Survey of Stock Market Prediction Using Machine Learning Approach

ASHISH SHARMA
HOD Govt. Women's Polytechnic, Indore
sharmaashish_gwpci@yahoo.co.in

DINESH BHURIYA
Govt. Women's Polytechnic, Indore
dineshbhuriya2006@gmail.com

UPENDRA SINGH
upendrasingh49@gmail.com

Abstract: Stock market is basically nonlinear in nature and the research on stock market is one of the most important issues in recent years. People invest in stock market based on some prediction. For predict, the stock market prices people search such methods and tools which will increase their profits, while minimize their risks. Prediction plays a very important role in stock market business which is very complicated and challenging process. Employing traditional methods like fundamental and technical analysis may not ensure the reliability of the prediction. To make predictions regression analysis is used mostly. In this paper we survey of well-known efficient regression approach to predict the stock market price from stock market data based. In future the results of multiple regression approach could be improved using more number of variables.

Keywords: Stock Market, Prediction, Data Mining, Multiple Regression, polynomial regression, linear regression.

I. INTRODUCTION:

Stock market plays a very important role in fast economic growth of the developing country like India. So uscountry and other developing nation's growth may depend on performance of stock market. If stock market rises, then countries economic growth would be high. If stock market falls, then countries economic growth would be down [1][2]. In other words, we can say that stock market and country growth is tightly bounded with the performance of stock market. In any country, only 10% of the people engaging themselves with the stock market investment because of the dynamic nature of the stock market [2]. There is a misconception about the stock market i.e. buying and selling of shares is an act of gambling. So this misconception can be changed and bringing the awareness across the people for this. The prediction techniques in stock market can plays a crucial role in bringing more people and existing investors at one place. The more promising results of the prediction methods can be change the mindset of the people. Data mining tools also helps to predict future trends and behaviors; helping organizations in active business solutions to knowledge driven decisions [3][4]. Intelligent data analysis tools produce a data base to search for hidden information that may be missed due to beyond expert's predictions. Extraction which was previously unknown, implicit and potentially useful information from data in databases, is an effective way of data mining. It is commonly known as knowledge discovery in databases (KDD) [5]. Although data mining and knowledge discovery in databases (or KDD) both are used as similar often, Data mining is actually part of knowledge discovery [3][4][5]. Data mining techniques play important role in stock market which can search uncover and hidden patterns and increasing the certain level of accuracy, where traditional and statistical methods
are lacking. There is huge amount of data are generated by stock markets forced the researchers to apply data mining to make investment decisions. The following challenges are addressed by data mining techniques in stock market analysis [2][6].

II RELATED WORKS

Prediction of stock prices is very challenging and complicated process because price movement just behaves like
a random walk and time varying. In recent years various researchers have used intelligent methods and techniques in stock market for trading decisions. Here, we present a brief review of some of the significant researchers. A Sheta [7] has used Takagi-Sugeno (TS) technique to develop fuzzy models for two nonlinear processes. They were estimated software effort for a NASA software projects and the prediction of the next week S&P 500 for stock

market. The development process of the TS fuzzy model can be achieved in two steps 1) the determination of the membership functions in the rule antecedents using the model input data; 2) the estimation of the consequence parameters. They used least-square estimation to estimate these parameters. The results were promising. M.H. FazelZarandiet al. [8] have developed a type-2 fuzzy rule based expert system for stock price analysis. Interval type-2 fuzzy logic system permitted to model rule uncertainties and every membership value of an element was interval itself. The proposed type-2 fuzzy model applied the technical and fundamental indexes as the input variables. The model can be tested on stock price prediction of an automotive manufactory in Asia. Robert K. Lai et al. [9] have established a financial time series-forecasting model by evolving and clustering fuzzy decision tree for stocks in Taiwan Stock Exchange Corporation (TSEC). The forecasting model integrated a data clustering technique, a fuzzy decision tree (FDT), and genetic algorithms (GA) to construct a decision-making system based on historical data and technical indexes. The set of historical data can be divided into k sub-clusters by adopting K-means algorithm. GA was then applied to evolve the number of fuzzy terms for each input index in FDT so the forecasting accuracy of the model can be further improved. S AbdulsalamiSulaiman Olaniyi et al [11] have proposed a linear regression method of analyzing coupled behavior of stocks in the market. The method successfully predicts stock prices based on two variables.

III PREDICATION METHOD

REGRESSION:

In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables . More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable;[1] for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.[2][3]

In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification.[4] The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems

3.1 POLYNOMIAL REGRESSION

In statistics, polynomial regression is a form of linear regression in which the relationship

between the independent variable x and the dependent variable y is modelled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y | x), and has been used to describe nonlinear phenomena such as the growth rate of tissues,[1] the distribution of carbon isotopes in lake sediments, and The progression of disease progression of disease epidemics.[3] Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(y | x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression. The predictors resulting from the polynomial expansion of the "baseline" predictors are known as interaction features. Such predictors/features are also used in classification settings.

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In simple linear regression, the model

$$y = a_0 + a_1 x + \varepsilon,$$

is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable x. In this model, for each unit increase in the value of x, the conditional expectation of y increases by a1 units.

In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improves by increasing amounts for each unit increase in temperature. In this case, we might propose a quadratic model of the form

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon.$$

In this model, when the temperature is increased from x to x + 1 units, the expected yield changes by a1 + 2a2x. The fact that the change in yield depends on x is what makes the relationship nonlinear (this must not be confused with saying that this is nonlinear regression; on the contrary, this is still a case of linear regression).

In general, we can model the expected value of y as an nth degree polynomial, yielding the general polynomial regression model

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon.$$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters a0, a1, ... Therefore, for least squares analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regression. This is done by treating x, x2, ... as being distinct independent variables in a multiple regression model.

3.2 RBF REGGRESSION

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\phi(\mathbf{x}) = \phi(\|\mathbf{x}\|)$; or alternatively on the distance from some other point c, called a center, so that $\phi(\mathbf{x}, \mathbf{c}) = \phi(\|\mathbf{x} - \mathbf{c}\|)$. Any function ϕ that satisfies the property $\phi(\mathbf{x}) = \phi(\|\mathbf{x}\|)$ is a radial function. The norm is usually Euclidean distance, although other distance functions are also possible. For example, using Łukaszyk–Karmowski metric, it is possible for some radial functions to avoid problems with ill conditioning of the matrix solved to determine coefficients we (see below), since the $\|\mathbf{x}\|$ is always greater than zero.[1]

Sums of radial basis functions are typically used to approximate given functions. This approximation process can also be interpreted as a simple kind of neural network; this was the context in which they originally surfaced, in work by David Broom head and David Lowe in 1988,[2][3] which stemmed from Michael J. D. Powell's seminal research from 1977.[4][5][6] RBFs are also used as a kernel in support vector classification.[7]

3.3 SIGMOID REGRESSION

A sigmoid function is a mathematical function having an "S" shape (sigmoid curve). Often, sigmoid function refers to the special case of the logistic function shown in the first figure and defined by the formula

$$S(t) = \frac{1}{1 + e^{-t}}.$$

Other examples of similar shapes include the Gompers curve (used in modeling systems that saturate at large values of t) and the ogee curve (used in the spillway of some dams). A wide variety of sigmoid functions have been used as the activation function of artificial neurons, including the logistic and hyperbolic tangent functions. Sigmoid curves are also common in statistics as cumulative distribution functions, such as the integrals of the logistic

distribution, the normal distribution, and Student's t probability density functions.

3.4 LINEAR REGRESSION

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.[1] (This term should be distinguished from multivariate, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models.[3] Most commonly, linear regression refers to a model in which the conditional mean of given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[4] This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
- Given a variable y and a number of variables X1, ..., Xp that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the Xj, to assess which Xj may have no relationship with y at all, and to identify which subsets of the Xj contain redundant information about y.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

IV CONCLUSIONS

The aim of our research study is to help the stock brokers and investors for investing money in the stock market.The prediction plays a very important role in stock market business which is very complicated and challenging process dueto dynamic nature of the stock market.

REFERENCES

1. Eugene F. Fama “The Behavior of Stock Market Prices”, the Journal of Business, Vol 2, No. 2, pp. 7–26, January 1965.
2. Ambika Prasad Das “Security analysis and portfolio Management”, I.K. International Publication, 3rd Edition 2008.
3. Introduction to Data Mining and Knowledge Discovery (1999), Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.).180 SachinKamley, ShaileshJaloree, R. S. Thakur
4. Larose, D. T. (2005), “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2,John Wiley & Sons, Inc.
5. Dunham, M. H. & Sridhar S. (2006), “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition.
6. Stock market challenges from http://www.google.com.
7. A Sheta, “Software Effort Estimation and Stock Market Prediction Using Takagi-Sugeno Fuzzy Models”, InProceedings of The IEEE International Conference on Fuzzy Systems, pp.171-178, Vancouver, BC, 2006.
8. M.H. FazelZarandi, B. Rezaee, I.B. Turksen and E. Neshat, “A type-2 fuzzy rule-based expert system model forstock price analysis”, Expert Systems with Applications, Vol.36, No.1, pp. 139-154, January 2009.
9. Robert K. Lai, Chin-Yuan Fan, Wei-Hsiu Huang and Pei-Chann Chang, “Evolving and clustering fuzzy decisiontree for financial Time series data forecasting”, An International Journal of Expert Systems with Applications,Vol.36, No.2, pp. 3761-3773, March 2009.
10. Shyi-Ming Chen and Yu-Chuan Chang, “Multi-Variable Fuzzy Forecasting Based On Fuzzy Clustering andFuzzy Rule Interpolation Techniques”, Information Sciences, Vol.180, No.24, pp. 4772-4783, 2010.
11. S AbdulsalamSulaimanOlaniyi, Adewole, Kayode S., Jimoh, R. G,” Stock Trend Prediction Using RegressionAnalysis – A Data Mining Approach”, ARPN Journal of Systems and Software Volume 1 No. 4, JULY 2011,Brisbane, Australia.