Abstract—Stock market analysis and prediction has been one of the widely studied and most interesting time series analysis problems till date. Many researchers have employed many different models, some of them are linear statistic based while some non linear regression, rule, ANN, GA and Fuzzy logic based. In this paper we have proposed a novel model that tries to predict short term price fluctuation, using Candlestick Analysis. This is a proven technique used for short term prediction of stock price fluctuation and market timing since many years. Our approach has been hybrid that combines Self Organizing Map with Case Based Reasoning to indentify profitable patterns (candlestick) and predicting stock price fluctuation based on the pattern consequences.

Index Terms— Candlestick Analysis, Short Term Stock Prediction, Self Organizing Map, Case-Base Reasoning

I. INTRODUCTION

Stock data analysis has remained one of the challenging time series analysis problem over the years. Because of the complexity and intrications, the problem has received attention from many researchers. Owing to its high expected returns, there are many starting from local investor to a big finance houses who are interested in finding out how the stock market works? Many techniques have been derived over a period of time by many researchers and investors to predict the stock price behavior. These techniques are broadly classified into two categories Fundamental and Technical [13]. Fundamental analysis takes into account all those medium to long term factors like inflation, market capital, future business opportunities, EPS, dividends that affect the performance of a company and try to predict the future growth of the company, that in turn enables us to estimate future valuation and stock price of the company. This is rather a long term prediction where as Technical analysis try to predict short to medium term price movement of the stock based on the historical price data only and does not consider any environmental or fundamental factors.

This research suggests a technical analysis based method for predicting stock price fluctuation. It considers a special type of technical indicator known as candlesticks for predicting the short term (usually a day ahead) price movement of the stock price. Since its invention in 16th century (in Japan) candlestick has remain one of the proven trading indicator not just for stock but for any type of trading like commodity, crude just to name a few [12]. Unlike other technical indicators, which try to find statistical relation of current price with future price, candlestick indicator tries to find the investor’s sentiment on a given stock. The idea is simple “what will be the price of a stock tomorrow depends on what people think about that stock at the end of the day”. In order to indentify various candlestick patterns, we have used SOM and CBR as a pattern recognizer and pattern matching mechanism. Decision derived by the CBR (SOM) is in turn combined with other technical indicator like Stochastics and Volume, in order to improve the accuracy of prediction.

All available models for predicting stock price behavior can be divided into three category 1) Machine Learning based (Black box), 2) Expert knowledge based (White box) and 3) Combination of both (Hybrid). In Black box models [3][5] the system learns itself from previous experiences. Such a model does not require an expert, but the representation of the knowledge is complex and not verifiable (like ANN) where as in White box models [2][4][6] the representation of the knowledge is simple and verifiable but requires expert to impart knowledge in the system (like Fuzzy rule, CBR). System is only as good as the expert’s knowledge. Hybrid models [7][8] try to combine the benefits of both. In this paper we have proposed a hybrid model which combines SOM with CBR.

Rest of the paper is arranged as follow. II covers introduction to Candlestick analysis, III covers literature review, IV discuss the methodology, V discuss the proposed architecture, VI Analysis and Result, VII conclusions and VIII future works.

II. CANDLESTICK ANALYSIS BASED STOCK PREDICTION

A. Introduction to Candlestick Analysis

The data required to produce a standard candle chart includes the open, high, low, and close prices in a time period under study (usually one day). A candle pattern is composed by serial candle lines. The box that makes up the difference between the open and close is called body of candlestick. The height of the body is the range between the day’s open price and the day’s close price. When the body is black, it means that the closing price was lower than the opening price. When the closing price is higher than the opening, the body is white. The candlestick line may have small thin lines above and or
Candlestick pattern is sequence of three to five specific types of candlesticks occurring together. By the trading experience, the investor tries to identify the candle patterns to help themselves to make the decision such as to buy, sell, or hold the stock over short term period. There are many rules that have been defined to identify the candle patterns. Fig. 2 shows a candle pattern named “Evening Star”.

Rules for describing the above pattern are 1) upward trend is apparent 2) body of the first candle is white, continuing the current trend, second candle is with almost without body i.e. open price is same as close price and very small shadow (which indicates states of indecision) 3) third day candle is with black body, and closing almost half way down the first white candle [18].

B. Psychology behind Candlestick Patterns

The meaning of above pattern can be concluded as follow, as the upward trend is apparent number of buyers are more compare to number of sellers in the market since last few days, but formation of the second day pattern indicates that the buyers are almost same as sellers and therefore the opening and closing price of the stock remains almost same. For a rally to continue buyers should be in majority but the black candle on third day indicates that sellers have taken control over market and therefore the price is likely to go further down.

III. LITERATURE REVIEW

ABDELMOUEZ, SHERIF, AMIR [10] have done a comparison of linear methods, multiple regression, and neural networks on the problem of forecasting stock prices. The results indicate the superiority of neural networks, liner method [1], and then the multiple regression method. YANG [2] has proposed a Fuzzy Logic based method for the short term and long term stock price predictions. The system proceeds through two steps 1) Clustering of Input data 2) Specification of output. KOHA, SAKAKIBARA, NISHIWA [3] have proposed a neural network based methods for stock forecasting. They have used both back propagation network and simple recurrent network and found that simple recurrent network is better because of its time “capturing capabilities”. SENG-CHO, CHAU CHEN, CHI YUHANG [5] have propose an intelligent stock trading decision support system that can forecast the buying and selling signals according to the prediction of short-term and long-term trends using rule-based neural networks. PEI CHANN CHANG, CHIN-YUAN FAN and JYUN-JIE LIN [4] have presented financial time series-forecasting model by a case based clustering TSK fuzzy rule system [11] for stock price prediction. The model integrates a case based reasoning technique, and TSK Fuzzy Rule based system (TSK) to construct a prediction-system based on historical data and technical indexes. CHUNG-HON LEON LE, WENSUNG CHEN and ALAN LIU [6] have proposed a fuzzy logic based knowledge representation methodology to represent the candlestick pattern, and implement a prototype system to help the user to identify the candlestick patterns automatically. PEI-CHANN CHANG and CHIEN-YUAN LAI [8] have proposed a hybrid system to combine the self-organizing map (SOM) of neural network with case-based reasoning (CBR) method, for sales forecast of new released books. MURICA, RODELLAR, and KOLAKOWSKI [7] have described a soft computing approach to solve the damage identification problem using CBR and SOM.

IV. METHODOLOGY

This research uses Case based reasoning for matching the current candlestick pattern with standard candlestick pattern and thus predicting the future price movement based on the consequences of standard patterns. Instead of using convectional distance based case matching method, our approach uses Self organizing maps to match current case with the available cases based on case index.

A. Case Based Reasoning

Case based reasoning is one of very effective and widely used problem solving technique. It is very effective because it works very similar to human mind. It compares the new problem, also called new case, with already observed cases and uses this past experience to suggest the solution for current case. There are four basic stages of case base life cycle
Retrieve: In this stage the current case is compared with all standard cases available in case base and most matching case is retrieved.

Reuse: the solution of the current case is proposed based on solution of matching case.

Revise: the proposed solution is revised after the getting the actual results.

Retain: modified case is again stored in case base for later use.

B. Self Organizing Maps

Self organizing maps are special type of neural network that has only two layer input and output layer, the output layer of the self organizing map is arranged as a two dimensional grid while input layer is one dimensional vector [14]. The number of nodes in the input layer is equal to number of input where as the number of node in output depends on the problem at hand, normally a grid of 10X10 is found sufficient for most of the problems[9].

![Architecture of self organizing maps](image)

Fig. 3: Architecture of self organizing maps

One of the useful properties of SOM is that it maps N-Dimensional input vector into 2-Dimension point in the output space i.e. for a given input vector only one node is activated as a time in output grid. Therefore for each different input vector I_n, we have different output point C_n, this mapping is done according to topological neighborhood i.e. if two input vector I_1 and I_2 are similar then output points C_1 and C_2 corresponding to them are nearer to each other on the 2D grid.

C. Combining SOM with CBR

In this approach all standard candlestick patterns are modeled as cases as shown below.

![Modeling of “Evening start” candlestick pattern as a case](image)

Fig. 4: Modeling of “Evening start” candlestick pattern as a case

Here last three days open, high, low, close price information is used to define a candlestick pattern. Lowest price value of three days is consider as 0% and highest of three days is 100%, rest of other price value are model between 0 to 100% with respect to highest and lowest. A case is defined by open, high, low, close value of three days, for example above pattern is model as follow.

NAME (op1,hg1,lg1,cl1;op2,hg2,lg2,cl2;op3,hg3,lg3,cl3)

EVENING_STAR (20,70,0,60;80,100,70,80;70,80,25,30)

![Combining SOM with CBR for pattern identification and matching](image)

Fig. 5: Combining SOM with CBR for pattern identification and matching

Figure 5 shows the combination of SOM with CBR for candlestick pattern matching and recognition task [7]. The process is explained by following steps.

Step1: All standards candlestick pattern are model as observed cases (as shown before) and stored in pattern data base along with their possible consequences.

Step2: SOM is trained with all standard patterns available in pattern data base.

Step 3: Each pattern is fed to SOM one by one and point C(x, y), in 2D output space (known as case index) corresponding to each pattern is determined and stored in case base.

Step 4: For prediction, three days open, high, low, close price information of a given stock is selected and modeled as new case.

Step5: The new case is fed to SOM and output case index N(x, y) is determined.

Step6: The distance of new case index N(x, y) with respect to all available standard case indexes C(x, y) is computed and one with the minimum distance is selected as matching case.

Step7: The prediction for new case is done based on the consequences of standard matching case combined with other indicators like stochastics and volume.

Step8: Finally the predicted price movement is compared with actual price movement and if required consequences of matching case are revised and retained in pattern database for future use.
V. PROPOSED ARCHITECTURE

A. Database

Contains three types of data bases, 1) Stock Data which used to store historical values of open, high, low, and close price along with volumes for all stock listed in NSE from February 2000 to July 2007 2) Pattern Database used to store all standard candlestick patterns modeled as cases 3) Case base used to store mapping of pattern to unique case index assigned by SOM during training process.

B. Stock Data Normalization Module

Used to normalize the selected stock data and make them suitable for being used as input to various modules like SOM, Stochastics and Volume.

C. Stochastic Indicator Module

Generates a stochastic chart from the historical data and find out whether the current price is oversold or overbought region with respect to last 15 days price records.

D. Volume based Indicator Module

Generate a volume based chart from the historical data and find out whether current volume is high, low or medium with respect to last 15 days volume records.

E. SOM Training Module

Responsible for generating case base, first it will train SOM with all the patterns available in Pattern Database and then find out the case index C(x, y), which is a point in 2D space, corresponding to each pattern. These mapping of patterns with their case index are stored in case base.

F. Update Data Module

Allows the user to automatically update stock data using .xls or .csv files which are available from NSE Stock Exchange on daily basis;

G. Visual Pattern Editor Module

Allows the user to enter new candlestick pattern by simply drawing the patterns on a canvas, as the there are only fuzzy definition available for candlestick patterns; it will be advantageous to enter the patterns graphically rather them textually because the textual description may vary from user to user.

H. SOM Module

Takes normalize stock data for the given stock as new input case and feed it to Trained SOM in order to find output case index N(x, y) in the output space.

I. CBR Module

Responsible for finding out most suitable matching case for new input case by finding the distance between new case index N(x, y) and already available case indexes C(x, y) for all patterns. One with minimum distance is most matching case.

J. Decision Support Module

Combine the consequence of matching case with output of stochastics and volume module and derive a conclusion, for example if pattern indicates negative reversal and stochastics are in overbought area and volume is very high then it is very certain that the price will fall down over next few days.

VI. ANALYSIS AND RESULTS

The characteristic of proposed method is compared with existing methods and following facts can be derived from our analysis.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Linear</th>
<th>ANN</th>
<th>FUZZY</th>
<th>CBR</th>
<th>SOM-CBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self learning</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Statistics based</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Knowledge based</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Verifiable knowledge</td>
<td>NA</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Require expert</td>
<td>NA</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Require training</td>
<td>NA</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Adaptable</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Clustering ability</td>
<td>N</td>
<td>D</td>
<td>D</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>Extendable</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Non-linear</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Fuzzification required</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Computations efficiency</td>
<td>P</td>
<td>Y</td>
<td>A</td>
<td>A</td>
<td>V</td>
</tr>
<tr>
<td>Dependency on data set</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

*V- very good, A-average, P-poor, Y-Yes/Available, N-No/Not available
E- Easy, D-Difficult, NA- Not Applicable

Proposed system is expert knowledge based and does not depend on data. With help of visual editor tool even a normal
user can also enter profitable candlestick pattern as it appears on chart. It is possible to draw patterns on candlestick chart, thus verify the knowledge imparted in the system. It is possible to add new pattern in the system with help of pattern editor which makes it extendable. System keep on revising the pattern consequences by comparing proposed result with actual result thus it is also an adaptable and self correcting system. It is computationally more promising than normal CBR, because all standards N-D, cases are mapped as a point in 2-D space by SOM, so finding the matching case problem is reduced to finding distance between two points in 2-D space rather than N-D space. As all patterns are mapped as a point in 2-D space it is possible to cluster the available cases, which will reduce total number of cases and thereby reducing the possibility of too many matching cases. It does not require fuzzyfication-defuzzyfication of input data and the prediction is non-linear in nature. System requires training of SOM whenever a new pattern is added in to the system, so the SOM can indentify that pattern and able to map it on 2-D space.

VII. CONCLUSION

Because of advance in field of Data Mining and AI, soft computing based automatic analysis of data has received a considerable attention from research community. Automated analysis of stock data is one of such problems which are widely studied by many researchers. This research identifies capability of Candlestick analysis as one of the useful technical indicators for short term prediction of stock price fluctuations, and market timing. The capability of SOM /CBR as a pattern identifying and forecasting model are also studied and compared with the other existing models and a hybrid model for candlestick analysis based prediction of stock price fluctuation using SOM/CBR is proposed.

VIII. FUTURE WORK

In the next phase of the research the primary goal will be to implement the proposed system and evaluate the system with historical database. The accuracy of system for indentifying various candlestick patterns and efficiency of the system for predicting price fluctuation will be estimated. Based on the result and analysis new possibilities for further improvement will be explored.

REFERENCES