Market Impact Analysis via Sentimental Transfer Learning

Xiaodong Li*, Haoran Xie†, Tak-Lam Wong† and Fu Lee Wang‡
*College of Computer and Information, Hohai University, Nanjing, China. Email: xiaodong.c.li@outlook.com
†Department of Mathematics and Information Technology, The Education University of Hong Kong.
‡Caritas Institute of Higher Education.

***Abstract*—The problem that how to improve the market impact prediction performances of predictors that are trained based on stocks with few market news is studied in this preliminary work. We propose sentimental transfer learning to transfer the knowledge learned from news-rich stocks that are within the same sector to the news-poor stocks. News articles of both kinds of stocks are mapped into the same feature space that are constructed by sentiment dimensions. New predictors are then trained in the sentimental space in contrast to the traditional ones. Experiments based on the data of Hong Kong stocks are conducted. From the early results, it could be seen that the proposed approach is convincing.**

I. Introduction

Market impact analysis based on textual financial news articles has been studied for many years [1], [2]. One problem remains to be solved is how to improve the prediction performances for stocks that are lack of news reports. As the observations shown in Fig. 1, even within the market index HSI[1], the distribution of news reports is highly imbalanced, and predictors that are trained on stocks that have much fewer reports than the high-market-capital ones will have higher variance and thus worse prediction performances.

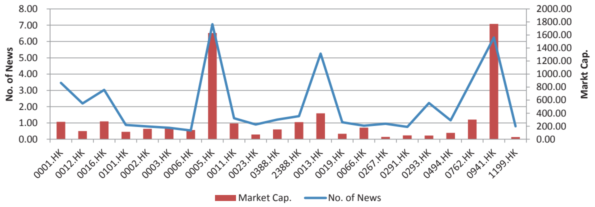

Fig. 1. Numbers of news reports among stocks of HSI.

In algorithmic trading, especially pair trading, stocks within the same sector are believed to have similar price movements. In this paper, following the idea, we propose sentimental transfer learning to solve the news-poor training problem, which is to transfer the knowledge learned from the news-rich stocks in the same sector to the news-poor ones. In sentimental transfer learning, the news articles from both news-rich and news-poor stocks are mapped into the same feature space that is constructed by sentiment dictionaries. Within the feature space, the instances of both kinds of stocks are used together to train the predictor. In the next step, the predictor is plugged into a framework of market impact analysis proposed in [3]. Preliminary evaluation based on Hong Kong stock prices and news articles are conducted, and the experimental results show that the proposed approach has convincing performance.

The rest of this paper is organized as follows. Section II firstly reviews the categorization of transfer learning, and secondly presents our proposed sentimental transfer learning. Section III reports the preliminary experimental results and gives some discussions.

II. Sentimental Transfer Learning

The purpose of the transfer learning is to transfer the knowledge from a source domain to a task in a target domain. In the survey by Pan and Yang [4], transfer learning approaches can be categorized into four groups, among which the idea of instance transfer technique is to reuse part of the instances in the source domain and help the learning task in the target domain, and the idea of feature transfer learning is to learn a feature representation to "better" represent the data in both the source domain and the target domain. In this paper, sentimental transfer learning is a combination of those two techniques: 1) it reuses the news articles of the news-rich stocks and 2) it constructs a sentimental feature space and maps the instances of both the news-rich and news-poor stocks into the same space.

The work flow of the sentimental transfer learning is illustrated in Fig. 2. In the first step, the news articles of the news-poor stocks are divided into two parts, the training part

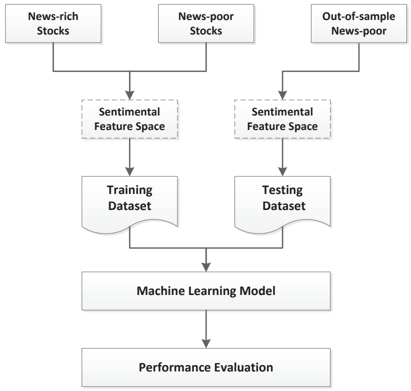

Fig. 2. The work flow of sentimental transfer learning.

[1] Hang Seng Index

is mixed with the news of the news-rich stocks and the other part is left for testing. In the second step, words in the news articles are searched in the sentiment dictionary and projected onto sentimental dimensions if they have any affective aspects. As the sentimental dimensions are fixed in the dictionary, each word can be represented by a sentiment feature vector of the same length. Thus, each news article can be represented by a sentiment feature vector by summing up all words' vectors. In the third step, all the preprocessed instances are fed into the machine learning model. After training and testing, the model will be evaluated in the final step.

III. Preliminary Experiments and Discussions

A news archive from FINET[2] is employed. The news archive contains both company-specific and market related news from Jan. 2003 to Mar. 2008. Each piece of news is tagged with a time stamp showing the time the news is released, which helps classify news by dates. Stock codes of companies that are mentioned in the news are listed at the end of the article, which helps establish the mapping from the news articles to stocks and vice versa. The data set is split into three parts for different purposes: 1) from Jan. 2003 to Dec. 2005 is the training data set; 2) from Jan. 2006 to Dec. 2006 is the validation data set; and 3) from Jan. 2007 to Mar. 2008 is the independent testing data set.

Stocks selected from HSI are investigated. There are 4 sectors in HSI, i.e., Commerce, Finance, Properties and Utilities. As shown in Fig. 1, 0941.HK (Commerce), 0005.HK (Finance), 0001.HK (Properties) and 0002.HK (Utilities) are with relatively more news reports, and in contrast, 0293.HK (Commerce), 0023.HK (Finance), 0012.HK (Properties) and 0006.HK (Utilities) are with fewer reports. We select those eight stocks and obtain their daily quotes (i.e., Open, High, Low, and Close prices) from Yahoo! Finance[3]. The daily OHLC data has the same period of data with the news. Each news is labeled by

$$y = \begin{cases} +1 & \text{if } r \geq \theta \\ 0 & \text{if } -\theta < r < \theta \\ -1 & \text{if } r \leq -\theta \end{cases}, \tag{1}$$

where θ is 0.5%, and r is simple return which is calculated based on Open and Close prices,

$$r = \frac{Close - Open}{Open}. \tag{2}$$

The prediction performance is measured by a standard metric, i.e. accuracy. Since there are three classes, we use a three-class version of accuracy. Loughran-McDonald financial sentiment dictionary (LMD) is adopted in the experiment, which is a manually constructed sentiment dictionary by Loughran and Bill McDonald [5]. The dictionary contains more than 3911 words and 6 sentiment dimensions. We use the version 2012.

The benchmark uses only the news articles from the news-poor stocks. The experiment results are shown in Table I. It is observed from the results that in Finance, Properties and Utilities sectors, with the help of sentimental transfer learning, the accuracy results are better than the benchmark. However, the result in Commerce sector does not have the same pattern. The reason is that stocks in Commerce have various businesses, e.g., 0293.HK is an airline company and 0941.HK is a telecommunication company. Although they are in the same sector, the selection method based on sector categorization is not sufficient for sentimental transfer learning, and more adaption factors, such as stock price correlations, need to be considered, which will be investigated and optimized in the future work.

TABLE I
Experiment results.

Validation	Commerce	Finance	Properties	Utilities
Benchmark	0.4218	0.3878	0.3567	0.4462
STL	0.3629	0.6437	0.3724	0.4303
Testing	Commerce	Finance	Properties	Utilities
Benchmark	0.3872	0.3503	0.3611	0.3796
STL	0.3548	0.5772	0.3631	0.4279

IV. Acknowledgments

This work is fully supported by the Fundamental Research Funds for the Central Universities 2016B01714, the National Natural Science Foundation of China (Grant No. 61602149), the Start-Up Research Grant (RG 37/2016-2017R), the Internal Research Grant (RG 66/2016-2017) of The Education University of Hong Kong, a grant from Research Grants Council of Hong Kong Special Administrative Region, China (UGC/FDS11/E06/14), and a grant from Research Grants Council of Hong Kong Special Administrative Region, China (UGC/FDS11/E03/16).

References

[1] X. Li, X. Huang, X. Deng, and S. Zhu, "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information," *Neurocomputing*, vol. 142, pp. 228–238, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2014.04.043

[2] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2014.04.022

[3] X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. L. Wang, "Does summarization help stock prediction? A news impact analysis," *IEEE Intelligent Systems*, vol. 30, no. 3, pp. 26–34, 2015. [Online]. Available: http://dx.doi.org/10.1109/MIS.2015.1

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.191

[5] T. LOUGHRAN and B. MCDONALD, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011. [Online]. Available: http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x

[2] http://www.finet.hk/mainsite/index.htm

[3] http://finance.yahoo.com/