COMPARITVE STUDY TO ANALYZE STRATEGIES FOR STOCK MARKET PREDICTIONS USING DATA MINING AND ARTIFICIAL NEURAL NETWORKS

Ankit Raj Saxena\(^1\), Aayush Gupta\(^1\)

\(^1\)Department of Computing Science and Engineering, VIT University.

Email: ankitrajsaxena@yahoo.com

Received on: 15.10.2016

Abstract

Data mining is based on one of the assumptions that historic data can be used to predict future decisions. Forecasting stock market has been an important topic that has had many scholars’ attention in the recent years. The hidden patterns in past data can be discovered using data mining techniques and used in financial markets to make intelligent investment decisions. This paper does a comparative analysis of two such methods.

Keywords: Data Mining, Neural Networks, Stock Market, Bollinger Signal, Typical Price, Chaikin Money Flow indicator, Stochastic Momentum Index, Relative Strength Index, Bollinger Bands, Moving Average.

1. Introduction:

The financial market is extremely tough to model, given its non-linear and non-parametric nature. The premise of stock market prediction is to use the publicly available data from the past and find patterns in it using data mining techniques to make effective predictions. The financial market contains a vast amount of data, which cannot be managed by traditional databases. The field of data mining will help to classify the relevant data from this repository. Taking after the suspicion of specialized examination that examples exist in value information, it is conceivable on a basic level to utilize data mining methods to find these examples in a robotized way. Once these examples have been found, future costs can be anticipated.

2. Literature Review:

The stock market has evolved in a lot of ways in the recent past - with the advancement in communication, companies can now attract investors worldwide. This has expanded the scope of the market to a global level. This in turn, has increased the amount of data to a large extent. Many authors have asserted the use of fundamental analysis – which is...
based on information about macroeconomics, industry and the company. In this paper, we compare two methods for classification of financial data using data mining. The first method uses decision trees for the classification of data. The second method uses five methods of stock prediction - Typical Price (TP), Chaikin Money Flow indicator (CMI), Stochastic Momentum Index (SMI), Relative Strength Index (RSI), Bollinger Bands (BB), Moving Average(MA) and Bollinger Signal.

3. Comparison and Description of Proposed Framework:

First Method

This method involves the classification of data using decision trees. The following steps are followed in the procedure:

2.1. Collecting The Data:

The data was collected from a few companies due to its vast extent. The data was collected based on 6 attributes - Previous day close price of the stock(Previous), Current day open price of the stock(Open), Current day minimum price of the stock(Min), Current day maximum price of the stock(Max), Current day close price of the stock(Last), The action taken by the investor on this stock(Action).

2.2. Preparing The Data:

The initial data gathered consisted of continuous numeric values. This data was transformed to discrete values by generalizing to a higher level. The criterion used to transform the data was based on the previous day’s closing price of the stock. If the values of the attributes min, max and open were more than this, they were assigned the high level value of Positive. If the values were lesser, the value of Negative was assigned. Finally, if the values were equal, then Equal was assigned.

2.3. Constructing The Tree:

The classification model is built by constructing the decision tree. The decision tree technique is used mainly because the construction of decision tree classifiers does not require any domain knowledge, thus it is appropriate for exploratory knowledge discovery. Also, the steps in constructing the tree are simple and fast. The decision tree is constructed by determining the most useful attribute, by using the information gain metric. The gain ratio was then used to rank the attributes and construct the tree. The tree was then traversed, and a set of classification rules were generated, with and without pruning.
2.4. Deploying The Model:

The classification rules generated from the previous step can be used in a learning system, such as an artificial neural network (ANN), so that it can predict the next day’s stock prices.

**Second method**

In this method, five algorithms for analyzing stocks were combined to predict the next day’s stock price.

2.5. Typical Price:

The typical price is the average of the high low and closing prices. It can be calculated using the following formula:

\[
TP = \frac{(H+L+C)}{3}
\]

*Fig.1: Formula for calculating average of the high and low closing prices.*

Where H=High, L=Low, C=Close.

2.6. Chaikin Money Flow Indicator:

The Chaikin Money Flow Indicator is based upon the principle of Accumulation/Distribution. The principle of Accumulation/Distribution is in turn, based on the mid-point of stock price on closing. If the stock closes above the value of its mid-point \([(\text{high}+\text{low})/2]\) for the day, then that day has an accumulation. Otherwise, there was a distribution on this day. The CMI (Chaikin Money Indicator) is calculated by adding the values of accumulation/distribution for thirteen periods divided by thirteen period sum of the volume.

2.7. Stochastic Momentum Index:

Before discussing the Stochastic Momentum Index (SMI), let us talk about Stochastic Oscillator. The Stochastic Oscillator calculates the value of the closing price of the stock relative to the high/low range. The SMI is quite similar to the Stochastic Oscillator. The main difference is that the SMI calculates the value of the closing price of the stock relative to the midpoint of the high/low range, i.e. \([(\text{high}+\text{low})/2]\). The SMI values vary from +100 to -100. Extremely high SMI values indicate overbought conditions. Similarly, extremely low SMI conditions indicate oversold conditions.

2.8. Relative Strength Index: This marker looks at the quantity of days a stock completes up with the quantity of days it completes down. It is computed for a specific time traverse more often than not somewhere around 9 and 15 days.
The normal number of up days is separated by the normal number of down days. This number is added to one and the outcome is utilized to gap 100. This number is subtracted from 100. The RSI has a range somewhere around 0 and 100. A RSI of 70 or above can demonstrate a stock which is overbought and due at a fall in cost.

\[
\text{RSI} = 100 - \left( \frac{100}{1 + \text{RS}} \right);
\]

\[
\text{RS} = \frac{\text{AG}}{\text{AL}}
\]

\[
\text{AG} = \frac{[\text{PAG} \times 13 + \text{CG}]}{14};
\]

\[
\text{AL} = \frac{[\text{PAL} \times 13 + \text{CL}]}{14};
\]

**Fig2: Formulae for calculating Relative Strength Index.**

- PAG = Total of Gains during past 14 periods/14
- PAL = Total of Losses during past 14 periods/14
- AG = Average Gain
- AL = Average Loss
- PAG = Previous Average Gain
- CG = Current Gain
- PAL = Previous Average Loss,
- CL = Current Loss

2.9. Bollinger Bands:

These consist of three bands – upper, middle and lower band. These bands are based on moving average. This is because moving average is used to calculate standard deviation. The upper band is two SDs above the moving average. The lower band is two SDs below the moving average. The middle band is the moving average itself. The upper and lower lines are plotted by market unpredictability.

At the point when the market is unstable the space between these lines broadens and amid times of less instability the lines come nearer together. The center line is the basic moving normal between the two external lines (groups). As costs draw nearer to the lower band the more grounded the sign is that the stock is oversold the cost ought to soon rise. As costs ascend to the higher band the stock turns out to be more overbought importance costs ought to fall.
2.10 Moving Average: The most famous marker is the moving average. This demonstrates the average cost over a timeframe. For a 30 day moving average you include the end costs for each of the 30 days and partition by 30. The most widely recognized averages are 20, 30, 50, 100, and 200 days. Longer time ranges are less influenced by every day value variances. A moving average is plotted as a line on a diagram of value changes. At the point when costs fall underneath the moving average they tend to continue falling. On the other hand, when costs transcend the moving average they tend to continue rising. In this calculation we are utilizing the ideas of various systems like SMI, RSI, CMI and Bollinger band. By utilizing the benefits of all the above systems we can make the net benefit as high. The purchase flag and offer signs can be delivered by utilizing the capacity Bollinger signals. By contrasting and moving average hybrid we can discover how powerful the new system is. We are keeping the moving average hybrid as the benchmark.

4. Comparison

In the first method, the classified data was fed into a K-means neural networking algorithm. In the network, sixty percent of the data was used for training the algorithm, and the rest was used for testing. After the testing, the accuracy of the model ranged from 44% to 55%. The reason for the low accuracy might be due to internal factors in the company that affect the stock prices such as news, performance etc. It may also depend on external factors such as political events and decisions.

In the second method, the advantages of various methods such as TP, CMI, RSI, Bollinger Bands etc. were combined to increase the efficiency of prediction. The resultant prediction had an accuracy of about 52.8%. Among the various methods employed, it was noted that Bollinger Bands had the highest profitable signal of around 84%. This calculation could maybe be utilized as a purchasing or offering sign or it could be utilized to offer certainty to a broker's forecast of stock costs.

5. Conclusion

Studying the two algorithms, we note that while the accuracy of the first method varied over a wide range, it can be drastically improved by considering more data, and training the neural network using an even more diverse data set. This will definitely increase the accuracy of the decision tree algorithm. However, the second method employs algorithms
which focus on short term data. As such, the accuracy of the second algorithm cannot be increased much further with any more data.

References

1. Qasem A. Al-Radaideh, Adel Abu Assaf, Eman Alnagi – “Predicting stock prices using data mining techniques”.


