STOCK MARKET PREDICTION USING KMP ALGORITHM FOR FAST NEWS SEARCH AND EFFECTIVE DATA MINING METHODS

FATEN ALZAZAH

Computer science and engineering Department, New York Institute of Technology, USA
E-mail: fatensubhi@gmail.com

Abstract- Before evolution of text mining techniques, data mining and statistical techniques were used to predict the stock movement based only on past prices, their weakness is that they depend heavily on structured data .Which neglects the important influence of unstructured information .In this paper, we will use the most important data mining techniques using Exponential smoothing and Linear Trend Line concurrently with text mining using fast String matching algorithm(KMP algorithm) for stock market prediction.

Keywords- KMP Algorithm, News Articles search, Data Mining, Exponential Smoothing, Linear Trend Line, Stock Market Prediction.

I. INTRODUCTION

Stock market prediction is the act of trying to identify the future movement of stock value of the company. The successful prediction of stocks movement will leads to more profit investors can made, this research aims to build a model that use one of the most efficient String Matching Algorithms to search for pattern in news articles, and we use two data mining techniques to study the past prices of stock market .The model seeks to extract useful information from data sources (e.g newspapers , web …..etc) through the identification and exploration of interesting patterns in the case of text mining. In this research we would develop a system which is able to use KMP (Knuth-Morris-Pratt) algorithm for text mining. For date mining we will use Exponential Smoothing concurrently with Linear Trend Line to model the movement of the stock market, and our concerning will be more on text mining techniques, since data mining techniques only based on studying of past prices of stock will not give accurate result in prediction.

II. LITERATURE REVIEW

Several papers related to the area of this research have been found and the most important ones are going to be discussed in this section. (Huang W, Nakamori Y, Wang S Y) investigate the predictability of financial movement direction with SVM by forecasting the weekly movement direction of NIKKEI 225 index. To evaluate the forecasting ability of SVM, they compare its performance with those of Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Back propagation Neural Networks. Further, they propose a combining model by integrating SVM with the other classification methods. Various stock market prediction system have been implemented that address different aspects of stock analysis and use different methods of prediction. For example, some used hybrid approach of classification rule of data mining and financial news articles effects. They presents the predicted system used data of BSE to predict the stock market trend. This system used hybrid approach of technical analysis and rule based classification of data mining. Simultaneously they check the effect of news article to daily stock price . First, they find out the pattern of incoming financial news and then calculate index value from formula (Soniya Dewangan1, S R Tandan2, Shagufta Farzana3, Rohit Miri).Another model, named Genetically optimized Neural Network (GNN), used different error metrics like Average Absolute Error (AAE), Mean Absolute Percentage Error(MAPE) ,Mean Square Error(MSE), Max AE, R-SQ ,during the study paper performs an experiment to forecast stock market movement in India using Artificial Neural Network (ANN) and Genetic Algorithm (GA) (Tuhin Mukherjee1, Aritra Banerjee) (Rupawari Jadhav1, M. S. Wakode) presents study on the survey of different methods that gave helps the prediction of stock market using financial news from Twitter data. It also discussed different classification algorithm for sentiment analysis and the comparison between them. Another system, built a stock domain specific dictionary via the NLP of 78,216 news articles taken from two different style media and showed the sentimental words, news sentiments and opinions calculated using the dictionary. Then, by conducting a stock prediction experiment, they found that the opinions extracted from the news could be useful in accurate predicting stock market movements (Yoosin Kim, Seung Ryul Jeong, Imran Ghani).

III. PROBLEM STATEMENT

Several papers related to the area of this research have been found and the most important ones are going to be discussed in this section. (Huang W, Nakamori Y, Wang S Y) investigate the predictability of financial movement direction with SVM by

forecasting the weekly movement direction of NIKKEI 225 index. To evaluate the forecasting ability of SVM, they compare its performance with those of Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Backpropagation Neural Networks. Further, they propose a combining model by integrating SVM with the other classification methods. Various stock market prediction system have been implemented that address different aspects of stock analysis and use different methods of prediction. For example, some used hybrid approach of classification rule of data mining and financial news articles effects. They presents the predicted system used data of BSE to predict the stock market trend. This system used hybrid approach of technical analysis and rule based classification of data mining. Simultaneously they check the effect of news article to daily stock price . First, they find out the pattern of incoming financial news and then calculate index value from formula (Soniya Dewangan1, S R Tandan2, Shagufta Farzana3, Rohit Miri).Another model, named Genetically optimized Neural Network (GNN), used different error metrics like Average Absolute Error (AAE), Mean Absolute Percentage Error(MAPE) ,Mean Square Error(MSE), Max AE, R-SQ ,during the study paper performs an experiment to forecast stock market movement in India using Artificial Neural Network (ANN) and Genetic Algorithm (GA) (Tuhin Mukherjee1, Aritra Banerjee) .(Rupawari Jadhav1, M. S. Wakode) presents study on the survey of different methods that gave helps the prediction of stock market using financial news from Twitter data. It also discussed different classification algorithm for sentiment analysis and the comparison between them. Another system, built a stock domain specific dictionary via the NLP of 78,216 news articles taken from two different style media and showed the sentimental words, news sentiments and opinions calculated using the dictionary. Then, by conducting a stock prediction experiment, they found that the opinions extracted from the news could be useful in accurately predicting stock market movements (Yoosin Kim, Seung Ryul Jeong, Imran Ghani).

IV. TIME SERIES PATTERNS

(David R. Anderson, Dennis J. Sweeney , Thomas A. Williams, Jeffrey D. Camm and Jeffrey W. Ohlmann, James J. Cochran and Michael J. Fry) A time series is a sequence of observations on a variable measured at successive points in time or over successive periods of time. The measurements may be taken every hour, day, week, month, or year, or at any other regular interval.1 The pattern of the data is an important factor in understanding how the time series has behaved in the past. If such behavior can be expected to continue in the future, we can use the past pattern to guide us in selecting an appropriate forecasting method.

4.1.Trend Pattern

A time series may also show gradual shifts or movements to relatively higher or lower values over a longer period of time. If a time series plot exhibits this type of behavior, we say that a trend pattern exists.

A trend is usually the result of long-term factors such as population increases or decreases, changing demographic characteristics of the population, technology, and/or consumer preferences. To illustrate a time series with a trend pattern, consider the time series of bicycle sales for a particular manufacturer over the past 10 years, as shown in Figure 1. Note that 21,600 bicycles were sold in year one, 22,900 were sold in year two, and so on. In year 10, the most recent year, 31,400 bicycles were sold. Visual inspection of the time series plot shows some up and down movement over the past 10 years, but the time series also seems to have a systematically increasing or upward trend.



Figure 1: bicycle sales for a particular manufacturer over the past 10 years

4.2. Seasonal Pattern

Seasonal patterns are recognized by seeing the same repeating patterns over successive periods of time. For example, a manufacturer of swimming pools expects low sales activity in the fall and winter months, with peak sales in the spring and summer months. Manufacturers of snow removal equipment and heavy clothing, however, expect just the opposite yearly pattern. Not surprisingly, the pattern for a time series plot that exhibits a repeating pattern over a one-year period due to seasonal influences is called a seasonal pattern. As an example of a seasonal pattern, consider the number of umbrellas sold at a clothing store over the past five years. Table 15.5 shows the time series and Figure 2 shows the corresponding time series plot.

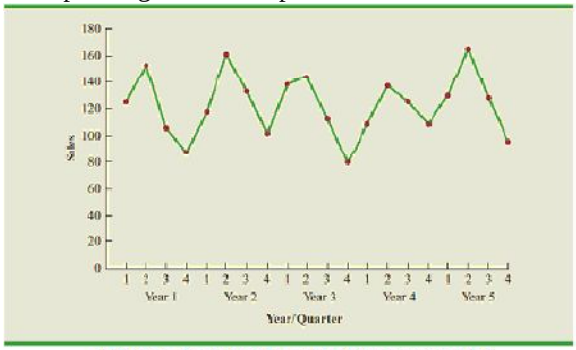


Figure 2: umbrella sales time series plot

4.3 Cyclical Pattern

Cyclical pattern exists if the time series plot shows an alternating sequence of points below and above the trend line lasting more than one year. Many economic time series exhibit cyclical behavior with regular runs of observations below and above the trend line.

Business cycles are extremely difficult, if not impossible, to forecast. As a result, cyclical effects are often combined with long-term trend effects and referred to as trend-cycle effects.

4.4 Irregular Pattern

These components are the random elements that are generally can't be predicted. such irregular pattern can be found by moving the trend ,cyclical and seasonal component.

V. TIME SERIES SMOOTHING METHODS

In cases in which time series is stable, and has no significant trend, seasonal, or cyclical effects, one can use smoothing methods to average out the irregular pattern, in our research we will use Linear Trend Line and Exponential smoothing to average out the irregular pattern of the time series.

5.1. Exponential Smoothing

The idea of exponential smoothing is to smooth the original series the way the moving average does and to use the smoothed series in forecasting future values of the variable of interest. In exponential smoothing, however, we want to allow the more recent values of the series to have greater influence on the forecast of future values than the more distant observations (Eva Ostertagova, Oskar Ostertag).

5.1.1 Forecasting With The Simple Exponential Smoothing (SES) Model

Let an observed time series y1,y2,…,yn . Formally, the simple exponential smoothing equation takes the form $\hat{y}_{i+1} = \alpha y_i + (1-\alpha)\hat{y}_i$

Where y_i is the actual known series for the period i, $y\hat{}_i$ is the forecast value of the variable Y for period i ,$y\hat{}i+1$ is the forecast value for time period i+1 and a is the smoothing constant. Smoothing constant (BROWN, R. G., MEYER, R. F.). Smoothing constant a is a selected number between zero and one

Example:

The simple exponential smoothing model can be illustrated by using data about number of personnel in the industrial production in Slovakia over the years 2001 – 2010 that are in the Table 1 below (Slovenskej republiky):

Tab. 1 Number of personnel in industrial production in Slovakia over the years 2001 – 2010

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
486459	488240	493259	497302	518142	510698	526075	532853	447685	430658

(Eva Ostertagova, Oskar Ostertag)to forecast the number of personnel in industrial production in Slovakia for the year 2011 that are in the table 2 below. By plotting of observed values we can see that we can use SES method. Since no forecast is available for the first period, then set the first estimate equal to the first observation. For illustration here try a= 0.4, a= 0.8 and a= 0.9.

.

Tab. 2 Exponential smoothing forecast using MATLAB

i	Yi	Y^i(0.4)	Y^i(0.8)	Y^i(0.9)
	1.0 e+005*			
1.0000	4.8646	NaN	NaN	NaN
2.0000	4.8824	4.8646	4.8646	4.8646
3.0000	4.9326	4.8717	4.8788	4.8806
4.0000	4.9730	4.8961	4.9218	4.9274
5.0000	5.1814	4.9268	4.9628	4.9685
6.0000	5.1070	5.0287	5.1377	5.1601
7.0000	5.2607	5.0600	5.1131	5.1123
8.0000	5.3285	5.1403	5.2312	5.2459
9.0000	4.1768	5.2156	5.3091	5.3203
10.0000	4.3066	4.9201	4.6433	4.5612
11.0000	NaN	**4.6747**	**4.3739**	**4.3320**
NaN	NaN	*0.3462*	*0.3148*	*0.3088*

This concept is illustrated in Figure 3, which shows a time series observed for periods 1 to10 and the forecast for period 11.

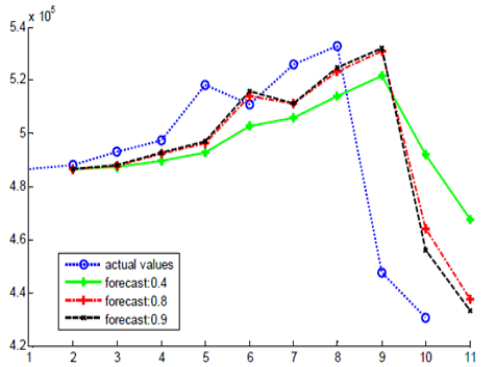


Fig 3 Plot for actual values and forecast values using $\alpha = 0.4$, $\alpha = 0.8$ and $\alpha = 0.9$

5.2 Linear Trend Line

(Dan Reid, Nada R. Sanders)Linear trend line is a time series technique that computes a forecast with trend by drawing a straight line through a set of data. This approach is a version of the linear regression technique, and is useful for computing a forecast when data display a clear trend over time. A linear trend line uses the following equation to generate a forecast:

$$Y = a + bX$$

where Y = forecast for period X
X = the number of time periods from $X = 0$
a = value of Y at $X = 0$ (Y intercept)
b = slope of the line

The steps for computing the forecast using a Linear trend line are as follows:

Step 1 Compute parameter b:

$$b = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

Step 2 Compute parameter a:

$$a = \overline{Y} - b\overline{X}$$

Step 3 Generate the linear trend line:

$$Y = a + bX$$

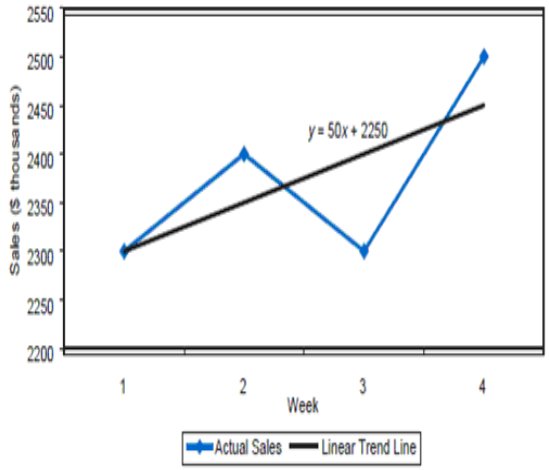
Step 4 Generate a forecast (Y) for the appropriate value of time (X)

Example: A manufacturer has plotted product sales over the past four weeks. Use a linear trend line to generate a forecast for week 5 .

	Weeks X	Sales Y	X^2	XY
	1	2,300	1	2,300
	2	2,400	4	4,800
	3	2,300	9	6,900
	4	2,500	16	10,000
Totals	**10**	**9,500**	**30**	**24,000**
	$\overline{Y} = 2375$		$\overline{X} = 2.5$	

Table 3, A manufacturer product sales over the past four weeks

figure 4, Linear Regression with Time as an Independent Variable



For the fifth week the actual sale will be y=2500. look at Table 3, and figure 4,

VI. NEWS BASED STOCK MARKET PREDICTION

The prediction of stock values using the news or personal opinions on the web constitutes an attempt to extract meaningful characteristics from texts using text mining; classify the news as good or bad for stock prices, simulate investment and predict price trends (Yoosin Kim, Seung Ryul Jeong, Imran Ghani).

News based stock market prediction can be considered as a text classification task. Generally the goal is to forecast some aspects of the stock market such as price or volatility based on the news content. Based on prediction goal a set of final classes are defined, such as "Up" (which means this news cause the prices to go up), "Down" (which means this piece of news is probable to causes decrease in prices) and etc. the prediction system is supposed to classify the incoming news into one of these classes (Azadeh Nikfarjam, Ehsan Emadzadeh, Saravanan Muthaiyah,). In our research we will use one of the most effective string matching algorithm which is KMP Algorithm for text mining using new articles to search for positive or negative news ,and we will prove that KMP algorithm is the fastest algorithm for fast search of pattern.

VII. STRING MATCHING ALGORITHMS

String matching is an important field of study. String matching consists in finding one, or more generally, all the occurrences of a string called pattern within another string called the text (Baljit Singh).Except for the Naïve brute-force algorithm each string matching algorithm perform some preprocessing based on the pattern and then find all valid shift; we will call this later phase "matching . The total running time of each algorithm is the sum of the preprocessing and matching times . The similar but much cleverer Knuth – Morris- Pratt (or KMP) algorithm is presented in Section 6.2; the KMP algorithm has the same O(n) matching time , and it reduces the preprocessing time to only O(m) .

7.1 KMP Algorithm

Knuth-Morris-Pratt algorithm is one of the efficient string matching algorithms. This algorithm examines for existences of a pattern p within a main text t by using the reflection that while matching, the mismatch occurs, the word itself represents satisfactory information to regulate where the next match can begin, thus avoiding the reexamination of formerly matched characters. The KMP algorithm uses a bit table to discover the mismatch of the pattern in an input text. This algorithm performs the comparison from left to right. It uses the bit table for the comparison, if match it returns the index of the text. Otherwise it checks the next bit (R. Janani , S. Vijayarani).

7.1.1 The pre-computed table

(F.Alzazah)TT-table tells how much of the beginning of the pattern has already been matched at any position in the pattern. However, this only affects the system when the pattern has repeated strings, a pattern that consists of different characters such as "abcdefg" would have no useful information in the TT-table, while "abaab" would have useful information because the beginning of the pattern shows up later in the pattern. The useful information

for such deductions can be pre-computed by comparing the pattern with itself. Using the pre-computed table next potential valid shift is at S=s+(q-TT[q]).

7.1.2 Knuth-Morris-Pratt Algorithm

KMP-Matcher(T, P)

```
1.   n ← length[T]
2.   m ← length[P]
3.   π ← Compute-Prefix-Function(P)
4.   q ← 0
5.   for i ← 1 to n
6.       do while q > 0 and P[q + 1] ≠ T[i]
7.              do q ← π[q]
8.          if P[q + 1] = T[i]
9.              then q ← q + 1
10.         if q = m
11.             then print "pattern occurs with shift" i – m
12.                 q ← π[q]
```

VIII. SYSTREM ARCHITECTURE

In this research we are going to study past prices figures in stock markets using Exponential Smoothing methods and Linear Trend Line to deal with structured data. These methods are used in cases in which the time series is fairly stable and has no significant trend, seasonal, or cyclical effects, to average out the irregular component of the time series .We do not neglect also the importance of unstructured textual data (news) and we use one of the String matching algorithms to help us find positive or negative news .In our research we use one of most effective string matching algorithms which is KMP algorithm which used in our research to search for positive or negative news articles by matching pre identified positive news key words with daily news articles then predict the future stock price movement based on these news. Figure 5 , shows the Architecture of the proposed system.

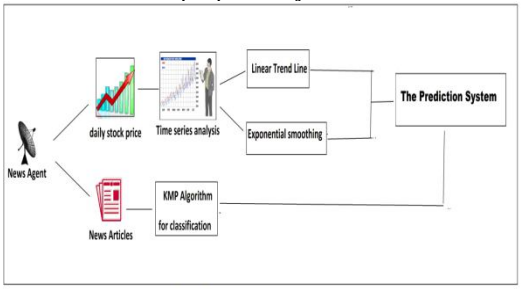


Figure 5, the Architecture of the proposed system

CONCLUSION AND FUTURE WORK

In this paper , we discussed a methodology of applying KMP algorithm for fast news articles search, in addition to two effective time series smoothing methods for stock market prediction. Future work includes creating a standard test data set for stock prediction proposed in this paper and applying it to a test environment.

REFERENCE

[1] Huang W, Nakamori Y, Wang S Y. Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 2005; 32(10): 2513-2522.

[2] Soniya Dewangan, S R Tandan, Shagufta Farzana, Rohit Miri. Prediction of Stock Market Using Financial News Analysis and Supervised Data Mining Technique. IJRASET,2016:541-546

[3] Tuhin Mukherjee , Aritra Banerjee, Prediction Through Genetic Algorithm: A Case Study in Indian Share Market. International Journal of Emerging Technology and Advanced Engineering,2013;215-218.

[4] Rupawari Jadhav , M. S. Wakode, Survey : Sentiment Analysis of Twitter Data for Stock Market Prediction. International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 3, March 2017,558-562.

[5] Yoosin Kim1 , Seung Ryul Jeong1 , Imran Ghani2, Text Opinion Mining to Analyze News for Stock Market Prediction, Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014,1-11

[6] Pegah Falinous ,Stock Trend prediction Using News Article :A text Mining Approach .Unpublished master thesis Lulea University of Technology ,2007.

[7] Vincent Wing-Sing Cho, Knowledge Discovery from Distributed and Textual Data, phd thesis Presented to The Hong Kong University of Science and Technology, Vincent W.S,1999.

[8] YW Seo, J Giampapa, K Sycara, Text Classification for Intelligent Portfolio Management.

[9] Technical report CMU-RI-TR 02-14, Robotics Institute,2002.

[10] David R. Anderson, Dennis J. Sweeney , Thomas A. Williams, Jeffrey D. Camm and Jeffrey W. Ohlmann, James J. Cochran and Michael J. Fry, An Introduction to Management Science: Quantitative Approaches to Decision Making, 14th Edition, 2014.

[11] Eva Ostertagová, Oskar Ostertag, THE SIMPLE EXPONENTIAL SMOOTHING MODEL, MODELLING OF MECHANICAL AND MECHATRONIC SYSTEMS , The 4th International conference,2011,380-384.

[12] BROWN, R. G., MEYER, R. F., The fundamental theory of exponential smoohing, Operations Research, 9, 1961, p. 673-685,

[13] Slovenskej republiky, Statistical yearbook, URL: http://www.statistics.sk,2016

[14] Dan Reid, Nada R. Sanders, Operations management an integrated approach R. 4th ed, Hoboken, NJ John Wiley 2010

[15] Yoosin Kim, Seung Ryul Jeong, Imran Ghani, Text Opinion Mining to Analyze News for Stock Market Prediction , Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014,1-12

[16] Azadeh Nikfarjam, Ehsan Emadzadeh, Saravanan Muthaiyah, Text mining approaches for stock market prediction, Computer and Automation Engineering (ICCAE), The 2nd International Conference ,Volume: 4, IEEE Xplore, 2010,256-260

[17] Baljit Singh ,anomaly detection using data mining ,unpublished master thesis. Department of computer

science and engineering , indian Institute of Technology, 2001.

[18] R. Janani , S. Vijayarani , An Efficient Text Pattern Matching Algorithm for Retrieving Information from Desktop, Indian Journal of Science and Technology, Volume 9, Issue 43, November 2016 ,1-11.

[19] F.Alzazah , Using KMP and Genetic Algorithm For Intrusion Detection, Proceedings of the Third International Conference on Emerging Research Paradigms in Business and Social Science,2015,540-553.

★★★