Neurocomputing 142 (2014) 228–238

Contents lists available at ScienceDirect

Neurocomputing

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

Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information

Xiaodong Li [a], Xiaodi Huang [b], Xiaotie Deng [c,e], Shanfeng Zhu [d,e,*]

[a] Department of Computer Science, City University of Hong Kong, Hong Kong
[b] School of Computing and Mathematics, Charles Sturt University, Albury, NSW 2640, Australia
[c] AIMS Lab., Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China
[d] School of Computer Science, Fudan University, Shanghai 200433, China
[e] Shanghai Key Lab. of Intelligent Information Processing, Fudan University, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form
29 January 2014
Accepted 3 April 2014
Communicated by P. Zhang
Available online 6 June 2014

Keywords:
Multiple kernel learning
Stock return prediction
News analysis

ABSTRACT

The interaction between stock price process and market news has been widely analyzed by investors on different markets. Previous works, however, focus either on market news purely as exogenous factors that tend to lead price process or on the analysis of how past stock price process can affect future stock returns. To take a step forward, we quantitatively integrate information from both market news and stock prices in order to improve the accuracy of prediction on stock future price return in an intra-day trading context. In this paper, we present the design and architecture of our approach for market information fusion. By means of multiple kernel learning, the hidden information behind the two sources is effectively extracted, and more importantly, seamlessly integrated rather than simply combined by a single kernel approach. Experiments on comprehensive comparisons between our approach and three baseline methods (which use only one type of information, or naively combine the two sources) have been conducted on the intra-day tick-by-tick data of the Hong Kong Stock Exchange and market news archives of the same period. It has been shown that for both cross-validation and independent testing, our approach is able to achieve the best results.

1. Introduction

Stock market is one of the most important and active parts of financial markets. The most important role of stock market is to determine the prices of stocks. Therefore, a fair and open price determination process is critical to the well functioning of the financial markets. While determining the *fair value* of a stock, investors have to analyze a large amount of information, which can be in the form of specific news about the underlying company or liquidity-specific data, such as recent trading activities. Among all the forms of information, intra-day market news and tick prices are critical to the trading decisions an investor makes. Tetlock [44] and Schumaker and Chen [37] show in their recent papers that market news has certain predictability on the stock price returns. It has also been shown by many market microstructure researchers that stock prices, especially the intra-day tick-by-tick prices, are closely related to the embedding of public and private information and the formation of prices [7].

With the advancement of the algorithmic trading [18,22], the reporting speed and the volume of market data have been increasing significantly.[1] In particular, more data on both news and stock prices make it increasingly difficult to process them manually. Therefore, how to use computer algorithms to process and model market information has become a challenging problem in the practice of both computer science and finance.

Computer science researchers have studied the problem by approaches that can be mainly classified as two categories: classification and regression. The classification approaches cast the problem as two-class or multi-class classification by making directional predictions in the form of a label that tags a market event[2] [12,13,35–38,49,50].

One major issue with the classification approach is *signal strength bias*, which affects how to quantify, interpret and compare

* Corresponding author at: School of Computer Science, Fudan University, Shanghai 200433, China.
E-mail address: zhusf@fudan.edu.cn (S. Zhu).

[1] Bloomberg (http://www.bloomberg.com/) and ThomsonReuters (http://thomsonreuters.com/) publish their real-time price tickers (quotes and trades) and news tickers (titles, bodies and tags) through networks to the world within a few milliseconds.

[2] Market events can be patterns that are extracted from information sources, such as a "trending" or a "reverting" in prices.

http://dx.doi.org/10.1016/j.neucom.2014.04.043
0925-2312/

the strength of the prediction labels. For example, for a specific data set, in one setting we may use $\pm 1\%$ simple return as two thresholds which determine three classes, i.e., *positive*, *neutral*, and *negative*, or in another setting we use $\pm 0.5\%$ and $\pm 1.5\%$ as four thresholds which determine five classes, namely, *extreme positive*, *positive*, *neutral*, *negative* and *extreme negative*. Thus, the following scenarios could happen:

- Assuming that in the first setting, we receive two output prediction labels that are eventually the same (e.g. *positive*), it is difficult to tell whether both of them indicate the same price return, or it predicts $+1\%$ in the first case and $+2\%$ in the second case.
- Assuming that we receive the same class label (e.g. *positive*, again) in both settings, it is hard to determine whether the label in the first setting is stronger than the one in the second setting or vice versa, especially when there are overlaps between the two labeling methods.
- The labeling method itself has largely determined the final accuracy of a system. In an extreme case, if we choose a large threshold, where nearly all of the simple returns do not exceed this threshold (e.g. 10 times return for a short period of time), then every sample would be labeled with *neutral*, and the final accuracy would therefore become 100% regardless of any classification models used.

Differing from the directional predictions by classification approaches, regression approaches make numerical forecastings [5,42,51]. Because regression does not have the signal strength bias, we adopt regression approaches rather than classification approaches. To be specific, we use regression models in our experiments to predict short-term stock price returns in order to overcome the signal strength bias.

Another issue with aforementioned approaches, which is also the main problem we deal with in this paper, is that models using either of the market information sources would lead to *information bias*. To formalize this, we denote the news information set as **N** and the price information set as **P**. Approaches with information bias model the problem as either of

$$\pi : \mathbf{N} \mapsto \mathbf{L} \quad \text{or} \quad f : \mathbf{N} \mapsto \mathbf{R},$$
$$\pi : \mathbf{P} \mapsto \mathbf{L} \quad \text{or} \quad f : \mathbf{P} \mapsto \mathbf{R}, \tag{1}$$

where **L** denotes a set of nominal labels and **R** is a set of estimated numerical values. Fig. 1 illustrates market scenarios that could happen. At time point t_0, the subsequent price movement is affected by both market news and short-term prices. However, if a model uses only partial information, it cannot explain the mapping from the input combinations to the outputs:

- If a model uses only **P**, a rational prediction in the left figure would be "trending down without reverting", and in the right figure it would be "trending up without reverting". Thus, "reverting and trending up" in the left figure and "reverting and trending down" in the right figure would increase the error rate of the model.
- If a model uses only **N**, derived f would not be able to explain why the price is still "trending down" when good news is released, as shown in the left figure, as well as why the price is still "trending up" when bad news is released, as shown in the right figure.

Considering **N** and **P** together, we believe that both information sources play important roles in driving the future trend of the prices.

In summary, to overcome both signal strength bias and information bias, we cast the problem of stock price prediction as

$$f : \{\mathbf{N}, \mathbf{P}\} \mapsto \mathbf{R}, \tag{2}$$

where both **N** and **P** are used. We adopt Multi-Kernel Support Vector Regression (MKSVR) as f in our system. The reasons are twofold:

1. MKSVR is a regression approach. The outputs of MKSVR are numerical values rather than user-defined *categories*. This would overcome the signal strength bias.
2. MKSVR integrates two information sources in an effective way. Since **N** and **P** have heterogenous features, simply combining those features is insufficient. MKSVR, which could have multiple sub-kernels, uses one sub-kernel to handle one of the information sources. MKSVR learns the weights of sub-kernels and determines which information source is more effective in prediction, where the weights could be interpreted as the extent to which one information source contributes to the prediction.

We design and implement a system for stock market predictions using MKSVR that combines both news articles and Hong Kong Stock Exchange (HKEx) tick prices. In particular, MKSVR has two sub-kernels: one is responsible for news articles, while the other accepts short-term historical prices. After learning the weight of each sub-kernel, the derived model makes numerical forecasting on stock short-term returns. Compared with three other baseline models implemented in the experiments, MKSVR has achieved the best performance with the smaller regression error.

The contributions of this paper are summarized as follows:

1. We build up a system with work flows that use MKSVR to integrate both news articles and stock tick prices, rather than only one of them, to quantitatively make intra-day short-term stock return predications.
2. Our approach by multiple kernel learning (MKL) can effectively integrate both news articles and stock tick prices, outperforming

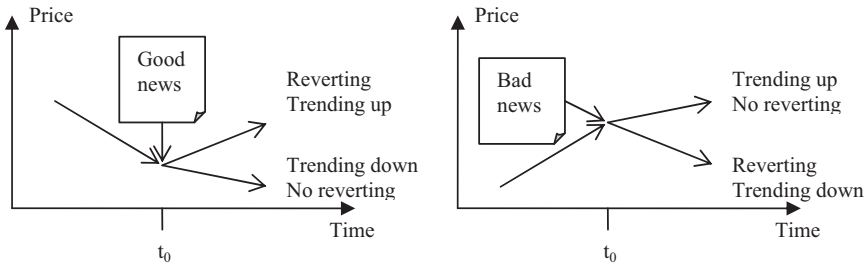

Fig. 1. Possible scenarios of price movements based on the impact of news articles and short-term prices. At time point t_0, the inputs are news article and short-term price trend before t_0. In the left figure, the outputs are either "reverting and trending up" or "trending down without reverting". In the right figure, the outputs are either "trending up without reverting" or "reverting and trending down".