Neurocomputing 115 (2013) 142–149

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Associating stock prices with web financial information time series based on support vector regression

Xun Liang [a,b,*], Rong-Chang Chen [c], Yangbo He [b], Ying Chen [b]

[a] School of Information, Renmin University of China, Beijing 100872, China
[b] Institute of Computer Science and Technology, Peking University, Beijing 100871, China
[c] Department of Distribution Management, National Taichung University of Science and Technology

ARTICLE INFO

Article history:
Received 24 June 2012
Received in revised form
21 January 2013
Accepted 27 January 2013
Communicated by: P. Zhang
Available online 26 February 2013

Keywords:
Web financial information time series
Stock price time series
Support vector regression

ABSTRACT

Over the past years, the effect of massive information from the web on the financial market has increased. How to process and utilize such information attracts both researchers and portfolio managers. In this paper, financial information obtained daily from the web is treated as a time series and then associated with stock price volatilities. First, six research topics on financial time series are outlined, namely, analyses of stock price time series ***P***, trading volume time series ***V***, web information time series ***W***, and relationship between ***P*** and ***V***, ***P*** and ***W***, as well as ***V*** and ***W***. Second, a model connecting ***P*** and ***W*** based on the support vector regression (SVR) is examined as an example of the six research topics. Third, given that a typically successful way of computer-based natural language processing is through the conduct of keyword analysis, the novel finance-computer time series ***W*** is explicitly defined in terms of financial keywords and is used in the present paper as the topic of investigation. The relationship between ***P*** and ***W*** is modeled using SVR. Because during the pre-web era people cannot manually and efficiently process image information from the newspapers and sounds from the television and radio over a longer period of time (e.g., a year), they were unable to obtain the time series ***W***. Therefore, it is the *web* that makes the research on the relationship between ***P*** and ***W*** in the meaning of quantity of ***W*** possible. Finally, experiments on the Shanghai and Shenzhen security markets revealed that the introduction of ***W*** helps improve model accuracies. As the web further develops, more and more ordinary people share their views on the web. The "long-tail" of massive financial information formed by these "grass roots" has a noticeable effect on the financial markets. In financial markets, those who quickly capture and interpret financial information have the potential to generate profits. With the use of the newly found model connecting ***P*** with ***W*** and fast decision making, financial market practitioners can be rewarded.

1. Introduction

1.1. Financial information plays an important role in security markets

The security market, which market participants often find to be complex, is one of the main venues for firms to generate funds. Over time, the security market experiences various volatilities that reflect negative and positive trends of stock prices of publicly traded firms; hence, as a whole, it is regarded as a national economic weatherglass. Consequently, a financial forecast is very important for finance-related government departments (e.g., the Securities and Exchange Commission), financial analysts, speculators, as well as investors.

There are many factors influencing the volatility of stock prices, including macroeconomic indices, national economic policy, and international financial climate among others. Stock prices also are influenced by microeconomic factors, such as firm behavior, earnings reports, and firm prospects. Various pieces of financial information jointly and to a large degree influence the judgments of stock market practitioners and the change in their buying and selling behaviors. Those who quickly capture massive financial information have the potential to generate profits. Thus, the function of financial information in a security market can never be overemphasized [1–3]. In the famous mixture distribution hypothesis [4], a very important assumption is that market information is the main reason behind price volatility and trading volume volatility [5–7]. According to the efficient market theory [8], financial information also plays a crucial role in affecting the volatility of the security market. Usually, on one hand, when the

* Corresponding author at: School of Information, Renmin University of China, Beijing 100872, China. Tel.: +86 10 82500549.
E-mail addresses: xliang@ruc.edu.cn, liangxun@pku.edu.cn, dsxl@yahoo.com (X. Liang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.011

volume of financial information is small, the volatility of stock prices is also small as well and the stock market is calm. On the other hand, when the stock volatility rate is large, the stock market often fluctuates [9]. In light of this, some scholars stated that the foundation of the security market is information.

Nevertheless, the stock market is a very complex system and changes in stock prices do not entirely depend on financial information. For example, behavioral finance has successfully developed a few hypotheses that topple the traditional efficient market theory [10]; notwithstanding that, in the daily stock market trading, financial information plays an indispensable role in moving the market.

The economic renovation in the 1980s gave birth to the first Chinese stock market in 1991. After two decades, the Chinese stock markets have advanced and matured through development and innovation. Now, investors more trade stocks on the basis of monitored information, such as macroeconomic indices, government policies, and market profits of firms, among others.

1.2. Sources of financial information

In terms of sources, financial information can be classified into television or radio financial information (sound carrier), newspaper financial information (paper carrier), web financial information (network carrier), and word-of-mouth financial information.

The functions of single financial information on television, radio and newspaper already have been widely investigated. For instance, some scholars investigated the effect of disclosure of profit information on stock markets, economic systems in different nations, and different microstructures of the stock market. In general, most articles involved a case study that explored the relationship between specific financial information and changes in stock prices. For example, in one case study, representative information on a group of similar events, such as data after a merger, was investigated. People did not go beyond the case study because it was impossible for them to watch *all* television programs, listen to *all* the radio programs or read *all* newspapers. In other words, what they studied was the relationship between information and market only in terms of *quality* of information. They did not probe on *quantity* of information.

1.3. Web financial information

With the rapid development of web technology and other communication techniques, data online has increased exponentially. Although the web is only one of the channels where investors can receive financial information, the important contents of the information are often reiterated across several media platforms.

Compared with traditional media, the web is a preferred tool to express ideas. Currently, more and more people around the world share information on the web. This makes up an undercurrent that may move the financial market. In particular, BBS, tweeter, blogs, and micro-blogs provide convenient ways for market practitioners to post opinions and influence each other. Because this information comes from a large volume of ordinary people or "grass roots," as opposed to the views of market authorities, it forms a noticeable "long-tail effect" on the market.

Unlike financial information obtained from conventional media platforms, web financial information possesses the following unique features. First, web content is real time. The webmaster can add information any time he or she wants. In contrast, most newspapers publish news on a daily basis, and many television or radio programs have pre-determined timetables. The web information can report the same information more rapidly than other channels. Second, the content range of the web is large. The web provides not only official news but also comments posted by a large number of market participants, such as the "grass roots" as well. Third, the web allows interactivity because people can influence each other's opinions. Fourth, the web can publish information simultaneously. Furthermore, using the web, readers can choose the information they want to read. They do not have to read information they are not interested in. This is in contrast with television and radio programs, from which viewers and listeners are forced to take even information they are not interested in. Fifth, the web has an immense and virtually unlimited space. This is in contrast with a newspaper that has limited space, and television and radio programs that have limited air times. Sixth, the web media offers volume variability. The volume of information on the web changes depending on the number of daily events. Comparatively, newspapers are constrained by column space, while television and radio programs are constrained by durations no matter how much volume of the news during the day is. Seventh, the web provides a feature of modifiability. On the contrary, published in newspapers and aired in television and radio programs cannot be modified. On the contrary, information on the web can be updated, as often as the webmaster wants. Eighth, the publication time interval granularity of daily newspaper is a day. However, the information on the web can be harvested with timestamps and thus in the following text mining the obtained information can further award us any time granularity theoretically such as a day, hour, or minute. So it is the *web* that makes possible that stock prices can be associated with web financial information time series in adjustable time granularities.

Recently, more and more research papers were published pertaining to web financial information mining. In one of those studies, articles on stock prices were being posted online on a daily basis. Results showed that the number of words of the articles was positively correlated with stock price movements [3].

1.4. Web financial information time series

Let $\boldsymbol{W}=\{W_t: t=0, 1, 2,\dots\}$, where W_t is an index reflecting the web financial information time series. In applications, W_t can be the number of financial keywords on the web at time t, web financial word score at time t, financial text sentiment value at time t, or entropy-based financial text sentiment value at time t. Clearly, this finance-computer time series $\boldsymbol{W}$ itself is a novel time series and is "aligned" with the corresponding stock price time series along the time axis.

Compared with the stock price time series $\boldsymbol{P}=\{P_t: t=0, 1, 2,\dots\}$ (where P_t is the stock price at time t), $\boldsymbol{W}$ has a large degree of self-correlation in volatility. For example, one piece of leading positive stock information tends to be followed by more positive stock information in the next few days. On the contrary, a large increase in stock price happening in one day does not necessarily lead to a large increase in the next day. In other words, $\boldsymbol{W}$ has more self-correlation than $\boldsymbol{P}$.

Manually collecting and processing a lot of financial information on the newspaper or television or radio on a daily or hourly basis over a longer period of time (e.g., a year) is difficult. To process financial information from a newspaper, the meaning of words for scanning should first be recognized and understood using image recognition technology. To process financial information on television or radio, speech recognition in acoustics must be first conducted. As a result, over the years, although scholars investigated the relationship between $\boldsymbol{W}$ and $\boldsymbol{P}$, the shortage of data on $\boldsymbol{W}$ has prevented the success of this endeavor.

With the aid of natural language processing technology in computer science, computers can recognize the meaning of news on the web to some extent. For computers, the complexity of

processing n pieces of information is almost the same as that of processing one piece of information. Over the past several decades, financial experts have investigated the relationship between stock prices and financial information. However, most of their efforts covered only single financial information, which was opposed to ***W***, because of the difficulty in obtaining sufficient data on ***W***.

Since the birth of the web, gathering data ***W*** has become easy. On the web, words are conveyed in electronic digits. Therefore, the above mentioned recognition work is no longer needed. A harvesting computer server can be hired to gather financial information directly from these websites. Conclusively, it is the *web* that allows people to conveniently investigate the quantity of financial information, such as the form of time series, as well as its relation with the stock market in a relatively longer period of time. In this paper, the relationship between ***W*** and ***P***, in terms of time series, is investigated.

1.5. Six research topics with financial information time series *W*

In theory, ***W***, ***P***, and $\boldsymbol{V}=\{V_t: t=0, 1, 2,\ldots\}$ (where V_t is the stock trading volume at time t) are the three important financial time series. In Fig. 1, there are six relationships illustrated by three solid lines and three dotted lines.

For the three solid lines, scholars have been studying the ***P***–***V*** relation, the properties of ***P***, and the ***V*** self-relation for many years, and have gathered many results [11–18]. For example, [12] demonstrated that from the point of view of information economics, the information of the stock market is positively asymmetrically related with prices. In [13], the ***P***–***V*** relation with a smaller hourly granularity has been investigated.

Because each of the three uncultured relations involves huge volumes of work, this paper addresses only the ***P***–***W*** relation among the three relations with dotted lines in Fig. 1.

Furthermore, because keywords serve as one of the most important texts for financial information ***W***, financial keywords are employed in this paper for simplicity. In particular, the information sentiment values [19] based on the above-mentioned financial keywords are used in forming ***W***.

1.6. Modeled with support vector regression

In the past, machine learning methods in computer technology have demonstrated great successes in forecasting time series [20–23]. Machine learning methods, which include neural networks, and support vector regression (SVR), normally have the capacity to implement any nonlinear relationships. In recent years, the SVR method has received increasing attention. Its advantages include the absence of local minima and the optimal separation or largest margin width between two clusters, which is achieved simply by solving linearly constrained quadratic programming problems. In this paper, the SVR-based tool is used to build the joint model of ***P*** and ***W***.

Being awarded with the tool that is incorporated within the new connection between ***P*** and ***W***, a financial market practitioner has the potential of making more profits in the market.

1.7. Organization of the paper

The remainder of the paper is organized as follows. Section 2 outlines the process of obtaining ***W*** and a forecast model based on SVR. Section 3 presents the conduct of experiments. Section 4 concludes the paper.

2. A model with financial information by SVR

2.1. Web information harvesting and preprocessing with word segmentations

Financial information in this study was collected from the main websites in China via special crawlers. These websites include http://finance.sina.com.cn, http://business.sohu.com, http://www.cnlist.com, http://www.cfi.net.cn, http://www.jrj.com.cn, http://www.hexun.com, http://www.cs.com.cn, http://www.szse.cn, http://www.sse.com.cn. Considering that the duplicated news contents on different websites increase the number of readers and may influence the market, the duplicated news was not removed.

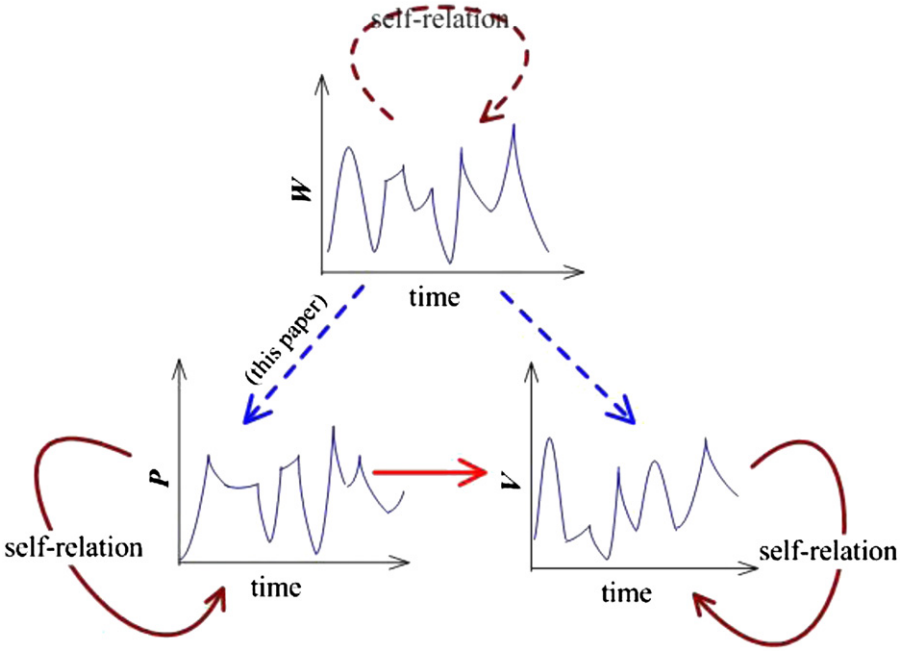


Fig. 1. The relationships among ***P***, ***V***, and ***W***.

The Chinese language has special characteristics that make its processing challenging and intellectually rewarding. While the written English text has word boundaries where all the sentences are based on the word as a unit, and words are separated by spaces, the Chinese text does not have word boundaries. In Chinese, all words are connected together without spaces. Moreover, in Chinese language processing, the Chinese word segmentation into word phrases is an unavoidable pre-processing step. The software we used is the ICTCLAS segmentation system, coded by the Chinese Academy of Sciences (http://www.cnblogs.com/riky/archive/2007/03/09/669340.html). Based on the word segmentation results, those function words and only the notional words, such as nouns, pronouns, adjectives, and adverbs, were kept. Assume that after this process, the word series is w_1, w_2, w_3,…, w_n with n as the number of words. The next steps are the same as in English text processing [2,3].

2.2. Automatically picking out keywords

After word segmentation, the value of $tfidf = tf \times idf$ is computed for every word [24,25], where $tf = freq(p, D)/\text{size}(D)$, $idf = -\log(df(p)/m)$, freq(p, D) denotes the occurring frequency of word p in webpage D, size(D) is the number of all the words in webpage D, $df(p)$ is the total number of web pages containing word p, and m is the total number of web pages.

The use of *tfidf* helps in the selection of the keywords from each article. The values of *tfidf* were placed in order from the largest to the smallest for all the n words, and the first n' words were obtained as the keywords of this article, w_1, w_2, w_3,…,$w_{n'}$, where n' is assumed to be the pre-assigned number of keywords. If the total number of words is smaller than n', all the words are regarded as keywords. For convenience, this smaller number is still written as n'.

2.3. Selecting financial keywords and assigning them with sentiment values

Next, we discover the finance-related keywords. We built a financial dictionary in advance by collecting about 4200 financial words manually. In the dictionary, each financial word has a sentiment value, represented by a number between -1 and 1. The above n' keywords, w_1, w_2, w_3,…,$w_{n'}$, were assigned sentiment values based on the dictionary and the words that were not in the dictionary were removed. As a result, the remaining keywords were financial keywords. Assume that after this step, the word group included w_1, w_2, w_3,…,$w_{n''}$ with n'' as the number of financial words in this article. Note that n'' may vary from article to article.

2.4. Determining the sentiment values of the article and the sentiment value of a day

For each article, all the sentiment values of the $w_{n''}$ financial keywords were added and the sentiment values for the article were obtained. Moreover, adding up the sentiment values of all the articles on day t led to the sentiment value W_t.

2.5. Associating stock price time series with financial keyword time series

P_t is modeled by

$$P_t = \sum_{i=1}^{r} a_{t-i} f(P_{t-i}) + \sum_{i=1}^{q} b_{t-i} g(W_{t-i}) \tag{1}$$

where a_t is the coefficient, r and q are the numbers of traced days or sliding window lengths, and f and g are nonlinear functions.

In this paper, the SVR model is employed for the implementation of the nonlinearities (see Fig. 2),

$$\max_{\alpha_1,\ldots,\alpha_l} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^{\sim})(\alpha_j-\alpha_j^{\sim})K(x_i,x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^{\sim}) - \sum_{i=1}^{l}P_i(\alpha_i-\alpha_i^{\sim}),$$

$$\text{s.t.} \sum_{i=1}^{l}\alpha_i = \sum_{i=1}^{l}\alpha_i^{\sim},$$

$$0 \le \alpha_i, \alpha_i^{\sim} \le C, \quad i = 1,\ldots,l, \tag{2}$$

where l is the number of training data, $K(\bullet, \bullet)$ is the kernel function, ε is a given small positive parameter, C is a given positive parameter, and α_i and $\alpha_i^{\sim}$ are the Lagrangian multipliers. In addition, the formula $x_i = (P_{i-r},\ldots,P_{i-1}, W_{i-q},\ldots,W_{i-1})^{\mathrm{T}}$ is set with $(\bullet)^{\mathrm{T}}$ as the transpose of $(\bullet)$. The regression function is

$$\sum_{i=1}^{s}(\alpha_i^{\sim *}-\alpha_i^{*})K(x_i,x)+b^{*} \tag{3}$$

where s is the number of support vectors, α_i^{*} and $\alpha_i^{\sim *}$ are the optimal Lagrangian multipliers and b^{*} is the optimal offset.

3. Experiments

3.1. Computing the sentiment value time series for financial information

In the step of finding n' keywords w_1, w_2, w_3,…,$w_{n'}$, we let $n' = 30$ in the experiments. If the keywords obtained were fewer than n', all the obtained keywords were used. By finding the sentiment value for each keyword in the financial dictionary and adding the values together, the sentiment value for each article was obtained. Finally, after combining all the sentiment values on day t, the overall sentiment value W_t was computed (see Table 1 for a portion of the series).

There are three steps for preprocessing W_t's.

First, if t is not a trading day, there is no P_t. In this case, we added W_t to the next nearest trading day after day t. Say that $t+j$ is such nearest trading day, then $W_{t'+j} = \sum_{k=t}^{t+j-1} W_k$, and $W_{k'} = 0$, $k = t,\ldots,t+j-1$. Moreover, the days without data in experiments were skipped. For convenience of narration, t=2010-01-04,…,2012-09-28 are still written as t=2010-01-01,…,2012-09-30.

Second, it was observed that there were some large absolute values of W_t', whereas most of W_t' were within a moderate region of values. To adequately lower these values, a function $(\bullet)^{1/3}$ was set for W_t'. That is, $W_{t''} = (W_{t'})^{1/3}$. Clearly, function $(\bullet)^{1/3}$ actually

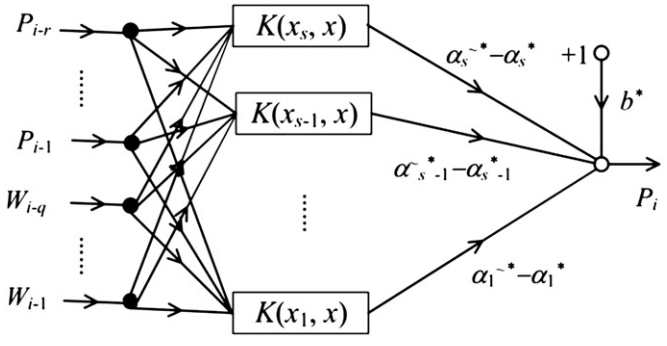


Fig. 2. The SVR architecture.

Table 1
An excerpt of the sentiment values for financial information from 2010-01-01 to 2012-09-30.

Date	Overall sentiment values
…	…
2012-06-29	−190.1
2012-06-28	−370.8
2012-06-27	−173
2012-06-26	312.1
2012-06-25	630.6
…	…
2011-12-20	110
2011-12-19	166.8
2011-12-18	13.4
2011-12-17	−25.4
2011-12-16	−134.8
2011-12-15	−102
…	…
2011-03-31	−133.9
2011-03-30	−270.6
2011-03-29	−536.2
2011-03-28	44.3
2011-03-27	9.2
2011-03-26	34.1
2011-03-25	−7.2
…	…
2010-01-22	−53.4
2010-01-21	77.8
2010-01-20	279.6
2010-01-19	20.2
2010-01-18	−119.2
…	…

sets a soft upper limit for W_t' (see Fig. 3). Evidently, without this step, a lot of small values of W_t' might be inappropriately ignored in computation because of several unusually large $|W_t'|$.

Third, normalization of $W_t'' \in (-\infty, +\infty)$ over the whole time horizon t=2010-01-01 to 2012-09-30 was performed, with $W_t''' = 2\ (W_t'' - W_{\min}'')/(W_{\max}'' - W_{\min}'') - 1 \in [-1, +1]$. For convenience, W_t''' is rewritten as W_t.

3.2. Shanghai and Shenzhen Securities Composite Indices

The closing prices of Shanghai Securities Composite Index from 2010-01-01 to 2012-09-30 (see Table 2 for a portion) were used for demonstration. The closing prices of Shenzhen Securities Composite Index from the same period were also studied.

As a preprocessing before training SVR, a normalization was also performed with $P_t' = (P_t - P_{\min})/(P_{\max} - P_{\min})$. When one needs the actual stock prices, a reverse operation can be conducted, $(P_{\max} - P_{\min})\, P_t' + P_{\min}$. For convenience, P_t' is rewritten as P_t in the following.

3.3. Training with LibSVM

The software used for training SVR is the LibSVM software platform [26]. LibSVM is a widely used software for SVR research. Apart from incorporating the efficient sequential minimal optimization, the developers have made many improvements, particularly in the past years when many more remarkable advances on this software have been achieved [27–30].

SVR allows the researchers to choose many different kernels, such as polynomial (POLY) kernels, and radial basis (RB) kernels. The following five kernels were experimented, namely, three POLY kernels $(a\, x_i^{\mathrm{T}} x + 1)^k$ with $(a, k) = (0.1, 2), (0.1, 3)$, and $(0.01, 2)$, and two RB kernels $\exp(-a||x_i - x||^2)$ with $a = 0.1$ and 0.01.

The training and predicting tasks were conducted on a rolling basis. In the early morning of day $t+1$ before the market opened, an SVR was trained using the l pairs of data, $\{(P_{t-r}, \ldots, P_{t-1}, W_{t-q}, \ldots, W_{t-1})^{\mathrm{T}}, P_t\}$, $\{(P_{t-r-1}, \ldots, P_{t-2}, W_{t-q-1}, \ldots, W_{t-2})^{\mathrm{T}}, P_{t-1}\}, \ldots, \{(P_{t-r-l+1}, \ldots, P_{t-l}, W_{t-q-l+1}, \ldots, W_{t-l})^{\mathrm{T}}, P_{t-l+1}\}$, and the trained SVR and input $(P_{t-r+1}, \ldots, P_t, W_{t-q+1}, \ldots, W_t)^{\mathrm{T}}$ were used to predict $P_{t+1}^{\sim}$. Then in the afternoon of day $t+1$ after the market closed, the actual price P_{t+1} on day $t+1$ was known. Therefore, when the researchers wanted to predict the values on day $t+2$, a new SVR training was initiated by including the actual price P_{t+1}. The new training used the new l pairs of data $\{(P_{t-r+1}, \ldots, P_t, W_{t-q+1}, \ldots, W_t)^{\mathrm{T}}, P_{t+1}\}$, $\{(P_{t-r}, \ldots, P_{t-1}, W_{t-q}, \ldots, W_{t-1})^{\mathrm{T}}, P_t\}, \ldots, \{(P_{t-r-l+2}, \ldots, P_{t-l+1}, W_{t-q-l+2}, \ldots, W_{t-l+1})^{\mathrm{T}}, P_{t-l+2}\}$, and should be performed in the early morning of day $t+2$ before the market opened. The newly trained SVR and input $(P_{t-r+2}, \ldots, P_{t+1}, W_{t-q+2}, \ldots, W_{t+1})^{\mathrm{T}}$ were used for predicting $P_{t+2}^{\sim}$.

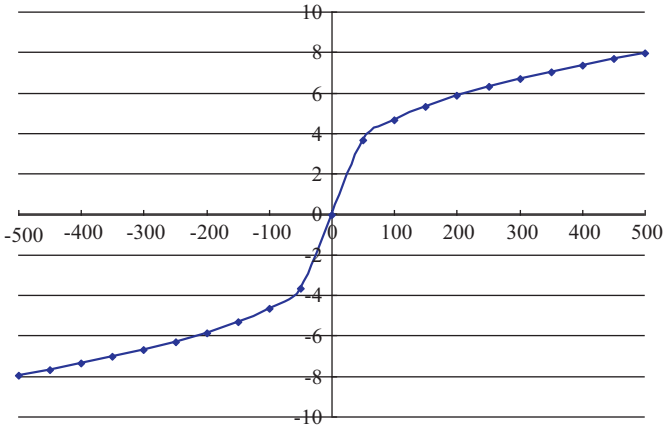


Fig. 3. The soft upper limit function $(\bullet)^{1/3}$.

With $r = 1, \ldots, 5$, $q = 1, \ldots, 4$, $l = 8, 9, 10, 11$ and with five different kernels, a total of 400 experiments were conducted.

Let $R_t = |P_t^{\sim} - P_t|/P_t$ be the relative predicting error. Subsequently, the mean predicting error μ_t and the standard deviation σ_t of R_t were evaluated for each experiment.

In the experimental studies, we noticed that the values with the information delay q=2 frequently provided the most satisfactory results. The mean predicting errors and standard deviations over different r, l, and kernels for the Shanghai Securities Composite Index are shown in Table 3. From Table 3, the two compromises were reached: (1) r=2, q=2, RB (a=0.1, k=3), and l=9 led to the smallest mean error 2.11% and moderate standard deviation 1.81%; and (2) r=2, q=2, RB (a=0.01), and l=9 also led to the same smallest mean error 2.11% and moderate standard deviation 1.25%. The mean predicting errors and standard deviations over different r, l, and kernels for the Shenzhen Securities Composite Index are shown in Table 4. From Table 4, a compromise, r=3, q=2, RB (a=0.1), and l=10, was reached and led to the smallest mean error 2.05% and moderate standard deviation 1.37%. Clearly, the prediction precisions were similar between the two markets. Additionally, the experimental results showing that q=2 frequently gave the best performances indicate that two days in the past seemed an adequate time delay for the information to influence the stock prices. The time delays shorter or longer than this period both produced larger mean predicting errors and standard deviations.

Predicting errors were also compared when SVR were trained with inputs only as past stock indices, namely, the l pairs of training data were $\{(P_{t-r}, \ldots, P_{t-1}), P_t\}$, $\{(P_{t-r-1}, \ldots, P_{t-2}), P_{t-1}\}, \ldots, \{(P_{t-r-l+1}, \ldots, P_{t-l}), P_{t-l+1}\}$. Similar experimental processes for the Shanghai Securities Composite Index showed that r=2, q=2, RB (a=0.1), and l=10 led to the best results with mean predicting error 3.08% and standard deviation 1.92%. For the Shenzhen Securities Composite Index, results showed that r=2,

Table 2
An excerpt of the Shanghai and Shenzhen Securities Composite Indices from 2010-01-01 to 2012-09-30.

Date	Shanghai Securities Composite Index					Shenzhen Securities Composite Index				
	Open	Close	Change (%)	Low	High	Open	Close	Change (%)	Low	High
2012-09-28	2042.89	2086.17	2.12	2040.35	2089.63	8422.29	8679.22	3.05	8415.04	8690.24
2012-09-27	2003.1	2056.32	2.66	2002.32	2068.07	8198.56	8486.27	3.51	8198.56	8520.19
2012-09-26	2027.86	2004.17	−1.17	1999.48	2033.04	8289.85	8193.37	−1.16	8178.67	8335.14
2012-09-25	2028.63	2029.29	0.03	2021.81	2039.1	8296.95	8284.25	−0.15	8256.62	8364.69
2012-09-24	2015.77	2033.19	0.86	2005.26	2040.85	8148.37	8310.94	2.00	8095.52	8364.31
…	…	…	…	…	…	…	…	…	…	…
2012-04-26	2408.56	2404.7	−0.16	2393.26	2414.75	10182.46	10201.65	0.19	10152.05	10258.57
2012-04-25	2382.21	2406.81	1.03	2376.63	2411.42	10012.85	10162.01	1.49	9993.69	10182.09
2012-04-24	2380.24	2388.83	0.36	2350.4	2415.75	10010.35	10041.89	0.32	9841.44	10180.08
2012-04-23	2403.52	2388.59	−0.62	2383.07	2411.51	10142	10076.05	−0.65	10034.25	10154.51
2012-04-20	2374.66	2406.86	1.36	2372.13	2407.29	10004.28	10131.04	1.27	9996.72	10144.54
…	…	…	…	…	…	…	…	…	…	…
2011-12-09	2315.51	2315.27	−0.01	2309.46	2331.26	9516.25	9480.27	−0.38	9439.84	9583.15
2011-12-08	2329.91	2329.82	0.00	2302.64	2346.65	9603.14	9580.52	−0.24	9477.13	9712.14
2011-12-07	2325.56	2332.73	0.31	2317.81	2339.85	9582.38	9606.25	0.25	9551.21	9656.23
2011-12-06	2326.66	2325.9	−0.03	2310.15	2331.89	9568.87	9586.94	0.19	9505.22	9618.41
2011-12-05	2363.11	2333.23	−1.26	2327.61	2363.13	9805.45	9586.08	−2.24	9572.02	9807.69
…	…	…	…	…	…	…	…	…	…	…
2011-06-14	2696.22	2730.04	1.25	2691.67	2735.55	11501.22	11732.04	2.01	11495.11	11765.12
2011-06-13	2687.65	2700.38	0.47	2669.4	2703.42	11518.06	11521.98	0.03	11427.15	11558.66
2011-06-10	2696.13	2705.14	0.33	2672.32	2707.74	11508.58	11593.38	0.74	11454.04	11596.58
2011-06-09	2743.54	2703.35	−1.46	2702.7	2747.33	11725.21	11513.2	−1.81	11508.39	11735.21
2011-06-08	2742.04	2750.29	0.30	2715.25	2754.12	11726.33	11761.1	0.30	11581.64	11787.12
…	…	…	…	…	…	…	…	…	…	…
2010-11-18	2855.61	2865.45	0.34	2827.18	2873.7	12056.84	12146.15	0.74	11926.15	12181
2010-11-17	2852.48	2838.86	−0.48	2824.12	2891.05	12040.66	11917.49	−1.02	11849.9	12196.6
2010-11-16	3007.64	2894.54	−3.76	2885.75	3007.64	12822.51	12217.27	−4.72	12146.66	12822.51
2010-11-15	2984.87	3014.41	0.99	2939.68	3016.13	12772.85	12808.09	0.28	12558.21	12877.38
2010-11-12	3121.92	2985.44	−4.37	2975.16	3150.15	13594.88	12726.54	−6.39	12702.95	13658.29
…	…	…	…	…	…	…	…	…	…	…
2010-01-08	3177.26	3196	0.59	3149.02	3198.92	13180.9	13267.44	0.66%	13040.34	13270.63
2010-01-07	3253.99	3192.78	−1.88	3176.71	3268.82	13514.64	13235.48	−2.07	13148.25	13571.35
2010-01-06	3277.52	3254.22	−0.71	3253.04	3295.87	13509.39	13505.18	−0.03	13456.29	13677.4
2010-01-05	3254.47	3282.18	0.85	3221.46	3290.51	13539.83	13517.38	−0.17	13324.56	13597.36
2010-01-04	3289.75	3243.76	−1.40	3243.32	3295.28	13766.1	13533.54	−1.69	13533.54	13782.81

Table 3
With $q=2$, the mean predicting errors μ_t and standard deviations σ_t with different r, l and kernels for the Shanghai Securities Composite Index.

r	Kernels	Mean predicting errors $\mu_t \pm$ standard deviations σ_t			
		$l=8$	$l=9$	$l=10$	$l=11$
1	POLY ($a=0.1$, $k=2$)	2.32% ± 1.22%	2.86% ± 1.89%	2.62% ± 1.91%	3.70% ± 2.10%
	POLY ($a=0.1$, $k=3$)	3.08% ± 1.94%	3.84% ± 2.60%	3.37% ± 1.72%	5.03% ± 2.65%
	POLY ($a=0.01$, $k=2$)	3.19% ± 1.66%	3.41% ± 1.66%	3.05% ± 1.98%	4.52% ± 2.08%
	RB ($a=0.1$)	2.48% ± 1.51%	2.54% ± 1.53%	2.54% ± 1.16%	2.31% ± 1.68%
	RB ($a=0.01$)	2.42% ± 2.12%	3.40% ± 1.05%	2.18% ± 1.23%	2.82% ± 1.52%
2	POLY ($a=0.1$, $k=2$)	2.13% ± 1.96%	2.50% ± 1.20%	3.08% ± 1.69%	2.31% ± 1.40%
	POLY ($a=0.1$, $k=3$)	2.61% ± 1.49%	**2.11**% ± **1.81**%	4.28% ± 2.66%	2.59% ± 1.68%
	POLY ($a=0.01$, $k=2$)	3.22% ± 2.09%	3.28% ± 1.30%	2.51% ± 1.39%	4.46% ± 2.48%
	RB ($a=0.1$)	2.86% ± 1.76%	2.46% ± 1.56%	3.13% ± 1.31%	2.95% ± 1.39%
	RB ($a=0.01$)	2.40% ± 1.74%	**2.11**% ± **1.25**%	2.27% ± 1.33%	2.37% ± 1.60%
3	POLY ($a=0.1$, $k=2$)	3.33% ± 2.10%	3.29% ± 1.59%	2.30% ± 1.26%	2.19% ± 2.22%
	POLY ($a=0.1$, $k=3$)	3.44% ± 1.31%	2.87% ± 1.58%	4.20% ± 2.07%	4.37% ± 2.60%
	POLY ($a=0.01$, $k=2$)	4.55% ± 1.39%	3.21% ± 1.97%	2.21% ± 1.06%	4.50% ± 1.33%
	RB ($a=0.1$)	3.03% ± 1.40%	2.55% ± 1.32%	3.00% ± 1.85%	3.67% ± 2.28%
	RB ($a=0.01$)	2.63% ± 1.27%	2.12% ± 1.72%	3.18% ± 1.84%	2.71% ± 1.62%
4	POLY ($a=0.1$, $k=2$)	5.32% ± 2.66%	4.96% ± 2.91%	5.13% ± 1.98%	4.72% ± 2.15%
	POLY ($a=0.1$, $k=3$)	5.81% ± 2.46%	5.83% ± 3.63%	5.54% ± 2.14%	5.86% ± 2.64%
	POLY ($a=0.01$, $k=2$)	3.43% ± 2.00%	3.42% ± 2.11%	2.92% ± 1.49%	2.46% ± 1.48%
	RB ($a=0.1$)	2.69% ± 1.66%	3.06% ± 1.74%	2.56% ± 1.84%	2.92% ± 1.88%
	RB ($a=0.01$)	2.28% ± 1.50%	2.40% ± 1.52%	2.52% ± 1.18%	3.26% ± 2.53%
5	POLY ($a=0.1$, $k=2$)	5.51% ± 4.07%	5.58% ± 3.50%	5.34% ± 2.88%	5.44% ± 2.53%
	POLY ($a=0.1$, $k=3$)	4.88% ± 3.73%	4.36% ± 1.52%	4.17% ± 2.95%	4.56% ± 2.49%
	POLY ($a=0.01$, $k=2$)	2.59% ± 2.41%	3.38% ± 1.86%	2.73% ± 1.63%	3.84% ± 2.10%
	RB ($a=0.1$)	2.37% ± 1.82%	3.23% ± 1.72%	2.33% ± 1.67%	3.02% ± 1.56%
	RB ($a=0.01$)	2.41% ± 1.80%	3.18% ± 1.86%	3.68% ± 1.94%	2.75% ± 1.20%

$q=2$, RB ($a=0.01$), and $l=11$ led to the best results with mean predicting error 3.91% and standard deviation 2.43%. An overall look at the results may lead to the conclusion that the predicting errors increased slightly compared with the counterparts in Tables 3 and 4.

A portion of values predicted purely through the use of stock index time series and the one that involved additional financial information time series can be visualized in Fig. 4. It may be concluded that the addition of the financial information time series generally improved the accuracy of prediction.

4. Conclusions

By defining a novel finance-computer time series ***W***, this paper describes a web information-based triangle framework for conducting studies on quantitative relations among ***P***, ***V***, and ***W***, and further explores the relationship between ***P*** and ***W*** quantitatively using a model based on nonlinear SVR. Comparing with the experimental results purely based on ***P***, the mean predicting errors forecasted by ***P*** and ***W*** were obviously reduced. The addition of ***W*** thereof enhanced the forecasting performance.

In theory, the investigation provides an insightful understanding of the functions of financial information time series from a new angle, and it can be theoretically treated as a new and promising way to probe into complex problems in security markets.

As more and more ordinary people around the world share information on the web, the "long-tail" formed by these "grass roots" has a noticeable effect on the market. The rapid responses of computer processing of financial information can help financial managers quickly find the aggregated effect of opinions from these "grass roots," as well as other enormous web financial information. The bridge we built in this paper between ***P*** and ***W*** containing the views of "grass roots" offers a novel clue to ***P*** from the ***W*** side for financial analysts, and enables them to rapidly

Table 4
With $q=2$, the mean predicting errors μ_t and standard deviations σ_t with different r, l and kernels for the Shenzhen Securities Composite Index.

r	Kernels	Mean predicting errors $\mu_t \pm$ standard deviations σ_t			
		$l=8$	$l=9$	$l=10$	$l=11$
1	POLY ($a=0.1$, $k=2$)	2.62% ± 1.40%	3.01% ± 2.11%	3.46% ± 2.09%	4.06% ± 2.44%
	POLY ($a=0.1$, $k=3$)	3.14% ± 2.12%	3.88% ± 2.65%	4.19% ± 1.78%	3.26% ± 2.11%
	POLY ($a=0.01$, $k=2$)	3.81% ± 1.92%	3.51% ± 1.67%	3.52% ± 1.79%	3.35% ± 2.35%
	RB ($a=0.1$)	3.34% ± 1.96%	2.90% ± 1.77%	3.02% ± 1.55%	2.73% ± 1.70%
	RB ($a=0.01$)	3.26% ± 2.27%	4.33% ± 1.33%	2.21% ± 1.55%	2.12% ± 2.33%
2	POLY ($a=0.1$, $k=2$)	2.49% ± 2.34%	3.40% ± 1.42%	3.13% ± 1.74%	2.58% ± 2.53%
	POLY ($a=0.1$, $k=3$)	2.98% ± 1.62%	2.69% ± 2.13%	3.26% ± 1.63%	3.04% ± 1.78%
	POLY ($a=0.01$, $k=2$)	3.33% ± 2.81%	3.61% ± 1.23%	4.39% ± 2.07%	3.28% ± 2.13%
	RB ($a=0.1$)	3.50% ± 2.91%	3.82% ± 1.79%	2.97% ± 2.19%	2.39% ± 1.98%
	RB ($a=0.01$)	3.87% ± 2.27%	3.09% ± 2.14%	2.19% ± 1.96%	2.70% ± 1.82%
3	POLY ($a=0.1$, $k=2$)	3.28% ± 1.83%	4.96% ± 2.78%	4.03% ± 2.16%	3.39% ± 2.39%
	POLY ($a=0.1$, $k=3$)	3.61% ± 1.34%	2.64% ± 1.72%	3.48% ± 2.14%	3.59% ± 1.66%
	POLY ($a=0.01$, $k=2$)	2.92% ± 1.97%	3.48% ± 2.11%	3.54% ± 1.95%	2.98% ± 1.73%
	RB ($a=0.1$)	3.02% ± 1.54%	2.88% ± 1.76%	**2.05**% ± **1.37**%	2.24% ± 1.65%
	RB ($a=0.01$)	3.27% ± 1.07%	3.80% ± 1.82%	2.71% ± 1.27%	2.39% ± 2.74%
4	POLY ($a=0.1$, $k=2$)	4.42% ± 1.76%	4.88% ± 1.87%	3.05% ± 1.21%	4.76% ± 2.64%
	POLY ($a=0.1$, $k=3$)	3.50% ± 2.85%	3.63% ± 2.23%	3.36% ± 2.78%	4.14% ± 2.20%
	POLY ($a=0.01$, $k=2$)	3.30% ± 2.01%	3.02% ± 2.05%	2.96% ± 2.09%	3.81% ± 2.02%
	RB ($a=0.1$)	2.58% ± 2.28%	3.28% ± 2.08%	3.10% ± 1.48%	3.22% ± 1.97%
	RB ($a=0.01$)	2.24% ± 2.40%	2.06% ± 1.37%	3.06% ± 1.51%	2.28% ± 1.60%
5	POLY ($a=0.1$, $k=2$)	3.51% ± 1.99%	2.88% ± 1.71%	3.32% ± 1.80%	4.70% ± 1.58%
	POLY ($a=0.1$, $k=3$)	3.39% ± 1.87%	3.06% ± 1.58%	2.35% ± 1.37%	2.92% ± 1.82%
	POLY ($a=0.01$, $k=2$)	2.86% ± 2.49%	3.44% ± 1.89%	2.98% ± 1.51%	2.40% ± 2.38%
	RB ($a=0.1$)	3.44% ± 1.42%	3.76% ± 1.76%	2.88% ± 2.04%	2.12% ± 2.98%
	RB ($a=0.01$)	2.83% ± 1.46%	3.45% ± 2.21%	2.69% ± 1.22%	2.34% ± 1.69%

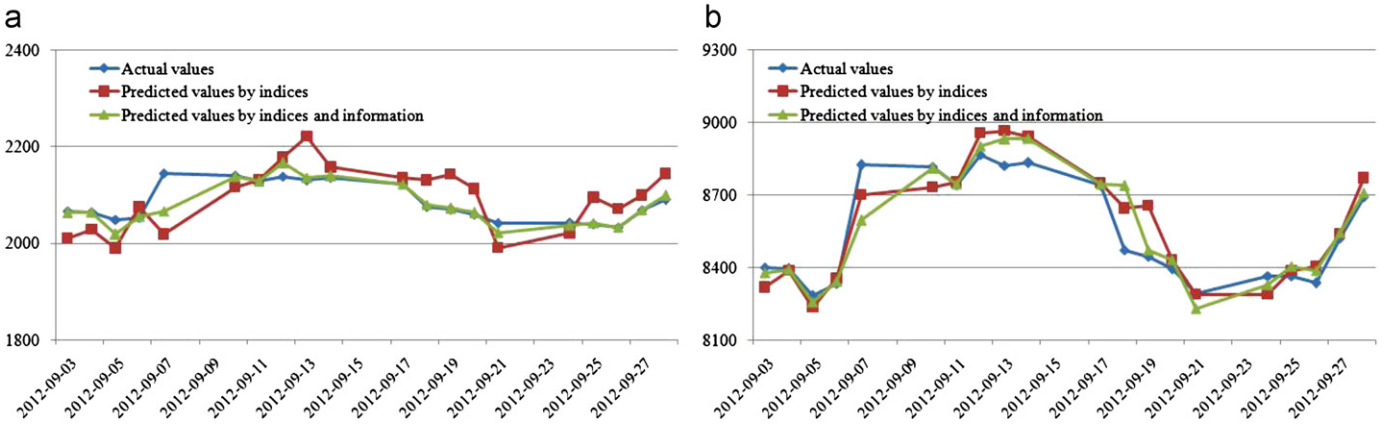


Fig. 4. The actual values, the predicted values given by the indices, and the predicted values given by a combination of indices and information from 2012-09-01 to 2012-09-30. (a) The Shanghai Security Market and (b) The Shenzhen Security Market.

control the risks of their portfolios against financial market and subsequent changes.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the paper greatly. The project was sponsored by the NSF of China under grant numbers 70871001 and 71271211.

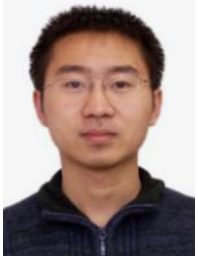
References

[1] W. Antweiler, M.Z. Frank, Is all that talk just noise? The news content of web stock message boards, J. Financ. 59 (2004) 1259–1294.
[2] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Making words work: using financial text as a predictor of financial events, Decis. Support Syst. 50 (2010) 164–175.
[3] S.W.K. Chan, J. Franklin, A text-based decision support system for financial sequence prediction, Decis. Support Syst. 52 (2011) 189–198.
[4] L. Harris, Transaction data tests of the mixture distribution hypothesis, J. Financ. 22 (1967) 127–141.
[5] O. Kim, R.E. Vernechia, Market reaction to anticipated announcements, J. Financ. Econ. 30 (1988) 213–309.
[6] F. Grundy, D. Bruce, Trade and the revelation of information through prices, Rev. Financ. Stud. 2 (1989) 495–526.
[7] T. Andersen, Return volatility and trading volume: an information flow interpretation of stochastic volatility, J. Financ. 51 (1996) 169–204.
[8] S.M. Keane, Efficient Market Hypothesis and the Implications for Financial Reporting, Van Nostrand Reinhold, New York, 1983.
[9] P.F. Pai, K.P. Lin, C.S. Lin, P.T. Chang, Time series forecasting by a seasonal support vector regression model, Expert Syst. Appl. 37 (2010) 4261–4265.
[10] A. Shleifer, Inefficient Markets: An Introduction to Behavioral Finance, Oxford University Press, 2000.
[11] G.D. Santis, S. Imrohoglu, Stock returns and volatility in emerging financial markets, J. Int. Money Financ. 16 (1997) 561–579.
[12] D. Easley, M. O'Hara, Price, trade size and information in securities markets, J. Financ. Econ. 19 (1987) 69–90.
[13] P. Jain, G. Joh, The dependence between hourly prices and trading volume, J. Financ. Quant. Anal. 23 (1988) 169–283.
[14] R. Rogalski, The dependence of prices and volume, Rev. Econ. Stat. 36 (1978) 268–274.
[15] H. Bessembinder, R.J. Seguin, Price volatility trading volume, and market depth: evidence from future markets, J. Financ. Quant. Anal. 28 (1993) 21–39.
[16] I.E. Blume, M. O'Hara, Market statistics and technical analysis: the role of volume, J. Financ. 49 (1994) 153–182.
[17] J. Lakonishok, Past price changes and trading volume, J. Portfolio Manage. 15 (1989) 18–24.
[18] D.B. Sicilia, The Greenspan Effect—Words That Move the World's Markets, McGraw-Hill, 1999.
[19] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inf. Syst. 26 (2008) 3.
[20] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.
[21] X. Liang, Neural network method to predict stock price movement based on stock information entropy, Lect. Notes Comput. Sci. 3973 (2006) 442–451.
[22] X. Liang, An effective method of pruning support vector machine classifiers, IEEE Trans. Neural Networks 21 (2010) 26–38.
[23] M. Orchel, Support vector regression based on data shifting, Neurocomputing 96 (2012) 2–11.
[24] S.M. Weiss, N. Indurkhya, T. Zhang, F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer, 2010.
[25] M.W. Berry, J. Kogan, Text Mining: Applications and Theory, Wiley, 2010.
[26] C.-C. Chang, C.-J. Lin, LibSVM: A Library for Support Vector Machines, [Online]. Available: 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.
[27] C.-J. Lin, Asymptotic convergence of an SMO algorithm without any assumptions, IEEE Trans. Neural Networks 13 (2002) 248–250.
[28] S.S. Keerthi, C.J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Comput. 15 (2003) 1667–1689.
[29] R.E. Fan, P.H. Chen, C.J. Lin, Working set selection using second order information for training SVM, J. Mach. Learn. Res. 6 (2005) 889–1918.
[30] M. Chang, C.J. Lin, Leave-one-out bounds for support vector regression model selection, Neural Comput. 17 (2005) 1188–1222.

Xun Liang received the B.Sc. and Ph.D. degrees (with honors) in computer engineering from Tsinghua University, Beijing, China, in 1989 and 1993, respectively, and the M.Sc. degree in economics and operations research from Stanford University, Stanford, CA, in 1999. He worked as a Postdoctoral Fellow at the Institute of Computer Science and Technology, Peking University from 1993 to 1995, and the Department of Computer Engineering, University of New Brunswick from 1995 to 1997. He was a recipient of the Alexander von Humboldt fellowship. He has worked as a software architect or CTO leading over ten intelligent information products in California from 2000 to 2007 and as Associate Professor at the Institute of Computer Science and Technology in Peking University from 2005 to 2009. Currently, he is a Professor at the Department of Economic Information Systems at the Renmin University of China. His research interests include neural networks, support vector machines, and financial information systems.

Rong-Chang Chen received the Ph.D. degree in mechanical engineering from National Chiao Tung University, in 1994. He has six-year working experience in the fields of electronic commerce, supply chain management, and data mining. Currently, he is an Associate Professor at National Taichung University of Science and Technology. His research interests are mainly in data mining, artificial intelligence, and decision support system.

Yangbo He received the Ph.D. degree in mathematics at the School of Mathematical Sciences of Peking University in 2004. He worked as a Postdoctoral Fellow at the Institute of Computer Science and Technology, Peking University from 2004 to 2006. He is currently an Associate Professor at the School of Mathematical Sciences, Peking University. His research interests include data mining, and financial information systems.



Ying Chen received a B.Sc. in mathematics at the School of Mathematical Sciences of Peking University in 2008. He is currently a Ph.D. student in computer science at the Institute of Computer Science and Technology of Peking University. His research interests include neural networks, support vector machines, and financial information systems.