Stock market movement direction prediction using tree algorithms

Gunter Senyurt, Abdulhamit Subasi
E-mail : *gsenyurt@ibu.edu.ba, asubasi@ibu.edu.ba*

Abstract

One of the highly challenging businesses today is the task of forecasting the market movements by examining the financial time series data as correctly as possible in order to hedge against the almost incalculable risk involved and to yield better profits for investors. If there was a highly credible estimation technique available giving better results than the traditional statistical tools for financial markets, it would be a great asset for trading decision makers of all kinds such as speculators, arbitrageurs, portfolio fund managers and even individual investors. In this study CART, C4.5 and Random Forest algorithms were used to predict the movement direction of a 10 year Istanbul Stock Exchange index (XU-100). Ten technical market indicators such as momentum, MACD and RSI were used in this study as the feature set.

Keywords: Price movement direction, CART, C4.5, Random Forest, forecasting, stock market.

1. INTRODUCTION

The complex dynamism of the markets is characterized by the nonlinearity and nonparametric nature of the variables influencing the index movement directions including human psychology and political events. The unpredictable volatility of the market index makes it a highly challenging task to accurately forecast its path of movement. On the other hand, it is crucial for investors to estimate the trend of the markets as precisely as possible in order to reach the best trading decisions for their investments, so in this context it is in the investor's best interest to use the most accurate time series forecasting model to maximize the profit or to minimize the risk. By means of this study, it is aimed at contributing to the demonstration and verification of the XU-100 index movement path predictability through some tree algorithms. The stochastic performance parameter is accuracy and it is defined as the ratio of the correctly classified instances divided by the number of all instances. The remaining part of this study is organized into four sections. The next section presents an overview of the theoretical literature while in section 3 the research data and the structures of tree algorithms CART, C4.5, Random Forest is described. In section 4, the reports and results of empirical findings from the comparative WEKA analysis are given. Finally, the last section contains the concluding remarks.

2. Literature Review

CART review

The classification tree analysis CART (classification and regression trees) is suggested first by Breiman (1984) and uses the predictor variables splitting rule to build a binary decision tree (Denison, Mallick and Smith, 1998). The CART method is experimented in the credit scoring area, retail lending and evaluation of insurance risks in workers' compensation showing better results than logistic regression and discriminant analysis (Friedman 1991, Devaney 1994, Lee 2006, Kolyshkina 2002).

C4.5 review

The C4.5 method is high in efficiency when used for inductive inference. Recent research has shown that this algorithm produces high accuracy in image segmentation (Polat and Gunes, 2009; Mazid, Ali and Tickle, 2010). In another work a hybrid approach including C4.5 is suggested with potentially high outcomes (Jiang and Yu, 2009; Mazid, Ali and Tickle, 2010). It is also used for classification of remote sensing data (Yu and Ai, 2009; Mazid, Ali and Tickle, 2010). Another variant of C4.5 successfully trimmed down the leaf node number and improved accuracy (Yang, 2009; Mazid, Ali and Tickle, 2010).

Random Forest review

High-dimensional classification and regression problems can be approached by using random forest algorithm that is extensively researched by Breiman (2001). Among the machine learning techniques used to predict markets random forest is quite successful (Dietterich, 2000). Though the practicality of random forest is excellent it is hard to interpret and clarify mathematically (Breiman, 2002; Lin and Jeon, 2006; Biau et al., 2008, Biau and Devroye, 2008).

3. Materials and Methods

CART method

CART constructs a tree where the data is separated into two parts by binary variable splits. The best divider variable and the best point to split is determined by variance minimization.

The CART algorithm can be viewed as a classification procedure consisting of four distinct parts:

Part 1: a variance criterion,

Part 2: the criterion how good it is split,

Part 3: the terminal node class assignments and estimates of resubstitution,

Part 4: determining the right tree complexity (Buyukbebeci, 2009).

The root node, internal nodes and leaf (terminal) nodes constitute the CART tree. Two child nodes follow each root and internal node. Each node contains and is defined by the subset of the original learning sample. The splitting of each node into child nodes is characterized by a certain rule depending on the chosen feature. The child nodes inherit subsamples with minimum variance that measures their heterogeneity from parent nodes (Iscanoglu, 2005).

The goodness of the splitting procedure is defined by an impurity function that is derived from a a variance function which is applied to each split point indicating the best point for splitting (Iscanoglu, 2005).

Gini, Entropy and Twoing are the main rules for binary recursive splitting that are derived from the impurity function (Breiman, Frydman, Olshen and Stone, 1984)

C4.5 Method

In doing classification with C4.5, the concepts of entropy and correlation coefficient need to be explained in brief. Entropy is a measure of uncertainty among random variables in a collection of data or in other words entropy provides information about the behavior of random processes used in data analysis. Correlation coefficient has its uses as a chief statistical tool in data analysis finding the relationship between variable sets. Different ways of calculations have been introduced to boost the efficiency of the correlation coefficient among which are Kendall, Pearson's and Spearman's correlation coefficients.There are several test options with WEKA providing data classification such as training set, supplied test set, percentage split and cross validation. In this paper, cross validation is chosen as the test option (Mazid M., Ali S. and Tickle K. (2010).

Random Forest method

Random forests are based on conjoining lots of binary regression trees. In the process of growing these large number of regression trees independent subsets of variables are used. Random forests randomly choose variables to split and a bootstrapped sample of the dataset builds the decision trees (Efron and Tibshirani, 1993). When K trees are aggregated the predicted decision is gained as the average value over these K trees. Marking each single tree predictors by $h_1, ..., h_K$, the final outcome is:

$$h(X) = \frac{1}{K} \sum_{k=1}^{K} h_k(x)$$

Research Data

In this study, all experiments were conducted on WEKA software using its tree classifiers built-in tool to make comparisons of prediction performances based on the chosen dataset. The dataset is comprised of 10 input variables with 2733 instances in total. These 10 input attributes are technical market indicators as used by Kara, Boyacioglu and Baykan (2010) which are 10-day moving average, 10-day weighted moving average, momentum, stochastic %K, stochastic %D, RSI (Relative Strength Index), MACD (moving average convergence divergence), Larry William's %R, A/D (Accumulation/Distribution) Oscillator and CCI

(Commodity Channel Index). The total number of cases or 2733 trading days have 1440 days with increasing direction (advances), while 1293 days show decreasing direction (declines). In the analysis, 10-fold cross-validation was used as the test option in WEKA.

4. Results and Discussion

The relevance and quality of the data, usually, has a big impact on the performance of the model used. Thus, the choice of data becomes the most important part in forecasting the markets. In this study, all series are real-valued and the input data spans from 02/01/1997 to 31/12/2007. For WEKA testing, the accuracy or correctly classified instances metric is utilized, showing the ability of the model to capture the data. The dataset with 10 features is tested using CART, C4.5 and Random Forest classifiers in order to see which tree algorithm has better predictive power over the others. The results of the tests can be seen in the Table 1 where CART and Random Forest classifiers have almost identical prediction power, whereas C4.5 has a little less prediction power compared to the other two tree algorithms.

Table 1. Tree Classifiers Test Results

	% Accuracy (correctly classified instances)
CART	78.05
C4.5	77.29
Random Forest	78.23

5. CONCLUSION

The issue of accurately predicting the stock market price movement direction is highly important for formulating the best market trading solutions. It is fundamentally affecting buy and sell decisions of an instrument that can be lucrative for investors. This study focuses on predicting the ISE National 100 closing price movement directions using tree algorithms based on the daily data from 1997 to 2007. Even though the prediction performance of tree classifiers such as CART, random forest and C4.5 do not really outperform studies alike in literature, it is still likely that the forecasting performance of the models can still be improved by doing the followings: Either the model parameters should be adjusted by thorough experimentation or the input variable sets need to be modified by selecting those input attributes that are more realistic in reflecting the market workings. (Kara, Boyacioglu, and Baykan, 2010) had already proved the significance of using ten particular technical market indicators which gave also about %78 accuracy in this study, as well. More appropriate

variables has to be found that may improve the forecasting performance of the models employed that can be a further subject of study for interested readers.

Acknowledgement : We sincerely deliver our special thanks to Assist. Prof. Melek Acar Boyacioglu for her graciousness in sharing her knowledge with us.

REFERENCES

Biau G., Devroye L., and Lugosi G. (2008), Consistency of random forests and other averaging classifiers, Journal of Machine Learning Research, 9, 2015-2033.

Biau G., Devroye L., and Lugosi G. (2008), On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, Technical report, Universite Paris 6.

Breiman, L., Frydman, H., Olshen, R.A., and Stone, C.J. (1984), Classification and Regression Trees, Chapman and Hall, New York, London.

Breiman, L. (2001), Random forests, Machine Learning, Kluwer Academic Publishers, 45, 5-32.

Breiman, L. (2002), Manual on setting up, using, and understanding Random Forests v3.1, Technical Report, http://oz.berkeley.edu/users/breiman.

Buyukbebeci, E. (2009), Comparison of MARS, CMARS and CART in predicting default probabilities for emerging markets, Master Thesis, METU, Ankara.

David G.T. Denison, Bani K. Mallick, Adrian F.M. Smith, Biometrika, Vol.85, No.2 (1998), 363-377.

Devaney, S. (1994), The Usefulness of Financial Ratios as Predictors of Household Insolvency: Two Perspectives, Financial Counseling and Planning, 5, 15-24.

Dietterich T.G. (2000), Ensemble methods in machine learning, Lecture Notes in Computer Science, Springer-Verlag, 1-15, 2000.

Efron B. and Tibshirani R. J. (1993), An Introduction to the Bootstrap, New York, Chapman and Hall.

Friedman, J.H. (1991), Multivariate adaptive regression splines, The Annals of Statistics, 19, 1 ,1-141.

Frydman, H., Olshen, R.A., and Stone, C.J., Classification and Regression Trees, Chapman and Hall, New York, London, 1984.

Iscanoglu, A. (2005), Credit Scoring Methods and Accuracy Ratio, Master Thesis, METU, Ankara.

Jiang S. and Yu W. (2009), A Combination Classification Algorithm Based on Outlier Detection and C4.5, Springer Publications.

Kara Y., Boyacioglu M.A., Baykan O.K., (2010). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications 38, 5311–5319.

Kolyshkina I. and Brookes R. (2002), Data Mining Approaches to Modeling Insurance Risk, Report, Price Waterhouse Coopers.

Lee T., Chiu C., Chou Y., and Lu C. (2006), Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Computational Statistics & Data Analysis, 50, 1113-1130.

Lin Y. and Jeon Y. (2006), Random forests and adaptive nearest neighbours, Journal of American Statistical Association, 101, 578-590.

Mazid M., Ali S. and Tickle K. (2010), Improved C4.5 Algorithm for rule based classification, Recent Advances in Artificial Intelligence, Knowledge Engineering and Data Bases, Australia.

Polat K. and Gune S. (2009), A novel hybrid intelligent method based on C4.5 decision tree classifier and one against-all approach for multi-class classification problems, Expert Systems with Applications, vol.36, 1587-1592.

Yang X.Y. (2009), Decision tree induction with constrained number of leaf node, Master Thesis, National Central University, Taiwan.

Yu M. and Ai T.H. (2009), Study of RS data classification based on rough sets and C4.5 algorithm, In Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.

Weka (1999-2010), Waikato Environment for Knowledge Analysis, Version 3.7.3, The University of Waikato Hamilton, New Zealand.