CAN WEB NEWS MEDIA SENTIMENTS IMPROVE STOCK TRADING SIGNAL PREDICTION?

Nana Lin, School of Information, Renmin University of China, Beijing, P.R. China, linnana@126.com

Wei Xu, Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), MOE, and School of Information, Renmin University of China, Beijing, P.R. China, weixu@ruc.edu.cn

Xinwei Zhang, School of Information, Renmin University of China, Beijing, P.R. China, xinweify.zhang@gmail.com

Siqi Lv, School of Information, Renmin University of China, Beijing, P.R. China, lvsiqiruc@163.com

Abstract

The technical analysis and machine learning have been integrated in stock trading signal forecasting. And it has been proved that there are some weaknesses in technical analysis because of the complex environment in the stock market. In our prediction system, web news media sentiment analysis is regarded as a supplementary way to cover the shortage of technical analysis. It is considered to bring the stock market sentiment which reflects the subjective information of investors into the prediction system. Web news media sentiment indicators (WNMS) are designed to bring the information about stock market sentiment in our system. The WNMS is generated by analyzing the variance of sentiment elements from the news in the Stock Timely Rain Sector of Sina Finance and Economics Website and it is imported into the prediction system as features combined with common feature indicators (CFI). GMKL is applied to establish the relationship between the trading signals generated by piecewise linear representation (PLR) and the features of the trading signals (SCFI). Comparative experiments are adopted in nine stocks from Shanghai and Shenzhen Stock Exchange to determine the effect of PLR-GMKL and WNMS in prediction. From the aspects of the prediction accuracy and the profit, the final comparative results show that the PLR-GMKL model performs better than the PLR-WSVM model. And the prediction system performs best when adding WNMS into features and using PLR-GMKL model.

Keywords: Stock Trading Signal Prediction, Web News Sentiment Analysis, PLR, GMKL, GA.
1 INTRODUCTION

Stock trading signals forecasting is regarded as one of the most popular but difficult research topics. The reason why trading signals are so difficult to predict is that stock price variations are subject to the problem of high nonlinearity and non-stationary, as a result of political and economic environment both impact stock price variations (Cao et al. 2005).

Usually, there are two approaches to stock trading signals prediction, including fundamental analysis and technical analysis (Mabu et al 2013). The fundamental analysis forecasts trend reversals and trades at the peak or valley of the stock price. Technical analysis is a major analytical technique based solely on price data. Analysing financial time series data involves looking for peaks, valleys, trends, and other factors (Bao et al 2008).

In recent years, with the development of web news media, more public information and news impacted the price in the stock market (Antweiler and Frank, 2004). Ederington (1993) mentioned that media wording, media coverage and other that has nothing to do with the fundamentals, such as the length of space form, had some impact on asset pricing, which seems contradictory with classic financial theory. Fand and Peress (2009) also confirmed that there was a significant media impact in the U.S. stock market. Besides, Simpson and Ramchander (2004) pointed out that the web news media played an important role in the stock market volatility. Driven by systemic cognitive biases, investors often over-react to both good and bad news. Das and Chen (2007) extracted sentiment from small talk on the web to predict financial market. At the same time, Tetlock (2007) also suggested that high web news media pessimistic sentiment predicts downtrend pressure on market prices, which consistent with theoretical models of noise and liquidity traders. Nowadays, many researchers are using different methods to study stock trading signals prediction. However, the related research paper which regarding the web news media sentiment as indicators to predict the stock trading signals is rare. Therefore, we conduct a deep analysis from this aspect. Xu et al. (2012) proposed a web mining approach for financial market prediction by sentiment analysis.

In this paper, we present a piecewise linear representation and generalized multiple kernel learning (PLR-GMKL) model. A piecewise linear representation (PLR) method is used to determine the turning points of stock price, and then they are transformed into stock trading signals. Stock trading signals, traditional technical indicators and web news media sentiment indicators are used as input data for GMKL (Yang, H. et al 2011). This paper also used genetic algorithm (GA) to choose the optimal indictors which influence prediction decisively. In addition, GMKL algorithm is applied to predict stock trading signals.

This paper is organized as follows: Section 2 provides related theory background and literature review in the field of stock trading signal prediction. Section 3 describes PLR-GMKL model and analyses the choice of web news media sentiment indexes in detail. Section 4 presents the experimental tests on 9 shares to illustrate the performance of PLR-GMKL. Finally, the conclusions and future work are given in Section 5.

2 LITERATURE REVIEW

As the stock market is affected by many factors, the noise level is very high. Therefore, the forecasting of stock trading signals has become an upmost research direction in the area of finance engineering. So, several auto-trading models/systems are further developed, and these auto-trading methods make full use of financial analysis and prediction techniques to support real-time decision-making.

Initially, several rule based models/systems have been developed to support auto-trading. For example, Dymova et al (2010) applied a rule-base evidential reasoning method to stock trading system, while Hsu et al. (2011) adopted extended classifier systems to discover knowledge rules. With the development of uncertain theories and evolutionary computation, some extended and optimized models have been suggested for designing auto-trading strategies. Huang et al. (2009) developed
hierarchical coevolutionary fuzzy system for predicting financial time series, and then proposed a prudent trading strategy based on the predicted results. Troiano and Kriplani (2011) suggested inverse fuzzy transform for building a new class of technical indicators to identify early trading signals. Meanwhile, some artificial intelligent tools such as neural networks (NNs) or support vector machines (SVMs) have been widely used in auto-trading strategy design. For example, Vanstone et al. (2012) proposed neural networks to support stock trading with fundamental variables. Also, Change et al. (2009) integrated a PLR a NN for stock trading point prediction.

In this paper, we combine the advantage of statistical model and intelligent tools, and proposal a novel piecewise linear representation and generalized multiple kernel learning (PLR-GMKL) model. A piecewise linear representation (PLR) method is used to determine the turning points of stock price, and GMKL is employed to predict stock trading signals.

3 A PLR-GMKL MODEL FOR TRADING SIGNAL PREDICTION

In this paper, the framework of model includes piecewise linear representation, genetic algorithm, generalized multiple turning points from the stock price data and convert them into trading signals. Second, GA is used again to optimize the selection of traditional technical indicators. The optimal technical indicators, web news media sentiment indicators and stock trading signals are used as dataset. Third, the whole history dataset is divided into some recursive training-testing sets. Finally, use GMKL to predict the stock trading signals. The framework of prediction system is illustrated in the Fig. 1.

![Diagram](image)

Figure 1. The Framework of PLR-GMKL Prediction System
### 3.1 Piecewise Linear Representation, PLR

Piecewise linear representation (PLR), is an important method for time series data analysis. Because time series is a kind of big data, using original data directly is inefficiency. To extract the effective feature of the time series, this research applies PLR to process the data based on the following steps.

#### 3.1.1 Data normalization

Different stocks have different price level. To avoid the influence of price level, all data should be standardized before PLR. Max-min normalization is applied to convert them into value between 0 and 1.

#### 3.1.2 Setup the best PLR threshold value ($\delta$) by GA

For PLR, different segmentation threshold will create different patterns. The threshold is smaller and the patterns are more sensitive. Therefore, it is important to choose a suitable threshold. We apply genetic algorithm to optimize the threshold value of PLR (Chang et al 2011). It can generate better segments and higher profits. The best value is 0.05.

#### 3.1.3 Use PLR to segment the stock data

The improved PLR contains four basic algorithms, the top-down, the bottom-up, sliding window, as well as the sliding window bottom-up algorithm. This paper adopts the top-down method to divide the time series recursively until the need of the research is met. Basic steps are as follows:

- Connect the starting point(S) and the end point(E) of the data segment, denoted as segment SE.
- Calculate the vertical distance of each point to SE.
- Choose the point X whose vertical distance is the largest, and let the distance be $d$.
- If $d$ is larger than the threshold, recursively partition SX and XE.
- If $d$ is less than the threshold, no need to segment.

#### 3.1.4 Transform segmentation results to trading signals

After the segmentation, the trading signals obtained by PLR are divided into two classes, buying point (BP) and selling point (SP).

### 3.2 Indicator Selection & GA Optimization of the Parameters

#### 3.2.1 Technical indicator

Technical index analysis is based on the analysis of the history of stock prices. We assume that all possible influence factors in the stock market can be included, and future stock price trend can be predicted.

At present, widely used technical indicators include moving average (MA), relative strength index (RSI), transaction volume (TV) and so on. In addition, we add some new technical indicators (Luo et al 2013) which can represent the features of the stocks as input variables.

To sum up, the technical indicators used in this paper are listed in Table 1.

<table>
<thead>
<tr>
<th>Technical indicator</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving average (MA)</td>
<td>Reflect the growing trend of the stock.</td>
</tr>
<tr>
<td>MA(5), MA(6), MA(10), MA(12), MA(50)</td>
<td></td>
</tr>
<tr>
<td>Bias (BIAS)</td>
<td>The difference between CP and moving average line.</td>
</tr>
<tr>
<td>BIAS(5), BIAS(10)</td>
<td></td>
</tr>
<tr>
<td>Relative strength index (RSI)</td>
<td>Reflect the relative strength of the stock.</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSI(6), RSI(12)</td>
<td>Determine the signals of over-purchasing, over-selling, or deviation</td>
</tr>
<tr>
<td>Random index(K,D)</td>
<td>K(9), D(9)</td>
</tr>
<tr>
<td>The moving average convergence/divergence(MACD)</td>
<td>MACD is the difference between a 26-period and 12-period Exponential Moving Average.</td>
</tr>
<tr>
<td>Williams%R (W&amp;R)</td>
<td>W&amp;R(12)</td>
</tr>
</tbody>
</table>
| Closing price(CP)               |交易量最近日均交易金额
| Transaction volume(TV)          |交易量
| The amplitude of the price movement (ALT)| $ALT = \frac{p_t(t) - p_{t-1}(t)}{p_{t-1}(t)}$ |
| The average transaction price(ATP)| $ATP = \frac{transaction \ money}{transaction \ volume} = \frac{TM}{TV}$ |
| The change rate of transaction money to previous trading day(CRTM)| $CRTM = \frac{TM(t) - TM(t-1)}{TM(t-1)}$ |
| The change rate of average transaction price to the previous trading day(CRATP)| $CRATP = \frac{ATP(t) - ATP(t-1)}{ATP(t-1)}$ |
| The difference of MA between the short run and the long run(DMA) | $DMA = MA(10) - MA(50)$ |
| The average of DMA(AMA)         |AMA is the average of DMA for 10 days.                                      |
| Differences of technical indicator(Δ)| $ΔMA(5), ΔMA(6), ΔMA(10), ΔMA(12), ΔMA(20), ΔMA(50), ΔBIAS(5), ΔBIAS(10), ΔRSI(6), ΔRSI(12), ΔK(9), ΔD(9), ΔW&R(12)$ |
| Differences of technical indicator between t day and t + 1 day. | $ΔMA(5), ΔMA(6), ΔMA(10), ΔMA(12), ΔMA(20), ΔMA(50), ΔBIAS(5), ΔBIAS(10), ΔRSI(6), ΔRSI(12), ΔK(9), ΔD(9), ΔW&R(12)$ |

Table 1. Technical Indicators

3.2.2 Web news media sentiment indicators

As the well-known indicator of macro economy, especially the indicator of stock prices, the volatility trend of stock price is affected by the macroeconomic and financial market. Hess et al. (2008) pointed that financial and economic news has a volatility effect to stock and other commodities futures. Therefore, we believe that the web news media information will has an extent influence on stock trading signal prediction. We choose the Sina Finance and Economics Website stock timely rain sector news for the following three considerations. Firstly, Sina Finance and Economics Website stock timely rain plate is one of the important ways to obtain information for Chinese investors because Sina is one of the biggest portal website in China. Secondly, Sina Finance and Economics Website has a large daily news number of news. There is more abundant information compared with other sites. Finally, the news in Sina Finance and Economics Website can be crawled easily and time series are relatively complete. So we choose Sina Finance and Economics Website stock timely rain plate as the source of web news media source. The time period of news has been chosen from 1st, February, 2010 to 28th, September, 2012. 550 days in total. We use LocoySpider to crawl 66751 pieces of news.

In order to validate the web news media sentiment indicators for stock trading signal prediction accuracy, we calculate the amount of negative sentiment words (NSN) and the amount of positive sentiment words (PSN). Because there is no delimiter between words in Chinese, a segmentation processing is needed. In our research, we use Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) which is developed by Institute of Computing Technology, Chinese Academy of Sciences to do the segmentation process. ICTCLAS will split the sentence into a word sequence which uses space character as delimiter. We pick up words from the word sequence generated in former step and match the words with sentiment dictionary. We counted the amount of words appeared in pessimistic sentiment dictionary and negative sentiment dictionary. For Chinese we use Tsinghua Sentiment Dictionary v1.0.

Considering the degree of importance and the emotion attitude are kinds of relative concept, using the absolute numerical value loses its characteristics. We use max-min normalization to convert them into value between 0 and 1. By further calculation, we get positive sentiment variability index (ΔPS) and negative sentiment variability index (ΔNS), which represents as web news media sentiment indicators (WNMS).
3.2.3 Genetic algorithm

Genetic algorithm is widely used to solve optimization and search problem. First of all, code the parameters into chromosome. The GA algorithm models the natural selection by iterative selection, crossover and mutation. Finally, the chromosome which is in accordance with the best fitness will be generated. The basic steps of algorithm are shown in Fig.2.

![Figure 2. The Structure of Genetic Algorithm](image)

3.2.4 Use GA to optimize the technical index selection

Technical index analysis is considered as one of the most reliable techniques in stock trading. Some indicators may be useless in the stock prediction. Therefore, the indicators which are considered as the reference of stock trading should be selected carefully.

This paper uses GA to select the technical index. In the process of GA, the code is represented in binary as strings of 0s and 1s initially. 0 denotes the indicator is not chosen and 1 denotes the indicator is chosen. The fitness of the chromosome is represented by the forecast accuracy of the PLR-GMKL. The higher the accuracy is, the higher the fitness is. Initial population is randomly generated 0 and 1 binary string.

For instance, as for parameters A, B, C, D, E, the binary string '11100' denotes that 'the index A, B, C are adopted, and the index D, E are not adopted'.

For each stock, algebra is set up previously. In the final offspring, all the chromosomes whose fitness is relatively high are used as the reference to select indicators. The process will be expounded at the section 4.

3.3 GMKL-Framework

A brief introduction of Generalize Multiple Kernel Learning (GMKL), the extension of $L_1$-MKL and $L_2$-MKL, is shown here. It was proposed to improve the capability of classification of $L_1$-MKL and $L_2$-MKL (Yang, H. et al 2011).
MKL is a kind of supervised learning method. Given a data set \( \{x_i, y_i\}_{i=1}^{N} \), \( x_i \in \mathbb{R}^n \), \( y_i \in \{+1, -1\} \). The objective of supervised learning is to find a decision function \( f \), based on the input data.

In the MKL, \( Q \) basic kernels are given. Each kernel \( K_i(i=1,\cdots,Q) \) defines a feature mapping from original space to feature space \( F \), \( \phi : X \rightarrow F_i(i=1,\cdots,Q) \). We define \( \phi_{\theta} = \sqrt{\theta_1} \phi_1 \times \cdots \sqrt{\theta_Q} \phi_Q : X \rightarrow F \) as a combination mapping from input data space to feature space \( F \). \( \theta_i(i=1,\cdots,Q) \) is the weight of kernel, which should be learned from data.

Then the decision function of MKL with weight and \( b \) is represented as:

\[
f_{\alpha,\beta,\rho}(x) = \phi^\top \phi_{\theta}(x) + b = \sum_{i=1}^{Q} \sqrt{\theta_i} \alpha_i \phi_i(x) + b
\]

where \( \phi \) is \( Q \)-dimensional weight vector, consisting of \( \omega_i \). \( \omega_i \) represents a \( d(F_i) \)-dimensional weight vector. After making the variable transformation \( v_i = \sqrt{\theta_i} \omega_i \), \( \hat{v} = (v_1^T, \cdots, v_Q^T)^T \) and \( b \) could be attained by solving the following optimization model:

\[
\min_{\alpha,\beta,\rho} C \sum_{i=1}^{N} \left( \sum_{q=1}^{Q} v_q^i \phi_q(x) + b, y_i \right) + \frac{1}{2} \sum_{i=1}^{Q} v_q^i \theta_q
\]

\[s.t. \quad \Phi(\theta) \leq 1 \]

where \( \Phi(\theta) \) defines the regularizer of \( \theta \) and \( \hat{v} = (v_1^T, \cdots, v_Q^T)^T \). \( R \) is a convex function which makes the constraint convex. The Objective of MKL is to find the optimal weight vector \( \hat{v} = (v_1^T, \cdots, v_Q^T)^T \).

GMKL was proposed in order to capture the orthogonal information and create sparse solutions at the same time, based on \( L_1 \)-MKL and \( L_p \)-MKL. More specifically, GMKL combines the \( L_1 \)-MKL and \( L_2 \)-MKL. It imposes the composite constraint of \( L_1 \)-norm and \( L_2 \)-norm constraint. The model is shown as follow:

\[
\min_{\alpha,\beta,\rho} C \sum_{i=1}^{N} \left( f_{\alpha,\beta,\rho}(x_i), y_i \right) + \frac{1}{2} ||\alpha||^2
\]

\[s.t. \quad v ||\theta|| + (1-v) ||\theta|| \leq 1 \]

\[0 \leq v \leq 1 \]

This model has been proven to take advantage of both \( L_1 \)-MKL and \( L_2 \)-MKL. It obtains sparse solutions and shows high grouping effect on data with noise.

The dual form of Eq.3 is:

\[
\min_{\alpha,\beta,\rho} \frac{1}{2} \sum_{i=1}^{Q} \left( \sum_{q=1}^{Q} v_q \phi_q(x_i) \right) (\alpha_i y_i) - L_v \alpha
\]

\[s.t. \quad \alpha^\top y = 0, \quad \alpha \leq C1_y \]

where \( C \) is the parameter that balances the capability of generalization and the empirical risk. In this paper, the parameter \( v \) is set as 0.5.

### 3.4 Construction of trading signals prediction system by GMKL

In the prediction system, first we obtain the trading signals by PLR form the history time-series. Transforming trading signals to be the label set, 1 represents buying point (BP) and -1 represents
selling point (SP). Then feature set is generated by the indicators selected by GA. A two-class classification problem is set up on the train-testing dataset.

The whole input dataset is partitioned recursively to get the training set and testing set. For the first time, the 40% of the whole dataset is used as the training set and the following 5% of the dataset was used as the testing-set. Then for each time, the training-testing set move forward by 5% until the whole dataset was traversed.

GMKL mode is used to solve the two-classification problem on every training-testing dataset. Decision function is generated after the training dataset including label dataset and feature dataset was inputted into GMKL. Decision function then models the label of testing dataset. The label obtained by decision function is regarded as the prediction result. After recursively traversing the whole dataset, some measurements including accuracy and trading strategy profit based on the prediction are used to measure the effect of the prediction. This will be expounded in the section 5.

3.5 Investment Strategy performance

Although the forecast accuracy is a measure of the predicting model, but the more important target for trading signal prediction is to make profit. Inspired by the Modern Portfolio Theory, we assume that investors use their own money to buy all of the nine stocks in order to minimize the investment risk. That means one ninth cash-in-hand is used to buy each stock. Let the strategy be Modern Portfolio Strategy (MPS). We can calculate the MPS profits for the two rules to measure the effect of our prediction system. Two trading rules based on the trading signals are shown in the next section. The initial investment cash is 90 units, which means the initial investment for each stock is 10 units.

3.5.1 Rule One: Radical Trading Rule

For each stock, we use the whole investment for each stock (10 units at the first time) to buy when meeting the buy signal. After that, sell all the stock when meeting the sell signal. This kind of all-in and all-out trading operations are executed at the trading signals. If there is successive appearance of buy signal, we could only buy stock at the first appearance.

3.5.2 Rule Two: Conservative Trading Rule

We buy each stock only one unit if receive buy signal. When a sell signal is received, sell all the stocks. If buy signals appear successively, we keep the buying operation. If the cash-in-hand isn’t enough, we can add some to the total investment amount.

There are two differences between the two rules. First, the total investment amount of Rule One is fixed, while it is variable for Rule Two. Second, Rule One always uses the all cash to buy stocks, while Rule Two doesn’t.

4 EXPERIMENTS AND RESULTS

In this paper, nine stocks are chosen randomly according to the three trends (uptrend, steady, downtrend) from Shanghai and Shenzhen Stock Exchange. The codes are 000852, 600059, 600850, 600998, 600697, 600167, 600736, 600488 and 600019. The time span is from 2010.2.1 to 2012.9.28. For each stock, there are 550 days of data. From start to finish, the stocks whose closing price increased or decreased more than 10% are adjudged as uptrend or downtrend. If the floating rate of stock closing price is less than 10%, the stock is judge for a steady trend. Therefore, 000852, 600059, and 600850 are uptrend; 600998, 600697 and 600167 are steady; 600736, 600488 and 600019 are downtrend.

For each stock, the top ten chromosomes will be chosen by GA and the indicators which occur more than eight times are defined as Individual Feature Indicator (IFI) of the specific stock. Synthesizing the IFI of nine stocks and the frequency of occurrence of each indicator, those indicators whose probability of occurrence is larger than 60% will be defined as Common Feature Indicator (CFI).
Based on CFI and IFI, forecast accuracies can be calculated by PLR-GMKL and they are demonstrated in Table 2.

<table>
<thead>
<tr>
<th>CODE</th>
<th>INDIVIDUAL FEATURE INDICATOR</th>
<th>IFI ACC</th>
<th>ALL ACC</th>
<th>CFI ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>600059</td>
<td>ΔD(9) ΔMA(10) ΔMA(6) ΔRSI(6) ΔRSI(12) CRATP ALT AMA BIAS(5) MA(10) RSI(6)</td>
<td>69.05</td>
<td>52.38</td>
<td>60.90</td>
</tr>
<tr>
<td>000852</td>
<td>MA(12) MACD RSI(6) MA(50) BIAS(10) DMA AMA W&amp;R(12) CRTM CRATP ΔD(9)</td>
<td>61.54</td>
<td>44.23</td>
<td>50.92</td>
</tr>
<tr>
<td>600850</td>
<td>CP RSI(12) MA(6) BIAS(5) ALT CRTM CRATP ΔMA(10) ΔMA(12) NINEPERMA AMA</td>
<td>70.00</td>
<td>40.00</td>
<td>55.00</td>
</tr>
<tr>
<td>600167</td>
<td>MACD NINEPERMA MA(10) MA(12) K(9) D(9) DMA AMA CRATP ΔMA(6) ΔD(9)</td>
<td>72.31</td>
<td>55.38</td>
<td>43.08</td>
</tr>
<tr>
<td>600697</td>
<td>RSI(6) DMA W&amp;R(12) ALT ΔRSI(12) ΔRSI(6) ΔMA(5) ΔMA(12) ΔK(9) TV MA(50)</td>
<td>60.41</td>
<td>47.92</td>
<td>50.00</td>
</tr>
<tr>
<td>600908</td>
<td>RSI(12) ALT CRTM CRATP ΔRSI(12) ΔK(9) AMA ΔMA(6) ΔMA(20) NINEPERMA</td>
<td>63.59</td>
<td>29.17</td>
<td>47.83</td>
</tr>
<tr>
<td>600736</td>
<td>NINEPERMA RSI(6) MA(5) MA(10) MA(50) K(9) CRATP ΔMA(12) ΔBIAS(5) ΔW&amp;R(12)</td>
<td>64.58</td>
<td>37.50</td>
<td>50.00</td>
</tr>
<tr>
<td>600488</td>
<td>ΔW&amp;R ΔBIAS(10) ΔBIAS(5) ΔMA(20) ΔMA(6) W&amp;R K(9) MA(5) RSI(6) ΔMA(12) CRTM</td>
<td>63.08</td>
<td>53.33</td>
<td>52.31</td>
</tr>
</tbody>
</table>

Table 2. Prediction Accuracy for Different Indicator Selection

For each stock, the forecast accuracy of CFI (CFI ACC) is lower than the accuracy of IFI (IFI ACC) because IFI for a particular stock contains all the individual and common characteristics. CFI lose the individual information so the accuracy goes down. However, CFI ACC wins the accuracy of all indicators (ALL ACC) on seven stocks; loses on two. The average of CFI ACC is 52.24% while it is 46.40 for ALL ACC. So GA optimization can be used to extract CFI that reflect the common characteristic. CFI obviously reduces the noise and improve the accuracy in prediction.

In addition, to examine the effect of WNMS and GMKL algorithm in trading signal prediction, four kinds of tests were carried out. We respectively use GMKL and WSVM to classify the trading signals based on CFI and SCFI. SCFI means than the used indicators to predict include CFI and WNMS. The results were shown in Table 3.

<table>
<thead>
<tr>
<th>CODE</th>
<th>MODEL</th>
<th>ACC</th>
<th>Rule 1 profit</th>
<th>Rule 2 profit</th>
<th>BHS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SCFI</td>
<td>CFI</td>
<td>SCFI</td>
<td>CFI</td>
</tr>
<tr>
<td>600098</td>
<td>WSVM</td>
<td>25</td>
<td>33</td>
<td>1.82</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCFI</td>
<td>CFI</td>
<td>SCFI</td>
<td>CFI</td>
</tr>
<tr>
<td>600098</td>
<td>GMKL</td>
<td>55</td>
<td>47.83</td>
<td>-1.14</td>
<td>-1.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCFI</td>
<td>CFI</td>
<td>SCFI</td>
<td>CFI</td>
</tr>
<tr>
<td>600098</td>
<td></td>
<td>-20.26</td>
<td>-0.34</td>
<td>-12.71</td>
<td></td>
</tr>
</tbody>
</table>
We first compare the performance of GMKL and WSVM using the prediction results based on SCFI. There are 6 out of all 9 stocks whose prediction accuracies of GMKL are higher than that of WSVM. The average prediction accuracy of GMKL is 56.43%, and it is much higher than the average prediction accuracy of WSVM, 43.61%. As for the profits, there are 6 stocks whose profits based on the prediction result of GMKL algorithm are higher than profits based on the prediction result of WSVM in both two rules. When we apply Rule One on trading, the Modern Portfolio Strategy (MPS) profit of GMKL is 1.77%, and the MPS profit of SVM is -4.98%. The former performance is 6.75% higher than the latter one and wins the BHS strategy with 8.53%. When Rule Two is applied to MPS, the profit for GMKL is 0.72%, while it is -15.16% for WSVM. The Rule Two MPS is higher than that for BHS strategy by 7.48%.

When the prediction is based on CFI, there are 6 out of all 9 stocks whose prediction accuracies of GMKL outperform that of the WSVM. The average prediction accuracy of GMKL is 52.24% which exceeds the accuracy of WSVM (47.46%) by 4.78%. When the prediction result is applied to trading, GMKL performs better in 5 stocks. Based on Rule One, the MPS profit for GMKL is -1.64%, outperforming that of WSVM, which is -6.34%. And it is higher than the profit of BHS strategy with 5.12%. When we use Rule Two to trade, the MPS profit of GMKL is -0.72%, which is higher than that of WSVM (-18.20%).

We could also observe the effect of the WNMS to stocks prediction. We compare the prediction results based on GMKL to determine whether WNMS is counts as a predicting trading signal. After adding WNMS to the prediction process, prediction accuracies of 6 stocks improve. The average prediction accuracy of SCFI is 56.43% which is higher than CFI with 4.19%. When we apply the Rule One to trade, the MPS profit for SCFI, 1.77%, outperforms the MPS profit for CFI, -1.64%. Under the guidance of Rule Two, the MPS profit for SCFI is 0.72%, which is higher than the MPS profit of CFI by 15.95%. The MPS profitability of Rule One and Rule Two surpasses that of BHS strategy.

In addition, the effect of WNMS could be examined by comparing the trading signals. Figure 3 shows the comparison of the prediction results of trading signals for the 9 stocks based on Rule One. We can see that the WNMS is helpful for prediction, especially at the great turning. The prediction system with WNMS always captures the highest price to sale and then lowest price to buy, which indicates that the web news media sentiment could bring the information about the markets.
shortage of technical analysis which only takes information of the stock price into consideration. We take the 600697 for example. At the 63rd turning point in dataset, only the prediction system with WNMS can generate the buy point, while the prediction without WNMS can’t. This accounts for the phenomenon that the WNMS can bring higher profit.
5 CONCLUSION

We propose PLR-GMKL trading signals prediction system in this paper. It has been proved that this model effectively improves the prediction accuracy and increases the profit, comparing to other relative research. The innovation of this paper lies in as follows. First, through web mining tools, the web news media sentiment index is used as indicators to predict trading signals in the prediction system. The results show that the web news media sentiment indicators are beneficial to improve the prediction accuracy. Second, GMKL algorithm is used in the trading signals prediction for the first time. Comparing with WSVM, GMKL overcomes its weakness. And GMKL algorithm performs better in the prediction. Third, our paper applied the Modern Portfolio Strategy to the trading, based on the signals predicted by the system. The profit of the strategy is a better measurement for the prediction systems and more convincing, compared with the measurement where comparing the profit of investing on a single stock.

Our research can be improved in several aspects. In the analysis of web sentiment, we merely used WNMS. Without meticulous classification of sentimental degree, the research is confined to certain extent. In the following study, more related research could be carried out. Also, the stocks selected in the study are limited in both quantity and diversity. In the future, we could expand the set of stock that we observe and analyse. Moreover, this research is constrained to domestic stock market and the only news media source is Sina Finance and Economics Website stock timely rain sector news. In subsequent research, we could extend to the foreign financial markets.

Acknowledgement

This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 71001103, 91224008, 91324015), Beijing Social Science Fund (No. 13JGB035), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No. Z131101000413058), and Program for Excellent Talents in Beijing.

References


