASIANTILES	CARERATING	DEEPIND	NAVKARCORP	NIFTYBEES
	PFC				

VII. Observation

Table I, II, III, IV, V and VI shows cluster analysis result i. e. cluster wise list of like performing stocks for N (N=10)days. In stock market stock price fluctuate in minute due to supply and demand. Likewise, many factors affects the stock price fluctuation namely political climate, epidermic outbreaks (Covid-19), dividends, economy, interest rate etc.

From table I, II, III, IV, V and VI we observed that few stocks like MRF, NIFTYBEES and LIQUIBEES are grouped with the

TABLE III
Stock list for the month of March for N= 10 days

Clusters	Stock List				
1	AMBIKCO	APARINDS	BANCOINDIA	BERGEPAINT	CIPLA
	FIEMIND	IRB	KOTAKGOLD	MRF	NILKAMAL
	ZEEL				
2	AGCNET	APLLTD	BHARATIWIN	BRITANNIA	BRNL
	COLPAL	CPSEETF	DABUR	GESHIP	GITANJALI
	GOLDBEES	GOLDIWIN	GREAVESCOT	HAVELLS	HDFCMFGETF
	HDFCNIFETF	ICRA	IDBIGOLD	INFIBEAM	INGERRAND
	INSECTICID	JUNIORBEES	KIRLOSENG	L&TFH	LIQUIDBEES
	PIDILITIND	POLYMED	QGOLDHALF	RALLIS	SETFGOLD
	SETFNN50	SUNTECK	TATAINVEST	UTINEXT50	VOLTAS

TABLE IV
Stock list for the month of October for N = 10 days

Clusters	Stock List				
1	21STCENMGM	INDRAMEDCO	SJVN	TATAINVEST	
2	3MINDIA	ACC	AIFL	BHARATFORG	DENABANK
	EIHOTEL	ENIL	GUJALKALI	HAL	ICICILOVOL
	JAGRAN				

TABLE V
STOCK LIST FOR THE MONTH OF NOVEMBER FOR N = 10 DAYS

Clusters	Stock List				
1	AMBER	DBCORP	HINDUNILVR	RELCONS	
2	AIFL	BAJAJHLDNG	HDFCNIFETF	ICICIB22	IGARASHI
	KOTAKNIFTY	KSK	LIQUIDBEES	LIQUIDETF	M50
	NITINFIRE	OMAXE	SABEVENTS	SANOFZ	SETFNIFBK
	TATACOFFEE				

TABLE VI
STOCK LIST FOR THE MONTH OF DECEMBER FOR N = 10 DAYS

Clusters	Stock List				
1	ABB	AXISGOLD	GNFC	HDFCBANK	MIDHANI
	QGOLDHALF				
2	AARTIDRUGS	BANKBEES	CARERATING	CERA	CONCOR
	DEN	GOLDBEES	HATHWAY	HDFCMFGETF	HDFCNIFETF
	ICICIGOLD	ICICILIQ	ICICINV20	ICICINXT50	KOTAKGOLD
	LIQUIDBEES	LIQUIDETF	MHRIL	MONTECARLO	PETRONET
	RELDIVOPP	SETFGOLD	TATAINVEST	TCS	TRITURBINE
3	INDOSOLAR	RADIOCITY			
4	EIMCOELECO	NIACL			

different stocks of different sectors in the month of January, February, March and December. These altogether divert stocks are appeared to be as a group which otherwise would have never been even thought of. The significance of clustering is apparent from it.

From figure 4, we observed that the average prediction accuracy (five months) obtained without augmenting cluster information to candlestick method is 50.99% and with augmenting cluster information is 72.99%. It is clearly observed that if we augment cluster information to the prediction model, it enhances the prediction accuracy.

The N+1th day stock movement prediction accuracy for a maximum number of identical stocks for the first and last quarter of the year 2018 for N=10 days is given in table VII. From table VII we observed that our proposed clustering driven candlestick method gives better prediction accuracy than conventional candlestick based prediction for a considered time window.

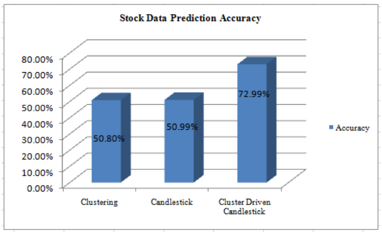


Fig. 4. Average Stock Data Prediction Accuracy

TABLE VII
MONTH WISE PREDICTION ACCURACY

Month	Prediction Accuracy				
	Number of Stocks	Clustering	Candlestick	Clustering Driven Candlestick	% Increase compare to Candlestick
1	14	42.58%	57.14%	64.28%	7.14%
2	33	78.12%	65.62%	81.25%	15.63%
3	10	60%	30%	80%	50%
4	14	42.85%	50%	78.57%	28.57%
5	24	30.43%	52.17%	60.86%	8.69%

VIII. CONCLUSION

In this paper, we have presented the use of rough set based clustering technique for forming a predictive trading strategy for NSE stock data. We have designed a cluster driven candlestick based predictive model to predict next day movement of like performing stocks. Clustering groups the stocks with similar behavior and hence, it is used as one of the phases in developing the proposed predictive model. We have compared candlestick prediction with clustering driven candlestick prediction and observed that cluster driven candlestick based prediction outperforms conventional candlestick prediction. The unique feature of our proposed cluster-based predictive model is to provide prediction for a group of stocks belongs to varied sectors, whereas other techniques provide a prediction for a given stock (symbol) at a time.

Our proposed cluster-based predictive model can help swing traders to take trading decisions for more than one stock. We observed that obtained clustering result if augmented with a candlestick, can improve the overall stock movement prediction accuracy.

More granular is the data better is the result. So we plan to carry out the same experiment for more granular data i.e. for hourly trading.

REFERENCES

[1] B. Dragut, Andreea. (2010). Stock Data Clustering and Multiscale Trend Detection. Methodology and Computing in Applied Probability. 14. 87-105. 10.1007/s11009-010-9186-7.

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

[3] A. Haldane, "Patience and finance," Oxford China Business Forum, Beijing, September 2010.

[4] Craighead, Steven and Klemesrud, Bruce. (2002). Stock Selection Based on Cluster and Outlier Analysis.

[5] Vivek Rajput and Sarika Sanjay Bobde, Stock market prediction using hybrid approach, International Conference on Computing, Communication and Automation (ICCCA) 2016, pp. 82-86.

[6] Renugadevi, T and R R, Ezhilarasie and Sujatha, M and Umamakeswari, A. (2016). Stock Market Prediction using Hierarchical Agglomerative and K-Means Clustering Algorithm. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i48/108029.

[7] B.S. Bini, Tessy Mathew, Clustering and Regression Techniques for Stock Prediction, Procedia Technology, Volume 24,2016, Pages 1248-1255, ISSN 2212-0173.

[8] J. T. Lee, Anthony and Lin, Ming-Chih and Kao, Rung-Tai and Chen, Kuo-Tay. (2010). An Effective Clustering Approach to Stock Market Prediction. PACIS 2010 - 14th Pacific Asia Conference on Information Systems. 54.

[9] Yogita S. Patil, Manish R. Joshi (2018), "Enhancing BIRCH Clustering with Rough Set Principles", International Journal of Computer Engineering and Applications, Volume XII, Special Issue, May 18, www.ijcea.com ISSN 2321-3469.

[10] G. L. Morris. Candlestick charting explained: Timeless techniques for trading stocks and futures. 3rd Ed. McGraw-Hill, New York, NY, 2006.

[11] Candlestick trading forum. pcf (personal criteria formulas) for telechart software. available at:http://www.candlestickforum.com/PPF/Parameters/16_263_/candlestick.asp

[12] M. M. Goswami, C. K. Bhensdadia and A. P. Ganatra, "Candlestick Analysis based Short Term Prediction of Stock Price Fluctuation using SOM-CBR," 2009 IEEE International Advance Computing Conference, Patiala, 2009, pp. 1448-1452.