Expert Systems with Applications 41 (2014) 3651–3661

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Automated trading with performance weighted random forests and seasonality ☆

Ash Booth [a,*], Enrico Gerding [a], Frank McGroarty [b]

[a] Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom
[b] Southampton Management School, University of Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Random forests
Ensemble learning
Stock price prediction
Machine learning
Algorithmic trading

ABSTRACT

Seasonality effects and empirical regularities in financial data have been well documented in the financial economics literature for over seven decades. This paper proposes an expert system that uses novel machine learning techniques to predict the price return over these seasonal events, and then uses these predictions to develop a profitable trading strategy. While simple approaches to trading these regularities can prove profitable, such trading leads to potential large drawdowns (peak-to-trough decline of an investment measured as a percentage between the peak and the trough) in profit. In this paper, we introduce an automated trading system based on performance weighted ensembles of random forests that improves the profitability and stability of trading seasonality events. An analysis of various regression techniques is performed as well as an exploration of the merits of various techniques for expert weighting. The performance of the models is analysed using a large sample of stocks from the DAX. The results show that recency-weighted ensembles of random forests produce superior results in terms of both profitability and prediction accuracy compared with other ensemble techniques. It is also found that using seasonality effects produces superior results than not having them modelled explicitly.

1. Introduction

A particularly active area of research among both academics and industry professionals over the last two decades has been the construction of systems that autonomously trade in securities and currencies. The goal of such endeavours is to produce artificial intelligence methods that can be employed to construct systems that are better than or at least as good as their human counterparts in recognising investment opportunities. Many such systems take as inputs past market prices and attempt to generate a trading signal indicating the direction (and sometimes magnitude) of movement of a given security relative to another.

To this end, we describe an automated trading system that, for the first time, uses performance-weighted ensembles of random forests to predict the price return during well documented seasonality events. Although these seasonal regularities are consistent enough to be used as a stand-alone strategy, rates of return are highly volatile. To address this problem, we propose an expert system that captures the current state of the market in its input and uses this information to both predict the profitability of a seasonality trade and act upon this information. In doing so, the system performs risk management while opening and closing trading positions in an attempt to improve the profit and reduce the volatility over a basic seasonality trading strategy.

Before describing such a system, we first investigate the DAX for the following regularities: upward biases at the turn-of-the-month and over exchange holidays, as well as a downward bias over the weekend. Given our findings, we develop an autonomous system for the trading of stocks over seasonality events. To that end, a recency-biased performance-weighted ensemble of random forests is used to predict the expected profit of a seasonality trade given the prevailing market conditions. The random forests are trained and added to the ensemble over time, in an online fashion, so as to capture various phases of the market.

While others have investigated the performance of ensemble systems and random forest algorithms for predicting using unfiltered daily stock prices, as yet, no study has explored the use of ensembles of random forests for building expert systems for trading empirical regularities in stock markets. In more detail, this paper makes the following contributions to the state of the art:

- We present the first ensemble learning system for trading seasonal trends and demonstrate its effectiveness on equity market data.

☆ This work was supported by an EPSRC Doctoral Training Centre Grant (EP/G03690X/1).

* Corresponding author. Tel.: +44 (0)23 8059 4510.
E-mail addresses: ash.booth@soton.ac.uk (A. Booth), eg@ecs.soton.ac.uk (E. Gerding), f.j.mcgroarty@soton.ac.uk (F. McGroarty).
URL: http://www.ashbooth.com (A. Booth).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.009

- The highly active fields of online ensemble generation and random forest predictors are fused to produce a novel expert system for stock trading formed from performance weighted ensembles of random forests.
- Using experiments based on real data, we show that our recency biased performance-weighted ensembles of random forests, used for trading seasonal events, outperforms other ensemble methods. These include cumulative weighting systems, simple averaging of regressors as well as a number of non-ensemble regression techniques.

The paper is structured as follows. Section 2 gives an overview of the relevant literature. Section 3 presents an analysis of the seasonality and empirical regulations found in the equity data and describes the features that we use as inputs to the learning system while Section 4 describes the trading algorithm itself. In Section 5 the results are summarised while Section 6 gives concluding remarks and discussed potential future work.

2. Literature

Many attempts have been made to devise a consistently profitable autonomous trading system. Inspiration for such trading systems comes from a variety of fields, ranging from fundamental analysis and econometric modelling to evolutionary computation, machine learning and even news mining (Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2013). A considerable number of these approaches use machine learning algorithms to generate trading rules. These are typically based on so-called technical analysis, which uses statistically notable short-term opportunities captured by technical indicators, such as momentum and trending. However, the absence of a solid mathematical foundation for technical analysis has meant that its presence in the academic literature is very limited. Furthermore, during the 1960s, trading rules based on technical indicators were studied and found not to be profitable (Fama & Blume, 1966; Alexander, 1961). It was this work that led notable academics to dismiss technical analysis and support the Efficient Market Hypothesis (Fama, 1970). However, the major problem with the early studies of technical analysis was the ad hoc specifications of the trading rules that suggested the use of data dredging.[1] Despite this, profitable technical trading strategies for inter-day trading in the S&P 500 have been found using genetic algorithms, although the strategies did not fair any better than simple buy and hold strategies (where the asset is bought at the start of the test period and sold at the end) when presented with out of sample data (Allen & Karjalainen, 1999). More recently, there has been a surge in work generating trading strategies by using technical indicators as inputs to machine learning models.

To this end, Artificial Neural Networks (ANNs) were first applied to stock prediction in 1988 when a feed forward network was used to analyse daily stock returns of IBM (White & Diego, 1988). Since then, a plethora of research has attempted to find predictive rules for US (Refenes, Zapranis, & Francis, 1995; Enke & Thawornwong, 2005), Japanese (Kamijo & Tanigawa, 1990) and many other stock prices as well as a multitude of other financial assets (Chen, Leung, & Daouk, 2003; Cheng, Wanger, & Lin, 1996). Although the studies mentioned above indicated that they outperform the relevant benchmarks, many showed that ANNs had some limitations in learning due to the vast noise and complex dimensionality of stock market data.

In addition, Support Vector Machines (SVMs) have been used to forecast stock market movements. The SVM is a non-parametric technique that was developed by Cortes and Vapnik (1995). It has since been used in many studies to predict the daily stock price movements and has been shown to outperform neural networks (Tay & Cao, 2001; Kim, 2003; Huang, Nakamori, & Wang, 2005; Chen & Shih, 2006). A variety of other machine learning approaches including reinforcement learning (O, Lee, & Zhang, 2006), evolutionary bootstrapping (Lebaron, 1998) and principle component analysis (Towers & Burgess, 1999) have been used to forecast financial markets. Although many of these studies indicate that they outperform their benchmarks, we have found that most approaches exhibit large drawdowns[2] in profits, and large and excessive switching behaviour resulting in very high transaction costs. These behaviours have often been attributed to overfitting. As we show in this paper, this problem can be overcome using random forest methods.

2.1. Random forests and financial prediction

Random forests (RFs) are a nonparametric and nonlinear classification and regression algorithm first proposed by Ho (1995) and further developed by Breiman (2001). They have been shown to always converge such that overfitting is not a problem (Breiman, 2001) and, as such, have proved successful in a number of domains including image classification (Bosch, Zisserman, & Munoz, 2007), ecological prediction (Prasad, Iverson, & Liaw, 2006) and microarray data classification (Díaz-Uriarte & Alvarez De Andrés, 2006).

In particular, Creamer and Freund (2004) used this technique to successfully predict the performance of companies and to measure corporate governance risk in Latin American banks. In their study, the performance of random forests was compared to logistic regression and Adaboost, finding that random forests consistently produced superior results. The merits of random forests in financial prediction were also demonstrated by Lariviere and Vandenpoel (2005) who showed that random forest regression could be used for exploring both customer retention and profitability. They analysed a sample of 100,000 customers using data from a large European financial services company finding that random forests techniques provided better prediction results for the validation and test samples compared to linear regression and logistic regression models.

Recently there has been a surge in interest in he use of random forests for stock market prediction. Maragoudakis and Serpanos (2010) used a method called Markov Blanket random forest to make predictions on the direction of stock markets. They report their proposed strategy to outperform a simple buy-and-hold investment strategy by an average of 12.5% to 26% for the initial period and from 16% to 48% for the remainder, as well as outperforming linear regression, SVMs and ANNs. More recently, Qin, Wang, Li, and Ge (2013) used gradient boosted random forests to make predictions on the direction of the Singapore Exchange. Using boosting to weight the individual trees of the forest and a forgetting factor to address market changes, their empirical results showed that their proposed methods were able to generate excess returns compared to a buy-and-hold strategy. Also, Zbikowski and Grzegorzewski (2013) used a novel online-adaptation method to allow random forests to adapt to non-stationary financial time series, while Xu, Li, and Luo (2013) demonstrated the random forest algorithms's ability to select features for trend prediction in stock prices.

[1] Data dredging represents the statistical bias that arises from the misuse of statistics. It implies the deliberate inappropriate use of statistics to uncover misleading relationships in data.

[2] We define drawdowns as the peak-to-trough decline over the duration of a strategy's test period. Drawdowns will be measured as the percentage change from the time a retrenchment begins to when a new high is reached.