Research on the Applications of Data Mining in Financial Prediction

ZHOU Yingying[1, a]

[1]NINGBO DAHONGYING UNIVERSITY, Ningbo 315175,China

[a]zhouyingying@126.com

Keywords: Data Mining and Storage; Financial Prediction; Risk Analysis; Real-World Applications.

Abstract. Time series is an important data, the prevalence of time series to make use of data mining technology can effective access to information and knowledge discovery. In the process of time series data mining, feature representation and similarity measure is an important basic work, the smooth implementation of for other data mining task provides a good data processing method and technical support. Based on information entropy, this paper puts forward a kind of neural network classification method based on statistical learning theory, is different with the traditional neural network method. The proposed methodology is verified through the numerical analysis and experimental simulation. Further research interests are also discussed.

Introduction

With the development of computer technology, multimedia technology and financial technology, people are increasingly exposed to a lot of financial data. Time series is a kind of common and associated with the time of high-dimensional data, also is the main research object in the area of data mining, widely exists in financial, medical, weather, and in the field of network security. In recent years with the development of social economy and information technology, the amount of time series data growth is faster and faster. Accordingly, the use of data mining technology in the time series database found potential useful information and knowledge, also more and more attention by researchers, and the research results are widely used in economic, financial, electronic information, medical and health care, education, and in industrial engineering and other fields. Therefore, how to get from a large number of time series data mining valuable information and knowledge and able to serve the society is the one of the main research direction in the area of data mining. Time series data mining and traditional data mining, can find potential from the class data contain valuable information and knowledge, the final feedback and applied to the production practice in the society. Time series data is usually a high dimension and data changes over time and the produce process easily affected by environmental factors, and there is a noise. For such complex data, research how to effectively obtain information and knowledge, for social production practice and scientific research has very important theoretical research value and practical significance. Due to the high dimension characteristic of time series itself, in the practical application, usually need to characteristics of time series of local feature extraction or global decomposition, reduce the dimension of the original time series, and the combination of time series similarity measure method is more efficient to time series data mining, and then from time series data to extract valuable information and knowledge [1-3].

Data classification is based on the characteristics of data sets to construct a classifier, using the classifier to the unknown categories of samples given category of a technology. The process of constructing classifier is generally divided into training and test of two steps. In the training phase, the analysis of the characteristics of the training data set, for each category to produce an accurate description of the corresponding data set or model. In the testing phase, using the description of the category or model classifying test, test its classification accuracy. There are many different data classification algorithms and models, such as Bayesian classification, decision tree learning, statistical method, and neural networks. Characteristics of the said method can not only in the high-dimensional space time sequence is mapped to a low dimensional feature space, and implement the data dimension reduction, also can effectively reflect the time sequence, the basic form and important information to lay a good foundation to improve the efficiency of the time series data

© 2015. The authors - Published by Atlantis Press

mining. Similarity measure method, meanwhile, is another important in time series data mining process, is also a time series is one of the basic and key problem in data mining. Most of the time series data mining technology of the initial work requires similarity comparison, such as clustering, classification, interest models found, abnormal findings and time series visualization, etc [4].

Based on information entropy, this paper puts forward a kind of neural network classification method based on statistical learning theory, is different with the traditional neural network method. Its classification algorithm is based on the network all the neurons in the voting results, there is no convergence problem or not. It scalable network structure and connection mode is suitable for programmable hardware implementation, is advantageous to the characteristics of high dimensional data extraction and analysis, have huge amounts of data to solve the high dimension, high correlation between financial problems has important practical significance. Based on the classification of several representative financial problems experiment, shows that the statistical learning algorithm based on information theory makes this method on the analysis of the financial problems accuracy is superior to the existing majority of artificial neural network. The detailed general discussion will be conducted in the following sections.

Our Theoretical Analysis

The Neural Network Theory. In recent years, domestic scholars put the improved variety of neural network for data classification which obtained better than the traditional neural network classification of test results. This article discusses the neural network and BP network, is also a kind of feedforward structure. It on the network structure and the traditional neural network has certain similarities, but its internal organization and the learning mechanism of data have great difference with the traditional neural network. The figure 1 shows the structure of the network [5].

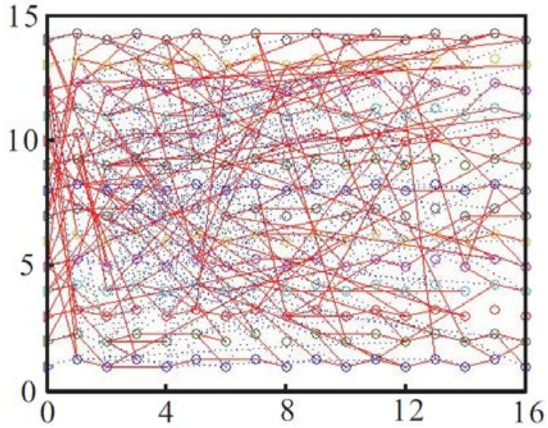

Figure 1.The Structure of Traditional Network

Circle represents the neuron has the basic function of the same computing power unit. Each neuron by its only determines the rows and columns. Each neuron in addition to the data input and data output and contains control of a binary input and two output control. Each neuron corresponds to its data input and control input is an event-driven processor. When the control signal is high, the neurons are activated, only the selected data on its input space for simple arithmetic or logic operation to reduce the information entropy value of the input data, so as to obtain the data information. Under the activated state, various statistical information and control the output neurons are obtained from calculation. The sub-space is defined in the formula 1.

$$\begin{aligned} S_{ot} &= \left\{ x \in S_{k,l} : \Phi_{k,l}\left(x\right) > \tau \right\} \\ S_{oti} &= \left\{ x \in S_{k,l} : \Phi_{k,l}\left(x\right) > \tau \right\} \end{aligned} \tag{1}$$

Neurons and its nearby neurons randomly is linked together, constitute a mesh structure. Each neuron from connected with all the connections in the choice of its input data sources, and arithmetic or logical on the selected data transformation, the results output to the next level of neurons, the connections between neurons is loose, and it is random, the farther the distance, the lower the connections between neurons the probability of each other. In addition to the connection of the data line, and connections between neurons, and control the connection also follow the principle of locality and loose. Each neuron cables and control cables have the same number of data. The traditional neural network connection number increases with the number of neurons in a cubic increasing, and the number of wires in this structure linearly increased with the increase of the number of neurons, so it will greatly save the practical application of software and hardware resources. Usually, the number of neurons in each layer is take dimension is greater than or equal to the input data. Based on the self-organizing characteristics of network structure, network layer can be in the learning process of data by the learning algorithm to automatically determine. For the structure of neural network, the data of the learning process is the process of network structure optimization. Through the study of the input data, the internal functions of the connections between neurons and the neurons can be refined and optimized [6]. The instructions are shown in the figure 2.

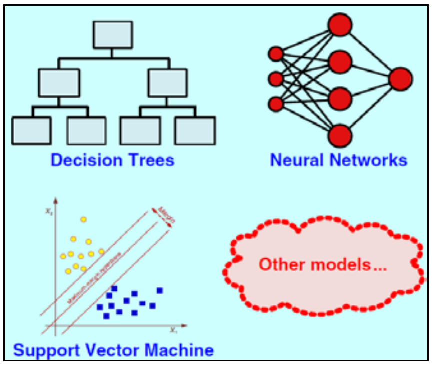

Figure 2.The General Instructions for the Model

Statistical Learning Theory. Network neurons synchronized to the input data in the study, by calculating the information entropy to choose the optimal transformation function and threshold value to obtain the largest amount of information. At the end of the study, the optimal information entropy and the corresponding input data selection, transformation function and threshold value are preserved. In the training phase, the network connection and each neuron adjustments according to the input data information and tend to be a stable adaptive structure. When the system is in the testing phase, the selected control input will determine whether a neuron is activated and vote for the classification of a test sample, the classification of the final result will be voting for activation of neurons by all. Neurons in the training phase, the selection of input data and by transform function to transform as the output data, and by setting the threshold value of comparison, the input data can be sure belong to two subspaces by threshold segmentation. The quality of learning by the network is to obtain information from the data to determine. In order to calculate the amount of information from data, need to be calculated according to the training sample probability as the following formulas. Using the information entropy concept, through different kinds of training data on the probability of each subspace, defines the amount of information obtained from the input data, namely.

$$I = 1 - \frac{\Delta E_s}{E_{\max}} = 1 - \sum_{sc} P_{sc} \log\left(P_{sc}\right) - P_s \log\left(P_s\right) / P_c \log\left(P_c\right) \qquad (2)$$

Information is input space degree of measurement for segmenting neurons, represents the input data contains the degree of access to information. For each neuron, choose a different input data transformation function and the combination of the threshold will get different information. In the training phase, each neuron to selection of input data, actively adjust the transformation function and threshold value to obtain maximize the amount of information. When a learning process for all sample data network, each neuron input connection, transformation function and threshold value are determined according to the information contained in the training data and, namely the network according to the information from the training data, the self-organization optimization. After finishing training phase of the learning process, the test sample data can be input to the neural network, the network will be classified according to the results of training samples of the test. Each neuron in the stage of training records for each type of data in the output space recognition probability. The recognition probability of individual neurons classification can't reflect the real situation, here, a voting mechanism of MRC weight function is used to classify the input data.

$$B_c = 1 - 1 / \left\{ 1 + \sqrt{\sum_{i=1}^{n} \left(1 / \left\{ \frac{1}{P_{cci}} - 1 + \varepsilon \right\} \right)^2} \right\} \tag{3}$$

In addition, because the initial network connection is random, each with different initial connection, which may lead to the differences of each network which will be a performance. In order to balance the performance differences, can produce multiple neural network at the same time, synchronous training to them, the final classification result by the collective vote to determine all network. Experiments show that collective voting mechanism by multiple neural networks can improve the accuracy of classification result.

The Time Series Analysis. Refers to the time series data of adaptive method translates into another data space, and the conversion process and characteristic coefficient choose independently of the data itself; Data adaptive method not only depends on the local data values of a single time series, and the influence of time series data set all the data objects, such as singular value decomposition method to increase or delete the data set arbitrary object will influence the final characteristics of results. Model based method is to suppose beforehand by some time series model, by establishing the model, finally using coefficient of model parameters or to represent the time sequence method. In recent years, the characteristic of time series representation and similarity measure research has made some progress, and is widely applied in various fields. Patient groups in the medical and health care, for example, anomaly detection, the financial data of individual stocks of similarity and abnormal findings and monitor consumer spending fraud, etc. With the deepening of the research, however, there are some issues worthy of research and attention. However, due to the high order polynomial fitting time series is easy to appear over fitting phenomenon, and higher order orthogonal polynomial base vector corresponding to the coordinate coefficient is far less than that of low time orthogonal polynomial base vector corresponding coordinate coefficient, if using the Euclidean distance to measure the coordinate coefficient sequence similarity, is easy to overlook the coordinates of numerical coefficient implied important information of the original time series, and the corresponding distance metric function has not been proved in theory satisfies the requirement of lower bound, omission phenomenon could occur in the similarity search. Therefore, it is necessary to research a kind of meet the requirements of lower bound and measure quality higher similarity measure method, to improve the orthogonal polynomial regression model in the time series data mining application effect. Time series piecewise aggregate approximation and symbolic representation method is now more popular time series feature representation, it has been widely in the field of time series data mining application. At the same time, the corresponding similarity measure method meets the demand of distance lower bounds, avoids the production of similarity search omission. But as a result of piecewise aggregate approximation and considering the average segment sequence information, only for the area of data distribution without further consideration, which ignores the local form of information and data distribution uncertainty, not well to compare has obvious form of the distribution of time series. So the comprehensive consideration or aggregation of data distribution uncertainty means and form approximate method and symbolic representation method also has the very vital significance and broad application prospects.

The In-Depth Mathematical Analysis. Support Vector Machines (SVMs) belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. The mapping function in SVMs can be either a classification function (used to categorize the data, as is the case in this study) or a regression function (used to estimate the numerical value of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable.

The mathematical formulation according to the literature reviews [7-10] could be summarized into the following formulas:

$$k\left(x,t\right)\left(\frac{x}{t}\right)^{\frac{1+\varepsilon}{l}}=\frac{\left(mt\right)^{\frac{\lambda-1}{2}}}{\left(\max\left\{m,\mathrm{t}\right\}\right)^{\lambda}}\left(\frac{x}{t}\right)^{\frac{1+\varepsilon}{l}} \tag{4}$$

$$c_t^i=\left(\sum_{j=1}^{m}w_{cij}\times L_{ij}\left(c_t^j\right)\right)/\sum_{j=1}^{m}w_{cij} \tag{5}$$

Dynamic time warping method is a kind of good scalability time series similarity measure method, it can be said without time series characteristics under the condition of directly comparing similarity effectively. Compared with Euclidean distance measure method, not only has no sensitivity for abnormal points, but also can realize the distance between the different lengths of time series measurements. The experiment will be conducted later [11-12].

Experiment and Simulation Result

In order to validate the accuracy of the model, simulation by MATLAB software, the three financial sectors is a typical problem is analyzed. In this experiment, the first is based on financial ratios and net asset value of index system, using the proposed neural network to predict focused company. It and the traditional neural network has a certain similarity in the structure, but the behavior of individual neurons and similar to the SVM algorithm. The structure and the characteristics of the algorithm has the more powerful than the traditional neural network to the vast and complex data processing ability, especially for the large amount of data and complex financial problem. The loose connections and extensible rules of network structure is suitable for hardware implementation, can help to improve the practical application of operation speed, increased the possibility of application in practice. The experimental result is shown in the following figures.

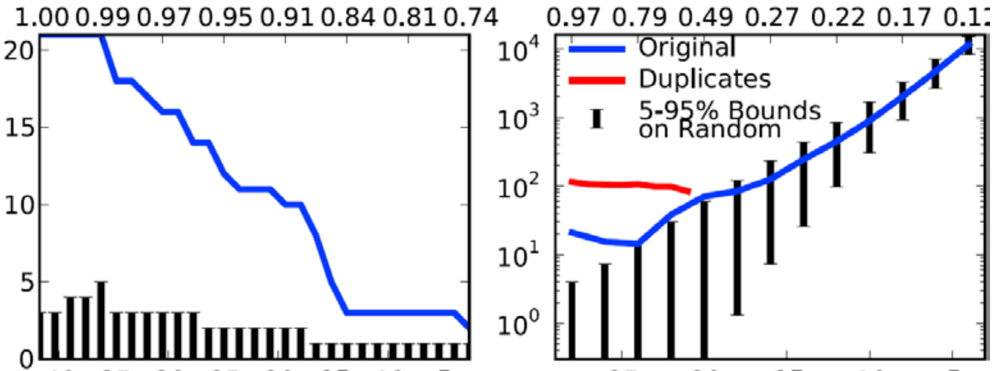

Figure 3.The Experimental Result for Our Approach

Conclusion and Summary

With the development of computer technology, multimedia technology and financial technology, people are increasingly exposed to a lot of financial data. Time series is a kind of common and associated with the time of high-dimensional data, also is the main research object in the area of data mining, widely exists in financial, medical, weather, and in the field of network security. Time series is an important data, the prevalence of time series to make use of data mining technology can effective access to information and knowledge discovery. In the process of time series data mining, feature representation and similarity measure is an important basic work, the smooth implementation of for other data mining task provides a good data processing method and technical support. Based on information entropy, this paper puts forward a kind of neural network classification method based on statistical learning theory, is different with the traditional neural network method. Its classification algorithm is based on the network all the neurons in the voting results, there is no convergence problem or not. It scalable network structure and connection mode is suitable for programmable hardware implementation, is advantageous to the characteristics of high dimensional data extraction and analysis, have huge amounts of data to solve the high dimension. The experimental analysis reacts to the fact that the proposed method is feasible.

Acknowledgements

Research of Creativity and rule matching with science and technology resources allocation.
G12 - XK25.

References

[1] Yukselturk, Erman, Serhat Ozekes, and Yalın Kılıç Türel. "Predicting dropout student: An application of data mining methods in an online education program." European Journal of Open, Distance and e-Learning 17.1 (2014): 118-133.

[2] Ismail, Leila, Mohammad M. Masud, and Latifur Khan. "FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing." Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014.

[3] Alglave, Jade, Luc Maranget, and Michael Tautschnig. "Herding cats: Modelling, simulation, testing, and data mining for weak memory." ACM Transactions on Programming Languages and Systems (TOPLAS) 36.2 (2014): 7.

[4] Chen, Bing, et al. "Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel." Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, 2014.

[5] Hu, Xiaolin, Ming Lu, and Simaan AbouRizk. "BIM-based data mining approach to estimating job man-hour requirements in structural steel fabrication." Proceedings of the 2014 Winter Simulation Conference. IEEE Press, 2014.

[6] Vaidya, Jaideep, et al. "A Random Decision Tree Framework for Privacy-preserving Data Mining." Dependable and Secure Computing, IEEE Transactions on 11.5 (2014): 399-411.

[7] Xu, Biao, Xu-Huan Wang, Wei Wei, and Haoxiang Wang. "On reverse Hilbert-type inequalities." Journal of Inequalities and Applications 2014, no. 1 (2014): 1-11.

[8] John, Fritz. "Extremum problems with inequalities as subsidiary conditions." In Traces and Emergence of Nonlinear Programming, pp. 197-215. Springer Basel, 2014.

[9] Lederer, Johannes, and Sara Van De Geer. "New concentration inequalities for suprema of empirical processes." Bernoulli 20, no. 4 (2014): 2020-2038.

[10] Chauvel, Louis. Intensity and shape of inequalities: The ABG method for the analysis of distributions. No. 609. LIS Working Paper Series, 2014.

[11] Ben-Tal, Aharon, Dick Den Hertog, and Jean-Philippe Vial. "Deriving robust counterparts of nonlinear uncertain inequalities." Mathematical Programming 149, no. 1-2 (2015): 265-299.

[12] Mackenbach, Johan P., Ivana Kulhánová, Gwenn Menvielle, Matthias Bopp, Carme Borrell, Giuseppe Costa, Patrick Deboosere et al. "Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries." Journal of epidemiology and community health (2014): jech-2014.