Optimized Stock Factors Prediction Model Based on LSA and Multiple SVM Models

Wang Peng[1], Kiyoshi Izumi[12]

*1 Department of Systems Innovations,School of Engineering,the Univ. of Tokyo
*2 CREST, JST

Abstract: We put forward an optimized method of stock factors prediction model which can be easily extended to realtime prediction system., based on the new version of Yahoo Financial text board of 2012.11~2013.6 with about 4000 companies. Preceding studies have verified that BBS text can be used to forecast trade volume and return. On this basis, LSA (Latent Semantic Analysis) and multi-SVM model are put forward in our framework to improve the accuracy of natural language processing and the prediction.

1 Introduction

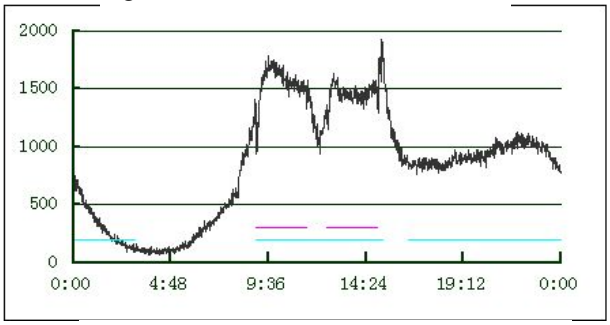
Individual investors present and share their own suggestions or opinions in the stock text board is pretty common. Antweiler & Frank‘s[1] experiments show that this kind of text has interaction with the market of America. The further, Maruyama[2] through comparisons of several previous researches show that it is difficult to find a correlation between the amount of posting messages and stock trade volume. On the other hand, stock return has been proved that has correlation with message content. Hirohiko[3] also give a similar conclusion by using a factor model. We also agree with it especially for the stock that is traded by a large number of individual investors.

However, there are some analyticalmethods can be optimized. First of all, experimental stock and period selection method can be improved, because they haven’t considered the habit of individual investors. In natural language processing of Maruyama’s framework, only top 500 word in TF-IDF is used, which may lead to information missing. Therefore, we put forward an optimized framework with multiple SVM models and smaller time scale. And then we use it find the strong correlation with the stock that always be traded by individual investors.

2 Feasibility Analysis

Figure1 present the distribution of posting frequence in Yahoo! stock board during 2012.11~2013.6, from which we can find out two apparent posting peaks. According to the pink timeline, user activities in stock text board is the most active. This frequence line is also at a high level during Nikkei 225 Futures Trading hours(blue timeline). Because of this distribution shape, it will retain the vast majority of information about active users if we shorten the test period to trading hours.

Figure1 : Text Distribution in BBS



3 Experiment

3.1 LSA

LSA is an effective method for analyzing the relationships between documents and words. For instance, X is a matrix that is built by a document set D, which has i documents and t different terms. Then it can be decomposed by SVD(singular value decomposition) as following equation:

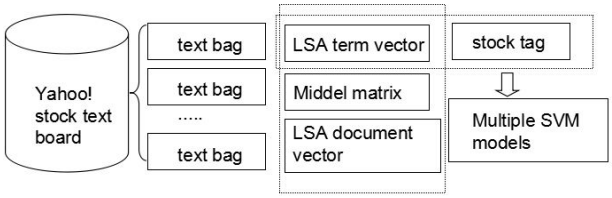
$$X = U\Sigma V^T$$

V^T is the final terms matrix, from which we can capture the eigenvectors. LSA can decompose the original term matrix to a brief and smaller operational matrix. The features of origianl matrix are retained.

3.2 Framework

In the first step, we deliver text data to a text bag per 10 or 20 or 30 minutes (time scale). In natural language processing of every bag, pronouns, numerals, adverbs, articles, and so on are all be deleted. Segmentation process is based on Mecab in hadoop cluster.

Figure2：Framework Processing



Then All of the text bags with same time scale are transformed to a large sparse matrix. Through LSA this matrix decomposed into LSA term vertor matrix, SVD transform matrix, LSA document vector matrix. Decomposed term matrix and stock price data are used to train svm model. As a result this models can be the basis Multi-model prediction. For the accuracy of train model, 10-folds cross validation has been implemented.

3.3 Experiments

The identical text database as previous researchers, Yahoo! Financial text board is the original corpus, which is the most widespread stock BBS. We use Nikkei Thesaurus[4] as dictionary, which is made for news search index and contains about 13,000 words by different fields. After segmentation processing, 37,954-dimensional sparse matrix is generated for LSA processing. We sampled five stocks in the Nikkei 225,which is on the top of frequent trading list.They are 6758 Sony, 8396 The Bank of Tokyo-Mitsubishi UFJ, 8411 Mizuho, 8604 Nomura HD, 9501 Tokyo Electric Power Co. Regression model is built by SMO (SVM by Sequential Minimal Optimization) regression. As a comparative test, linear regression is tested as well.

3.4 Result Analysis

Average correlation coefficient and root relative squared error in five each model are shown in Graph3 and Graph4 (y axis). The x axis of both graphs is the time scale.

It is clear that the pink line of SMO is always on the top in Gryaph 3, while linear regression always on the bottom. But their position is exchanged in the Graph4. Along with the increasing of time scale correlation coefficient are growing steady. However, root relative squared errors of these models go down at the same time. It suggests us that it is not a good decision for real time to use corpus as big as ones can or the text data long ago.

Figure3：Correlation Coefficient

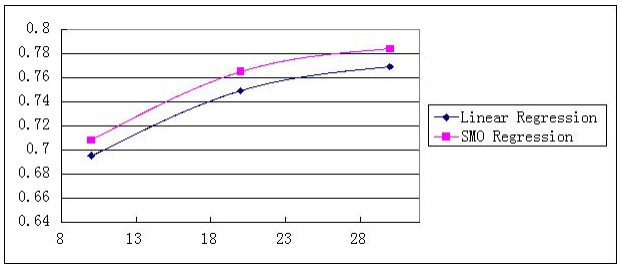
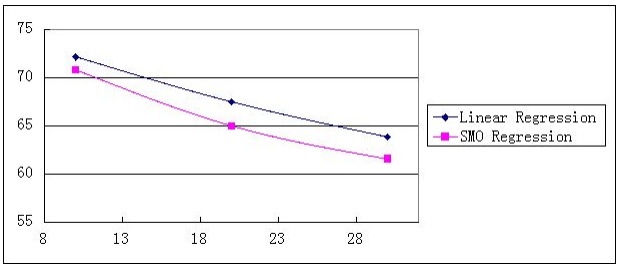


Figure4：Root Relative Squared Error



4 Conclusion

Through this framework, we obtain Multi-SVM models. If we set the accuracy as their weights, it is possible to build a prediction system. Besides individual stock factors, we will also test the correlation with the factors of Nikkei 225 futures in the future work.

Finally, sparse matrix calculation by LSA takes too long, we consider using hadoop to over come it.

Reference

[1] Antweiler, W. and Frank, M. Z., “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards,” *Journal of Finance*, Vol. 59, No. 3, pp.1259-1294, 2004.

[2] 丸山健，梅原英一，諏訪博彦，太田敏澄，インターネット株式掲示板の投稿内容と株式市場の関係，証券アナリストジャーナル，第 46 巻第 11 ・ 12 号，pp.110-127，2008.

[3] 山下一雄，石上隆達，佐藤 哲也，インターネット掲示板にみる社会的関心と株価変動の関係，日本社会情報学会第 20 回全国大会研究発表論文集，pp.237-240, 2005.

[4] 日経テレコン 21 ホームページ日経シソーラス http://t21.nikkei.co.jp/public/help/contract/price/01/help_kiji_thes_field.html