Expert Systems With Applications 97 (2018) 60–69

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A novel data-driven stock price trend prediction system[☆]

Jing Zhang[a,*], Shicheng Cui[a], Yan Xu[a], Qianmu Li[a], Tao Li[b]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing 210094, China
[b] School of Computer Science, Florida International University, 11200 SW 8th Street, Miami, FL 33199, USA

ARTICLE INFO

Article history:
Received 8 September 2017
Revised 12 December 2017
Accepted 13 December 2017
Available online 13 December 2017

Keywords:
Feature selection
Morphological pattern recognition
Random forest
Stock price prediction

ABSTRACT

This paper proposes a novel stock price trend prediction system that can predict both stock price movement and its interval of growth (or decline) rate within the predefined prediction durations. It utilizes an unsupervised heuristic algorithm to cut raw transaction data of each stock into multiple clips with the predefined fixed length and classifies them into four main classes (*Up, Down, Flat*, and *Unknown*) according to the shapes of their close prices. The clips in *Up* and *Down* can be further classified into different levels reflecting the extents of their growth (or decline) rates with respect to both close price and relative return rate. The features of clips include their prices and technical indices. The prediction models are trained from these clips by a combination of random forests, imbalance learning and feature selection. Evaluations on the seven-year Shenzhen Growth Enterprise Market (China) transaction data show that the proposed system can make effective predictions, is robust to the market volatility, and outperforms some existing methods in terms of accuracy and return per trade.

1. Introduction

Stock price trend prediction is a classic and interesting topic that has attracted many researchers and participants in multiple disciplines such as economics, financial engineering, statistics, operations research, and machine learning. Although a lot of efforts have been paid during the past several decades (Abarbanell & Bernard, 1992; Adam, Marcet, & Nicolini, 2016; Adebiyi, Adewumi, & Ayo, 2014; Blume, Easley, & O'hara, 1994; Göçken, Özçalıcı, Boru, & Dosdoğru, 2016), accurate forecast of the stock price, even its movements, is still not easy to achieve hitherto, though some advanced machine learning techniques have been utilized. For instance, Kim (2003) used support vector machines to predict the direction of the daily socket price movements in Korea, obtaining a hit rate 56%. Schumaker and Chen (2009) included the text mining technique into socket price forecast, achieving a hit rate 57%. Tsai and Wang (2009) combined the decision tree and neural networks to make prediction to Taiwan stock market. The accuracy of the hybrid model achieves around 70%. However, their test data sets were relatively small, only including dozens of stocks. According to a recent empirical study (Gerlein, McGinnity, Belatreche, & Coleman, 2016), the prediction accuracies of several machine learning models (such as C4.5, K*, logistic model tree, etc.) are in the range of 48%～54%.

Traditional technical analysts have developed many indices and sequential analytical methods that may reflect the trends in the movements of the stock price. However, technical analysis contradicts with the efficient-market hypothesis but they cannot make generalised inferences regarding the accuracy. For example, the efficient-market hypothesis states that as long as the market is weak-form efficient, the price of a stock follows the random walk model (Fama, 1995) and cannot be predicted by analyzing prices from the past. Meanwhile, the prices are affected by many macroeconomical factors, fundamental factors of companies and the involvement of public investors. Therefore, some criticism of technical analysis is that it only considers transactional data of stocks and completely ignores the fundamental factors of companies (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014; Patel, Shah, Thakkar, & Kotecha, 2015) which might be helpful, if the market is in weak-form efficiency.

The fundamental factors of a company cover many aspects such as basic financial status, marketing and development strategies, political events, general economic conditions, commodity price indices, interest rate changes, movements of other stock markets, expectations and psychology of investors, and so on. Comprehensively figuring out the impact of these compound factors on the movement of the stock price is obviously out of the capability of human analysts. Researchers have begun to develop some text-mining based methods that can automatically analyze some

[☆] The source code of the system is available at https://sourceforge.net/p/xuanwu/svn/.

[*] Corresponding author.
E-mail addresses: jzhang@njust.edu.cn (J. Zhang), qianmu@njust.edu.cn (Q. Li), taoli@cs.fiu.edu (T. Li).

https://doi.org/10.1016/j.eswa.2017.12.026
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

of these fundamental factors (Nassirtoussi et al., 2014). For example, Schumaker and Chen (2009) extracted information from the breaking financial news to increase the accuracy of prediction. Bollen, Mao, and Zeng (2011) analyzed the mood of investors from twitter to reveal the sentiments of investors to some stocks. Ruiz, Hristidis, Castillo, Gionis, and Jaimes (2012) analyzed the correlations of financial time series from micro-blogging activities. Si et al. (2013) proposed a technique to leverage topic based sentiments from Twitter to facilitate the prediction of the stock market. Even the up-to-date deep learning techniques have been introduced to conduct event-driven stock market prediction, where events are extracted from news (Ding, Zhang, Liu, & Duan, 2015). However, the automatic fundamental factor analysis may be of some weakness. First, even though the messages or reports are released by the companies, public media or some third-party institutes, it still cannot be guaranteed that there is no misleading information. Second, it is not very clear how strong the correlation is between the released information and the stock price movement. Third, when the market is in semi-strong-form and strong-form efficiencies, the fundamental factor analysis even cannot bring excess returns (Timmermann & Granger, 2004).

Fortunately, in today's big data age, above issues could be bypassed, as a new train of thought, saying "let the data speak for themselves", has been proposed and drawn more attention. Unlike the information obtained from newspapers, micro-blogging and twitter, the everyday transaction data taking place in trade systems are absolutely realistic. The rapid development of machine learning provides a lot of new opportunities to utilize these transaction data to predict the trend of the stock price movement. In fact, applying machine learning to stock prediction has been studied for over thirty years. The early studies in 1990s mainly focused on using Neutral Networks to make prediction (Schöneburg, 1990; Zhang, Patuwo, & Hu, 1998), which partially refuted the validity of efficient market hypothesis (Lawrence, 1997). For example, Tsibouris and Zeidenberg (1995) utilized neural networks to predict stock price only based on past stock prices. The performance of these early methods usually was not good because of the size limitation of the neural networks. To address this issue, some recent studies resort to fusion or combination of models (Hadavandi, Shavandi, & Ghanbari, 2010; Tsai & Hsiao, 2010) and ensemble learning (Ballings, Van den Poel, Hespeels, & Gryp, 2015; Barak, Arjmand, & Ortobelli, 2017; Tsai, Lin, Yen, & Chen, 2011). All above studies have a common weak point that their practical availability is still questionable. In their studies, a small amount of carefully selected and labeled stock data were used to train and test models. Since the data do not cover all stocks and their movements in a stock market, the generalisation capabilities of the models are reduced in the real applications.

A real stock market carries out huge amount of transactions every day. We cannot expect that a real-world computer-aided decision system heavily relies on humans selecting and labeling the data used for model training. Unsupervised pattern recognition becomes more and more important in today's big data age (Wu, Zhu, Wu, & Ding, 2014). If the problem of automatic data preprocessing cannot be solved, the system will hardly to be pushed into a real usage, even if the learning algorithms inside are advanced. In this paper, we propose a novel data-driven stock price trend prediction system *Xuanwu*[1] The contribution of *Xuanwu* is three-fold. (1) it introduces unsupervised pattern recognition methods to generate training samples from raw transaction data without any human intervene; (2) it is a system for a real usage, in which multiple learning models are trained to meet the prediction goals derived from actual user requirements, and its application interface is of the maximum availability that is suitable for any stock and any prediction duration; (3) it provides a simple and easy-to-test framework in which different supervised learning models and feature selection methods can be easily integrated. Experimental results show that the proposed system outperforms some state-of-the-art methods in stock movement prediction even though the models of the compared methods are trained with carefully human-labeled samples.

The remainder of the paper is organized as follows. Section 2 describes the requirements from the aspect of users. Section 3 illustrates the architecture of the proposed system. Section 4 describes our unsupervised method to generate training samples. Section 5 presents our learning method in details. Section 6 addresses experimental results and discussions on these results. Section 7 concludes the paper and points out some future work.

2. Requirements

Xuanwu follows an assumption that the actual investment activities which are carried out based on the prediction results of the system do not have a far-reaching impact on the movements of the stock prices in the future. Therefore, it is specially suitable for the small startup investment companies that collect money from a small population and return the profits in fixed contract periods, whose investment volumes usually do not cause obvious market fluctuation. Otherwise, the prediction will be inaccurate. Moreover, these companies are unlikely to trade a stock very frequently (e.g., daily), nor do they hold a stock for a long time (e.g., more than three months) without make a deal. We outline the key points of user requirements in this section.

2.1. Prediction granularity

Nowadays, stock trades can take place in very high frequency when the market is open. The prediction granularity can be various such as second, minute, hour, day and even a fix investment period. *Xuanwu* chooses *trade day* as the prediction granularity. That is, it predicts the trend of a stock in a predefined period measured by *trade day*. Short-term stock prediction is also interesting (Lin, Yang, & Song, 2009) but not suitable for startup investment companies because of the constraints in the capital volumes and transaction costs. The standard prediction durations of *Xuanwu* (refer to Section 4.1) are 10, 15, 20, 30, 40, 50, and 60 trade days, which spans two weeks to three months.

2.2. Automatic pattern discovery

In the era of big data, continuous growth of generated data requires that the learning models update accordingly within short productive periods (Chen & Zhang, 2014; Sakurai, Matsubara, & Faloutsos, 2015). Obviously, it is no longer possible that training samples are still selected and labeled by humans. *Xuanwu* aims to get through all machine learning processes from generating training samples from the raw transaction data to building the prediction models without any human intervene. All that users need to do is to prepare a copy of original transaction data and then click to start the learning progress. All patterns that we are interesting in are extracted from every stock in the market. Then, the pieces of interested patterns are transformed into training samples.

2.3. Subdivision of classes

Since *Xuanwu* aims to predict the trend of a stock price movement by the end of a predefined period, it defines four main

[1] Xuanwu (Black Tortoise in English) is one of the Four Symbols of the Chinese constellations, usually depicted as a tortoise entwined together with a snake. The creature was thought to have spiritual power to predict the future.