IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Financial Stock Price Forecast Using Classification

Dr. Swapna Borde[1], Austrin F. Dabre[2], Harsh M. Kamdar[3], Rahul Y. Purohit[4]

Abstract: Data mining and predictive technologies developed using computer automated programs do a fair amount of trade in the market. Historic data holds the essential memory for predicting the future direction, is a well-founded theory about data mining. One way of predicting if future stocks prices will increase or decrease is Data Analysis. This technology is designed to help investors discover hidden patterns from the historic data that have probable predictive capability in their investment decisions. A challenging task of financial time series prediction is the prediction of the pricess of financial stock markets. Five methods of analyzing stocks were combined to predict if the day's closing price would increase or decrease. These methods were Typical Price (TP), Bollinger Bands, Relative Strength Index (RSI), CMI and Moving Average (MA). This paper discussed various techniques which are able to predict with future closing stock price will increase or decrease better than level of significance. Also, it investigated various global events and their issues predicting on stock markets. It supports numerically and graphically.

Keywords: Data mining, Data analysis, TP, RSI, CMI, MA.

I. INTRODUCTION

In many fields including trading, finance, statistics and computer science, forecasting the direction of stock prices is a persistent and widely studied topic. The motivation for which is naturally to predict the direction of future prices such that stocks can be bought and sold at profitable positions. To analyze stocks and make investment decisions professional traders use fundamental or analytical financial tools. It is nowadays a common notion that vast amounts of capital are traded through the Stock Markets all around the world. National economies are strongly linked and heavily influenced by the performance of their Stock Markets.

The uncertainty is the characteristic that all Stock Markets have in common, which is related with their short and long term future state. Such a feature is undesirable for the investor but it is also unavoidable whenever the Stock Market is selected as the investment tool. The best that one can do is to try to reduce this uncertainty. Stock Market Prediction (or Forecasting) is one instrument in this process. The researchers and academics of Stock Market Prediction task are divided into two groups those who believe that we can devise mechanisms to predict the market and those who believe that the market is efficient and whenever new information comes up the market absorbs it by correcting itself, thus there is no space for prediction.

II. REVIEW OF LITERATURE

Data mining can be described as "making better use of data". Data mining is a key research area because unmanageable amounts of data are faced by each and every human in the current technology scenario; hence, data mining or knowledge discovery apparently affects all of us. Ideally, we would like to develop techniques for "making better use of any kind of data for any purpose". However, we argue that this goal is too demanding yet. Over the last three decades, increasingly large amounts of historical data have been stored electronically and this volume is expected to continue to grow considerably in the future. Yet despite this wealth of data, many fund managers have been unable to fully capitalize on their value. This paper attempts to determine if it is possible to predict if the closing price of stocks will increase or decrease on the following day. The approach taken in this paper was to combine five methods of analysing stocks and use them to automatically generate a prediction of whether or not stock prices will go up or go down.

Data mining involves six common classes of tasks:[1]

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

III. EXISTING SOLUTION

A cluster is an ordered list of objects, which have some common characteristics. So a cluster is the collection of

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

objects which are similar and are different from the objects that belong to other clusters. Base objective of clustering is to find out the inherent grouping in a set of unlabelled data. There is no standard to find the best clustering algorithm which is independent of the dataset. It depends on user who must supply the criterion in such a way that the result of clustering will suits their needs.

Clustering algorithms can be applied in many domain like in marketing to find groups of customers with similar behaviours and their buying habits, in biology for classification of plants and animals, or in library for ordering books Therefore a system based on clustering could only specify the objects in our case the shares or stocks into the category where we could group them into the shares which previously had comparatively low closing prices and higher closing prices, but it was not possible to give result as a possible outcome predicting the rise or fall in the prices of the stock prices in the future.

IV. PROPOSED SYSTEM

The proposed system is basically a research based system which classifies the given database using output extracted form stock indicators. At present, data mining is a new and important area of research, and classification itself is very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The combination of data mining method and efficient data using classification model can greatly improve the efficiency of data mining methods, and it has been widely used.[4].

A descendant of CLS and ID3. Like CLS andID3, C4.5 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible rule set form [3].

A. Decision trees

B. An overview of stocks.

A) Typical price

The Typical Price indicator is calculated by adding the high, low, and closing prices together, and then dividing by three. The result is the average, or typical price.[2] The Typical Price indicator is simply an average of each day's price. The Median Price and Weighted Close are similar indicators

Algorithm:

1. Inputting High, Low, Close values of the daily share
2. Take an output array and add the values of H,L,

TP =[H+L+C] where H=High; Low; C=Close, where
3

The TP greater than the bench mark we have to sell or to buy.

B) Chaikin money flow indicator

Chaikin's money flow is based on Chaikin's accumulation/distribution. Accumulation/distribution in turn, is based on the premise that if the stock closes above its midpoint [(high+low)/2] for the day, then there was accumulation that day, and if it closes below its midpoint, then there was distribution that day.

Chaikin's money flow is calculated by summing the values of accumulation/distribution for 13 periods and then dividing by the 13-period sum of the volume.[2]

The Following formula was used to calculate CMI:

$$CMI = \frac{sum(AD,n)}{sum(VOL,n)}; \quad AD = VOL\ (\frac{CL-OP}{HI-LO})$$

AD stands for Accumulation Distribution, Where n=Period;

CL = today's close price; OP =today's open price;

HI = High Value; LO = Low value

C) Stochastic momentum index

Stochastic momentum index helps you see where the current close has taken place relative to the midpoint of the recent high to low range is based on price change in relation to the range of the price, which is better than the stochastic oscillator which used the relative close price . The SMI has a normal range of values between 100 to -100. When the present close price is higher than the median ,or mid-point , value of the high/low range, the resulting value is positive .when current closing price is lower than that of the midpoint of the high/low range, the SMI is negative value. Like the stochastic oscillator , the SMI is primarily used by traders or analysts to indicate overbought or oversold conditions in a market. Traders alsi use the SMI as a general trend indicator , interpreting values above 40 as indicative of a bullish trend and negative values greater than -40 as showing a bearish trend.

$$100x[\frac{[MOV[MOV[C-[5x[HHV(H,,13)+LLV(L,13]]]25,E,2,E]]}{[5x[MOV[MOV[[HHV(H,13+LLV(L,13)]]]],25,E],2,E]}$$

Where HHV=Highest high value.

LLV = Lowest low value.

E = exponential moving avg.

Using the following formula, exponential moving average was calculated.

$$EMA = [[(\ Price(i)-prevMVG)x(\frac{2}{N+1})]] + prevMVG$$

D) Relative strength index

The relative strength index is a technical momentum indicator that compares the magnitude of recent gains to recent losses in an attempt to determine overbought or oversold conditions of an asset. This indicator compares the number of days a stock finishes up with the number of days it finishes down. The average number of up days is divided by the average number of down days. This number is added to one and the result is used to divide 100. This number is subtracted from 100. The RSI has a range between 0 and 100. A RSI of 70 or above can indicate a stock which is overbought and due for a fall in price. When the RSI falls below 30 the stock may be oversold and is a good they can vary depending on whether the market is volatile.

RSI=100–(100/1+RS); RS=AG/AL

AG=[(PAG)x13+CG]/14; AL=[(PAL)x13+CL]/14

PAG = Total of Gains during past 14 periods/14

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

PAL = Total of Losses during past 14 periods/14
Where AG=Average Gain, AL=Average Loss PAG=Previous Average Gain, CG=Current Gain PAL=Previous Average Loss, CL=Current Loss
The following algorithm was used to calculate RSI:
Upclose = 0
DownClose = 0
Repeat for nine consecutive days ending today
If (TC > YC)
UpClose = (Upclose + TC)
Else if (TC < YC)
DownClose = (Down Close + TC)
End if

$$RSI=100-\frac{100}{1+[\frac{upclose}{downclose}]}$$

E) Bollinger bands
Bollinger Bands are based upon a simple moving average. This is because a simple moving average is used in the standard deviation calculation. The upper band is two standard deviations above a moving average; the lower band is two standard deviations below that moving average; and the middle band is the moving average itself**.** When the market is volatile the space between these lines widens and during of less volatility the lines come closer together. The middle line is the simple moving average between the two outer lines (bands).We receive the Bollinger signals from Bollinger signals.

$$stdDev = \sum_{i=1}^{N} (price(i)\text{-}MA(N))^2$$

$$Uppperband=MA+D\sqrt{\sum_{i=1}^{N} \frac{(price\ (i)-MA)2}{N}}$$

$$Lowerband=MA-D\sqrt{\sum_{i=1}^{N} \frac{(price\ (i)-MA)2}{N}}$$

Where D=No. of standard deviations applied.

V. IMPLEMENTATION METHODOLOGY

Here we will study how the entire project idea was implemented into a run able code which produces a tangible output, i.e. a running version of software. Hence in this chapter we will learn implementation of the project. Implementation is vital part in software building and designing. Our project is implemented in Java.

We have used NetBeans to build our Project. We are going to use MySQL database to store data. WampServer will serve as an interface between the program and database.

A) Gathering the input Data.
The input data for different stock indicators are the open, close, high, low prices etc. The data have been classified according to the capitalization of their respective share prices. We create separate databases for separate types of shares. So we can have the output for different types of shares and also the combined prices can be used for observing the output. The data has to be put in a particular format for predictive analysis tool. So we have to create separate database for the use of indicators and our predictive analysis tool.[3][8]

B) Data Pre-processing for data mining.
Even though there are defined stages – selection and generation where the data may have already been pre-processed to enhance its class predictive power, the choice of constructed the classification technique to be used may call for further adjustment. Different forms of normalisation are usually required (Witten and Frank, 1999). Data transformation could be in different forms.[4]

C) Decision Trees: DT is well-known to be one other effective classification technique in several domains (Chau et al., 2001). It is a way of representing series of rules that lead to a class or value. DT models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions. The prediction could be to predict categorical values (classification trees) when instances are to be placed in categories or classes. We use C4.5 a classification tree in which we give the input as yes or no and high ,mid or low values and get the gain ratio for the given set of data.

D) Use of IBM WEKA for data mining.
Real-time classification of data, the goal of predictive analytics, relies on insight and intelligence based on historical patterns discoverable in data. These patterns are presumed to be causal and, as such, assumed to have predictive power. That predictive power is used to compare the accurate results of IBM weka with the results of our own program and observe the accuracy.

Classification methods address these class prediction problems. The most familiar of these is probably the logit model taught in many graduate-level statistics courses. The example in this article will use the J48 classifier, included in Weka. The J48 is a tree-based classifier that considers a given set of features at each branch. It has few options, so it is simpler to operate and very fast.[4][2]

E) C4.5 algorithm.
if T is NULL then
return failure;
end if
if S is NULL then
return Tree as a single node with most frequent classlable in T;
end if
if [S] = I then
return Tree as a single node S;
end if
set Tree = {}
for o € S do
set Info(a,T) = o,andSplit Info(a,T) = o;
ComputeEntropy(a);
for a € values(a,T) do
setTa,v as the subset of T with attribute a= o;
Info(a,T)+ = | Ta.v| Entropy(a,v);
T,a
SplitinInfo(a,T) + = - | Ta,v| log | Ta,v |
T,aT,a
end for
Gain(a,T) = Entropy(a) – In fo(a,T)
Gain Ratio(a,T) = Gain(a,T)
SplitInfa(a,T)
end for
setabest = argmax{Gain Ratio(a,T)}

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

attachabest into Tree
for a € values(abest,T) do
call (4.5(Ta,v)
end for
return Tree

G) Flowchart

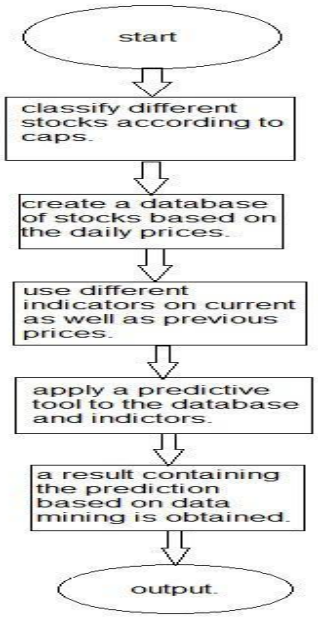


Fig 2

VI. RESULT

A. Screenshot

The result for moving average was calculated.

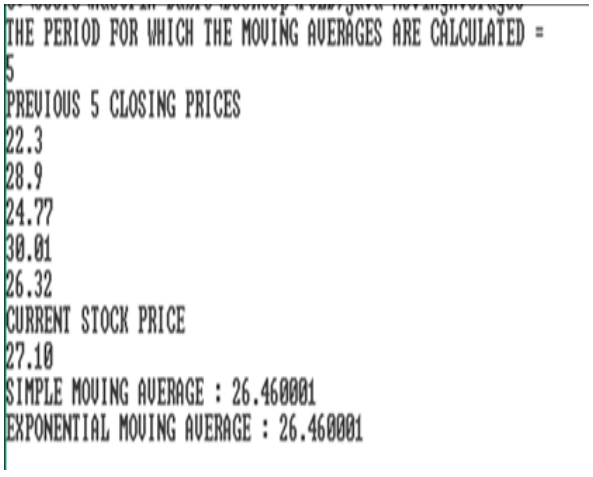
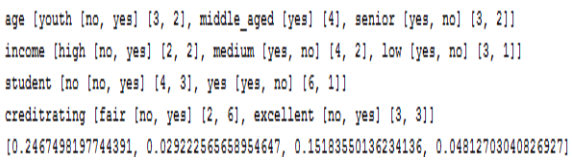


Fig.3



The result for information gain can be seen in the above figure. Fig.4

VII. CONCLUSION

In this paper we present a classification tool approach for classifying different stocks based on their historic data. For this approach we propose different phases.

Currently we have finished with:

- Design the financial indicators.
- Design of the classification algorithm.
- Database of stocks for classification.

We will implement the following modules:

- Database integration.
- Classification algorithm output.
- Comparing the output with IBM WEKA output.

REFERENCES

[1]. Financial Stock Market Forecast using Data Mining Techniques K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam.
[2]. https://www.ibm.com/developerworks/library/os-weka2/
[3]. http://iknowfirst.com/the-big-data-solution-for-wall-street
[4]. http://www.comp.leeds.ac.uk/scsod/MSc%20Dissertation.pdf
[5]. http://documents.software.dell.com/statistics/textbook/data-mining-techniques
[6]. http://www.salford-systems.com/products/cart
[7]. https://en.wikipedia.org/wiki/Category:Classification_algorithms
[8]. http://www.investopedia.com/terms/m/marketcapitalization.asp