Decision Trees in Stock Market Analysis: Construction and Validation

Margaret Miró-Julià, Gabriel Fiol-Roig, and Andreu Pere Isern-Deyà

Math and Computer Science Department,
University of the Balearic Islands
07122 Palma de Mallorca, Spain
{margaret.miro,biel.fiol}@uib.es

Abstract. Data Mining techniques and Artificial Intelligence strategies can be used to solve problems in the stock market field. Most people consider the stock market erratic and unpredictable since the movement in the stock exchange depends on capital gains and losses. Nevertheless, patterns that allow the prediction of some movements can be found and studied. In this sense, stock market analysis uses different automatic techniques and strategies that trigger buying and selling orders depending on different decision making algorithms. In this paper different investment strategies that predict future stock exchanges are studied and evaluated. Firstly, data mining approaches are used to evaluate past stock prices and acquire useful knowledge through the calculation of financial indicators. Transformed data are then classified using decision trees obtained through the application of Artificial Intelligence strategies. Finally, the different decision trees are analyzed and evaluated, showing accuracy rates and emphasizing total profit associated to capital gains.

Keywords: Data Mining, Decision Trees, Artificial Intelligence, Stock Market Analysis.

1 The Problem

The problem considered in this paper deals with the application of decision making techniques to stock market analysis [1]. The core of the problem is the generation of automatic buying and selling orders in the erratic stock market. The decision making system can be conceived from a data mining point of view [2]. In this sense, the system's design has three main phases as indicated in Figure 1: the data processing phase, the data mining phase and the evaluation phase. The main elements of the system are the original data base, the Object Attribute Table of transformed or processed data, the patterns or models of classified data and the knowledge extracted from the data.

The data processing phase selects from the raw data base a data set that focuses on a subset of attributes or variables on which knowledge discovery has to be performed. It also removes outliers and redundant information, and uses financial indicators to represent the processed data by means of an Object Attribute Table (OAT). The data mining phase converts the data contained in the OAT into useful patterns, in

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part I, LNAI 6096, pp. 185–194, 2010.

particular decision trees are found [3]. The evaluation phase proves the consistency of the pattern by means of a testing set. The positively evaluated decision system can then be used in real world situations that will allow for its validation.

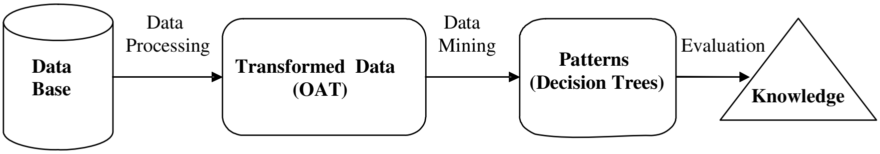

Fig. 1. The Decision Making System

The original data base is formed by financial information available in different web sites. In particular, data from the American stock market have been used. Out of all the variables originally considered the target data is formed by: date (D), opening price (O), closing price (C), daily high (H), daily low (L) and daily volume (V).

In order to obtain a decision making system based on classification techniques a particular company must be selected. This company must satisfy some restrictions: it must not be a technological security, in order to avoid using data from year 2000 economic bubble; it must not be a financial institution, due to the erratic behaviour of the past years; it must have a significant influence within the American stock market; it must be a reference company. These restrictions are appropriate to initiate the process however they might be reconsidered at a later point.

Due to these restrictions and after a careful study of the stock market, Alcoa was selected to complete the project. Alcoa is the world's leading producer and manager of primary aluminum, fabricated aluminum and alumina facilities, and is active in all major aspects of the industry. The targeted data includes daily information from the beginning of 1995 till the end of 2008.

2 The Object Attribute Table

The Data Processing phase transforms the targeted data into useful data written in terms of financial indicators [4]. The financial indicators considered are calculated using graphical methods (where price oscillation is studied) and mathematical methods (where mathematical equations are applied).

2.1 Financial Indicators

The *Simple Moving Average (SMA)* is the unweighted mean of the previous n data points. It is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends and is often used for the analysis of financial data such as stock prices (P), returns or trading volumes. There are various popular values for n, like 10 days (short term), 30 days (intermediate term), 70 days (long term) or 200 days depending on the number of days considered.

$$SMA_{n+1}(P) = \frac{P_n + P_{n-1} + \cdots + P_1}{n} \cdot \tag{1}$$

The moving averages are interpreted as support in a rising market or resistance in a falling market. The *SMA* treats all data points equally and can be disproportionally influenced by old data points. The exponential moving average addresses this point by giving extra weight to more recent data points.

The *Exponential Moving Average (EMA)* applies weighting factors which decreases exponentially, giving much more importance to recent observations while not discarding older observations totally. The degree of weighting decrease is expressed as a constant smoothing factor α, which can be expressed in terms of n

$$\alpha = \frac{2}{n+1} \cdot \tag{2}$$

And the exponential moving average is expressed as:

$$EMA_{n+1}(P) = \alpha P_n + (1-\alpha) EMA_n(P) \cdot \tag{3}$$

The moving averages are adequate indicators to determine tendencies. When a short term moving average crosses over a longer term moving average we have a rising tendency. Whereas when a short term moving average crosses under a longer term moving average we have a falling tendency. The moving averages have a drawback since the information provided has a time lag of several days. This can be avoided by using more powerful indicators.

The *Moving Average Convergence Divergence (MACD)* is a combination of two exponential moving averages with different number of data points (days). The *MACD* is calculated by subtracting the y-day *EMA* from the x-day *EMA*. A z-day *EMA* of the *MACD*, called the "signal line", is plotted on top of the *MACD*, functioning as a trigger for buy and sell signals. Variables x, y, and z are parameters of the *MACD*, usually values x = 12, y = 26 and z = 9 are considered.

The *MACD* is a trend-following momentum indicator that shows the relationship between two moving averages of prices. Its two-folded interpretation is as follows. If the *MACD* line and signal line are considered, when the *MACD* falls below the signal line, it may be time to sell. Conversely, when the *MACD* rises above the signal line, the price of the asset is likely to experience upward momentum, it may be time to buy.

The *stochastic (K)* is an indicator that finds the range between an asset's high (H) and low price (L) during a given period of time, typically 5 sessions (1 week) or 20 sessions (1 month). The current securities price at closing (C) is then expressed as a percentage of this range with 0% indicating the bottom of the range and 100% indicating the upper limits of the range over the time period covered.

$$K = 100\,\frac{C-L}{H-L} \cdot \tag{4}$$

An interesting interpretation for the stochastic can be obtained if the moving average associated with the stochastic *SMA(K)* is considered: the crossing points between *K* and *SMA(K)* trigger buy and sell signals. When *K* crosses over *SMA(K)* it may be time to buy, conversely when *K* crosses under *SMA(K)* selling might be convenient.

The *Bollinger Bands (BB)* consist in a set of three curves drawn in relation to security prices. The middle band M_B is a measure of the intermediate-term trend, usually a moving average (simple or exponential) over n periods. This middle band serves as the base to construct the upper U_B and lower L_B bands. The interval between the upper and lower bands and the middle band is determined by volatility, typically the interval is Q times the standard deviation of the same data that were used to calculate the moving average. The default parameters most commonly used are n = 20 periods and Q = 2 standard deviations.

95% of security prices can be found within the Bollinger Bands, the band represents areas of support and resistance when the market shows no tendencies. When the bands lie close together a period of low volatility in stock price is indicated. When they are far apart a period of high volatility in price is indicated.

Generally Bollinger Bands are used together with other indicators to reinforce its validity. However, on its own they can trigger buy and sell signals. When prices touch the lower band it might be convenient to buy, conversely when prices touch the upper band selling might be convenient.

2.2 Processing of Financial Indicators: Transformed Data

In order to construct the OAT of adequate data, three steps must be followed:

- Processing of the targeted data in order to generate financial indicators.
- Manipulation and discretization of financial indicators into binary attributes.
- Construction of the OAT [5, 6].

The financial indicators are calculated using R, a language for statistical computing and graphics [7]. R is applied mainly as a statistical package but it can be used in other areas of science such as numerical analysis, signal processing, computer graphics and so on because of the number of growing implemented libraries in R to these scientific areas. In particular, the financial indicators are calculated using R's TTR library. R is a popular programming language used by a growing number of data analysts inside corporations and academia. See [8] for details on the use of R in academia and other institutions. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSDband Linux, Windows and MacOS).

The indicators considered are the following: *MACD*, *EMA(C)*, *EMA(V)*, *K* and *BB*. They are calculated directly from the 3514 rows that form the daily targeted data. The resulting values are added as new columns in the targeted data table. The added columns are represented by means of eight binary attributes A1, A2, A3, A4, A5, A6, A7 and A8 in the following manner.

- The *MACD* indicator identifies buying (B) or selling (S) orders depending on its crossing with an intermediate term moving average *EMA(z)* called "signal line". If *MACD* crosses under the "signal line" a buying signal (B) is generated. On the other hand, if *MACD* crosses over the "signal line" a selling signal (S) is generated. If there is no crossing no signal is generated. These results are represented using 2 binary attributes, A1 and A2, as follows: when crossing occurs, if *MACD* > *EMA(z)* then A1 = 1; else if *MACD* < *EMA(z)* then A2 = 1. If there is no crossing (no signal is generated) then A1 = 0 and A2 = 0.

- The *EMA(C)* indicator represents the closing price's exponential moving average. When compared to the stock's actual price (*P*) two zones related with the strength-weakness of the asset are generated. These zones can be represented by a binary attribute A3 as follows: if $P > EMA(C)$ (strength) then A3 = 1, else if $P \leq EMA(C)$ (weakness) then A3 = 0.
- The *EMA(V)* indicator represents the volume's exponential moving average. When compared to the actual volume (*V*) two zones related with the trustworthiness of the asset are generated. These zones can be represented by a binary attribute A4 as follows: if $V > EMA(V)$ (trustworthy) then A4 = 1, else if $V \leq EMA(V)$ (non trustworthy) then A4 = 0.
- The *K* indicator also identifies buying (B) or selling (S) orders depending on the crossing of moving averages, in particular of *K* and *SMA(K)*. If *K* crosses under *SMA(K)* a selling signal (S) is generated. On the other hand, if *K* crosses over the *SMA(K)* a buying signal (B) is generated. If there is no crossing no signal is generated. These results are represented using 2 binary attributes, A5 and A6, as follows: when crossing occurs, if $K < SMA(K)$ then A5 = 1; else if $K > SMA(K)$ then A6 = 1. If there is no crossing (no signal is generated) A5 = 0 and A6 = 0.
- The *BB* indicator generates 4 zones which represent areas of support and resistance. When compared with the asset's actual price *P*, these 4 zones can be represented by two binary attributes A7 and A8 as follows: if $P < L_B$ (weak asset), then A7 = 0 and A8 = 0; if $L_B \leq P < M_B$, then A7 = 0 and A8 = 1; if $M_B \leq P < U_B$, then A7 = 1 and A8 = 0; if $U_B \leq P$ (strong asset), then A7 = 1 and A8 = 1.

The resulting table must be filtered. First of all, the original targeted data can be eliminated since they are only used to calculated financial indicators, which constitute the relevant data. Secondly, repeated and irrelevant rows are eliminated. The resulting table, the final Object Attribute Table, is formed by 8 columns of binary attributes, A1, A2, …, A8 and 61 rows as illustrated in Table 1.

Table 1. The Object Attribute Table with 3 classes

A1	A2	A3	A4	A5	A6	A7	A8	class
0	0	0	0	1	0	1	0	S
0	1	0	0	0	0	0	1	B
0	0	1	0	0	1	0	1	N
0	0	0	1	0	1	1	0	N
1	0	1	0	0	0	0	1	S
1	0	1	1	0	0	1	0	S
…	…	…	…	…	…	…	…	…

The next step is to assign the class to each of the remaining rows taking into account the values of the attributes as indicated by the financial expert. Three different classes are used depending on whether buying (B), selling (S) or no action (N) must be taken, as shown in the last column of Table 1.

In order to better evaluate the results of this paper, a second OAT with only two classes, buying (B) and selling (S) has also been considered. In this second OAT the no action class (N) has been eliminated and replaced by B and S classes according to the expert's criteria. Table 2 shows the 2-class OAT.

Table 2. The Object Attribute Table with 2 classes

A1	A2	A3	A4	A5	A6	A7	A8	class
0	0	0	0	1	0	1	0	S
0	1	0	0	0	0	0	1	B
0	0	1	0	0	1	0	1	S
0	0	0	1	0	1	1	0	B
1	0	1	0	0	0	0	1	S
1	0	1	1	0	0	1	0	S
…	…	…	…	…	…	…	…	…

The similarities between Table 1 and Table 2 are evident, since the only changes are the replacement of the N class in Table 1 by a B or S class in Table 2. The results obtained using the 3-class OAT is consistent with a more conservative behavior, whereas the 2-class OAT offers a risky strategy.

3 The Decision Trees

In the data mining phase, the data contained in the OAT is converted into useful patterns. In particular, a decision tree will be obtained. The decision tree is one of the most popular classification algorithms. A decision tree can be viewed as a graphical representation of a reasoning process and represents a sequence of decisions, interrelationships between alternatives and possible outcomes. There are different trees available depending on the evaluation functions considered. In this project, the ID3 algorithm is used due to its simplicity and speed [9]. ID3 is based on information theory to determine the most informative attribute at each step of the process. In this sense, it uses information gain to measure how well a given attribute separates the training examples into the classes. The ID3 algorithm is applied to the OATs represented in both Table 1 and Table 2.

The decision tree for 3 classes, see Figure 2, has a total of 37 nodes, 18 corresponding to the 8 attributes used and 19 leaf nodes or branches. The average branch length is 6. Attributes A2, A1 and A5 appear at the root of the tree suggesting the importance of indicators *MACD* and *K* in the classifying process. This is in accordance with the experts that consider *MACD* and *K* as key indicators in the buying-selling decisions, whereas *EMA(C)*, *EMA(V)* and *BB* are used as reinforcement indicators.

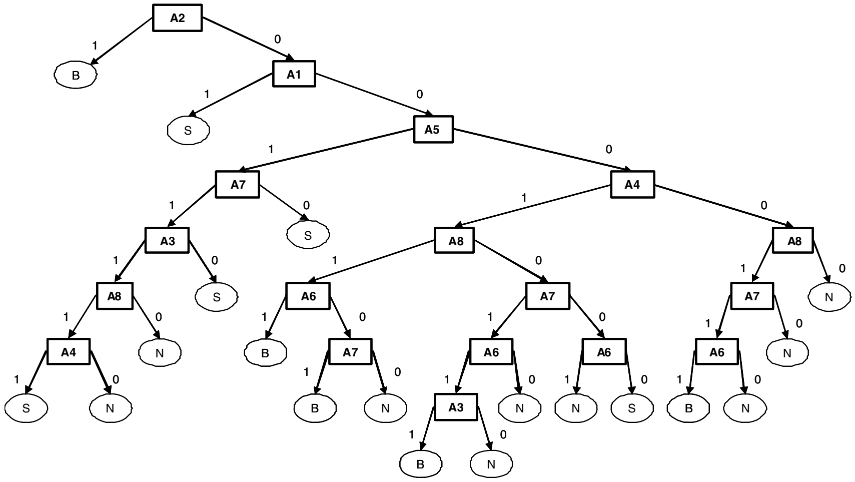

Fig. 2. The Decision Tree for 3 classes

Similarly, the 2-class decision tree, see Figure 3, has a total of 51 nodes, 25 corresponding to the inner nodes and 26 leaf nodes or branches. The average branch length is also 6. Note that when data is classified in 2 classes, the number of borderline data has increased and a larger number of attributes is needed to perform the classification. Attribute A2 appears at the root of the tree suggesting the importance of indicator *MACD* in the classifying process. Attribute A6 also appears at upper levels of the tree indicating the importance of *K* as key indicator in the buying-selling decisions.

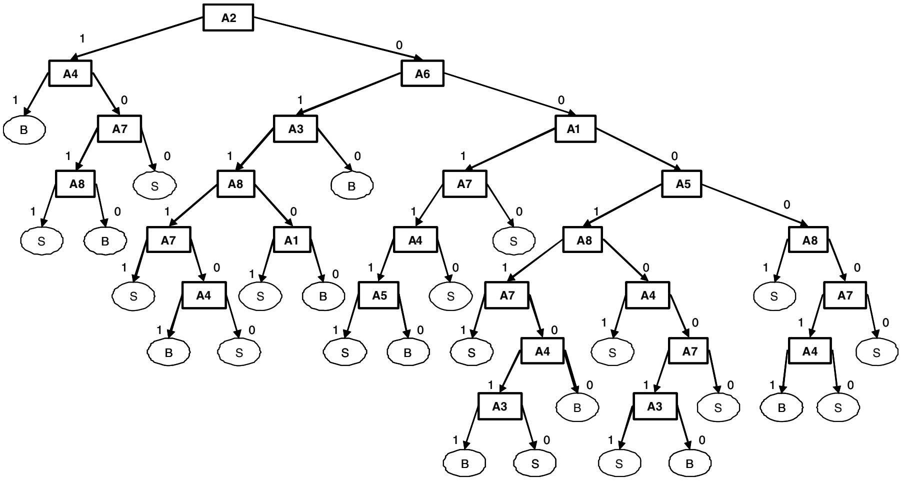

Fig. 3. The Decision Tree for 2 classes

4 Validation of the Decision Trees

The evaluation phase proves the consistency of the ID3 decision tree by means of a testing set. The testing set is formed by the stock market prices of a different asset, of similar characteristics to the one used to train the tree. This raw data are transformed into a testing OAT using the previous indicators and attributes. Now the decision tree is applied to the testing OAT and all instances are classified. The classes obtained with the tree are compared with the classes corresponding to the instances of the testing OAT as suggested by the expert.

The evaluation of the method provides the following results: a) for the 3-class decision tree, Instances = 289; Variables = 8; Correctly classified instances = 113; Incorrectly classified instances = 176; b) for the 2-class tree, Instances = 613; Variables = 8; Correctly classified instances = 300; Incorrectly classified instances = 313. These results are shown on Table 3.

Note that the number of classified instances varied due to the elimination of the N class in the 2-class tree.

Table 3. Evaluacion of the testing set

Parameters	3-class tree	2-class tree
Correct decision	113	300
Incorrect decision	176	313
Correct decision (%)	39.10	48.93
Incorrect decision (%)	60.90	51.06

These results might seem poor. However, stock market analysis is an unpredictable field and even though accuracy is desired, profit associated to capital gain is crucial. Therefore, classification accuracy is not the only parameter that must be evaluated.

The decision tree's behavior respect to profit will be evaluated by an automatic system. This system evaluates the testing set taking into account the following:

- It considers 30-day indicators;
- The system buys only when so indicated by the decision tree;
- The bought stock package is kept until the tree generates a selling order;
- All buying orders generated between the first buying order and the selling order are ignored;
- Once the selling order is generated by the tree, the package will be sold and all selling orders generated before the next buying order will be ignored.

Figure 4 illustrates the process carried out by the automatic system. Point 1 indicates a buying order; Point 2, also a buying order, will be ignored; Point 3 indicates a selling order; Point 4 corresponds to a buying order; Points 5 and 6 (buying orders) will be ignored; Point 7, indicates a selling order and Point 8 (selling order) will be ignored.

The system's performance can be evaluated taking into account the following parameters:

- Total win (%) if correct decision is made;
- Total loss (%) if incorrect decision is made;

- Average win (%) if correct decision is made;
- Average loss (%) if incorrect decision is made;
- Performance (%);
- *Win/Loss* ratio

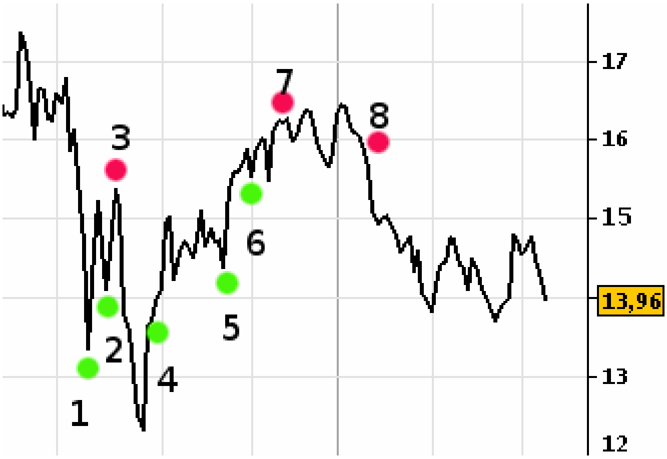

Fig. 4. Automatic System's Behaviour

The method provides the results shown on Table 4, the total yield for the 3-class decision tree is of 118.15%, whereas for the 2-class decision tree the yield is 87.89% which are acceptable results.

Table 4. Performance of the Automatic System

Parameters	3-class tree	2-class tree
Win when correct decision (%)	579.71	610.00
Loss when incorrect decision (%)	-461.56	-522.12
Average win when correct decision (%)	5.13	2.03
Average loss when incorrect decision (%)	-2.62	-1.67
Performance (%)	118.15	87.89
W/L ratio	1.26	1.17

5 Conclusions and Future Work

The stock market is an unpredictable, erratic and changeable domain. Nevertheless, tools that predict the stock market's behaviour exist. This paper considers a method based on financial indicators to generate decision trees that classify buying-selling orders. A decision tree in Artificial Intelligence is equivalent to an automatic investment system in the stock market analysis. The procedure developed here is long and costly, but once the tree is generated, it's automatic and applicable to any financial asset.

The initial raw data are daily stock market values of Alcoa, these data are transformed into an OAT and used as a training set to construct decision trees by means of the ID3 algorithm. Decision trees with 3 and 2 classes have been generated. These decision trees have been tested and evaluated. The results for the 3-class tree show a 39% accuracy percentage and a 118% profit gain, whereas for the 2-class tree the results show a 49% accuracy and a 88% profit gain. These results are considered satisfactory. If both trees are compared, the 3-class tree has a lower percentage accuracy but the profit gain is higher.

This paper offers a limited vision of one of the many solutions available. The following aspects can be considered as future work. Other different financial indicators may offer better results and should be studied, also other decision trees can be obtained using different learning algorithms and should be evaluated. The stock market is a dynamic world, and the decision tree could present time-related disturbances. Therefore, an automatic system using a threshold should be considered in order to avoid strong losing trends.

Acknowledgments. The authors want to thank Damià Vaquer, Gerard Draper and Jaume Jaume for their helpful comments and suggestions.

References

1. Fiol-Roig, G., Miró-Julià, M.: Applying Data Mining Techniques to Stock Market Analysis. Accepted for publication in the 8th International Conference on Practical Applications of Agents and Multi-Agent Systems. Trends and Strategies on Agents and Multiagent Systems (2010)
2. Fayyed, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, AI Magazine Fall 96, 37–54 (1996)
3. Fiol-Roig, G.: UIB-IK: A Computer System for Decision Trees Induction. In: Raś, Z.W., Skowron, A. (eds.) ISMIS 1999. LNCS, vol. 1609, pp. 601–611. Springer, Heidelberg (1999)
4. Weinstein, S.: Stan's Weinstein's Secrets For Profiting in Bull and Bear Markets. McGraw-Hill, New York (1988)
5. Miró-Julià, M.: Knowledge discovery in databases using multivalued array algebra. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) EUROCAST 2009. LNCS, vol. 5717, pp. 17–24. Springer, Heidelberg (2009)
6. Fiol-Roig, G.: Learning from Incompletely Specified Object Attribute Tables with Continuous Attributes. Frontiers in Artificial Intelligence and Applications 113, 145–152 (2004)
7. The R project, `http://www.r-project.org/`
8. `http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html`
9. Quinlan, J.R.: Induction of decision trees. Machine Learning 1, 81–106 (1986)