International Journal of Finance and Investment
ISSN: 3078-6002 (Print), ISSN: 3078-2295 (Online) | Vol. 3, No. 2, 2025

Stock Analysis Based on Decision Tree

Ganzhou Wu

School of science, Guangdong University of Petrochemical Technology, Maoming 525000, China

Abstract: It is selected the data of liquor leader Kweichow Moutai (600519) from January 1, 2020 to March 24, 2023, and is included the most representative and well-known opening price, closing price, highest price of the day, lowest price of the day. There are five items of trading volume as horizontal indicators. The 10-day moving average (MA10) is used as a longitudinal indicator. In the Adaboost regression, the rmse values for the training and test sets were 1.392 and 16.507, respectively. That is to say, after adding the 10-day moving average (MA10) of longitudinal data, the rmse values of the training set and the test set decreased by 0.058 and 0.559 respectively. Among them, 80% of the data is randomly selected as the training set, and 20% of the data is used as the test set.

Keywords: Decision tree, AdaBoost, Stock.

1. Introduction

Wavelet neural network is a new affine wavelet neural network first proposed by American scholars Pati and Shnaprasad in 1992. It is a method that uses wavelet to map neurons. After that, Morgan and Stannlog developed an Automated Investor system, which helps investors find the best investment plan in the securities market through clustering analysis technology, visualization methods and other technologies [1]. In addition, Binoy, Dafini and Emerhandas have also established an automated decision tree adaptive neural network-fuzzy hybrid system that can predict the stock market trend. First, the decision tree method is used to extract characteristic variables, reduce their dimensions, and apply them to the neural network to achieve the next day's stock market trend prediction [2]. On this basis, Penman H established an automatic prediction model of "Pr-measure", which can predict the annual profit change of enterprises. In short, through the analysis of the stock market trend, investors can be provided with more information, so as to make better investment decisions. In the work of Nortola, Condamin, Naim et al., they followed the development of machine learning and used feedback-based artificial neural network roadshow algorithms. They use decision tree technology to extract rules from them, and then randomly select a thousand companies in the same industry, and use their balance sheets, organizational structures, business activities, etc. as input variables [3]. On this basis, this paper also puts forward a method of "corporate financial status measurement" to select company stocks with good foundation [4]. This survey shows that if opinions are presented according to the decision tree model, the average annual return rate will reach 18.7%. In addition, they also combined rough sets with decision trees, and proposed a new decision tree model. The method uses rough set theory to analyze incomplete or imprecise data, and uses decision tree to classify and predict them, so as to achieve the purpose of making more accurate investment decisions for investors. In the stock market, decision tree algorithm is a promising algorithm, and its prediction accuracy is high. This paper proposes a stock price forecasting method based on SVM regression, which predicts the stock price, and its results are better than BP neural network [5]. Based on the improvement of the least squares support vector machine (LS-SVM) model, the Nasdaq index was predicted by using the improved least squares support vector machine (LS-SVM) model [6]. In the forecast of Nasdaq index, the LS-SVM model has obtained better forecast results [7].

Guo Gang, based on the basic principle of stock price K-line, proposed a method based on the combination of cognition and fuzzy to achieve multi-step prediction of stock price changes over time. On this basis, a new multi-step method of stock price based on chaos theory is proposed. This method only needs to consider whether the system belongs to the chaotic type, and based on the chaos theory, it gives a clear maximum time scale for multi-step prediction of stock prices, and in the multi-step prediction of complex nonlinear systems, this method can also obtain good prediction results [8]. She uses data mining technology to analyze and predict from two perspectives: fundamental and technical aspects of individual stocks in the A-share market with the fundamental and technical indicators of individual stocks as the main research objects, and uses neural network algorithms, logistic regression models, time model analysis, clustering methods, association rule analysis methods, decision tree algorithms and other methods. By comparing and evaluating the models established by various algorithms, the feasibility of various data mining methods in stock prediction is proved [9].

2. Decision Tree and AdaBoost Algorithm

2.1. Decision Tree

A decision tree is a tree-like structure that tests the sample from the root node. Suppose we have a batch of samples of known categories, start from the root node, test all the samples, and take the test results as a new sample, and then divide the samples into different sample subsets according to the test results, each sample subset is a sub-node. This is the process of classifying data with a set of rules. Decision trees provide a way to approximate rules for what value will be obtained in which case.

Decision trees are divided into two categories: one is the decision trees generated for discrete data, and the other is the decision trees generated for continuous data. The decision tree is regarded as a tree, in which the root node is taken as the node, and the node is taken as the starting point, and it is divided into more than two sub-nodes. Each leaf node is a

separate record. When establishing a decision tree, first find a branch of the initial value; All training sets are generated by decision trees, and each training set has to be classified. How to determine which attribute fields (Field) are currently the best classification indicators. The usual method is to use all the attribute fields for classification, and then quantify the classification results, so as to get the best classification results. For attribute domain segmentation, different algorithms use different criteria. Second, repeat the above steps until the data in each leaf node is the same type, and grow it into a complete tree. The goal of establishing a decision tree is to find the relationship between attributes and classification, and to predict the classification of unknown classifications in the future. The decision tree generation process is essentially a process of continuously generating training samples. With the continuous combination of data, each branch of the decision tree will gradually grow. In the process of decision tree generation, the most critical technology is the selection of attributes. The earliest ID3 algorithm was based on entropy in information theory, using a new method called the "gain criterion".

The C4.5 algorithm uses gain ratio as the attribute selection criterion when selecting attributes at all levels of the decision tree.

$$SplitInformation(A,S) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2(\frac{|S_i|}{|S|}) \quad (1)$$

$$GainRatio(A,S) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (2)$$

In SLIQ, SPRINT and PUBLIC algorithms, gini index is used instead of Information as the criterion of attribute selection. The performance of gini index is better than that of information quantity, and it is convenient to calculate. A dataset S, gini (S) that contains n classes for the dataset is defined as:

$$gini(S) = 1 - \sum p_j * p_j \quad (3)$$

In the equation, p_j is the frequency of the j data in S. The smaller the gini, the larger the Information Gain.

2.2. AdaBoost Algorithm

AdaBoost is a machine learning method proposed by Yoav Freund and Robert Schapire. It trains multiple weak classifiers through iterations, and combines these weak classifiers to form a strong classifier. The feature of AdaBoost algorithm is to train the next classifier by using the wrong samples of the previous classifier, so that it has the ability of adaptive improvement. Although it is sensitive to noisy data and abnormal data, compared with other learning algorithms, it is not easy to produce overfitting, so it has great advantages in solving some problems. In terms of implementation, AdaBoost adopts an iterative algorithm, adding a new weak classifier in each cycle until the bit error rate drops below the preset value. Each training sample has a weight value, which indicates the probability that the sample will be selected by a certain classifier. When a certain point is correctly classified, the possibility of the sample point being selected will decrease when constructing the next training set; On the contrary, when the sampling points are not correctly classified, their weights can be increased. In this way, AdaBoost can "focus" on indistinguishable (informative) samples, thus improving the accuracy of the classifier. In the specific implementation process, we can select the training samples and weak classifiers according to the weight of the samples, and combine multiple weak classifiers into a strong classifier by weighting and summing.

The basic idea of AdaBoost algorithm is to obtain a series of basic classifiers after repeated calculations, and then combine these basic classifiers together to form a strong classifier. The specific steps are as follows:

Let x_i and y_i represent the sample points and their class labels of the original sample set D. Let $W_k(i)$ represent the weight distribution of all samples at the kth iteration. This gives the AdaBoost algorithm as follows:

Initialization: input parameters are training set $D = \{x1, y1, \ldots, xn, yn\}$, maximum number of cycles k_{max}, sampling weight $W_k(i) = 1/n$, $i = 1, \ldots, n$; Iteration counter k is assigned 0; Counter k self-increments by 1;

Weak learner C_k was trained using $W_k(i)$ sampling weights; The training results of the weak learner C_k are evaluated and recorded in the error matrix E_k.

$$\alpha_k \leftarrow \frac{1}{2} ln \frac{1-E_k}{E_k} \quad (4)$$

$$W_{k+1}(i) \leftarrow \frac{W_k(i)}{Z_k} \times f(x) = \begin{cases} e^{-\alpha_k}, & if\ h_k(x_i) = y_i \\ e^{\alpha_k}, & if\ h_k(x_i) \neq y_i \end{cases} \quad (5)$$

Stop training when $k = k_{max}$.

3. Experimental Results

First of all, we must choose a suitable data sample, and the selection of samples has a direct connection with the rationality and effectiveness of the results. Since there are great differences in economic policies, company conditions, industry demand, etc. in each field, in order to improve the rationality and scientificity of the forecast, this article selects the liquor leader Kweichow Moutai (600519) from January 2020 Data from January 1 to March 24, 2023, and select the most representative and well-known opening price, closing price, highest price of the day, lowest price of the day, trading volume, and 10-day moving average (MA10), each as descriptive Statistical indicators are analyzed and discussed. The input variable is the above 6 indicators of the company every day, and the output variable is the closing price of the next day. Data All From Nuggets Data (juejinshuju.com).

In the traditional stock analysis algorithm, the selected indicators are generally: trading volume, the lowest price of the day, the highest price of the day, the opening price, and the closing price. If only these five indicators are selected, only the horizontal changes of the data are considered, that is, only the situation of the stock on the day can be seen, and the connection between a certain stock before or in the future and the present cannot be reflected. Therefore, in addition to considering the above five indicators in this paper, we have added a new feature on the basis of the original features: the 10-day moving average (MA10). The six features after adding the additional features are defined as follows:

$$F_2 = (f_1, f_2, f_3, f_4, f_5, f_6) \quad (6)$$

Where f_1 represents the opening price, f_2 represents the closing price, f_3 represents the highest price of the day, f_4 represents the lowest price of the day, f_5 represents the trading volume, and f_6 is the 10-day moving average (MA10).

Some of the data collected is as follows:

Table 1. Partial data

pub_date	open	high	low	close	volume	MA10
2020-01-15	1096.205	1108.65	1092.241	1099.289	2602911	1090.089
2020-01-16	1105.951	1105.951	1089.849	1094.218	2319165	1087.815
2020-01-17	1097.184	1099.932	1088.297	1094.713	2347275	1090.676
2020-01-20	1099.022	1099.022	1069.507	1078.403	3539130	1091.962
2020-01-21	1068.518	1074.449	1059.919	1062.884	3287405	1090.061
2020-01-22	1057.646	1071.484	1043.194	1063.092	3620004	1088.812
2020-01-23	1063.576	1063.576	1025.027	1040.644	5346843	1083.88
2020-02-03	973.6269	999.0104	968.6846	992.3284	12344288	1073.147
2020-02-04	1003.281	1044.796	999.3366	1026.025	6262418	1064.621
2020-02-05	1037.877	1041.83	1021.102	1037.867	4741824	1058.946

For the data that has been obtained, necessary preprocessing is required to ensure that the data obtained is complete and accurate. When building the model, due to the large demand for training samples, 80% of the obtained samples are used as training samples and the remaining 20% are used as test samples.

Table 2. AdaBoost Regression parameter setting

Parameter name	parameter value
Training time	0.204s
Data segmentation	0.8
Data shuffle	no
Cross validation	no
Number of base classifiers	100
Loss function	Linear
Base classifier	Decision tree classifier
Learning rate	1

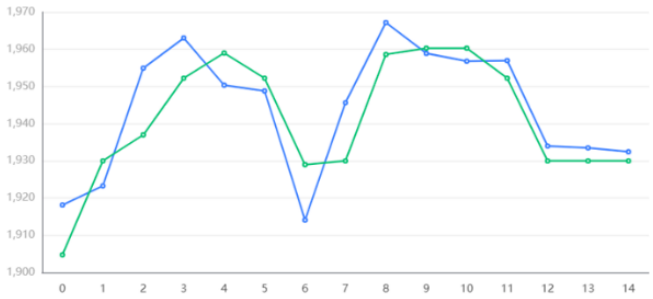

Figure 1. Data prediction

As can be seen from the figure above, modifying the basic prediction method used in the model and adding additional longitudinal input features can improve the accuracy of the prediction. When forecasting the stock price trend, this method has a better fitting effect than the decision tree method. Moreover, the error in trend forecasting is also gradually reduced. To some extent, it can be seen that the selected methods and additional indicators have a certain positive significance in improving the accuracy of experimental forecasting.

Table 3. AdaBoost Regression parameter setting

Numble	MSE	RMSE	MAE	MAPE	R^2
training set	6560.207	80.995	58.484	3.808	0.937
Test set	7315.536	85.531	65.931	3.695	0.651

Therefore, it can be judged that adding longitudinal indicators can reduce the mean square error and increase the accuracy of prediction. In the test samples of this experiment, because the standard square error is not small enough, there is a trend of uncertainty or even some errors in the prediction of stock trends. But this trend is not obvious in Adaboost decision tree.

4. Summary

The experimental results show that the Adaboost algorithm based on the decision tree can effectively reduce the mean square error by adding the 10-day moving average (MA10) as an input variable to construct a prediction model, thereby better improving the accuracy of the prediction results.

References

[1] Cao Ying, Miao Qiguang, Liu Jiachen, et al. Progress and Prospect of AdaBoosting Algorithm Research [J]. Journal of Automation, 2013, 39 (06): 745.758.

[2] Li Yijing, Guo Haixiang, Li Yanan, et al. Classification of a Boosting-based ensemble learning algorithm in unbalanced data [J]. Systems Engineering Theory and Practice, 2016, 36 (01): 189-199.

[3] Nottola, Condamin, Naim. Application study of BP neural network and decision tree on stock market prediction [C]. Ninth International Conference Oil Hybrid Intelligent Systems, 1991, (8): 174-178.

[4] Yuan Senmiao, Cheng Xiaoqing. Research on Decision Tree Algorithm in Quantitative Association Rule Discovery [J]. Journal of Computer Science, 2009, 123 (8): 866-871.

[5] KJ Kim. Financial time series forecasting using support vector machines [J]. Neurocomputing, 2003, 55 (1-2): 307-319.

[6] ZG Xiao, HS Xu, N Yao, et a1. Nuclear physics programs at HIRFL [J]. Air Conference, 2010, 1235 (11: 159-164.

[7] Guo Gang. Research and Application of Stock Intelligent Forecasting Decision-making [D]. Xi'an: Northwestern Polytechnical University, 2020.

[8] Tao Yuyu. Application of Decision Tree and Neural Network Algorithm in Stock Classification Prediction [D]. Hangzhou Dianzi University. 2023.

[9] Wang Yu. Stock Forecasting Model Based on Cart Tree and Boosting Algorithm [D]. Harbin University of Science and Technology, 2018.