Data mining technologies used in Stock Market Forecasting

YI TU

G35001246

INTRO TO BIG DATA AND ANALYSIS

2017 FALL

Abstract

As we all know, predicting something is difficult, let alone predicting the values of stock market which is too uncertain in terms of the financial time series and the instant changes of the prices of stock. When the features (inputs) and labels (outputs) are not linear-relational, it will be of great trouble to make a prediction precisely. There is no doubt that the prediction of stock market is a challenging task, but still in the last decade, there are methodologies and models developed to forecast the values of the stock market. The differences between these strategies and models mainly lie in the probabilities of making profits in the stock investments in addition to the accuracy of stock market forecasting. In the daily life, you can not do it one hundred percent right just because the stock market itself is of huge fluctuation, let alone predicting the absolute value of the stocks in terms of the daily basis. And you can not take all the factors which affects the stock prices into considerations just like inflation and economic growth. It is very hard for an individual to analyze all the information of large amount. Even if you are a super machine, you can do that, the predictions of the models or methodologies are still approximate because of the uncertain (random) factors. But as a quantity of data processing methods, data mining and the stock markets are linked tightly. Data mining can deal with these data well which are non-stationary, non-normal, high-noise. Moreover, data mining can learn new sample tp update or improve the model online. That's a very important advantage for a new situation.

1 Introduction

Forecasting stock investment return is an important financial issue that has been given a lot of at- tentions (Matas & Reboredo, 2012).

Stock market predictions need the professional knowledge of the variables of the dominant market which explain stock market behaviors which is both dynamic and volatile.

The strategies and models are online applications for buying the shares and selling the shares, web applications for buying and selling in addition to prediction systems which study the database of shares.

In this paper, we will explore the common technologies of data mining applied in the models or methodologies. Nowadays, we can have a chance to deal with massive information with the help of data mining technique. In terms of the prices of stock, we can use the data mining to extract significant and valuable information from massive data which pose influence on the prices of the stock. The combination of computing synergistically rather than exclusively helps reveal the essence of the forecasting of stock market.

2 Background

Financial markets are the sum of supply and demand and trading mechanisms formed by using financial assets as a trading instrument. In short, they are trading places for financial products. With the improvement of the marketization of economy and the continuous improvement of the system and mechanism of market operation, the development of the stock market will gradually step into the range of rational operation, and the irrational shocks will be substantially reduced or disappeared. The stock market plunged in 2008 and investors' confidence was frustrated. Both the factors of full cir-

culation of stock ownership and the fluctuations of the economic situation as well as the factors of international financial turmoil were also taken into consideration. However, there are also factors that cause China's special national conditions. After all, In terms of market economy, it is still not profound. Our lack of equity culture can not make up for a few years.

2.1 Purpose and significance

The stock market, an important feature of the market economy, has been the heart of tens of millions of investors since its inception. High risk and high return are characteristics of the stock market, so investors are always concerned about the stock market, the stock market analysis, trying to predict the development trend of the stock market. For over a hundred years, some analytical methods have been gradually improved with the emergence and development of the stock market, such as Dow analysis, K-line graph analysis, histogram analysis, point graph analysis, moving average, and morphology Analysis method, trend analysis method, angle analysis method, mysterious series and golden section ratio spiral calendar, four degrees space method, with the popularization and application of computer technology in the field of securities analysis, we have introduced new methods of index analysis. However, strictly speaking, these methods are merely analytical methods and can not directly predict the dynamics of the stock market. In addition, people also try to use regression analysis and other statistical means to establish a model to predict the stock market. However, one of the most fundamental difficulties in predicting the stock market using traditional forecasting techniques is the huge amount of data to be processed. As the stock market is subject to political, economic and other factors, its internal laws are very complex, some of the law of change cycle may be a year or even years, so the

need to pass through a large amount of data analysis can be obtained, while the traditional Prediction technology prediction effect is not satisfactory.

In recent ten years, great progress has been made in the research of data mining technology. The application of various data mining technologies has greatly promoted people's ability to analyze and process large amounts of data and brought good economic benefits to people. Therefore, it is foreseeable that data mining technology will have great potential in the stock market forecast.

2.2 Literature research

Technical analysis has developed rapidly in recent years, especially with the popularity of computers, a variety of analytical methods more and more. In view of the different characteristics of the stock market, domestic and foreign scholars have put forward a variety of analysis and forecasting methods. The following methods are commonly used in the analysis and forecasting.

1. securities investment analysis method. This is a common method used by market analysts.

2. time series analysis. This method primarily predicts future changes by establishing ARMA and ARIMA.

3. Other prediction methods. Such as expert assessment and market research methods such as qualitative methods, seasonal change method, Markov chain and discriminant analysis and other quantitative prediction methods.

4. neural network prediction method. Neural network is a new method of time series analysis.

In 1987, for the first time, Ledes and Farbor (Ledes & Farbor, 1987) introduced neural networks into the field of forecasting, both ideologically and technically as a kind of broadening and breakthrough. It solves the problem that the traditional prediction model is difficult to deal with high-dimensional nonlinearity, emphasis on quantitative indicators, difficult to deal with qualitative indicators, lack of adaptive and self-learning ability prediction. Subsequently, RefeneS et al. Compared the application of neural network forecasting method and multiple linear regression in stock market forecast, and pointed out that the smooth interpolation property of neural network makes it better to fit the data and can better pan The prediction accuracy is greatly improved than the statistical forecasting method. In 1992, Pati and Krishnaprasad (Pati & krishnaprasad, 1992) in the United States formally proposed the wavelet neural network for the first time. By using the wavelet function as a mapping function of neurons, taking full advantage of the local characteristics of the wavelet base and adaptive time-frequency characteristics, a discrete affine wavelet neural network ".

In the recent two years, with the maturing of data mining technology, more and more scholars have adopted data mining methods to predict the stock market trend. Commonly used such as: time series analysis, independent component analysis, artificial neural network. Some improvements in the analysis of the algorithm, such as discrete wavelet learning algorithm, BP algorithm, orthogonal least squares method, decision tree algorithm, rent rough set algorithm. Some models are put forward, such as building a hybrid neural network model which Royal, multivariate function estimation wavelet network and fuzzy wavelet network model, using the principle of minimum description of the selection of significant factors to establish the model.

"Atsalakis et al. (2011) adopted the Elliot wave theory and a neuro-

fuzzy approach. They presented the Wave Analysis Stock Prediction system, which was based on the neuro-fuzzy architecture that utilized the Elliott Wave Theory. The system showed a tendency to achieve hit rates in the 60% mark which was significantly better than forecasting with the help of a coin."

"Chen, Su, Cheng, and Chiang (2011) explored pattern recognition and time series forecasting. Theirs was a novel price-pattern detection method that looked for certain price-patterns (?price trend? and ?price variation?) contained in the time series variables that can be used to forecast the stock market."

"Zuo and Kita (2012) presented a Bayesian network technique to predict the up/down analysis of the daily stock indexes and the result were compared with the psychological line and trend estimation technical analyses. The average correction rate of their algorithm was almost 60%, which is almost equal to or higher than the technical psychological line (50?59%) and the trend estimation (50?52%)."

One of the most influential method is the pretreatment of time series first, and then extracted from the key forecasting attributes, which is proposed by Last. These attributes have a greater impact on the development trend of the time series, will form a set of attributes, these predictive attributes are characterized Time series of a feature, this feature has nothing to do with the time, so you can use common static data mining tools to predict the behavior of the time series forecast. In addition, some experts also introduce the theory of the situation into the trend forecasting of stocks. According to the theory of data fusion in the military field, the observed long-range power distribution is organically related to the current securities investment environment, the counterparty's investment operation intention and the operational maneuverability. , Analysis of the reasons for the ups and downs of

the securities, to be on the individual stocks, plates and the broader market estimates, the final formation of the securities composite trend chart.

3 Data mining technologies

In recent years, with the extensive use of databases and computer networks, coupled with the use of advanced automatic data generation and collection tools, the amount of data that people have has dramatically increased. The contradiction between the rapid increase of data and the lag of data analysis methods is becoming more and more prominent. People hope to make scientific research based on the large amount of existing data to make full use of huge amounts of data. However, it is difficult for data analysis tools currently in place to process the data in an in-depth way. Data mining is precisely to solve the shortcomings of the traditional analysis methods, and for large-scale data analysis and treatment appeared. Data Mining extracts useful information hidden behind data from a large amount of data. It is adopted by more and more fields and has achieved good results, which has greatly helped people to make correct decisions.

At present, data mining technology has been widely used in commercial fields such as banking, telecommunications, insurance, transportation and retail, and is also the same in stock market prediction. With the gradual opening up of the domestic securities industry policy, the competition in the securities industry is becoming more and more fierce, and the dependence and sensitivity of the stock market forecast to data are getting higher and higher. ” Data mining is the discovery of knowledge from the data, mining, development and utilization of these data can make the securities industry the most suitable position, will enable enterprises to accumulate long-term

accumulation to give full play to establish a competitive advantage. As a tool for analysis and decision support, data mining technology has gained more and more attention from domestic securities firms.

3.1 Basic idea

Data mining, as its name implies, is to dig out useful information from a large amount of data, that is, from a large number of incomplete, noisy, vague, random practical application data found hidden, regular, people unknown in advance But non-trivial processes that are potentially useful and ultimately understandable information and knowledge. The unknown information means that the information is not expected in advance, or novelty. The pattern of discovery of novelty requirements should have never been known before, and the information was previously unforeseen. Data mining is to discover information or knowledge that can not be found by intuition, or even counterintuitive information or knowledge.

Data mining is an interdisciplinary discipline that integrates theories and techniques in many fields such as databases, artificial intelligence, machine learning, statistics, etc. " Data Mining The use of various analytical tools to discover the relationships between models and data in massive data that can be used to make predictions that help decision makers find numbers. The potential linkages and the finding of neglected factors are considered to be an effective way of addressing the information deficit that the data explosion of today is facing.

3.2 Classification

Data mining involves many disciplines, including the three major database technology, statistics and machine learning. According to the type of database,

mining objects, mining tasks, mining methods and techniques, as well as applications and other aspects of classification.

1. Database classification by database type is mainly in the relational database mining knowledge. With the continuous increase of database types, data mining of different databases gradually emerged. The existing data mining types such as relational database mining, fuzzy data mining, historical data mining, spatial data mining and many other databases.

2. Data mining by object classification. In addition to mining the main object of the database, there are text data mining, multimedia data mining, Web data, these are unstructured data.

3. According to the task of data mining. Task classification data mining related analysis, timing patterns, clustering, classification, bias detection, prediction and so on. According to the task classification are: association rule mining, sequence pattern mining, clustering data mining, classification data mining, analysis of deviation analysis and forecasting data mining types.

4. According to data mining methods and technical: (1) inductive learning class: This class is divided into information-based method of mining methods and based on set theory method to mine classes. Based on the information theory method is to find a large amount of information in the database attributes to establish the attribute decision tree. Set-based method is based on the relationship between the tuples of attributes in the database to establish the rules between attributes. Various types include a variety of methods, mainly for classification. (2) Biomimicry class: this class is divided into neural network method

class and genetic algorithm class. The neural network method is based on the liver mathematical model and the Hebb learning rules which are simulated by human brain neurons. A series of algorithm models are put forward to solve the practical problems of recognition, prediction, association, optimization and clustering. Genetic algorithms simulate biological genetic processes and establish mathematical operators for the process of selection, crossover and mutation. Mainly used for problem optimization and rule generation. (3) Formula discovery class: In the scientific experiment and engineering database, it has aroused people's concern to find and discover the relationship between continuous attributes by artificial intelligence and to establish the formula between variables. There are many kinds of data mining in this class Methods, such as BACON and FDD. (4) statistical analysis categories: statistical analysis is an independent discipline, due to the data in the database to find a variety of different statistical information and knowledge, it also constitutes a data mining category methods. (5) fuzzy data categories: fuzzy data class is a way to reflect people's thinking. The fuzzy mathematics is applied to all tasks of data mining, forming fuzzy data mining class, such as fuzzy clustering, fuzzy classification, fuzzy association rules.

3.3 The process model

Data mining is a need to go through repeated multiple processes. As the role of software engineering in software development, the data mining process model provides macro guidance and engineering methods for data mining. Reasonable process model can combine the various processing stages organically to guide people to better develop and use data mining systems. From

data mining into the field of application, some people summarize and summarize the process of data mining, put forward different data mining processing model. UsalnaM. The multi-stage model given by Fayyad Gergory Piatetsky - Shapiro et al is a general model and the most widely accepted one. In 1996, Brachman and Anand, through understanding the problems encountered by many data mining users in practical work, found that a large part of the workload of users was in database interaction. They analyzed the data mining process from the user's point of view, It is believed that data mining should be more focused on the whole process of knowledge discovery for users, not just at one stage of data mining, and then put forward a user-centered process model. The model pays special attention to the interaction between user and database. Based on the data in the database, the user proposes a hypothetical model, then selects the data for knowledge mining and continuously adjusts and optimizes the model data. Brachman and Anand used this user-centered process model in the IIdACS (Interactive Marketing Analysis And Classification System), which is a data mining system they developed.

George H., Stanford University, 1997 John presents another data mining process model in his doctoral dissertation. The model emphasizes data mining by the data mining staff and field experts to participate in the whole process. Domain experts are very aware of the problems to be solved in this field. During the definition stage of the problem, domain experts explain to the data mining staff that the data mining staff introduce the technologies used in data mining and the types of problems that can be solved to field experts. Through mutual understanding, both parties have unanimous handling of the issues to be resolved, including the definition of the issue and the way the data are handled.

In 1999, Dr. Zhu Yan-shao, from the Institute of Computing and Com-

puter Science, Chinese Academy of Sciences, held that the aforementioned model did not support repetitive learning and multi-objective learning in the process of knowledge discovery. That is, a certain knowledge discovery algorithm was used to determine a batch of related data. When using other algorithms, Invalid, the data must be extracted and preprocessed. Therefore, in his doctoral dissertation, he proposes a data mining processing model that supports the multi-dataset multi-learning objectives and tries to separate the data and learning algorithms as far as possible so as to make data mining more suitable for practical work and enable the end user and data mining personnel The impact is as small as possible to improve learning efficiency. For the purpose of separating data from learning algorithms, the model uses the concept of data sets. Data set refers to the data extracted from the database in order to complete a learning task. The description of the dataset includes a description of the data and how the training data and test data were generated. Data sets are not specific to a learning algorithm, but are defined for a particular type of problem, which gives the data involved in the problem. In the specific algorithm for data processing, the data must be simple screening and processing to eliminate redundant data.

3.4 The process and tools

Data mining is an iterative process that usually involves multiple interrelated steps such as defining and analyzing topics, preprocessing data, selecting algorithms, extracting rules, evaluating and interpreting results, formulating patterns into knowledge, and finally applying. And with different application requirements and data bases, the steps of data mining may also be different.

1. Problem definition

For data mining, we must first analyze the application areas, including the application of a variety of knowledge and application goals. The problem definition stage is to understand the relevant areas of the situation, familiar with the background, to understand the user requirements. After the user's needs are identified, existing resources, such as existing historical data, should be evaluated to determine if the user's needs can be determined through data mining techniques. The goals of data mining and the data mining plan will then be further defined.

2. Data Preparation

   Data mining processing data sets usually have not only massive data, and there may be a lot of noise data, redundant data, sparse data or incomplete data. Data preparation includes data extraction, cleaning, transformation, and loading, including steps such as data cleaning, integration, selection, transformation, protocol, and data quality analysis.

3. Modeling

   Data mining modeling is the use of known data and knowledge to establish a model that can effectively describe the known data and knowledge, hope that the model can be effectively applied to unknown data or similar Situation. In data mining, many different models can be used: association rule model, decision tree model, neural network model, rough set model, mathematical statistics model, time series analysis model.

4. Evaluation

   Data mining model may have no practical or practical value, it may not accurately reflect the true meaning of the data, and in some cases

is contrary to the fact that the results of the data mining needs to be assessed. Determine whether there is deviation of data mining, mining results are correct, and determine which is valid and useful model, whether to meet user needs.

One method of evaluation is to directly use the data from the previously established mining database for testing, and to find new test data and test it. The other method is to use the current data in the actual operating environment for testing. It is the huge commercial potential of data mining technology, has attracted many companies engaged in data mining system research and development, and some have been commercialized. Not long ago most data mining tools were only controlled by professional technicians. However, more companies now offer more sophisticated data mining systems that are also available to non-professionals.

4 Comparison

The figure below shows the advantages and disadvantages of different methodologies used in forecasting the stock.

Paper ID	Methodology	Advantage	Disadvantage
[1]	Genetic Algorithm, Support vector machines.	SVM transform the inputs into decision classes. There is correlation between prices of certain stocks. Considering closing, opening, mean, standard deviation and number of days for which correlation is found is considered	Various political, economic factors, company policy decide trends of markets are not considered while calculation.
[2]	Sentiment Analysis, Trading model.	They collected aggregating information from multiple online sources. They performed sentiment analysis on given data and filtered out dataset as a result of sentiment analysis and they found the ratio of sentimental signals. Based on this, they created on trading model to predict stock prices and trend of market.	It is necessary to analyze effects of applying different sentiments analysis methodology.
[3]	ANN(artificial neural network), Back-propagation algorithm	It can be used in field where accurate mathematical model cannot be produced, for example stock market It can deal with noisy data.	Designing is challenging as it requires tedious trial and error process. Selection of data set is complex.
[4]	Linear regression, Data mining.	Linear regression is used to perform operation data set where target values. It establishes relation between target values and predicted values. Data mining technique have more successful performance in predicting various fields as it uses hidden knowledge of data.	Calculations using linear regression are very complex. In linear regression, Accuracy is low.
[5]	Linear regression, Neural networks, Genetic Algorithm, Support Vector Machine, Case based reasoning.	It is used to find accurate results among them. Helpful for gathering financial data. They are helpful to map the relations among financial product and financial news.	Depend on sentiments and opinion over news content and global events.
[6]	Typical price, Chwkin money flow indicator, Relative Strength Index.	It calculates the high, low and close value of the market. Also, it tends to give mid value so that customer can buy and sell share according to the values given	Problem is determining the probability that the relationships are not random at all market condition
[7]	Data collection, Feed-forward neural network.	Several machine learning techniques are used in parallel to predict most optimal stock market price. The main advantage is that it provide a very systematic approach and its ability to predict changes before they show up on the chart.	Requires large amount of historical data. It has very high time consuming factor depends on the accuracy of the data provided.
[8]	NewsCAT, Text preprocessing, Automatic text categorization.	It automatically analyzes and categorizes press releases derive stock trading recommendation from them. It can significantly outperform old trading strategies like buying and shorting stocks randomly immediately after press release.	Selection of categorization is poor. NewsCAT engine needs to be enhanced.
[9]	Data Mining.	It helps to find hidden pattern in from historic data that have probable predictive capability. It uses real time news to predict its effect on stocks.	It requires large amount of historic data. Large amount of data processing is required.
[10]	Artificial neural network(ANN)	It helps to build relation between non-linear input and output. It is very intelligent system and works like human brain.	ANN have not been fully explored. Prediction is satisfactory but still lot of improvement is needed.

4.1 Support vector machines (SVM)

In machine learning, SVM, also called support vector networks are supervised learning algorithm which is used for classification and regression analysis. The SVM method is based on the VC dimension theory of statistical learning theory and the minimum principle of structural risk. Based on the limited sample information, the complexity of the model (ie, the learning precision of a specific training sample) and the learning ability (ie, the ability to identify any sample to seek the best compromise with no error, in order to obtain the best promotional capacity.

Given a set of data, which can be marked as belong to one or the other of two categories, the SVM algorithm can be trained to develop a SVM model which is a non-probabilistic binary linear classifier to determine a new example to which category. The SVM model divide the points in space which corresponds to the examples by a gap which is as wide as possible.

By using a kernel trick which means mapping the data in low dimension to high dimensional feature spaces, SVM can efficiently deal with the non-separable data to do a classification.

More formally, a support vector machine constructs a hyperplane, either in a high or infinite dimensional space, which can be used for hyperplanes set in classification, regression, or other tasks. Intuitively, in a general realization of a hyperplane that separates the maximum distance through the nearest training data point to any class (so-called functional headroom), the classifier's generalization error due to the larger margin. The original problem may be described in a finite dimensional space, which often occurs to identify whether the set is linearly separable in the space. For this reason, it has been suggested that mapping the original finite dimensional space to a much higher three-dimensional space presumably makes it easier to separate

in space. Keeping the computational load reasonable, mappings using support vector machine programming are designed to ensure easy calculation in terms of variables that can be in the original space by the dot product, by defining the calculation of the kernel function k (x, y) chosen among them to accommodate problem.

The hyperplane in a high-dimensional space is defined as the dot product of a set of points and the vector in the space is constant. The defined hyperplane vector can be chosen to be linearly combined with the image of the feature vector based on the data occurring in the parameter $alpha_i$. This choice of a hyperplane, the point of the x's feature space is mapped to the hyperplane is defined by the relationship: $font\ sum_i\ alpha_i k(x_i, x) = mathrmconstant.$ Note that if k (x, y) becomes smaller as x further increases in y, the degree of the corresponding data base point x_i of the closeness of the test point x is measured at each of the sums. In this manner, the sum above the kernel can be used to measure the relative proximity of the data points of the various test points to be identified in one or the other set to be identified. Note the fact that the mapping of the setpoint x to any hyperplane can be quite convoluted, making the set not complex in the original space to be much more discriminatory.

SVM maps vectors to a higher-dimensional space where a maximum-interval hyperplane is established. There are two hyperplanes parallel to each other on either side of the data-separated hyperplane. Establishing the proper separation hyperplane maximizes the distance between two parallel hyperplanes. It is assumed that the greater the distance or difference between parallel hyperplanes, the smaller the total error of the classifier.

SVM transform the inputs into decision classes. In the case of SVM applied in stock market prediction, there is correlation between prices of

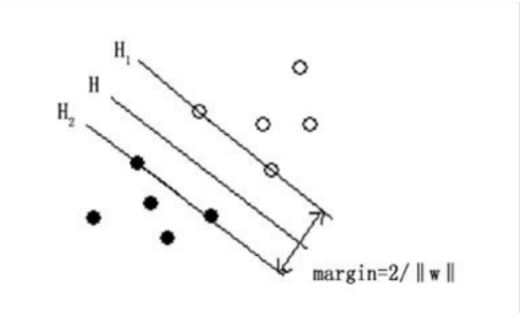

Figure 1: SVM

certain stocks. Considering closing, opening, mean, standard deviation and number of days for which correlation is found is considered.

But using SVM in forecasting still has some drawbacks. For instance, various political, economic factors, company policy decide trends of markets are not considered while calculation.

4.2 Artificial neural network(ANN)

In the field of machine learning and cognitive science, artificial neural network (ANN), or neural network (NN) or neural network, is a kind of artificial neural network Central nervous system, especially the brain), used to estimate or approximate the function. Neural networks are calculated by a large number of artificial neurons. In most cases, artificial neural network can change the internal structure based on external information. It is an adaptive system. Modern neural networks is a nonlinear statistical data modeling tool.

A typical neural network has the following three parts:

1. Architecture

   Architecture specifies the variables in the network and their topological relations. For example, variables in neural networks can be the weights of neuronal connections and the activities of the neurons.

2. Activity Rule

   Most neural network models have a short time scale dynamical rule that defines how neurons change their stimuli based on the activity of other neurons. The general incentive function depends on the weights in the network (ie the parameters of the network).

3. Learning Rule

   Learning rules specify how weights in the network adjust over time. This is generally seen as a long time scale kinetic rules. In general, learning rules depend on the neuron's stimulus value. It may also depend on the target value provided by the supervisor and the value of the current weight. For example, a neural network for handwriting recognition has a set of input neurons. Input neurons will be inspired by the data of the input image. After the stimulus values are weighted and passed through a function (as determined by the designer of the network), the stimulus values of these neurons are passed on to other neurons. This process is repeated until the output neurons are excited. Finally, the output neuron's stimulus value determines which letter is recognized.

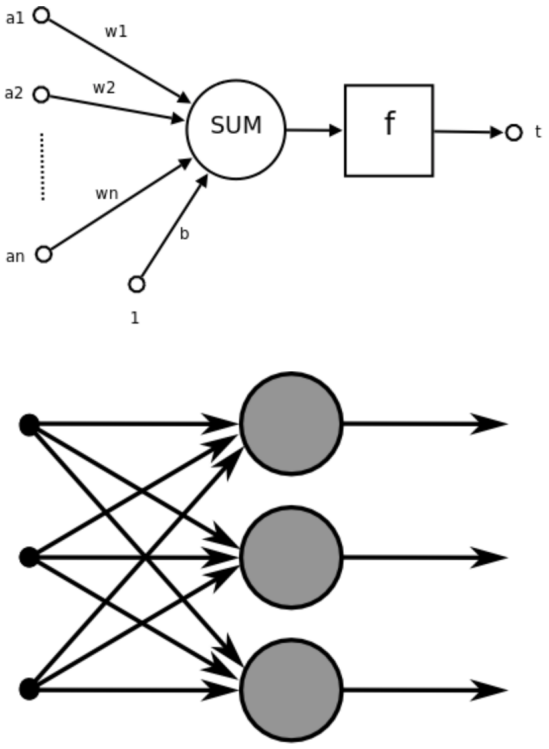

Figure 2: ANN

In general, an artificial neural network consists of a multi-layer neuron structure. Each layer of neurons has an input (its input is the output of a previous neuron) and the output. Each layer Do) Layer (i) is composed of Ni (Ni represents N on the i-th layer) network neurons, and each neuron on each Ni takes the neuron output corresponding to Ni-1 as its input , We call the connection between the neurons and their corresponding neurons with the biological name Synapse. In the mathematical model, each synapse has a weighted value, which we call weight, Then we need to calculate the potential energy of one neuron on the i-th level equal to each weight multiplied by the output of the corresponding neurons on the i-1 layer, and then the

summation of all the neurons on the i-th level The resultant potential energy is then controlled by the activation function on the neuron (often a sigmoid function) because it is differentiable and continuous, making it easy to work with differential rules rule. Obtain the output of the neuron, note that the output is a non-linear value, that is, the value obtained by the excitation function according to the limit to determine whether to activate the neuron, in other words we are not interested in the output of a neural network is linear or not.

The idea of constructing a neural network is inspired by the functioning of the neural network of creatures (humans or other animals). Artificial neural network is usually optimized by a learning method based on mathematical statistics type. Therefore, artificial neural network is also a practical application of mathematical statistics method. Through statistical standard mathematical method, we can get a large number of Using the function to express the local structural space, on the other hand, in artificial intelligence field of artificial perception, we can do the problem of artificial perception decision by the application of mathematical statistics (that is to say, through statistical methods, artificial neural network Can have simple decision-making ability and simple judgment ability like human beings), this method has more advantages than the formal logical reasoning calculation.

Artificial neural networks are similar to biological neural networks in that they can calculate parts of a function collectively and in parallel without the need to describe the specific tasks of each unit. The biologically inspired approach to modern software implementation of neural networks has largely been abandoned, replacing them with more pragmatic methods based on statistics and signal processing. In some software systems, a portion of a neural network or neural network (eg, an artificial neuron) is a part of a

large system. These systems combine adaptive and non-adaptive elements. Although this more general approach used by such systems is better suited to solve real-world problems, it is no longer relevant to traditional connectionist artificial intelligence. But they also have something in common: non-linearity, distributed, parallelism, local computing, and adaptability. From a historical point of view, the application of neural network model marks the beginning of the late 1980s from highly symbolic artificial intelligence (represented by expert systems using conditional rules) to low-symbol machine learning Dynamic system of parametric expression of knowledge as the representative) change.

It helps to build relation between non-linear input and output. And It is very intelligent system and works like human brain. It is observed that in most of the cases ANN models give better results than other methods (Guresen, Kayakutlu, & Daim, 2011).

But ANN have not been fully explored. In terms of the stock market forecasting, prediction is satisfactory but still lot of improvement is needed.

4.3 Data collection, Feed-forward neural network

Feedforward neural network is an artificial neural network.In this network, there are no loops. And the information moves only forward from the input nodes through the hidden nodes to the output nodes in the network. It includes two kinds of perceptron: single-layer perceptron and multi-layer perceptron

Several machine learning techniques are used in parallel to predict most optimal stock market price. The main advantage is that it provide a very systematic approach and its ability to predict changes before they show up on the chart. But, It requires large amount of historical data, and it is time

consuming depending on the accuracy of the data provided.

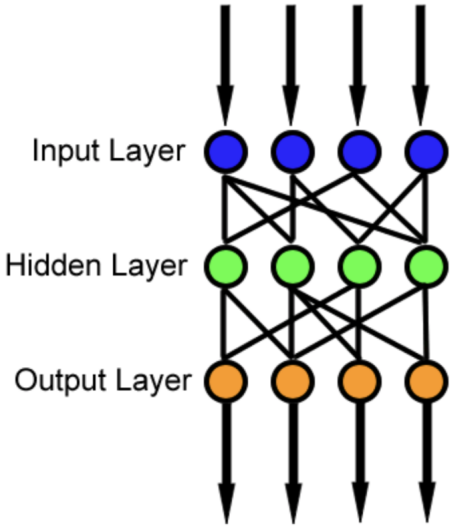

Figure 3: layer

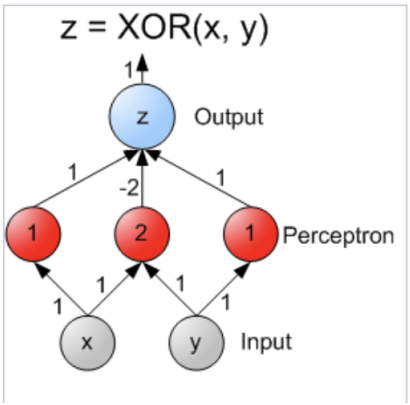

Figure 4: multilayer

4.4 Sentiment Analysis, Trading model

Sentimental analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

In recent years, sentiment analysis has become a hot topic among re-

searchers attempting to automatically extract and quantify the opinions and sentiments embedded in news headlines, financial microblogs and Twitter tweets.

Sentiment analysis is itself is a data mining technique, the primary goal being to automatically extract the news? attitude towards the subject of analysis without having to manually read content. The first step of analysis is cleaning up the text in question. News on webpages is generally in HTML format, so HTML tags and irrelevant items must be removed. After converting the text into a clean format, different pre-processing techniques are applied. One of the early-stage techniques is to identify the ?units? of the text; these may be words, sentences, phrases or n-grams. Lemmatization or stemming are also used to reduce the number of words by simplifying them to their common root. Additional techniques include removing capital letters, identifying the language of the text, removing stop-words, etc.

The most basic form of sentiment analysis is performed by counting the positive and negative words in the text, then using the resulting proportion to arrive at a sentiment score.

After the selection and preliminary processing of the text units, it is imperative to choose the right dictionary for the sentiment analysis. A given word can express a positive attitude in the field of finance but a negative attitude in some other field. For example, if we look at the sentiment of the word ?rise? in the financial sector, it usually means a positive thing: rise of prices. If we examine the same word in the context of healthcare, it may have negative connotation, such as rise of blood pressure. For this reason, it is imperative to select the sentiment dictionary that best fits the domain of the analysis.

Besides the application of sentiment dictionaries, statistical methods that

identify the sentiment of a given article can be employed. The most popular statistics-based solutions incorporate the nave Bayes or Maximum Entropy classifiers or Support Vector Machines, but neural networks can also be used for this task. The implementation can be done by using either supervised or unsupervised learning methods. For the supervised learning, a training set consisting of articles and their sentiment is required. Most of the time, the training set is created manually with the contributions of experts in the researched field. The shortcomings of this method are the amount of time needed to classify the news in the training set and the subjective nature of sentiment; different people may express a variety of opinions regarding the same article. In unsupervised learning, the stock prices are usually used to train the classifier. If stock prices go up, the articles for the day should express positive sentiment, and vice versa if stock prices fall.

To predict the stock prices from news article we use the concept of sentimental analysis. Prediction of stock prices is based on current stock prices and polarity of news articles. Polarity of text can be negative, positive or neutral.

To exactly predict the stock price is very complex task till the date. Here we are proposing to make a prediction based on news articles using one of the Text Mining concepts like sentiment analysis. We would like to make the prediction system for Indian Stock market. Implementation steps to be followed to make a prediction system are:

1. Gathering of news articles
2. Perform sentiment analysis on news articles
3. Get Polarity of the text
4. Make a prediction based on current stock price and calculated polarity

of the text

They collected aggregating information from multiple online sources.

They performed sentiment analysis on given data and filtered out dataset as a result of sentiment analysis and they found the ratio of sentimental signals.

Based on this, they created on trading model to predict stock prices and trend of market.

It is necessary to analyze effects of applying different sentiments analysis methodology.

4.5 ANN(artificial neural network), Back-propagation algorithm

Back-propagation algorithm calculates the error contribution of each neuron after a batch of data is processed. It is used in artificial neural networks, dealing with the data from image recognition or multiple images.

Back-propagation algorithm requires a already known output that is available for each input to calculate the loss function gradient. Therefore, it is generally considered a supervised learning method, although it is also used in unsupervised networks such as automatic encoders. It is a generalization of the Delta rules for multi-layer feedforward networks, and the chain rule can be used to calculate the gradient for each iteration. Back-propagation algorithm also requires that the excitation function of an artificial neuron (or "node") to be differentiable.

There are two phases: propagation and weight update. In the first phase, the algorithm calculates the cost (error term). In phase weight update, the weight must be updated in the opposite direction, "descending" the gradient.

```
initialize network weights (often small random values)
do
   forEach training example named ex
      prediction = neural-net-output(network, ex)  // forward pass
      actual = teacher-output(ex)
      compute error (prediction - actual) at the output units
      compute Δw_h for all weights from hidden layer to output layer  // backward pass
      compute Δw_i for all weights from input layer to hidden layer   // backward pass continued
      update network weights // input layer not modified by error estimate
until all examples classified correctly or another stopping criterion satisfied
return the network
```

It can be used in field where accurate mathematical model cannot be produced, for example stock market. In addition, this algorithm can deal with noisy data. However, the designing of the language is too challenging with respect to the fact that it requires tedious trial apart from the error process. Also, the selection of the data set might be too complex.

4.6 Linear regression, Data mining

In statistics, linear regression is a regression analysis that models the relationship between one or more independent and dependent variables. It uses the least-squares function known as a linear regression equation. This function is a linear combination of one or more model parameters called regression coefficients. The case of only one independent variable is called simple regression, and the case of more than one independent variable is called multiple regression. (This, in turn, should be dictated by multiple linear regression predictions from multiple dependent dependent variables rather than a single scalar variable.)

In linear regression, the data is modeled by using a linear prediction function, and unknown model parameters are also estimated from the data. These models are called linear models. The most commonly used linear regression model is the affine function of the conditional mean of y for a given x-value for y. Less generally, the linear regression model can be a median

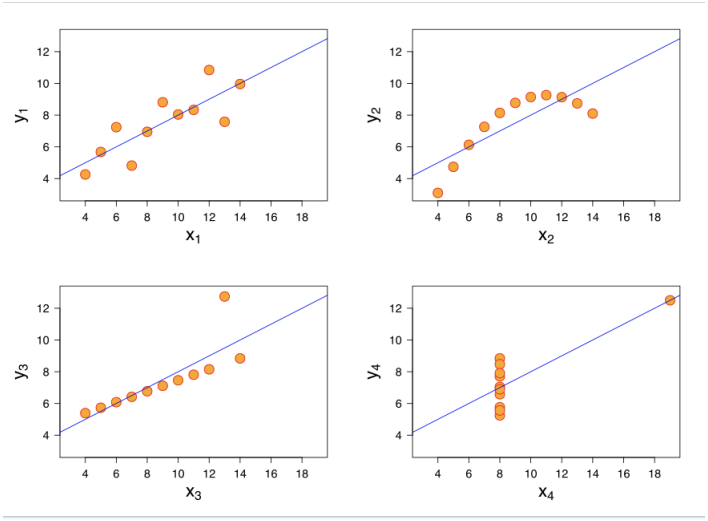

Figure 5: Linear Regression.

or some other quantile of the conditional distribution of y for a given X as a linear function of X. Like all forms of regression analysis, linear regression also focuses on the conditional probability distribution of y for a given value of X, rather than the joint probability distribution of X and y (multivariate analysis).

Linear regression is the first type of regression analysis that has been rigorously studied and widely used in practical applications. This is because a model that is linearly dependent on its unknown parameters is easier to fit than a model that is nonlinearly dependent on its position parameters and the resulting statistical properties are also easier to determine.

It is used to perform operation data set where target values and establishes relation between target values and predicted values. In these years, many facts have proved that data mining technique have more successful performance in predicting various fields as it uses hidden knowledge of data.

Linear regression is used to perform operation data set where target values. It establishes relation between target values and predicted values. Data

mining technique have more successful performance in predicting various fields as it uses hidden knowledge of data.

4.7 Hidden Markov model

Hidden Markov Model (HMM) is a statistical model used to describe a Markov process with hidden unknown parameters. The challenge is to determine the hidden parameters of the process from the observable parameters. Then use these parameters for further analysis, such as pattern recognition.

In a normal Markov model, the state is directly visible to the observer. The conversion probability of such a state is all the parameters. In the hidden Markov model, the state is not directly visible, but some of the variables that are affected by the state are visible. Each state has a probability distribution on the possible output symbols.

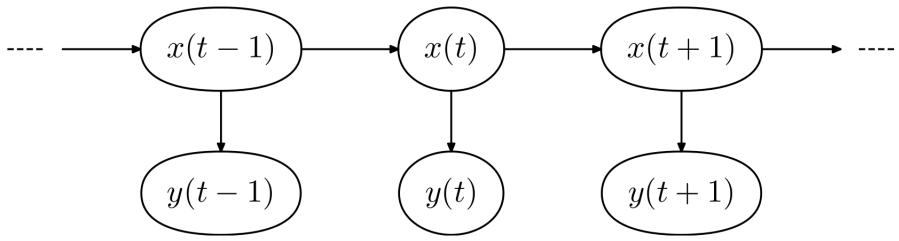

Figure 6: HMM

Therefore, the sequence of output symbols can reveal some information of the state sequence. A HMM can be presented as the simplest dynamic Bayesian network.

There are three types of related tasks about the application of the HMM:

1. Filtering

Given parameters apart from a sequence of observations, the problem is to calculate the distribution over hidden states of the last latent variable at

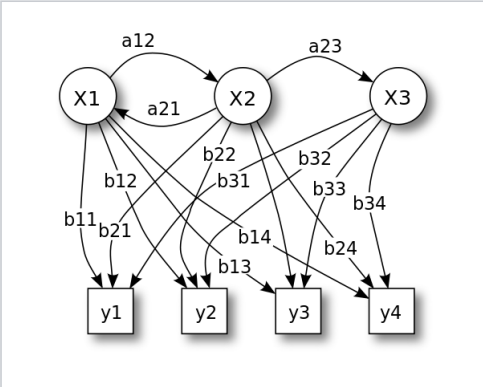

Figure 1. Probabilistic parameters of a hidden Markov model (example)
X — states
y — possible observations
a — state transition probabilities
b — output probabilities

Figure 7: HMM

the end of the sequence.

2. Smoothing

This task asks about the distribution of a latent variable in the middle of a sequence. It needs to find the maximum over all possible state sequences.

3.Most likely explanation

This task focus on the joint probability of the entire sequence of hidden states that generated a particular sequence of observations.

References

[1] Lapedes, A., & Farber, R. (1987). Nonlinear signal processing using neural networks: Prediction and system modelling (No. LA-UR-87-2662; CONF-8706130-4).

[2] Pati, Y. C., Krishnaprasad, P. S., & Peckerar, M. C. (1992). An analog neural network solution to the inverse problem of 'early taction'. IEEE

transactions on robotics and automation, 8(2), 196-212.

[3] Matas, J. M., & Reboredo, J. C. (2012). Forecasting performance of nonlinear models for intraday stock returns. Journal of Forecasting, 31, 172?188. doi:10.1002/for.1218

[4] Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 38, 10389?10397. Retrieved from http://dx.doi.org/10.1016/j. eswa.2011.02.068

[5] Chen, T. L., Su, C. H., Cheng, C. H., & Chiang, H. H. (2011). A novel price-pattern detection method based on time series to forecast stock markets. African Journal of Business Management, 5, 5188?5198.

[6] Zuo, Y., & Kita, E. (2012). Up/down analysis of stock index by using Bayesian network. Engineering Management Research, 1, 46?52. doi:10.5539/emr.v1n2p46

[7] Atsalakis, G. S., Dimitrakakis, E. M., & Zopounidis, C. D. (2011). Elliott wave theory and neuro-fuzzy systems, in stock market prediction: The WASP system. Expert Systems with Applications, 38, 9196?9206. doi:10.1016/j. eswa.2011.01.068

[8] Huang, C. Y., & Lin, P. K. (2014). Application of integrated data mining techniques in stock market forecasting. Cogent Economics & Finance, 2(1), 929505.

[9] Bognr, E. K. (2016). Applying big data technologies in the financial sector-using sentiment analysis to identify correlations in the stock market. Computational Methods in Social Sciences, 4(1), 5.

[10] Abraham, A., Nath, B., & Mahanti, P. K. (2001). Hybrid intelligent systems for stock market analysis. In V. Alexandrov, J. Dongarra, B. Juliano, R. Renner, & C. J. K. Tan (Eds.), Computational science?ICCS 2001 (Vol. 2074, pp. 337? 345). Heidelberg: Springer.