Application Research of Data Mining Technology for Financial Prediction

SU Wei[1, a]

[1] Qinghai university of finance and economics college, Qinghai Xining 810001, China

[a]suweiqh@163.com

Keywords: Financial prediction; Data mining; Financial time series; Clustering analysis; Support vector machine

Abstract. Financial prediction is an important research direction of financial data mining. In addition to general common characteristics, nonlinear, non-stationaryand dynamic, financial time series is also of some other characteristics, such as high noise, non-normal, rush thick tail, etc. As a result, the financial forecast is more challenging, and has broad application value and market prospects. This paper mainly studies the application of fuzzy correction model and hybrid model based on clustering analysis and neural network in the field of financial forecast.

Financial prediction model

Introduction

Financial forecast refers to a large amount of historical data on the financial markets use data mining method to predict the future behavior of the market. Financial prediction has a broad application value and market prospect, and therefore attracts many researchers. The structure can be seem as in figure 1.

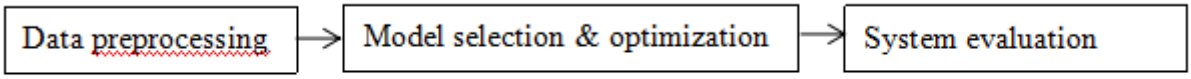

Fig 1 Prediction system structure

Data preprocessing

Before input to the algorithm, the data must be collected, observed, cleaned and selected. Because even the best prediction system may fail because of bad data form, so the data preprocessing is critical.

1) Linear change

Because the original time series volatility is too big, not suitable for forecasting model for fuzzy (including mixed forecasting model), generally the raw data down to a smaller range, this paper USES the linear transform the original data compression to [0.1, 0.9] interval, linear transformation formula is:

$$\tilde{x}_i = \frac{0.9-0.1}{\max(x_k)-\min(x_k)} x_i + \frac{0.1*\max(x_k)-0.9*\min(x_k)}{\max(x_k)-\min(x_k)}$$

2) Data filtering

Due to the stock price time series high peak, noise, filtering pre-treatment is necessarybefore the data is applied to the model data. In nonlinear time series prediction, pattern matching and other applications for data filtering is widespread and commonly used filter estimators are: orthogonal sequence extension, near the most estimates, the average wechat business estimates, etc. [1]This article uses the Nadaraya Watson kernel estimator.

$$\tilde{x}_i = \frac{\sum_{k=1}^{T} K_h(i-k)x_k}{\sum_{k=1}^{T} K_h(i-k)}$$

In which, the Kernel estimate function is Gaussian kernal:

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}}$$

Prediction model selection

Modeling is the most important process in the process of data mining. As an applied research branch of data mining, financial prediction is not exceptional.Model selection and optimization is the key in the process of prediction system design and development.

The traditional forecast models are autoregressive model, moving average model and autoregressive moving average model. Parameter estimation uses least mean square (LMS) general to estimate. Difference autoregressive moving average model and the autoregressive conditional heteroscedasticity model is an important tool, widely used in economic and financial field. But these are all prediction model based on mathematical statistics. In recent years, with the rapid development of computer science, artificial intelligence and machine learning research breakthroughscreated a new data mining research, as well as research in the field of financial forecast opens up a new way.

Statistical learning and support vector machine (SVM)

Machine learning theory

1) Machine learning problems were reviewed

People can acquire knowledge from the practical example, the analysis of the known facts and sums up the rule, predict the fact that cannot be observed directly. [2]Using a computer to imitate people's Learning behavior is known as based on the data of Machine Learning, Machine Learning). Machine learning is the purpose of the training sample estimate system based on the given dependencies between input and output, as accurately as possible to the unknown output forecast, as shown in figure 2.

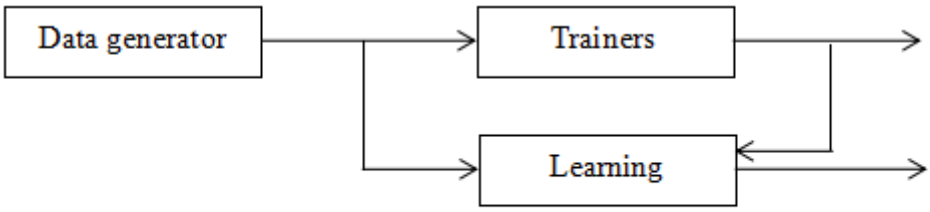

Fig 2 The structural model of machine learning

Machine learning are defined as follows:

Define one: suppose have a set of input and output data $(z^1, y_1),...,(z^n, y_n)$. These data are from G space, calculatedby F (z, y). z^i belongs to R_n, the training sample, and the output data to the corresponding input data is represented as y_i. In order to make the function as the independent variable of the function R (w) numerical minimum, must obtain optimization function $f(z, w_0)$.

$$R(w) = \int L(y, f(z,w))dF(z,y)$$

Definition two: suppose have a set of input and output data $(z^1, y_1),...,(z^n, y_n) \in R_n$, these data is from G space, through the calculation of F (z, y). z^i is training. regression data representationcorresponds to with data is y_{it}. Using the training sample set for function $f: z \to y$, which can accurate calculation numerical.

If y_i obtained by numerical z^i is unknown, so the learning is not in the condition of regulation; If y_i obtained by numerical y_i is known, the study is in a state of half-regulation.

In terms of regression for meter, the loss function of expression is:

$$L(y, f(z,w)) = (y - f(z,w))^2$$

In return for meter problem, with the help of a function $f(z,w)$, $R(w)$ is minimum. If the

function set has:

$$f(z,w)=\int ydP(y|z)$$

2) Statistical learning theory

All kinds of problems of machine learning in the processing mostly relies on the statistics. Past statistical research is gradual theory, namely sample scope of statistical properties of the nearly infinite. With Fisher's point of view, the statistics of the main research goal, can through the family of a model of a pre-determined model for project construction and determined. Fisher called the training model to estimate clusters parameter as learning process.

In the actual operations, due to the high cost for training model, in most cases will not be able to obtain sufficient training model. Sometimes because of views on specific problems restricted by various factors, it is difficult to know whether the training sample wood can fully meet the needs of machine learning. Normally in the process of concrete the sample amount is not sufficient. Statistical learning theory correctly treat training model insufficient, explore training model insufficient ideas to ensure that the generalization ability of learning algorithm. [3]It is four big discoveries in the last century 60s form the basis of statistical learning theory:

Support vector classification machine theory

Support vector classification machine is a way appeared in recent year that can effectively solve the curse of dimensionality and overfitting problem in pattern recognition and machine learning. It was firstly from Boser, b., Guyon, i. e. m., and Vapnik, V.N. presented at an international conference. Support vector machineis a kind of widely used calculation method of the study, which was based on the rule of SRM, its guiding ideology is to build an optimal hyperplane, plane and sample input space or the length of the feature space between maximum, to obtain the best level of generalization. The process of the construction of the optimal hyperplane is actually quadratic programming process of numerical problems.

1) Basic theory of linear support vector classification machine

Assume that regulation problem of training sample set for in advance

$$S=\{(z^1,y_1),(z^2,y_2),...,(z^l,y_l)\}$$

$$W^T+b=0$$

$$s.t.W^Tz_i+b>0$$

$$W^Tz_i+b>0$$

In order to achieve the classification hyperplane to be in better generalization level, when the training set of linear can be divided, must be fully bound W and b. Class samples z^i must meet $W^Tz_i+b\geq l$, the offosite samples of z^i must fit $W^Tz_i+b\leq l$. In addition to this, must widening hyperplanelength between $W^Tz_i+b=1$ and $W^Tz_i+b=-1$, and try to narrow the VC d. Seek to meet the requirements of all the above classification hyperplane, must take the Quadratic programmingvalues:

$$Minimize:\Phi(W)=12W^{\tau}W$$

Subject to: $y_i(W^{\tau}z^i+b)\geq l, i=1,2,...,n$

2) Basic theory of support vector regression machine

Input space is represented by R^n, at the same time x belongs to R. Φ is for nonlinear mapping, with the aid of R, x can be mapped to feature space. In this space, it can be expressed as the linear measure function:

$$y=f(x,W)=W^{\tau}\phi(x)+b$$

Based on the framework of the lowest risk criteria, through the data without interference to find the function f. Vapnik proposed makes the lowest risk with the help of a regular function:

$$\frac{1}{2}W^TW+\frac{C}{l}\sum_{i=1}^{l}|y_i-f(x^i,W)|_{\delta}$$

Among them, $C>0,\delta>0$, introduce the normal number of type c is a kind of punishment of

violation of constraints. Therefore need to set the penalty parameter C in advance.

Data mining technology in the stock

Concept of data mining

With data acquisition and storage technology ascension, all areas of human life have caused a large number of large database. Such as super market transaction data, the use of credit card records, phone records of communications industry, the trade data of the stock market, etc. [4]How to deal with the huge amounts of data, how to store a lot of data from these databases to extract useful information for us is a major problem we face. Data mining technology arises at the historic momentfor this need .

Main steps of data mining

(1) data collection

A large number of data is the precondition of data mining, and without the data, also cannot take up a data mining. Therefore, the data collection is the first step of data mining. Data can be from the existing transaction processing systems, it can also be derived from the data warehouse.

(2) data sorting

Data mining is a necessary part of data mining. Data which are obtained by the data collection phase, there may be some "pollution", which mean the data may be inconsistent, or the presence of missing data, etc., so the sorting of data is a must. At the same time, by data processing, can conduct simple generalization of data processing, thus get more abundant informationon the basis of the original data , thus to facilitate the next step of data mining.

(3) data mining

According to the target of miningdetermined, choose suitable mining model and algorithm for mining, sort data in early work, adjust the parameters of the model can use of several mining model, and then analyze the results. Generally it need to create a model with the training sample first, then use the test samples to test the model.

(4) purpose of data mining

The auxiliary decision is the ultimate goal of data mining. Decision makers canadjust the competitive strategy according to the results of data mining, combined with the actual situation. Sometimes, the process of data mining need to cycle of repeated times, can it been possible to achieve the desired effect.

Summary

As a great attractive area of research full of challenges in modern society, financial forecastis not only the economic and financial research topic, it also is the important direction of computer science research. Time series analysis is an important basis of modern economics, and in the study of economics, the tools and models tend to be statistical model and regression method. As a research direction of computer science, it need to be more focused on from the perspective of data mining and artificial intelligence algorithms to study.

References

[1] Andrew W.Lo, Harry Mamaysky, Jiang Wang. Foundation of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation[J]. NBER 1050 Massachusetts Avenue Cambridge, MA 02138 March 2000.

[2] Sun H, Wang S, Jiang Q. FCM-Based Model Selection Algorithms for Determing the Number of Cluster[J]. Pattern Recognition, 37(2004) 2027~2037.

[3] Dae-Won Kim, Kwang H.Lee, Doheon Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters[J]. Pattern Recognition 37(2004) 2009~2025.

[4] Zhnag Ling. The theoyr of SVM and Programrning based learning algorit hms in neural netwokrs. Chinese Journal of Computers,2001,24(2):113-118.