Expert Systems with Applications 42 (2015) 7046–7056

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Evaluating multiple classifiers for stock price direction prediction

Michel Ballings [a,*], Dirk Van den Poel [b], Nathalie Hespeels [b], Ruben Gryp [b]

[a] The University of Tennessee, Department of Business Analytics and Statistics, 249 Stokely Management Center, 37996 Knoxville, TN, USA
[b] Ghent University, Department of Marketing, Tweekerkenstraat 2, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Available online 14 May 2015

Keywords:
Ensemble methods
Single classifiers
Benchmark
Stock price direction prediction

ABSTRACT

Stock price direction prediction is an important issue in the financial world. Even small improvements in predictive performance can be very profitable. The purpose of this paper is to benchmark ensemble methods (Random Forest, AdaBoost and Kernel Factory) against single classifier models (Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbor). We gathered data from 5767 publicly listed European companies and used the area under the receiver operating characteristic curve (AUC) as a performance measure. Our predictions are one year ahead. The results indicate that Random Forest is the top algorithm followed by Support Vector Machines, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression. This study contributes to literature in that it is, to the best of our knowledge, the first to make such an extensive benchmark. The results clearly suggest that novel studies in the domain of stock price direction prediction should include ensembles in their sets of algorithms. Our extensive literature review evidently indicates that this is currently not the case.

1. Introduction

Predicting stock prices is an important objective in the financial world (Al-Hmouz, Pedrycz, & Balamash, 2015; Barak & Modarres, 2015; Booth, Gerding, McGroarty, 2014), since a reasonably accurate prediction has the possibility to yield high financial benefits and hedge against market risks (Kumar & Thenmozhi, 2006). A great point of discussion in literature is whether stock price behavior is predictable or not. For a long time investors accepted the Efficient Market Hypothesis (EMH) (Malkiel & Fama, 1970). This hypothesis states that no abnormal returns can be obtained by studying the evolution of stock prices' past behavior (Tsinaslanidis & Kugiumtzis, 2014; Yeh & Hsu, 2014). In the beginning of the 21st century however, some economists indicated that future stock prices are at least partially predictable (Malkiel, 2003). Therefore a lot of prediction algorithms have been explored and showed that stock price behavior can indeed be predicted (Huang, Yang, & Chuang, 2008; Ou & Wang, 2009). However predictable, it remains hard to forecast the stock price movement mainly because the financial market is a complex, evolutionary, and non-linear dynamical system which interacts with political events, general economic conditions and traders' expectations (Huang, Nakamori, & Wang, 2005).

Different techniques have already been explored for stock price direction prediction. One of the best performing algorithms in the financial world appears to be Support Vector Machines (SVM) (Huang et al., 2005; Kim, 2003; Lee, 2009). Other well-known techniques are Neural Networks (Kim & Chun, 1998), Decision Trees (Wu, Lin, & Lin, 2006), Logistic Regression (Brownstone, 1996), Discriminant Analysis (Ou & Wang, 2009) and K-Nearest Neighbors (Subha & Nambi, 2012). However, most studies ignore ensemble methods in their benchmarks. To the best of our knowledge Kumar and Thenmozhi (2006), Rodriguez and Rodriguez (2004), Lunga and Marwala (2006) and Patel, Shah, Thakkar, and Kotecha (2015) are the only four studies in the domain of stock price direction prediction that use an ensemble method. This is an important gap in literature because ensemble methods have been proven to be top performers in many other areas such as customer churn behavior (Ballings & Van den Poel, 2012), social media analytics (Ballings & Van den Poel, 2015) and unsupervised word sense disambiguation (WSD) (Brody, Navigli & Lampata, 2006).

In our study we will therefore include several ensemble methods such as Random Forest (RF) (Breiman, 2001), AdaBoost (AB) (Freund & Shapire, 1995) and Kernel Factory (KF) (Balings & Van den Poel, 2013) in our benchmark. While others conduct discrete analyses to predict exact stock prices, we focus on classification models (Leung, Daouk & Chan, 2000). Literature shows that forecasting the direction is enough to execute profitable trading strategies (Cheung, Chinn, & Pascual, 2005; Pesaran & Timmerman, 1995). Hence, we predict the direction of stock prices instead of absolute stock prices. The main contribution of this study is an

* Corresponding author.
E-mail addresses: Michel.Ballings@utk.edu (M. Ballings), Dirk.VandenPoel@UGent.be (D. Van den Poel), Nathalie.Hespeels@UGent.be (N. Hespeels), Ruben.Gryp@UGent.be (R. Gryp).

http://dx.doi.org/10.1016/j.eswa.2015.05.013
0957-4174/

extensive benchmark comparing the performance of ensemble methods (RF, AB and KF) and single classifier models (Neural Networks (NN), Logistic Regression (LR), SVM, K-Nearest Neighbors (KNN)) in predicting the stock price direction. We hypothesize that, given their superiority in other domains, ensemble methods will outperform the single classifier methods.

The remainder of this paper is structured as follows. In Section 2 we will review the literature on which algorithms have been used for stock price direction prediction. Section 3 details our methodology for benchmarking the ensemble methods against other algorithms. Section 4 discusses the results. Section 5 concludes this study and Section 6 describes limitations and avenues for future research.

2. Literature review

The use of prediction algorithms is in contradiction with one of the basic rules in finance, the Efficient Market Hypothesis (EMH) (Malkiel & Fama, 1970). This hypothesis states that if one can get an advantage from analyzing past returns, the entire financial market will notice this advantage and as a consequence the price of the share will be corrected. This means that no abnormal returns can be obtained by examining past prices and returns of stocks. Although the EMH is generally accepted, it was initially based on traditional linear statistical algorithms (Malkiel & Fama, 1970). Many researchers have already rejected the hypothesis by using algorithms that can model more complex dynamics of the financial system (Lo, Mamaysky, & Wang, 2000; Malkiel, 2003). Since methods handling the complex and non-linear financial market are yielding positive results, researchers still try to invent better techniques.

There are three major methodologies to predict the stock price behavior: (1) technical analysis, (2) time series forecasting and (3) machine learning and data mining (Hellström & Holmströmm, 1998). The first category uses charts and plots as a principal tool. Analysts use these plots to make a buy or sell decision. The second category aims at predicting future stock prices by analyzing past returns on stock prices. Common methods are the autoregressive method (AR), the moving average model (MA), the autoregressive-moving average model (ARMA) and the threshold autoregressive model (TAR). The third category, data mining, is "the science of extracting useful information from large data sets or databases" (Hand, Manilla & Smyth, 2001). The popularity of data mining in the financial world has been growing since the main problem with predicting stock price direction is the huge amount of data. The data sets are too big to handle with non data mining methods such that they obscure the underlying meaning and one cannot obtain useful information from it (Fayyad, Shapiro & Smyth, 1996; Widom 1995).

Several algorithms have been used in stock price direction prediction literature. Simpler techniques such as the single decision tree, discriminant analysis, and Naïve Bayes have been replaced by better performing algorithms such as Random Forest, Logistic Regression and Neural Networks. General-purpose solvers such as Genetic Algorithms (Kuo, Chen, & Hwang 2001) have also been used but generally perform worse and are computationally more expensive. The majority of stock price direction prediction literature has focused on Logistic Regression, Neural Networks, K-Nearest Neighbors and Support Vector Machines. Ensemble methods such as Radom Forest, (Stochastic) AdaBoost and Kernel Factory are still very unexplored in the domain of stock price direction prediction.

In Table 1 we provide an overview of those algorithms used for predicting stock price direction in literature (we excluded single Decision Trees, Naïve Bayes, Discriminant Analysis and Genetic Algorithms because they have been superseded by newer and better methods discussed above). LR stands for Logistic Regression, NN stands for Neural Networks, KN stands for K-nearest neighbors, SVM stands for Support Vector Machines, RF stands for Random Forest, AB stands for AdaBoost and KF stands for Kernel Factory. It is clear from Table 1 that our study is the first to include all seven algorithms in one benchmark. This is important if we want to find, globally, the best algorithm. Using suboptimal algorithms may hinder scientific progress in that important patterns in the data might be missed.

In our study we will benchmark ensemble methods against single classifier models. The ensemble methods mentioned above all use a set of individually trained classifiers as base classifiers. We believe that the ensemble methods will outperform the individual classification models because they have proven to be very successful in several other domains such as face recognition (Tan, Chen, Zhou, & Zhang, 2005), gene selection (Diaz-Uriarte & de Andres, 2006), protein structural class prediction (Ballings & Van den Poel, 2015) and credit scoring (Paleologo, Elisseeff, & Antonini, 2010). In stock price direction prediction literature both Support Vector Machines (SVM) and Random Forest (RF) have proven to be top performers (Kumar & Thenmozhi, 2006; Patel et al., 2015). However, there is no consensus on which algorithm is best with SVM outperforming RF in Kumar and Thenmozhi (2006) and vice versa in Patel et al. (2015). AdaBoost has also been shown to perform well, albeit not as well as Random Forest (Rodriguez & Rodriguez 2004). In an effort to help provide clarity in which algorithm is best, this study will benchmark SVM, AB, RF and four other algorithms.

Table 1
Algorithms for stock price direction prediction used in literature.

	Prediction method						
	LR	NN	KN	SVM	AB	RF	KF
Schöneburg (1990)		x					
Bessembinder and Chan (1995)	x						
Brownstone (1996)	x	x					
Saad, Prokhorov, and Wunsch (1996)		x					
Kim and Chun (1998)		x					
Saad, Prokhorov, and Wunsch (1998)		x					
Kim and Han (2000)		x					
Kuo et al. (2001)		x					
Kim (2003)		x		x			
Kim and Lee (2004)		x					
Rodriguez and Rodriguez (2004)	x	x			x	x	
Huang et al. (2005)		x		x			
Kumar and Thenmozhi (2006)	x	x		x		x	
Lunga and Marwala (2006)					x		
Wu et al. (2006)							
Wang and Chan (2007)	x						
Huang et al. (2008)	x	x	x	x			
Senol and Ozturan (2008)	x	x					
Lai, Fan, and Chang (2009)							
Lee (2009)		x		x			
Ou and Wang (2009)	x	x	x	x			
Kara, Boyacioglu and Baykan (2011)		x		x			
Wei and Cheng (2011)		x					
Subha and Nambi (2012)	x		x				
Lin, Guo, and Hu (2013)				x			
De Oliveira, Nobre, and Zárate (2013)		x					
Chen, Chen, Fan, and Huang (2013)		x					
Rechenthin et al. (2013)			x	x	x		
Ji, Che, and Zong (2014)		x					
Bisoi and Dash (2014)		x					
Zikowski (2015)				x			
Hafezi, Shahrabi, and Hadavandi (2015)		x					
Patel et al. (2015)		x		x		x	
This study	x	x	x	x	x	x	x