Belief Fusion of Predictions of Industries in China's Stock Market

Yongjun Xu[1], Lin Wu[1,2], Xianbin Wu[1,2], and Zhiwei Xu[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
[2] University of Chinese Academy of Sciences, Beijing, 100049
{xyj,wulin,wuxianbin,zxu}@ict.ac.cn

Abstract. This contribution presents the application of Dempster-Shafer theory to the prediction of China's stock market. To be specific, we predicted the most promising industry in the next month every trading day. This prediction can help investors to select stocks, but is rarely seen in previous literatures. Instead of predicting the fluctuation of the stock market from scratch all by ourselves, we fused ratings of 44 industries from China's securities companies using Shafer's evidence theory. Our preliminary experiment is a daily prediction since 2012-05-02 with ratings published 10 days before that day. Our predicted industries have an average rank of 19.85 in earnings, 11.8% better than random guessing (average rank is 22.5). The average rise rate of predicted industries in a month is 0.59%, 0.86% higher than overall (which is -0.274%), and nearly 0.7% higher than simple voting (which is -0.097%). Our predictions are posted on Weibo every day since 2014-04-28.

Keywords: Belief Fusion, Stock Market, Dempster-Shafer Theory, Prediction.

1 Introduction

A prediction of the most promising industry in the medium term can help investors select stocks or industry index funds[1] if the accuracy is better than random guessing. But this kind of prediction is rarely seen in literatures.

Previous prediction methods include statistics, technical analysis [5], fundamental analysis [5], and linear regression [12]. State-of-the-art in stock prediction techniques is surveyed in [1]. Instead of predicting the fluctuation of stock market from scratch all by ourselves, we fuse beliefs from experts — China's securities companies. These securities companies can be seen as soft data sensors [4] that observe related phenomena, and give ratings for some industries as the output of analysis. With their ratings as input, we try to predict the most promising

[1] "An index fund (also index tracker) is a collective investment scheme (usually a mutual fund or exchange-traded fund) that aims to replicate the movements of an index of a specific financial market, or a set of rules of ownership that are held constant, regardless of market conditions. As of 2007, index funds made up 11.5% of equity mutual fund assets in the US." — http://en.wikipedia.org/wiki/Index_fund.

F. Cuzzolin (Ed.): BELIEF 2014, LNAI 8764, pp. 348–355, 2014.

industry in the next month, which is expected to have the highest average rise in stock price.

As [1] states, "information regarding a stock is normally incomplete, complex, uncertain and vague". A securities company usually publishes ratings on less than 2 industries every trading day. These ratings consist of "buy", "overweight" and "neutral", representing different degrees of belief about the investment value. For industries with no ratings, its investment value is ignorant other than neutral. To combine ratings from different securities companies, we use the Dempster-Shafer theory [8,9], which is a powerful tool for uncertainty reasoning. In this theory, different sources express their uncertainty about the question of interest with belief functions. Then those functions are fused by Dempste's rule to arrive at the final degree of belief.

This paper is structured as follows. In Sect. 2, we present our model and formalization. We also describe our methods to predict. In Sect. 3, we introduce our experiment. Section 4 demonstrates our results. In Sect. 5, we draw the conclusion and discuss future work.

2 Problem Formalization and Methods

In this section, we describe the formalization and methods of our prediction of the most promising industry.

2.1 Question of Interest and Frame of Discernment

Under the framework of Dempster-Shafer theory, the answer to the question of interest is one element of a finite set called Frame of Discernment. It's composed of an exhaustive list of mutually exclusive answers.

In our problem, the question of interest is: "Which is the most promising industry in the medium term". There are 44 possible answers according to East-Money[2], which is one of China's largest financial website. Let's denote the frame of discernment by Ω. Then

$$\Omega = \{\text{electricity, electronic, culture and media, pharmaceuticals}\ldots\} \tag{1}$$

$$|\Omega| = 44 \tag{2}$$

2.2 Evidence and Basic Belief Assignment (BBA)

As far as we know, there is no direct answer to the question of interest. But ratings of industries published by securities companies can be seen as evidence for this question, which is available on the website [3].

[2] http://www.eastmoney.com/

[3] http://data.eastmoney.com/report/hyyb.html

A securities company may rate an industry as "buy", "overweight" or "neutral". There is no rating "sell" on the website. This special situation maybe results from government policies and China's special culture. As a securities company publishes ratings of limited industries on a single day (usually less than two industries on average), we combine its ratings from 10 consecutive trading days as a single report. From such a report, we can get 4 sets, which usually have intersections:

$$S_{\mathrm{buy}} = \{\text{industry } I | I \text{ is rated as "buy"}\} \tag{3}$$

$$S_{overweight} = \{\text{industry } I | I \text{ is rated as "overweight"}\} \tag{4}$$

$$S_{\mathrm{neutral}} = \{\text{industry } I | I \text{ is rated as "neutral"}\} \tag{5}$$

$$S_{\mathrm{others}} = \Omega - S_{\mathrm{buy}} - S_{\mathrm{overweight}} - S_{\mathrm{neutral}} \tag{6}$$

When securities companies give different ratings for a particular industry during these 10 days, only the latest one is adopted.

We haven't taken the freshness of data into consideration, for the purpose of reduction in computation burden. In our method, all the industries with the same rating from a securities company are in a single set. So the number of focal elements is only 4 in a securities company's report. After combination, the number will grow exponentially. To reduce it, we discard those focal elements whose masses are under a threshold, which is set experientially as 0.0001.

Our subjective judgment is that, S_{buy} is very likely to contain the most promising industry, $S_{overweight}$ is also likely to contain it, and S_{neutral} is not likely to contain the most promising industry. According to this judgment, basic belief assignment is calculated as follows when S_{neutral} is not empty:

$$m(S_{\mathrm{buy}}) = \frac{44 - |S_{\mathrm{others}}|}{44} * \frac{4 * |S_{\mathrm{buy}}|}{4 * |S_{\mathrm{buy}}| + |S_{overweight}| + |\overline{S_{\mathrm{neutral}}}|} \tag{7}$$

$$m(S_{overweight}) = \frac{44 - |S_{\mathrm{others}}|}{44} * \frac{|S_{overweight}|}{4 * |S_{\mathrm{buy}}| + |S_{overweight}| + |\overline{S_{\mathrm{neutral}}}|} \tag{8}$$

$$m(\overline{S_{\mathrm{neutral}}}) = \frac{44 - |S_{\mathrm{others}}|}{44} * \frac{|\overline{S_{\mathrm{neutral}}}|}{4 * |S_{\mathrm{buy}}| + |S_{overweight}| + |\overline{S_{\mathrm{neutral}}}|} \tag{9}$$

$$m(\Omega) = \frac{|S_{\mathrm{others}}|}{44} \tag{10}$$

When S_{neutral} is empty, $\overline{S_{\mathrm{neutral}}}$ equals Ω, basic belief assignment is calculated as follows:

$$m(S_{\mathrm{buy}}) = \frac{44 - |S_{\mathrm{others}}|}{44} * \frac{4 * |S_{\mathrm{buy}}|}{4 * |S_{\mathrm{buy}}| + |S_{overweight}|} \tag{11}$$

$$m(S_{overweight}) = \frac{44 - |S_{\mathrm{others}}|}{44} * \frac{|S_{overweight}|}{4 * |S_{\mathrm{buy}}| + |S_{overweight}|} \tag{12}$$

$$m(\Omega) = \frac{|S_{others}|}{44} \tag{13}$$

These formulas are representations of evidence. We'll modify them according to statistics in the future.

After obtaining the basic belief assignment of each securities company, we combine them one by one with Dempster's combination rule [8]:

$$m(A) = \frac{\sum_{A_1 \subseteq \Omega, A_2 \subseteq \Omega} \{m_1(A_1)m_2(A_2)|A_1 \cap A_2 = A\}}{\sum_{A_1 \subseteq \Omega, A_2 \subseteq \Omega} \{m_1(A_1)m_2(A_2)|A_1 \cap A_2 \neq \varnothing\}} \tag{14}$$

In the above equation, m_1 and m_2 correspond to the evidence from the two securities companies that are being combined, see Fig. 1. The prominent advantage of Dempster's combination rule over others [13,14] is its simplicity, which is important in our scenario with 44 singleton sets and 25 data sources. Besides, the result shouldn't be influenced when we change the order of combination. So we need a rule with features of associativity and commutativity.

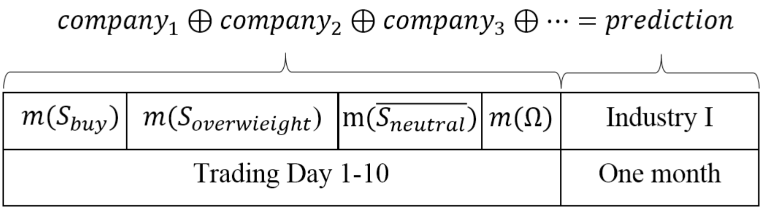

Fig. 1. Prediction of the most valuable industry in the following 30 days by combining ratings from each securities company

2.3 Decision

There is no agreement about how to make decision with final belief functions. Candidate methods include choosing the hypothesis with the maximum BBA, belief, plausibility or pignistic probability.

Smets proposed making decisions based on maximizing pignistic probabilities [11]. Pignistic probability is derived from BBA by pignistic transformation [11]:

$$BetP(\omega) = \sum_{W \subseteq \Omega, \omega \subseteq W} \frac{1}{|W|} \frac{m(w)}{1 - m(\phi)}, \forall \omega \subseteq \Omega \tag{15}$$

This method is widely used in literatures [2,6,7,10].

In our experiment, we found that decisions based on maximum beliefs have the best performance and we adopted it in our paper.

3 Experiment

There are about 50 to 150 ratings published on the website[4] every trading day since 2012-04-25. On other days, there are usually less than 10 ratings every day. We have collected all of these historical ratings, and made historical predictions every day, supposing that we were in the past. Besides, we are making new predictions every trading day, which are posted on Weibo[5](China's Twitter).

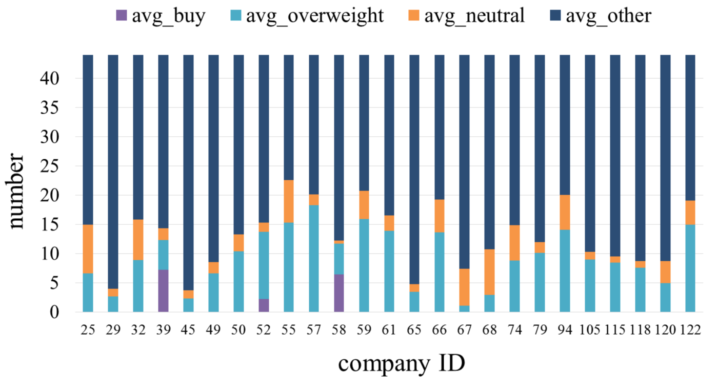

Fig. 2. Average number of industries with ratings "buy", "overweight", "neutral" and without ratings for each of the selected 25 companies in 10 consecutive trading days during the last two years

There are 78 securities companies who have published their ratings in the last two years. But some of them are not active, we manually selected the top 25 active companies as our data sources, see Fig. 2.

To evaluate our predictions, we supposed that an investor would spend 1 dollar to buy the stocks of each industry every trading day, and sell them after 30 days. And we supposed that money spent for each stock is proportional to the trading volume of it on that day (This is similar to index funds). Then we calculated the earnings for each industry in the next month and got the rank of our predicted industries. At the same time, we summed up the monthly rise rates of predicted industries on every trading day during the last two years and compare it with the average of all industries. The historical data of all the stocks (more than 2,000) in the last two years was provided by Shanghai Wind Information Co., Ltd.

We also compared our methods with simple voting: if an industry is ranked as "buy", "overweight" or "neutral", it gets 4, 1 or -1 vote respectively. If more than one industries has the largest votes, we chose one of them randomly.

[4] `http://data.eastmoney.com/report/hyyb.html`
[5] `http://weibo.com/u/3915945698`

4 Results

4.1 Computation Burden

The number of focal elements after the combination of 25 companies is usually below 1000, see Fig. 3. We ran our python program on a computer with a 4-core 2.4GHz CPU and 2GB RAM. The combination of 25 data sources took 1.98 seconds on average.

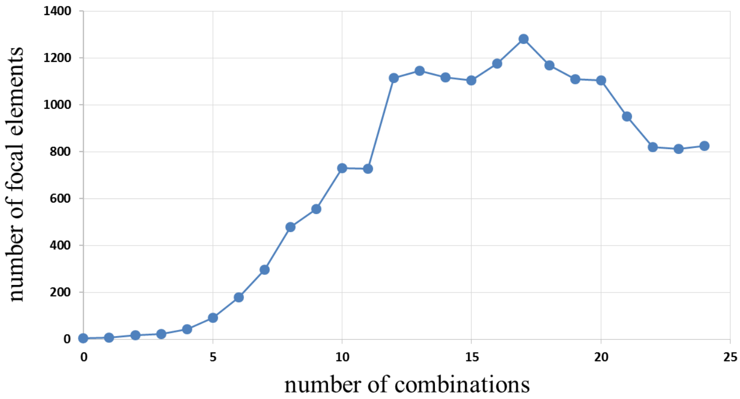

Fig. 3. The trend of the number of focal elements after each combination. The declination results from the threshold of minimum mass, which we set as set experientially as 0.0001.

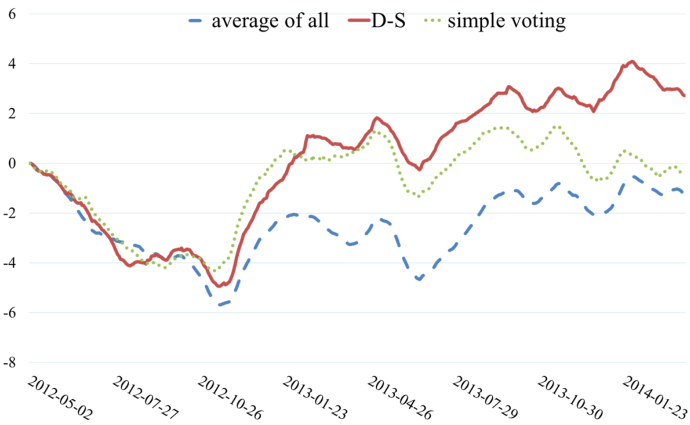

Fig. 4. Accumulated monthly rise rate of predicted industries by D-S and simple voting, and the average of all the industries

4.2 Accuracy

Every day, we ranked industries according to their rise rate in stock price in the previous month. The average rank of our predicted industries is 19.85 in the last two years, 11.8% better than random guessing (average rank is 22.5). The average rise rate of predicted industries in a month is 0.59%, 0.86% higher than overall (which is -0.274%), and nearly 0.7% higher than simple voting (which is -0.097%). The accumulated monthly rise rate day by day is shown in Fig. 4. To put it simple, this figure shows how many dollars you would have earned if you bought 1 dollar of the predicted industry's index fund every trading day and sold it after 30 days since 2012-05-02. From this figure, we can see that on 2014-04-28, you would have earned 2.715 and -0.45 dollars if the industry is chosen according to D-S and simple voting respectively.

5 Conclusion and Future Work

We have introduced the application of Dempster-Shafer theory to the prediction of industries in China's stock market in this paper. Instead of forecasting stock market by technical analysis or fundamental analysis, we fused output of soft data sensors C reports from securities companies. Our predicted industries have an average rank of 19.85 in earnings, 11.8% better than random guessing (average rank is 22.5). The average rise rate of predicted industries in a month is 0.59%, 0.86% higher than overall (which is -0.274%), and nearly 0.7% higher than simple voting (which is -0.097%). Our predictions are posted on Weibo every day since 2014-04-28.

In the future, we will continue collecting data for further evaluation and improvement of our methods. The basic belief assignment is determined according to our subjective judgment at present. When more data is available, we plan to adjust it on the basis of statistical information, such as using the probabilistic representation of evidence presented by [3].

Acknowledgments. This paper is supported in part by National Natural Science Foundation of China (NSFC) under grant No.(61173132).

References

1. Agrawal, J., Chourasia, V., Mittra, A.: State-of-the-art in stock prediction techniques. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2, 1360–1366 (2013)
2. Fiche, A., Martin, A., Cexus, J.-C., Khenchaf, A.: A comparison between a bayesian approach and a method based on continuous belief functions for pattern recognition. In: Denœux, T., Masson, M.-H. (eds.) Belief Functions: Theory & Appl. AISC, vol. 164, pp. 53–60. Springer, Heidelberg (2012)
3. Halpern, J.Y., Fagin, R.: Two views of belief: belief as generalized probability and belief as evidence. Artificial Intelligence 54(3), 275–317 (1992)

4. Jenkins, M.P., Gross, G.A., Bisantz, A.M., Nagi, R.: Towards context aware data fusion: Modeling and integration of situationally qualified human observations to manage uncertainty in a hard+ soft fusion process. Information Fusion (2013)
5. Kara, Y., Acar Boyacioglu, M., Baykan, M.K.: Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. Expert Systems with Applications 38(5), 5311–5319 (2011)
6. Karem, F., Dhibi, M., Martin, A.: Combination of supervised and unsupervised classification using the theory of belief functions. In: Denœux, T., Masson, M.-H. (eds.) Belief Functions: Theory & Appl. AISC, vol. 164, pp. 85–92. Springer, Heidelberg (2012)
7. Senouci, M.R., Mellouk, A., Oukhellou, L., Aissani, A.: Using the Belief Functions Theory to Deploy Static Wireless Sensor Networks. In: Denœux, T., Masson, M.-H. (eds.) Belief Functions: Theory & Appl. AISC, vol. 164, pp. 425–432. Springer, Heidelberg (2012)
8. Shafer, G.: A mathematical theory of evidence, vol. 1. Princeton University Press, Princeton (1976)
9. Shafer, G.: Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning 4(5), 323–362 (1990)
10. Shoyaib, M., Abdullah-Al-Wadud, M., Zahid Ishraque, S.M., Chae, O.: Facial Expression Classification Based on Dempster-Shafer Theory of Evidence. In: Denœux, T., Masson, M.-H. (eds.) Belief Functions: Theory & Appl. AISC, vol. 164, pp. 213–220. Springer, Heidelberg (2012)
11. Smets, P.: Decision making in the tbm: the necessity of the pignistic transformation. International Journal of Approximate Reasoning 38(2), 133–147 (2005)
12. Upadhyay, A., Bandyopadhyay, G., Dutta, A.: Forecasting stock performance in indian market using multinomial logistic regression. Journal of Business Studies Quarterly 3(3) (2012)
13. Yang, J.B., Xu, D.L.: Evidential reasoning rule for evidence combination. Artificial Intelligence 205, 1–29 (2013)
14. Yong, D., WenKang, S., ZhenFu, Z., Qi, L.: Combining belief functions based on distance of evidence. Decision Support Systems 38(3), 489–493 (2004)