American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-5, Issue-12, pp-207-212
www.ajer.org
Research Paper Open Access

A Study on Technical Indicators in Stock Price Movement Prediction Using Decision Tree Algorithms

J Sharmila Vaiz[1], Dr M Ramaswami [2]

[1]*Ph.D. Research Scholar, Dept. of Comp. Applns. Madurai Kamaraj University, Madurai, India*
[2]*Associate Professor, Dept. of Comp. Applns, Madurai Kamaraj University, Madurai, India*

ABSTRACT: *Predicting stock price movement is a highly challenging task as the nature of the stock prices are quite noisy, dynamic, complicated, non-parametric, chaotic and non-linear. The toughest job in stock market is to examine the financial time series data and make decisions which improve the investment returns and to minimize the loss incurred. Technical analysis is a trading tool that evaluates securities and attempts to forecast their future movement by analyzing price and volume data. Many traditional statistical tools are available for the investors for making decision in financial market. Many technical indicators such as Moving Averages (SMA, EMA, WMA, VWMA, DEMA), Trend Indictors (MACD, ADX, TDI, Aroon, VHF), Momentum indicators (Stochastic, RSI, SMI, WPR, CMO,CCI), Volatility indicators (BBands, ATR, Dochain Channel) and Volume indicators(OBV, MFI, CMF) are available to analyze the stock price movement. In this study, decision tree classification method is used to analyze the role of these technical indicators in predicting stock price movement of six high market capitalization companies of NSE.*

Keywords: *Technical indicators, Decision tree analysis, ROC curve, AUC*

I. INTRODUCTION

Stock market trend prediction is the act of determining the future value of the stock traded on an exchange. A successful prediction capitulate significant profit to the investors. Prediction methodologies fall into two categories. They are fundamental analysis and technical analysis. Fundamental Analysis is concerned with the company that underlies the stock itself. They evaluate a company's past performance as well as the credibility of its accounts. Many performance ratios are created that aid the fundamental analyst with assessing the validity of a stock, such as the P/E ratio. Warren Buffett is the most famous of all Fundamental Analysts.

Technical analysts or chartists are not concerned with any of the company's fundamentals. The technical analysis, on the other hand, is the evaluation of securities/assets by means of studying statistical indicators generated by market activity, such as past prices and volume. Therefore, technical analysis is the analysis of human mass psychology and it is also called behavioral finance. The technical analysis paradigm is thus that there is an inherent correlation between price and company that can be used to determine when to enter and exit the market.

With the advent of the digital computer, stock market prediction moved into the technological realm. In recent times, many financial analysts and stock market investors seem convinced that they can make profits by employing one technical analysis approach or another to predict stock market by data mining approach. Some of the functionalities of data mining are the discovery of concept or class descriptions, classification, associations and correlations, clustering, prediction, outlier and deviation analysis and similarity analysis. Data classification can be done by many methods like decision trees, rough set approach and artificial neural network have been applied to this domain. A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. The goal is to create a model that predicts the value of a target variable based on several input variables. Some of the predominant decision tree algorithms which includes: (i) ID3 (Iterative Dichotomiser3), (ii) C5.0 (successor of C4.5), and (iii) CART (Classification And Regression Tree). All these rule based forecasting models receive major technical indicators as an input and to predict trend of share-price either in upward direction or downward direction.

Technical analysis is a trading tool employed to evaluate securities and attempt to forecast their future movement by analyzing statistics gathered from trading activity, like price movement and volume. Based on the assumptions of technical analysis is that price movement and volume essentially contains certain patterns over period of time. Data mining techniques would also be used to discover these patterns in an automated manner and subsequently discovered patterns enables to predict future trend of stock price. In this study we use one of the efficient data mining techniques called tree-based classification is being used with 22 technical indicators derived from the OHLCV historical data in order to help the investors to build their decision when to buy or sell.

II. LITERARY REVIEW

Many types of rule-based decision tree models have been widely proposed for the prediction of financial time series. Qasem A AL Radaideh, Adel Abu Assaf and Eman Alnagi[1] studied the effect of decision tree based classification models of ID3 and C4.5 with six response variables attributes viz; Previous Close, Open, Min, Max, Last and a predictor variable Action(Buy/Sell). They transformed each attribute numeric value to discrete values (Positive/Negative/Equal). They analyzed the trading action of the investors of three major companies of Amman Stock Exchange (ASE). Their findings revealed that classifier accuracy of these models were in between 44% - 54%.

Gunter Senyurt, Abdulhamit Subasi[13] studied the impact of decision tree models and forecasting the movement direction of the 10 year Istanbul Stock Exchange Index(XU-100) with ID3, C4.5 and Random Forest methods. They deployed ten technical indicators - 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 and concluded that decision tree provides about 77%-78% accuracy in classifying the instances correctly.

Sadegh Bafandeh Imandoust, Mohammed Bolandraftar[2] attempted to develop three models and compared their performance in predicting stock price movement direction in Tehran Stock Exchange (TSE) Index. They used Decision tree, Random forest and Naïve Bayesian Classifier classification techniques with ten microeconomic variables and 3 macroeconomic variables as input. The ten microeconomic variables are: 10-day SMA, 10-day WMA, 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. Since oil, gold and USD/INR play prominent role in the Iranian economy, these variables were considered as fundamental indicators. The experimental results showed that Decision tree model (80.08%) surpass other two models Random Forest (78.8%) and Naïve Bayesian Classifier(73.8%).

Subsequently, Lamartine Almeida Teixeira a, Adriano Lorena Inacio de Oliveira[3] proposed a method for automatic stock trading which combines technical analysis and nearest neighbor classification algorithm. Stop loss, stop gain and RSI filter was used as technical analysis tool. The proposed method gives higher profitability and reduces the risk of market exposure.

Thus most of the researchers used five to ten technical indicators alone in their study. Based on the survey of the research work carried out by various authors, we observed that limited technical indicators alone were used in their research study and the role of other reliable stock market technical indicators are not discussed so for. In order to enhance visibility of research study we inducted more technical indicators and its effects were studied with three decision tree algorithms ID3, C5.0 and CART. The efficacy of all three models were discussed with various predictive measures including predictive accuracy, F-Measure, Receiver Operating Characteristics (ROC) and Area Under Curve (AUC).

III. MATERIALS AND METHODS

This study aims at discovering the relationship of technical indicators using decision tree, a powerful and easily interpretable algorithm, to predict the trend of a security. Technical indicators are broadly classified into four groups such as trend indicators, momentum indicators, volume indicators and volatility indicators. For example, if the RSI value is above 70 then it indicates the security is overbought and it signals a downtrend market, if RSI value is below 30 then the security is oversold and it signals an uptrend market. We took long time to analyze effect of more superior technical indicators such as moving averages, MACD, RSI, WPR, Aroon etc to conclude about the market behavior. But with the help of decision trees, it is very easy to build the strategy to find the movement of the trend.

3.1. Research Data

In this paper, we have used various datasets consists of 22 technical indicators which are derived from price and volume (OHLCV) daily data of six high market capitalization companies of NSE of India. The daily details of historical data for four years from January 2012 to December 2015 of six companies are retrieved from yahoo finance data. The dataset is divided with a probability of 70% as training set and 30% as testing set. The six high market capitalization companies of NSE are: (i) Tata Consultancy Services Ltd. (TCS) (ii) Reliance Industries Ltd. (RIL) (iii) Housing Development Finance Corporation Ltd. (HDFC) (iv) Hindustan Unilever Ltd. (HUL) (v) Sun Pharmaceutical Industries Ltd. (SPIL) and (vi) I.T.C Ltd. (ITCL)

3.2. Technical indicators

The 22 technical indicators used in this study in predicting the direction of the predicting the direction is listed in Table 1.

Table 1: List of Technical Indicators

S. No.	Variables	Technical Indicators
1	t_sma <- SMA(t)-Op	Simple Moving Average (SMA)
2	t_ema <- EMA(t)-Op	Exponential Moving Average (EMA)
3	t_wma <- WMA(t)-Op	Weighted Moving Average (WMA)
4	t_dema <- DEMA(t)-Op	Double Exponential Moving Average (DEMA)
5	t_vwma <- VWMA(t)-Op	Volume Adjusted Moving Average (VAMA)
6	macd_tzc,macd_tsc	Moving Average Convergence/Divergence (MACD)
7	ADX, ADX_Dir	Average Directional Movement Index (ADX)
8	TDI, TDI_Di	Trend Detection Index(TDI)
9	aroon_up,aroon_down, aroon_osc	Aroon Indicator(Aroon)
10	VHF	Vertical Horizontal Filter (VHF)
11	RSI	Relative Strength Index (RSI)
12	stoch_fastK,stoch_fastD,stoch_slowD	Stochastic Oscillator(stoch)
13	SMI,SMI_signal	Stochastic Momentum Index(SMI)
14	WPR	William%R– Williams Percentage Range(WPR)
15	CMO	Chande's Momentum Oscillator (CMO)
16	CCI	Commodity Channel Index(CCI)
17	Bbands_pctB	Bollinger Bands(BBands)
18	dc_high,dc_low	Donchain channel
19	ATR	Average True Range (ATR)
20	CMF	Chaikin Money Flow(CMF)
21	OBV	On Balance Volume (OBV)
22	MFI	Money Flow Index(MFI)

3. 3. Classification Algorithm

Classification of data objects is a data mining and knowledge management technique and is used to segregate similar data objects together. There are many classification algorithms available in literature but decision tree is the most commonly used because of its ease of execution when compared to other classification algorithms. In this study we use ID3, C5.0 and CART decision tree algorithms to predict the trend of the security.

3.3.1. Iterative Dichotomizer (ID3)

ID3 is one of the classification algorithms in data mining and was developed by Ross Quinlan[8]. ID3 builds a decision tree from a fixed set of examples. The resulting tree is used as a model to classify future samples. The example has several attributes and belongs to a class (like yes or no). The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain [12] which is the measure of purity. Information Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. Entropy measures the amount of information in an attribute.

3.3.2. Classification and regression trees (CART)

CART is a non-parametric decision tree learning technique and was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone[9]. It produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric. CART uses Gini Impurity[8] instead to help it decide which attribute goes into a decision node. Gini Impurity is a measure of the homogeneity (or "purity") of the nodes. If all data points at one node belong to the same class then this node is considered "pure". So by minimizing the Gini Impurity the decision tree finds the features the separate the data best.

3.3.3. C5.0

C5.0 is the most recent Quinlan iteration[9]. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. C5.0 offers a number of improvements on C4.5 in terms of Speed, Memory usage, smaller decision tree, support for boosting, weighting and winnowing.

All the Decision tree algorithms executed and tested in R platform[11]. R is a fast and flexible technical analysis can be done with TTR package[7]. Decision tree algorithm can be implemented with tree, rpart and C5.0 packages in R. R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques, and is highly extensible. R Packages is also easy to learn and have all dependencies being installed automatically. The graphical capabilities of R are outstanding, providing a fully programmable graphics language that surpasses most other statistical and graphical packages. Because R is open source, unlike closed source software, it has been reviewed by many internationally renowned statisticians and computational scientists.

4. Building the model

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. At the end of pre-processing of collected data, the next step is to build the classifier models using decision tree algorithms. A binary classification variable, price change is used to predict the type of trading action (buy/sell). The difference between the close price and the open price is assigned to price change variable. If the difference is greater than zero, then the close price is above the open price and it indicates buy signal else it indicates sell signal. 22 technical indicators indicated in Table 1 are used to build the strategy using decision tree algorithms.

Three decision tree algorithms (ID3, CART, C5.0) are applied to six companies with the assistance of tree, rpart and C5.0 packages in R. From the decision tree graph we observed that the technical indicator -stoch is designated as the root attribute for five companies and whereas the technical indicator –WPR placed as the root attribute for Sunpharma. Stoch indicator provides high information gain in deciding the type of trading action of most of the securities. The process of building tree continues till it reaches the leaf node (buy/sell). Among all 22 technical indicators maximum of 9 indicators appeared in the decision tree model and these indicators alone used to forecast the trend. And also the indicators Stoch, Donchain Channel, BBands contribute more in deciding the trend. Reliance uses 9 technical indicators and HDFC uses 5 indicators on the path to the leaf node.

IV. RESULTS AND DISCUSSIONS

The performance indicators like classification accuracy, F-Measure, Sensitivity, Specificity and AUC for all three models are summarized in Table 2. Based on the results of ID3, C5.0 and CART algorithms depicted in Table 2, we revealed that all the Decision tree models achieves more than 80% predictive accuracy.

Table 2: Classification Accuracy, F- Measure and Area Under Curve(AUC)

Script	Algorithm	Accuracy	F-Measure	Sensitivity	Specificity	AUC
TCS	ID3	0.8433	0.7983	0.8108	0.8696	0.9042
	CART	0.8100	0.7765	0.8108	0.8522	0.8537
	C5.0	0.8433	0.8033	0.8486	0.8348	0.8676
RIL	ID3	0.8800	0.8548	0.8779	0.8750	0.9248
	CART	0.8033	0.7649	0.8081	0.8828	0.8757
	C5.0	0.8667	0.8413	0.9360	0.8203	0.9211
HDFC	ID3	0.8367	0.8032	0.7988	0.9008	0.8906
	CART	0.8333	0.7934	0.8343	0.8550	0.8882
	C5.0	0.8600	0.8292	0.8875	0.8473	0.9098
HUL	ID3	0.8400	0.8106	0.8293	0.8529	0.8802
	CART	0.7933	0.7578	0.8110	0.7721	0.8270
	C5.0	0.8000	0.7902	0.7682	0.8456	0.8305
SPIL	ID3	0.8067	0.7852	0.8166	0.8092	0.8649
	CART	0.7867	0.7355	0.7988	0.7634	0.8357
	C5.0	0.8133	0.7812	0.8521	0.7634	0.8540
ITCL	ID3	0.8267	0.8278	0.8742	0.8227	0.9066
	CART	0.8367	0.8393	0.8050	0.8582	0.8812
	C5.0	0.8800	0.8808	0.8428	0.9362	0.8934

C5.0 algorithm gives a maximum of 88% accuracy in predicting the best action of ITCL and ID3 gives a maximum of 88% accuracy in predicting the best action of RIL security. F-Measure is alternative measure of test's accuracy which is the harmonic mean of precision and recall. Precision is the number of correct positive values results divided by the number of positive results that should have been returned. The F-Measure can be interpreted as a weighted average of the precision and recall, where an F-Measure reaches its best value at 1 and worst at 0. The F-Measure of the three algorithms for six securities is listed in Table 2. C5.0 algorithm yields highest F-Measure value among three models which amounts to 0.8808 for ITC and ID3 outer performs other two models of RIL which gives F-Measure of 0.8548.

The predictive accuracy and F-Measure are good performance indicators to study the efficacy of different classifier models for two class problem. Apart from these evaluation measures, another performance indicator is Receiver Operating Characteristics (ROC). ROC is a graphical plot [5] of the true positive rate (Sensitivity) against the false positive rate (1-Specificity) for the different possible cut points of a diagnostic test. It shows the tradeoff between Sensitivity (True positive rate) and Specificity (1-False positive rate). Classifiers in (0, 0) and (1, 1) are called as default classifiers. The coordinates (0, 0) with TPR=0 and FPR=0 represents a classifier which never predicts positive class and it never misclassifies a negative instance as positive. Similarly, the coordinates (1, 1) classifies all as positives thus producing high number of false negatives. The perfect classifier is located at (0, 1) with TPR=1 and FPR=0. The best possible prediction method would yield a point in the upper left corner or in the co-ordinates (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The point (0, 1) is also called a perfect classification. ROC curve for six companies using three classification algorithms are shown in Fig. 1. Comparing pair of classifiers through ROC is a difficult task when no classifier dominates the other. From the graphs we find many overlapping ROC curves and it is hard to judge the classifier method which is nearer to the perfect classifier (0, 1).

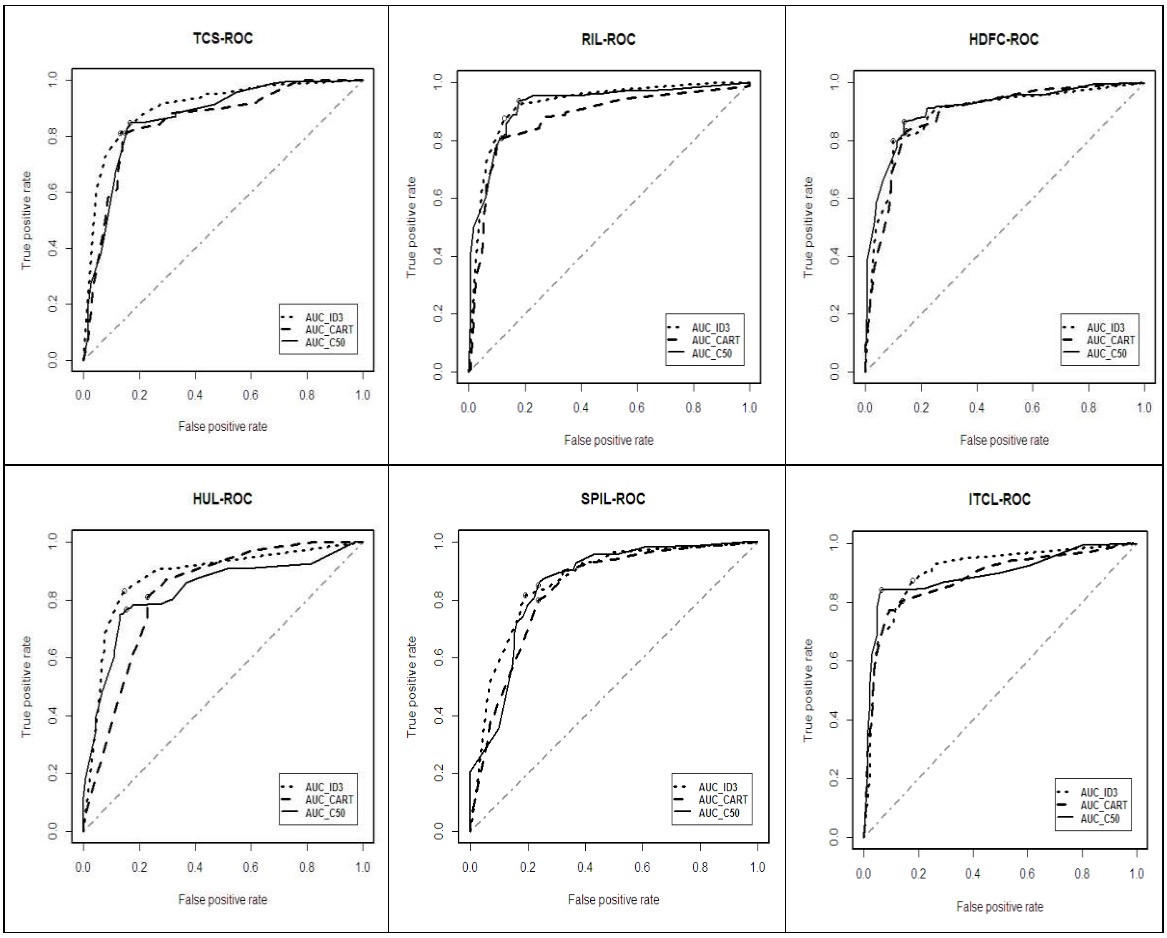

Figure. 1: ROC Curves for TCS, RIL, HDFC, HUL, SPIL and ITCL

In order to resolve this issue, we used another performance measure called Area Under Curve (AUC). The AUC value for each ROC curve is measured by using trapezoidal rule and high AUC value among them indicates the corresponding classification method is close to (0,1). The AUC values for each classification algorithm for all six companies are calculated and shown in Table 2. From Fig.2 we observed that ID3 classifier model surpass the performance of C5.0 and CART classifier models, since ID3 algorithm selects the best attributes based on the concept of entropy and information gain for developing the tree. Based on the AUC values (shown in Fig 2) for all six companies we noted that ID3 algorithm achieve more than 85% accuracy than C5.0 and CART algorithms. Moreover, the difficulty in analyzing the performance of classifier models using ROC graph has also resolved by using AUC value for binary, unbalanced datasets.

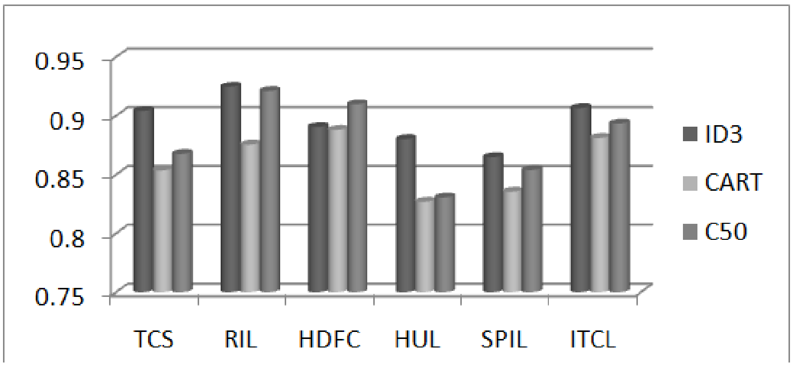

Figure 2: Area Under Curve for ID3, CART and C.50 of six companies

V. CONCLUSION

This study investigates the role of technical indicators in forecasting stock price movement with tree based classifier algorithms in data mining. Classification accuracy of classifier models ID3, C5.0 and CART reveals that technical indicators contributing more in prediction of stock movement and achieved about 85% of accuracy in predicting the market behavior. Performance of ID3, C5.0 and CART are analyzed with many predictive measures – predictive accuracy, F-Measure, ROC curve and AUC value. The investors risk exposure on market behaviour is reduced with the help of these decision tree classification techniques. From our experimental evaluation we concluded that tree based algorithms implicitly retains more relevant technical indicators and exclude redundant/ irrelevant technical indicators during tree construction process itself.

REFERENCES

[1]. Qasem A. Al-Radaideh Adel Abu Assaf and Eman Alnagi, Predicting Stock Prices using Data Mining Techniques, The International Arab Conference on Information Technology (ACIT'2013)

[2]. Sadegh Bafandeh Imandoust, Mohammed Bolandraftar, Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange, Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 4, Issue 6(Version 2), June 2014, pp.106-117.

[3]. Lamartine Almeida Teixeira a, Adriano Lorena Inacio de Oliveira, A method for automatic stock trading combining technical analysis and nearest neighbor classification, Expert Systems with Applications 37 (2010) 6885–6890

[4]. Foster Provost and Tom Fawcett, Analysis and Visualization of Classifier Performance under Comparison under Imprecise Class and Cost Distributions.

[5]. Matjaz Majnik, Zoran Bosni, ROC Analysis of Classifiers in Machine Learning, Technical report MM-1/2011, University of Ljubljana, Faculty of Computer and Information Science, Trzaska cesta 25, Ljubljana, Slovenia.

[6]. James A. Hanley, Ph.D. , Barbara |. McNeil, M.D., Ph.D., The Meaning and Use of the Area under a Receiver Operating Characteristic(ROC) Curve, Reprinted from Radiology, Vol. 143, No.1, Pages 29, 36. April 1982.

[7]. Joshua M Ulrich, Fast and Flexible Analysis with TTR,

[8]. Daniel T Larose, Chantal D Larose, Data mining and predictive analysis, Wiley series

[9]. Breiman, Friedman, Olshen, Stone, Classification and Decision Trees, Wadsworth 1984,

[10]. Tan, Steinbach, Kumar, Data mining Classification : Basic Concepts, Decision Tree and Model Evaluation,

[11]. Joao Neto, Classification and Regression Tree using R,

[12]. Entropy and Information Gain - http://www.math.unipd.it/~aiolli/corsi/0708/IR/Lez12.pdf

[13]. Gunter Senyurt, Abdulhamit Subasi, Stock market movement direction prediction using tree algorithms.