Towards Stock Market Data Mining using Enriched Random Forests from Textual Resources and Technical Indicators

Manolis Maragoudakis[1,2] and Dimitrios Serpanos[2,3]

[1] Department of Information and Communication Systems Engineering,
University of Aegean,Samos, 82000, Greece
[2] I.S.I. - Industrial Systems Institute Patras
Science Park building Platani, PATRAS, Greece, 26504
[3]Department of Electrical and Computer Engineering,
University of Patras, Rion, 26500, Greece
{mmarag@aegean.gr,serpanos@ece.upatras.gr}

Abstract. The present paper deals with a special Random Forest Data Mining technique, designed to alleviate the significant issue of high dimensionality in volatile and complex domains, such as stock market prediction. Since it has been widely acceptable that media affect the behavior of investors, information from both technical analysis as well as textual data from various on-line financial news resources are considered. Different experiments are carried out to evaluate different aspects of the problem, returning satisfactory results. The results show that the trading strategies guided by the proposed data mining approach generate higher profits than the buy-and-hold strategy, as well as those guided by the level-estimation based forecasts of standard linear regression models and other machine learning classifiers such as Support Vector Machines, ordinary Random Forests and Neural Networks.

Keywords: Stock return forecasting; Data mining; Expert systems; Random forests; Markov blanket; Trading strategies.

1 Introduction

Stock market prediction has always gained certain attention from researchers. There is a controversy as regards to whether there is a method for accurate prediction of stock market movement, mainly due to the fact that modeling market dynamics is a complex and volatile domain. Stock market research encapsulates two main philosophical attitudes, i.e. fundamental and technical approaches [1]. The former states that stock market movement of prices derives from a security's relative data. Fundamentalists are of the belief that numeric information such as earnings, ratios, and management effectiveness could determine future forecasts. In technical analysis, it is believed that market timing is the key. Technicians utilize charts and modeling techniques to identify trends in price and volume. These latter individuals rely on historical data in order to predict future outcomes. However, according to several researchers, the goal is not to question the predictability of financial time series but to discover a good model that is capable of describing the dynamics of stock market.

There is a plethora of proposed methods in stock market prediction. The majority of them are strongly related to structured, numerical databases and domain expertise

rules. In the field of trading, most of decision support tools focus on statistical analysis of past price records. Nevertheless, throughout recent studies, prediction is also based on textual data, based on the rational assumption that the course of a stock price can be influenced by news articles, ranging from companies releases and local politics to news of superpower economy [2].

However, unrestricted access to news information was not possible until the early 1990's. Nowadays, news are easily accessible, access to important data such as inside company information is relatively cheap and estimations emerge from a vast pool of economists, statisticians, journalists, etc., through the World Wide Web. Despite the large amount of data, advances in Natural Language Processing and Knowledge Discovery from Data (also known as Data Mining) allow for effective computerized representation of unstructured document collections, analysis for pattern extraction and discovery of relationships between document terms and time-stamped data streams of stock market quotes.

Nevertheless, when data tend to grow both in number of records and features, numerous mining algorithms face significant complications, resulting in poor prediction ability. The aim of this study is to propose a potential solution to the problem, by considering the well-known algorithm of Random Forests [3] and altering their construction phase by utilizing a Markov Blanket approach which discards irrelevant features, thus improving classification results. The importance of this study lies to the fact that technical analysis contains the event and not the cause of the change, while textual data may interpret that cause. Certainly, as it is tedious for a human investor to read all daily news concerning a company and other financial information, a prediction system that could analyze such textual resources and find relationships with price movement at future time windows is beneficial.

The paper is structured as follows: section 2 provides an overview of literature concerning Stock Market prediction using Data Mining techniques. Section 3 describes the proposed Markov Blanket Random Forest utilization. Section 4 provides an overview of our experimental design and discusses the evaluation outcome.

2 Previous Work

Due to numerous studies in traditional technical analysis, we shall emphasize on researches that study the influence of news articles on stock markets. Chang et al., [4] were among the first to confirm the reaction of the market to news article. They had shown that economic news always has a positive or negative effect in the number of traded stock. They used salient political and economic news as proxy for public information. Klibannof et al., [5] deal with closed-end country fund's prices and country specific salient news. They stated that there is a positive relationship between trading volume and news. Similar to the aforementioned approach, Chan and Wei [6] founded that news that is placed in the front page of the *South China Post* increase the return volatility in the Hong Kong stock market. Mitchell and Mulherin [7] used the daily number of headlines of Dow Jones as a measure of public information. They mentioned the positive impact of news on absolute price changes. Mittermayer [8] proposed a prediction system called NEWSCATS, which provides an estimate of the price after the publication of press releases. Schumaker and Chen [9] examined three

different textual representation formalisms and studied their abilities to predict discrete stock prices 20 minutes after an article release.

3 Markov Blanket Random Forests

A problem arises when the number of possible features is vast and the percentage of actually informative features is small, i.e. the performance of the base classifiers degrades. This phenomenon is particularly present in financial data sets, where most attributes represent technical indicators with little or unknown certainty about their correlation to the true course of a stock. Technically, in the case of a Random Forest classifier, this problem arises due to the fact that, if simple random sampling is used for selecting the subset of m eligible features at each node, almost all these subsets are likely to contain a predominance of non-informative features.

The solution proposed in this paper is based on the notion of a feature selection and reasoning algorithm, i.e. the Markov Blanket of the class attribute. The identification of relevant variables is an essential component of construction of decision support models, and computer-assisted discovery. In financial decision systems for example, such as the task at hand, elimination of redundant features could increase the computational performance significantly. The problem of variable selection in financial domains is more pressing than ever, due to the recent emergence of many news portals, on-line financial services, etc. Similar cases are also common in biomedical engineering, computational biology, text categorization, information retrieval, mining of electronic medical records, consumer profile analysis, temporal modelling, and other domains [10]. Several researchers [11] have suggested, intuitively, that the Markov Blanket (MB) of the target variable t, denoted as $MB(t)$, is a key concept for solving the variable selection problem. $MB(t)$ is defined as the set of variables conditioned on which all other variables are probabilistically independent of t. Thus, knowledge of the values of the Markov Blanket variables should render all other variables superfluous for classifying t.

3.1 Bayesian Networks and Markov Blanket

In order to better capture the significant properties of a Markov Blanket, a brief introductory section of Bayesian networks is included. Bayesian networks graphically represent the joint probability distribution of a set of random variables. A Bayesian network is composed of a qualitative portion (its structure) and a quantitative portion (its conditional probabilities). The structure BS is a directed acyclic graph where the nodes correspond to domain variables $x_1, \ldots, x_n$ and the arcs between nodes represent direct dependencies between the variables. Likewise, the absence of an arc between two nodes x_i and x_j represents that x_j is independent of x_i given its parents in BS. Following the notation of Cooper and Herskovits [12], the set of parents of a node x_i in BS is denoted π_i. The structure is annotated with a set of conditional probabilities (BP), containing a term $P(x_i=X_i|\pi_i=\Pi_i)$ for each possible value X_i of feature x_i and each possible instantiation Π_i of π_i. A Markov Blanket of a node x_i, denoted as $MB(x_i)$, is a minimal attribute set, such that every other attribute is independent of x_i given its Markov Blanket. Mathematically, the above statement is translated into:

$$\forall x_i \in \{x_1,...,x_n\} \setminus MB(x_i) \cup \{x_i\}, x_i \vdots x_k \mid MB(x_i), \quad (1)$$

where $\vdots$ denotes the conditional independence of x_i with x_k given $MB(x_i)$.

Suppose B_i and B_j are two Bayesian networks that have the same probability distribution, then $MB_{Bi}(x_k) = MB_{Bj}(x_k)$ for any variable x_k. Certainly, MBs are not exclusive and may vary in size, but any given BN has a unique $MB(x_i)$ for any x_i, which is the set of parents, children and parents of children of x_i. In Fig. 1, a Bayesian network is depicted along with the Markov Blanket of a target node x, colored in blue. As regards to the dataset interpretation, feature x is independent of all other features given its $MB(x)=\{U_i,U_j,Y_k,Y_l,Z_{km},Z_{ln}\}$.

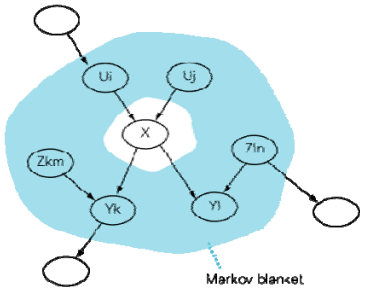

Fig. 1. An example of a Bayesian Network with the Markov Blanket of node x.

3.2 Random forests

Random Forests, in general, are a combination of decision tree classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Given a training set X comprised of N instances, which belong to two classes, and F features, a Random Forest multi-way classifier $\Theta(x)$ consists of a number of decision trees, with each tree grown using some form of randomization, where x is an input instance. The leaf nodes of each tree are labelled by estimates of the posterior distribution over the data class labels [13]. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions. The process is described in Fig. 2. Each tree is grown as follows:

- If the number of cases in the training set is N, sample N cases at random but with replacement, from the original data. This sample will be the training set for growing the tree.
- If there are F input features, a number $m<<F$ is specified such that at each node, m variables are selected at random out of the F and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible. Therefore, no pruning procedures are applied.

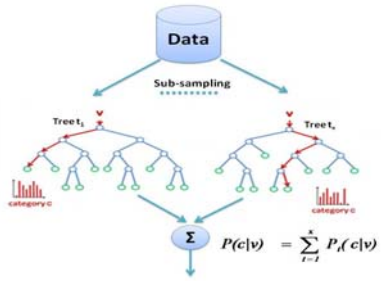

Fig. 2. *Hierarchical decomposition of a Random Forests classifier on a data set*

3.3 Markov Blanket Random Forests Implementation

Based on the existing implementations of Random Forests and taking our initial concerns on feature relevance into consideration, we propose a novel algorithm for classification using RF. The algorithm is entitled "*Markov Blanket Random Forests-MBRF*", since the danger of selecting irrelevant and misleading features is remedied by using the Markov Blanket of the class node to provide the best splitting criteria for each tree. By selecting random samples and obtaining the extracted *MB* of the target node, the probability of tree containing more informative features is increased. In case of high-dimensional datasets, the diversity of the ensemble is not compromised and is more robust that other, pre-filtering or weighting schemes. The algorithm is consisted of two distinct phases; the former regards the construction of the Markov Blanket and the latter deals with constructing the trees. Its basic procedure can be sketched in the following phases:

MBRF (Data D, Features F, n_{tree} trees, Target C)

1. Draw n_{tree} bootstrap samples from the original data D.
2. Build an unconstrained Bayesian network without learning the conditional probability table.
3. Obtain the MB of the class node C.
4. For each of the bootstrap samples, grow an un-pruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, use m_{try} of the Markov Blanket and choose the best split from among those variables.
5. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

4 Experimental Design and Evaluation

As mentioned earlier, articles containing financial news were combined with a plethora of technical indices in order to search for direct influence patterns of the former to the latter. More specifically, we focused on three heterogeneous stock securities from the Greek stock market (*Athens Stock Exchange, .ATG*), a major Greek bank (*Piraeus Bank, .TPEIR*), the main telecommunication provider of Greece (*OTE, .OTE*) and one of the biggest Greek airline companies (*Aegean, .AEGN*). We incorporated past data from the major European, Asian and American stock markets,

as well as data from energy and metal commodities. Finally, for each of the aforementioned three stock securities, a variety of major technical indices was utilized. News was automatically extracted from the electronic versions of the leading Greek financial newspapers, i.e. "*Naftemporiki*" (www.naftemporiki.gr) and "*Capital*" (www.capital.gr). The time period for all collected data was from November 2007 to January 2010. The technical indices were calculated using the *AnalyzerXL* tool. Table 1 tabulates data regarding the three benchmark stocks and their corresponding articles that were collected, while Table 2 contains data about historical data of other, main markets and commodities. Finally, Fig. 3 depicts a categorized list of the technical indices that were also taken into consideration.

Table 1. The benchmark tickers.

Name	Category	#Articles	#Days	Symbol
O.T.E.	Telephony	1779	519	.OTE
Bank of Piraeus	Bank	1607	514	.TPEIR
Aegean Airlines	Airline	308	547	.AEGN

Table 2. Market and commodities data.

Category	Description	#Days	Symbol
European Markets	^FCHI-CAC 40 Index	542	.FCHI
	^FTSE-FTSE 100 Index	569	.FTSE
	^GDAXI-Xetra Dax Index	537	.GDAXI
	^ATG-Athens Stock Exchange	511	.ATG
Asia/Pasific Markets	^HIS-Hang Seng Index	556	.HSI
	^AORD-All Ordinaries Index	569	.AORD
	^N225-Nikkei 225 Average Index	556	.N225
United States Markets	^GSPC-S&P 500 Index	558	.GSPC
	^IXIC-Nasdaq Composite Index	531	.IXIC
	^DJI-Dow Jones Industrial Average Index	539	.DJI
Energy	Brent DTD	566	BRT-
	WTI CUSHING	521	WTC-
Metals	Silver	537	XAG-HH
	Gold Bullion	537	XAU-B-HH

Group Name	Function or Indicator Name
Basic Functions	Median Price (AKA Typical Price Indicator)
Statistical Functions	Standard Deviation
Trend Indicators	MACD Indicator
	Simple Moving Average
	Exponential Moving Average
	Line Weighted Moving Average
Volatility Indicators	Average True Range
	Bollinger Band Width
Momentum Indicators	Williams %R
	TRIX Indicator
	Wilder RSI Indicator
	Chande Momentum Oscillator
	Price Rate-Of-Charge Indicator
	Cutler's Relative Strength Index
	DX (Directional Movement Indicator)
	Stochastic Oscillator
	Price Oscillator Percentage Difference
Market Strength Indicators	Chaikin A/D Oscillator
	Average of Volume ROC
	Market Facilitation Index (MFI)
Support and Resistance Indicators	Envelope

Fig 3. The technical indices considered.

Stock quotes are gathered on a per day basis and articles are aligned accoriding to their release date. In case an article was pubished on a Friday evening (after the closing of the Athens stock market) or during the weekend, it was considered as published on a Monday. The textual analysis phase consisted of three activities: (a) removal of stop words (i.e. articles, special characters, etc), (b) lemmatization of words using a Levenshtein distance based Greek lemmatizer [14], (c) removal of

terms appearing less than 30 times within the complete article corpus and taking the 150 most frequent of them. Upon completion of the aforementioned phases, we kindly asked a domain expert (financial journalist) to annotate terms according to their genre. More specifically, she annotated each word with a signed integer according to whether it encompasses a very positive (+2), positive(1), neutral(0), negative(-1) of very negative(-2) sense. Examples of such terms respectively are: *κερδοφορία* (*profitability*, +2), *ισχυρή* (*powerful*, +1), *πορεία* (*course*, 0), *υποχώρηση* (*downgrading*, -1) and *κρίση* (*crisis*, -2). The predicted class attribute contained three discrete values, namely *UP*, *STEADY* and *DOWN*, if the stock quote closed at a price more than 1%, between 1% and -1% and less than -1% in the following day respectively. A window of 5 days was used in order to predict the class, resulting in a high-dimensional dataset of more than 620 features. Article as well as stock quotes data was processed by our proposed methodology (MBRF), regular Random Forests (RF), Radial Basis Functions neural networks (RBF) and a derivative of Support Vector Machines, namely Sequencial Minimal Optimization (SMO) which can handle discrete values and acts similar to regression. Since the latter Machine Learning algorithms do not reduce features by default, in order to compare the MBRF technique against them, a PCA analysis approach was followed using the Nmath library for .NET platforms (http://www.centerspace.net/products/nmath).

Regarding the experimental design, two different approaches were followed. The former dealt with standard, 10 fold cross validation, classification in terms of stock quotes closeness, using datasets with articles and without articles, in order to evaluate the impact of articles on the predictability of a stock quote. We used the *F-measure* metric for evaluation, which acts as the harmonic mean of precision and recall. Table 3 tabularizes the *F-measure* score of all machine learining algorithms against linear regression (LR). From these outcomes, we could initially observe that combining information from both time series and textual data leads to improvement of the performance for all methodologies. Furthermore, by using only technical analysis data, SMO perform similar to MBRF and significantly outperform all other approaches, while when incorporating textual information, MBRF is noticeably the best classification approach, a fact that could be attributed to the dimensionality reduction when applying the Markov Blanket preprocessing step. According to Table 3, the performance of MBRF is one of the highest ever reported, with the drawback of a very time and recource consuming training phase.

Table 3. Classification performance in terms of F-measure.

Dataset	MBRF	RF	SMO	RBF	LR
No articles	64.23	55.25	63.56	53.21	47.54
Including articles	73.44	60.66	67.76	56.80	47.76

The latter experimental design developed was a simulated trading strategy, in an effort to further examine if the MBRF model could practically be applied to generate higher profits than those earned by employing the traditional regression model of by simply following a buy-and-hold (passive) investment strategy. The operational details of the trading simulation are explained as follows: The trading simulation assumes that the investor has 100,000€ to create a portfolio by selecting a balanced percentage of each of the three Greek stock quotes mentioned earlier. Each day, the investor could buy, sell or wait, according to the class prediction of the MBRF model.

We assume that transactional costs apply when buying or selling (0,335% and 0,35% respectively) and a random choice between 5% and 10% of the current portfolio can be traded each day. The time period was set to the last 35 weekdays of the aforementioned dataset. As Fig. 4 depicts, the dashed line, which represents the portfolio budget for the MBRF investing strategy is clearly outperforming the solid line of the buy-and-hold investment strategy by a mean factor of 12.5% to 26% for the first 2 weeks and from 16% to 48% for the remaining ones.

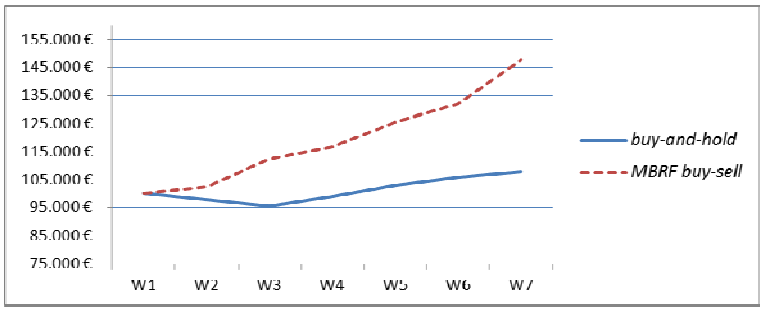

Fig. 4. Plot of portfolio outcomes using the two different trading strategies

References

1. Technical-Analysis. The Trader' s Glossary of Technical Terms and Topics, http://www.traders.com. 2005.
2. Ng, A. and Fu. A.W., 2003. Mining Frequent Episodes for Relating Financial Events and Stock Trends. In Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledgw Discovery and Data Mining, Lectures Notes in Computer Science, vol 2637, pp. 27-39, 2003.
3. Breiman, L., Random forests. Machine Learning Journal, 45:532, 2001.
4. Chung, F., Fu, T. Luk, R. and Ng. V., Evolutionary Time Series Segmentation for Stock Data Mining, In Proceedings of IEEE International Conference on Data Mining, pp. 83-91, 2002.
5. Klibanoff, P, Laymont, O. and Wizman, T.A. Investor reaction to Salient News in Closed-end Country Funds, Journal of Finance, 53(2), pp. 673-699, 1998.
6. Chan, Y. and John-Wei, K.C. Political Risk and Stock Price Volatility: The Case of Hong-Kong. Pacific-Basin Finance Journal, 4(2-3), pp. 259-275, 1996.
7. Mitchell, M.L. and Mulherin, J.H. The Impact of Public Information on the Stock Market, Journal of Finance, 49(3), pp. 923-950.
8. Mittermayer, M.A. Forecasting Intraday Stock Price Trends with Text Mining Techniques. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICS), IEEE Computer Society, 3(3), pp. 30064.2, 2004
9. Shumaker, R.P. and Chen, H. Textual Analysis of Stock Market Prediction Using Financial News Articles, On the 12th American Conference on Information Systems (AMCIS), 2006.
10. Díaz-Uriarte,R. and de Andrés,S.A. (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7,3.
11. Kohavi R. and John G. Wrappers for feature subset selection. In Artificial Intelligence journal, special issue on relevance, Vol. 97, Nos 1-2, pp. 273-324, 1997.
12. Cooper, G.F. & Herskovits, E., 1992: "A Bayesian Method for the Induction of Probabilistic Networks from Data." Machine Learning, Vol. 9, 309-347. Kluwer Academic Publishers, Boston.
13. Strobl,C. et al. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics, 8, 25.
14. Dimitrios P. Lyras, Kyriakos N. Sgarbas, Nikolaos D. Fakotakis, "Using the Levenshtein Edit Distance for Automatic Lemmatization: A Case Study for Modern Greek and English," Tools with Artificial Intelligence, IEEE International Conference on, pp. 428-435, 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2 (ICTAI 2007), 2007.