A SVM Approach in Forecasting the Moving Direction of Chinese Stock Indices

Zhongyuan Wei
Lehigh University

Follow this and additional works at: http://preserve.lehigh.edu/etd

Recommended Citation
A SVM Approach in Forecasting the Moving Direction of Chinese Stock Indices

by

Zhongyuan Wei

A Thesis
Presented to the Graduate and Research Committee
of Lehigh University
in Candidacy for the Degree of
Master of Sciences

in
Department of Industrial and Systems Engineering

Lehigh University
May 2012
Thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science in Department of Industrial and Systems Engineering.

A SVM Approach in Forecasting the Moving Direction of Chinese Stock Indices

Zhongyuan Wei

Date Approved

Thesis Director

(Name of Department Chair)
ACKNOWLEDGMENTS

First and foremost I would like to express my most sincere gratitude to my supervisor Prof. Dr. Katya Scheinberg who helped me develop knowledge in wide range of subjects from basic optimization to recently most popular machine learning algorithm support vector machines. She is not only a great teacher, but a wonderful mentor! Her patience, enthusiasm, superior knowledge and stimulating suggestions helped and encouraged me in all the time of writing this thesis.

I want to thank Prof. Tamás Terlaky for taking time out of his busy schedule to review my material and providing valuable suggestions that further improve this thesis.

I want to thank Prof. Tamás Terlaky, Prof. Eugene Perevalov, Prof. Robert H. Storer and Prof. George R. Wilson and other Professors in Department of Industrial and Systems Engineering for their great lectures. I also want to thank Mrs Rita Frey for all her kind and warm assistance on our daily lives.

I want to thank my best buddies Zheng Shi, Jingbo Li and Wenhao Han for all their help and support. I can’t imagine how my life in Lehigh would be if I didn’t know them.

Thank Yifei Liu for his help in teaching me the use of WIND and BLOOMBERG during my internship as an analyst in Guotai Junan Securities.

At last, I would like to give my special thanks to my mom Xiuqing Zheng and dad Fang Wei whose love enabled me to complete this work and encouraged me to go further.
# Table of Contents

List of Tables ........................................................................................................ vii

List of Figures ........................................................................................................ viii

Abstract .................................................................................................................... 1

Introduction .............................................................................................................. 3

Background .............................................................................................................. 7
  Stock index........................................................................................................... 7
  Support vector machine ...................................................................................... 9

Experiment Design ................................................................................................ 13
  Potential input features ....................................................................................... 14
  Data collection and preprocessing .................................................................... 15
  Feature selection ................................................................................................. 17
    Fisher Score (FS) ............................................................................................ 18
    Correlation-based Feature Selection (CFS) ..................................................... 19
  SVM parameter selection ................................................................................... 20
  Results comparison ............................................................................................ 21

Results ..................................................................................................................... 23
  Feature selection ................................................................................................. 23
  Parameter selection ............................................................................................ 24
  Results comparison ............................................................................................ 25

Conclusions ............................................................................................................. 31

Discussion .............................................................................................................. 33
  Features selected by the filter model .............................................................. 33
  Prediction accuracy ........................................................................................... 35
  Fundamental analysis vs. technical analysis .................................................... 38
  Is SVM really a good choice? .......................................................................... 39
Reference ...........................................................................................................41

Vitae ...................................................................................................................44
List of Tables

Table 1: Generally used technical indicators and their formulas. .......................... 5
Table 2: The description of stock indices in our analysis. ........................................ 9
Table 3: The initial 16 potential features that may be used for the prediction. .... 15
Table 4: Information about the experiment data ........................................................ 16
Table 5: Selected features and the final feature combination .................................. 23
Table 6: Parameters selected for each index feature combination ......................... 25
Table 7: The prediction accuracy of SVM and Neural Network. ......................... 26
Table 8: The coincidence matrix of models with selected features. ..................... 26
Table 9: Sensitivity and specificity of models with selected features. ................. 27
List of Figures

Fig. 1: The road map of the experiment. .................................................................13

Fig. 2: The gain charts for models with selected features. ...............................30

Fig. 3: The feedback loop between interest rate and real economy. ..........36

Fig. 4: Chinese inflation rate between Jan 2003 and Jan 2012 .....................37
Support vector machine (SVM) has been shown to be a reliable tool in prediction and classification using a convex objective function with constraints integrated in by Lagrange Multipliers and characterized by the involvement of kernel functions and the sparsity of the solution. In this paper, we build a series of SVM models based on the macroeconomic fundamental indicators to investigate the predictability of financial movement direction by forecasting the daily movement trend of main indices in Chinese Shanghai and Shenzhen stock markets: SSE 50 Index, SSE 180 Index, SSE Composite Index, SZSE 100 Index, SZSE Composite index and CSI 300 Index. To the best of our knowledge, this is the first application of SVM techniques to Chinese stock indices. After data cleaning and transformation, only a subset of potential input candidates is delivered to the next step through feature selection. Then two parameters of SVM are selected and tuned based on the selected subsets of features for SSE indices. To evaluate the forecasting ability of SVM, we compare its performance with Neural Networks. The experimental results show that SVM with carefully selected features performs
comparably to or better than SVM with the whole variable set. And in accordance with previous studies, SVM outperforms Neural Network in all indices. However, both models have low specificity values. The prediction accuracy of SVM for SSE 50 Index, SSE 180 Index, SSE Composite Index, SZSE 100 Index, SZSE Composite index and CSI 300 Index is 61.06%, 60.47%, 61.65%, 59.74%, 61.58% and 61.85%, respectively. These results support the claim that as an emerging market Chinese stock market is semi-strong form inefficient.

**Key words:** SVM; SSE Indices; SZSE Indices; CSI 300 Index; macroeconomic fundamental analysis; feature selection; parameter selection; Neural Network.
**Introduction**

Since first proposed in 1979 (Vapnik, 1979), Support Vector Machine (SVM) (its formula and deduction are described in *Background*) has been shown to be a useful technique for data classification (Burges, 1998), regression (Smola and Schölkopf, 2004) and prediction (Müller *et al.*, 1997; Kim, 2003). In many settings, such as pattern recognition and regression, SVM either matches or significantly outperforms competing methods with regard to the error rates on test sets (Burges, 1998).

The popularity and high performance of SVM could be explained by the specific formulation of a convex objective function with constraints integrated in by Lagrange Multipliers, and the characteristics such as the capacity control of the decision function, the involvement of kernel functions and the sparsity of the solution (Vapnik, 1998; Vapnik, 1999). Besides, in the context of financial data modeling, SVM is also appealing for the following reasons: (1) data could be performed without making strong assumptions; (2) many traditional Neural Network models had implemented the empirical risk minimization principle, while SVM is established on the structural risk minimization principle, which seeks to minimize an upper bound of generalization error, and is shown to be very resistant to the over-fitting problem; (3) SVM model is a linearly constrained quadratic program so that the solution of SVM is always globally optimal, while other Neural Network models may tend to fall into a local optimal solution (Kim, 2003; Huang *et al.*, 2005)

Compared with other fields, there were originally a few analyses for the applications of SVM in financial time-series forecasting. However, recently, SVM has become a popular tool and has exhibited excellent performance and promising results in financial prediction. Several papers have shown that SVM outperforms traditional-time-series methods, Back-propagation Neural
Network and autoregressive moving average models (ARIMA) on generalization error (Tay and Cao, 2001; Kim, 2003; Tay, 2003; Thissen et al., 2003).

Although SVM approach has been widely applied in financial forecasting, little analysis is extended into the stock market of China, which is the second biggest economic entity and also one of the most important emerging markets in the world. Additionally, while indices, the main indicators of countries’ economic condition, have been the efficient instrument for both hedgers and speculators in traditional and derivative market, less work is performed in this area than the analysis on individual stocks. Besides, model inputs in prior research typically involve price and volume data, and may also include a selection of well-known technical indicators (Table 1), but few papers apply fundamental indicators in the model. Compared to the technical analysis, which is the study of collective market sentiment mainly reflected in the price and volume, fundamental analysis focuses more on the intuitive physical interpretation and attempts to find the intrinsic value of the assets. Fundamental variables selected and included in the model generally have intuitive justification and have certain connection with the target; while it may be difficult to explain a technical analytical model.

The objective of this study is to explore the predictability of the movement direction of main indices for Chinese stock market with SVM using some macroeconomic fundamental indicators. Before we begin, we rely on several prior studies which help us to set the target accuracy of this experiment. For developed stock markets, the prediction power of SVM is generally high, for example, the hit ratio for NIKKEI 225 is 73% (Huang et al., 2005) and for S&P 500 is around 70% (R. Rosillo, D. de la Fuente, and J. A. L. Brugos, unpublished data); while for less developed stock markets, the prediction accuracy is lower, for example, the hit ratio for Korea composite stock price index is only 58% (Kim, 2003). Taking into consideration that the stock
market in China is not as mature as those in Japan and US, we set our expected accuracy within a range of 60%~65%.

The remainder of this paper is organized as follows. In Section 2, we will briefly explain the stock index and the theory of SVM. The methodology is described in Section 3. The experimental results are shown in Section 4. Some conclusions are drawn in Section 5. And finally, several interesting “hints” are discussed in Section 6.

**Table 1**: Generally used technical indicators and their formulas.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>%K</td>
<td>$(C_t - LL_{t-n}) / (HH_{t-n} - LL_{t-n}) \times 100$</td>
<td>Comparing a security’s price closed relative to its price range over n days. where $LL_t$ and $HH_t$ mean the lowest low price and highest high price during the last t days, respectively.</td>
</tr>
<tr>
<td>%D</td>
<td>$\sum_{i=0}^{n-1} %K_{t-i} / n$</td>
<td>Moving average of %K.</td>
</tr>
<tr>
<td>Slow %D</td>
<td>$\sum_{i=0}^{n-1} %D_{t-i} / n$</td>
<td>Moving average of %D.</td>
</tr>
<tr>
<td>Momentum</td>
<td>$C_t - C_{t-4}$</td>
<td>Measuring the amount change of a security’s price over four days.</td>
</tr>
<tr>
<td>ROC</td>
<td>$C_t / C_{t-n} \times 100$</td>
<td>Measuring the difference between the current price and the price n days ago.</td>
</tr>
<tr>
<td>William’s %R</td>
<td>$(HH_{t-n} - C_t) / (HH_{t-n} - LL_{t-n}) \times 100$</td>
<td>A momentum indicator measuring overbought/oversold levels.</td>
</tr>
<tr>
<td>Indicator</td>
<td>Formula</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>A/D Oscillator</td>
<td>$(H_t - C_{t-1}) / (H_t - L_t)$</td>
<td>A momentum indicator associating changes in price.</td>
</tr>
<tr>
<td>OSCP</td>
<td>$(MA_5 - MA_{10}) / MA_5$</td>
<td>The difference between two moving averages of a security’s price.</td>
</tr>
<tr>
<td>CCI (Commodity Channel Index)</td>
<td>$M_t = (H_t + C_t + L_t) / 3$&lt;br&gt;$SM_t = \frac{1}{n} \sum_{i=1}^{n} M_{t-i+1}$&lt;br&gt;$D_t = \frac{1}{n} \sum_{i=1}^{n}</td>
<td>M_{t-i+1} - SM_i</td>
</tr>
<tr>
<td>RSI (Relative Strength Index)</td>
<td>$100 - 100 / (1 + (\sum_{i=0}^{n-1} Up_{t-i} / n) / (\sum_{i=0}^{n-1} Dw_{t-i} / n))$</td>
<td>It is a price following an oscillator that ranges from 0 to 100.</td>
</tr>
</tbody>
</table>

where $Up_t$ and $Dw_t$ mean upward and downward price change at time $t$, respectively.
**Background**

**Stock Index**

Stock index is an important component of a security market. It reflects the market sentiment, measures market’s return and risk, and serves as the benchmark for index funds. Therefore, it not only provides an effective means for both individual and institutional investors to hedge against potential market risks, but creates new investment opportunities for market speculators and arbitrageurs. Given the rising popularity of index trading, forecasting stock market indices has profound implications and significance for researchers and practitioners alike.

Some results have shown that stock returns in well established stock markets, such as in the US, UK, France, Germany, and Japan, are to some extent predictable by applying various machine learning algorithms, in other words stock prices do not follow a random walk (Cao and Tay, 2000; Leung et al., 2000; Tay and Cao, 2001; Huang et al., 2005). In most cases, the degree of accuracy and the acceptability of certain forecasts are measured by the predictors’ deviations from the observed values and may involve certain subjective judgment (Tay and Cao, 2001; Huang et al., 2005; Chen and Shih, 2006). However, some studies suggested that an accurate prediction of the direction of the stock market movement could bring more profits than certain forecasts with small forecast errors (Aggarwal and Demaskey, 1997; Wu and Zhang, 1997). As Levich (Levich, 2001) said, the so-called useful forecasts predict the direction of price change and hence “... lead to profitable speculative positions and correct hedging decisions.” This means, in principle, the ability to forecast the sign of future returns should alone be sufficient to make profit, and compared to the efforts to predict the return, it would multiply the effectiveness. Therefore recent studies have tended to develop models to forecast market direction rather than returns.
Considering that the extent of predictability related to the emerging markets is seldomly analyzed, we chose Chinese stock indices as our targets and use SSE Indices, SZSE Indies and CSI 300 Index (Table 2) as the experiment data. As the most authoritative indicators that measure the performance of Chinese security market and serve as the weatherglass of the Chinese economy, SSE Indices, SZSE Indies and CSI 300 Index are compiled, calculated and published by Shanghai Stock Exchange and Shenzhen Stock Exchange, which are established on 26 Nov, 1990 and 1 Dec, 1990, respectively and the only two self-regulated legal entities under the supervision of China Securities Regulatory Commission (CSRC) in Mainland China.

In this analysis, the daily movement direction of three important SSE Indices is examined in order to cover different cross-sections of the Chinese industries. These indices are SSE Composite Index (a whole market index), SSE 180 Index (a performance benchmark index) and SSE 50 Index (an index for good quality, large scale stocks), which form a 3-level pyramid index structure for Shanghai stock market. Similarly, we chose SZSE 100 Index and SZSE Composite Index from SZSE indices. CSI 300 Index is the first cross-exchange equity index launched by the two exchanges together, and aims to reflect the price fluctuation and performance of China A-share market. Only the daily close prices of these indices were used in this experiment.
Table 2: The description of stock indices in our analysis.

<table>
<thead>
<tr>
<th>Index</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE 50</td>
<td>000016.SS</td>
<td>An index that selects 50 largest stocks of good liquidity and representativeness from Shanghai security market</td>
</tr>
<tr>
<td>SSE 180</td>
<td>000010.SS</td>
<td>A free-float weighted index that selects 180 largest stocks of good liquidity and representativeness from Shanghai security market</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>000001.SS</td>
<td>An index of all stocks (A-share and B-share) that are traded at the Shanghai Stock Exchange</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>399004.SZ</td>
<td>An index that selects 100 largest A shares of good liquidity and representativeness from Shenzhen security market</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>399106.SZ</td>
<td>An actual market-cap weighted index (no free float factor) that tracks the stock performance of all the A-share and B-share lists on Shenzhen Stock Exchange</td>
</tr>
<tr>
<td>CSI 300</td>
<td>000300.SS</td>
<td>A free float-weighted index that consists of 300 A-share stocks listed on the Shanghai or Shenzhen Stock Exchanges</td>
</tr>
</tbody>
</table>

**Support Vector Machine**

Let us consider the following setting: we want to separate two sets of separable data points, which are labeled as \( \{x_i, y_i\}, \; i = 1, 2, \ldots, N, \; y_i \in \{+1, -1\}, \; x_i \in \mathbb{R}^d \), by a linear predictor hyperplane.

Suppose we have some hyperplanes which separate these two set of points. The points \( x_i \) which lie on the hyperplane satisfy \( \mathbf{w}^T x_i + \beta = 0 \), where \( \mathbf{w} \) is normal to the hyperplane. Thus, we have
\[ w^T x_i + \beta \geq +1 \text{ for } y_i = +1, \]
\[ w^T x_i + \beta \leq -1 \text{ for } y_i = -1, \]
or simply
\[ y_i (w^T x_i + \beta) \geq 1, \]
\[ \forall i = 1, 2, \ldots, N. \]

Define the margin of a separating hyperplane to be the shortest (perpendicular) distance from the separating hyperplane to the closest positive point plus that to the closest negative point. For this case, we just simply need to look for the separating hyperplanes with the largest margin. The margin can be calculated as \( \frac{2}{\|w\|} \), where \( \|w\| \) is the Euclidean norm of \( w \). Maximizing \( \frac{2}{\|w\|} \) is equivalent to minimizing \( \|w\|^2 \). Thus this problem can be formulated as follows (adding \( \frac{1}{2} \) to the objective function is for calculation convenience):

\[
\min_{w, \beta} \frac{1}{2} w^T w \\
\text{s.t. } y_i (w^T x_i + \beta) \geq 1, \\
\forall i = 1, 2, \ldots, N.
\]

However, when we apply the above algorithm to the non-separable data, we cannot find feasible solutions. One method to extend its usage to handling non-separable data is to relax the constraints by introducing a positive slack variable \( \xi \) and to add a further cost \( C \) in objective function.

Thus, the adjusted model, called soft margin SVM (Cortes and Vapnik, 1995; Vapnik, 1998) solves the following primal problem:
\[
\min_{\xi, w, \beta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad y_i(w^T x_i + \beta) \geq 1 - \xi_i, \\
\xi_i \geq 0, \\
\forall i = 1, 2, \ldots, N,
\]

where \( x_i, i = \{1, \ldots, N\} \) are the N training points, \( y_i \) is the binary label of each point with values +1 or -1, and C is the penalty cost for those sample points that are not correctly classified by the SVM, a large C corresponding to a higher penalty to errors.

The Karush–Kuhn–Tucker (KKT) conditions for the Lagrange function \( L \) of the primal problem are:

\[
\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0, \\
\frac{\partial L}{\partial \beta} = -\sum_{i=1}^{N} \alpha_i y_i = 0, \\
\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \xi_i = 0, \\
y_i(w^T x_i + \beta) \leq -1 + \xi_i, \\
\xi_i \geq 0, \\
\alpha_i \geq 0, \\
\mu \geq 0, \\
\alpha_i \left[ y_i(w^T x_i + \beta) - 1 + \xi_i \right] = 0, \\
\mu \xi_i = 0,
\]

where \( \mu \) and \( \alpha \) are the non-negative Lagrange multipliers.

At optimality, we have

\[
w^* = \sum_{i=1}^{n} \alpha_i y_i x_i.
\]

And the dual problem is
\[
\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{s.t. } y^T \alpha = 0, \\
0 \leq \alpha \leq C,
\]

where \( e \) is the vector of all ones, \( C > 0 \) is the upper bound, \( Q \) is an \( N \) by \( N \) positive semi-definite matrix with the form of \( Q = \text{Diag}(y)x x^T \text{Diag}(y) \).

When we implement Kernel functions to construct \( Q \), \( Q_{ij} = y_i y_j K(x_i, x_j) \), and \( K(x_i, x_j) = \phi(x_i)\phi(x_j) \) is the kernel. Here training vectors \( x_i \) are mapped into a higher (maybe infinite) dimensional space (feature space) by the function \( \phi(x) \).

For a new data \( x \), the decision function is given by

\[
\text{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x_j) + b\right).
\]

Any function that satisfy Mercer’s conditions could be employed as the Kernel function (Vapnik 1995). In this study, we use radial basis function (RBF) as the Kernel function, which is expressed as

\[
K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0,
\]

where \( \gamma \) is the parameter of the Kernel.

And RBF tends to give good performance when additional information about the data is limited (Huang et al., 2005) and under general smoothness assumptions (Chen et al., 2006). Therefore, RBF is widely used in financial time-series analyses.
**Experiment Design**

Our experiment follows the basic data mining process demonstrated in the flowchart below (Fig. 1):

![Flowchart](image)

**Fig. 1:** The road map of the experiment. After collected, data was cleaned and transformed. Then it was split into two parts: training data (black arrow) and testing data (white arrow). At first, training data was used to perform feature selection using two filter models: Fisher Score (FS) and Correlation-based Feature Selection (CFS). Dimensionality reduction was then applied to both training and testing data. The SVM model was then trained on the training data and tested on the testing data later. The results comparison between SVM and standard 3-layer Neural Network was conducted as the last step.
**Potential Input Features**

We surveyed several academic papers that explored the cross-sectional relationship between macroeconomic input variables and the stock index by using the forecasting models. The potential macroeconomic inputs include term structure of interest rates (TS), short-term interest rate (ST), long-term interest rate (LT), consumer price index (CPI), industrial production (IP), government consumption (GC), private consumption (PC), gross national product (GNP) and gross domestic product (GDP) (Fama, 1988; Fama and French, 1988; Fama, 1992; Lakonishok, 1994; Leung, 2000). Unfortunately, CPI, IP, GC, PC, GDP and GNP are not included in this study, because the daily data is not available. Thus, among these indicators only LR and SR are viable for the experiment.

China is the biggest exporting country in the world. The economy growth has a strong correlation with China’s export. The largest 3 trading partners for China in recent 10 years are very stable: European Union (EU), the United States and Japan. Therefore, inspired by previous work on Japanese stock market index (Huang et al., 2005), we introduced several new features, which have strong intuitive correlation with the Chinese export, to reflect the USA’s, Japan’s and EU’s commercial influence on Chinese economy. These newly employed features are the daily closed value of S& P 500 Index, NIKKIE 225 Index and EURO STOXX 50 Index, which are the well-known market capitalization indicators of the USA, Japan and EU economic condition, respectively; and the daily midpoint value of the exchange rate of Japanese Yen (JPY), US Dollars (USD) and Euros (EUR) against Chinese Yuan (CNY). All candidate features and their descriptions are listed in the following table *(Table 3)*:
Table 3: The initial 16 potential features that may be used for the prediction.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Potential features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>Short-term and long-term interest rate</td>
<td>The time range including 0.5-, 1-, 2-, 3-, 5-, 7-, 10-, 15-, 20-, 30-year, a total of 10 features.</td>
</tr>
<tr>
<td>EC</td>
<td>EUR/CNY</td>
<td>Exchange rate of Euros against Chinese Yuan.</td>
</tr>
<tr>
<td>UC</td>
<td>USD/CNY</td>
<td>Exchange rate of Dollars against Chinese Yuan.</td>
</tr>
<tr>
<td>JC</td>
<td>JPY/CNY</td>
<td>Exchange rate of Japanese Yen against Chinese Yuan.</td>
</tr>
<tr>
<td>ST</td>
<td>EURO STOXX 50 (SX5E:IND)</td>
<td>A free-float market capitalization-weighted index of 50 European blue-chip stocks from those countries participating in the Europe Monetary Union (EMU).</td>
</tr>
<tr>
<td>NI</td>
<td>NIKKEI 225 (NKY:IND)</td>
<td>A price-weighted average of 225 top-rated Japanese companies listed in the First Section of the Tokyo Stock Exchange (TSE).</td>
</tr>
</tbody>
</table>

**Data Collection and Preprocessing**

Data of Chinese stock indices, IR, EC, UC and JC were collected via WIND; data of SP, ST and NI were collected via Bloomberg. *Wind* and *Bloomberg* are the professional data-feed software widely used in financial companies in China. Data from different inputs is merged by the key “*date*” in IBM SPSS® Modeler. In the data cleaning step, not all the instances (rows) contain
complete data in every fields (columns): some inputs have missing data. In this case, these instances are deleted, and only complete data was harnessed in the following steps. The data was then transformed. First, the “difference of natural logs” transformation was performed on the raw data, because such transformation does not display trends, nor extends the drifts from the mean (Huang et al., 2005). Then the whole data base was divided into two parts in the order of the date: the first 80% is the training data, which was used for training the model and choosing the parameters; and the last 20% as the testing data, which was used for the out-of-sample evaluation. Therefore, we used the earlier data to train the model and the latest data to test it. The details are listed in Table 4.

### Table 4: Information about the experiment data

<table>
<thead>
<tr>
<th>Index</th>
<th>Time Range</th>
<th>Number of Training Examples</th>
<th>Number of Testing Examples</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE 50</td>
<td>Jan 2004 – Sep 2011</td>
<td>1349</td>
<td>339</td>
<td>1688</td>
</tr>
<tr>
<td>SSE 180</td>
<td>Jan 2004 – Sep 2011</td>
<td>1349</td>
<td>339</td>
<td>1688</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>Jan 2004 – Sep 2011</td>
<td>1349</td>
<td>339</td>
<td>1688</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>Jan 2003 – Sep 2011</td>
<td>1521</td>
<td>380</td>
<td>1901</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>Jan 2003 – Sep 2011</td>
<td>1521</td>
<td>380</td>
<td>1901</td>
</tr>
<tr>
<td>CSI 300</td>
<td>Jul 2005 – Sep 2011</td>
<td>1081</td>
<td>271</td>
<td>1352</td>
</tr>
</tbody>
</table>
FEATURE SELECTION

Feature selection is a process of selecting a subset of original features according to certain criteria. It is a key constituent of data mining which is frequently used to reduce the number of features, remove irrelevant, redundant, or noisy data, and bring beneficial effects for applications. First, the implicit regularization achieved by feature pruning generally improves performance such as model generalization and prediction accuracy. Second, removing irrelevant features also considerably speeds up the data mining computing time. Third, too many features may make the convergence impossible and lead to random classification decisions. Fourth, it facilitates the comprehensibility of the result.

So in conclusion reasons for performing feature selection (Han and Kamber, 2006) include:

- Improving prediction accuracy.
- Reducing computational time.
- Reducing the measurement and storage requirements.
- Facilitating data visualization and model understanding.

Typically, there are supervised, unsupervised and semi-supervised feature selection algorithms, corresponding to labeled, unlabeled or partial labeled training data (Han and Kamber, 2006). Also, based on the design strategy, feature selection algorithms could be classified into three categories: filter, wrapper and embedded models (Han and Kamber, 2006). The filter model uses the general characteristics of data to evaluate features without the cooperation of the learning algorithm. The wrapper model uses the performance of a predetermined learning algorithm as evaluation criterion to select features. The embedded model incorporates feature selection as a part of the training process, which is reflected analytically in the objective of the learning model. Filter model and embedded model may return either a subset or the weights of all features.
In this experiment, we focus on filter-based methods for supervised feature selection. We set two different filter models to perform feature selection and select subset of features by using a MATLAB-based feature selection software package downloaded from http://featureselection.asu.edu. These models, namely Fisher Score and Correlation-based Feature Selection, are able to use their own particular emphases to mutually complement each other. After comparing the results of the two models, we chose the proper combination of features, which have intuitive correlations with the target.

**Fisher Score (FS)**

Fisher Score (Duda *et al.*, 2001) is an effective supervised feature selection algorithm. It selects features with the following characteristics: they assign similar values to the samples from the same class, while assigning different values to samples from different classes. The evaluation criterion of Fisher Score can be formulated as:

\[
SC_f(f_i) = \frac{\sum_{j=1}^c n_j (\mu_{i,j} - \mu_i)^2}{\sum_{j=1}^c n_j \sigma_{i,j}^2},
\]

where \( \mu_i \) is the mean of the feature \( f_i \), \( n_j \) is the number of samples in the \( j \)th class, and \( \mu_{i,j} \) and \( \sigma_{i,j} \) are the mean and the variance of \( f_i \) on class \( j \), respectively.

Fisher Score has been widely applied in many fields due to its generally good performance. However, Fisher Score is a univariate model, which evaluates features individually. Therefore it cannot solve feature redundancy problem.
Correlation-based Feature Selection (CFS)

CFS (Hall and Smith, 1999) is thus employed to handle feature redundancy. It uses a correlation based heuristic "merit" to evaluate the features based on the hypothesis that good feature subsets contain features highly correlated with (have high prediction power of) the class, yet uncorrelated with each other. This heuristic is built from two aspects: the usefulness of individual features for prediction and the level of inter-correlation among them. It can be stated as:

\[
Merit_s = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}},
\]

where \( Merit_s \) is the "merit" of a feature subset \( S \) containing \( k \) features; \( \bar{r}_{cf} = \sum_{f_i \in S} \frac{1}{k} \sum_{j \neq i} (f_i, c) \) is the mean feature-class correlation (\( f \in S, \text{and} \ c \text{ is the class} \)), an indication to how easily a class could be predicted based on the feature; and \( \bar{r}_{ff} \) is the average feature inter-correlation between the features which indicates the level of redundancy between them.

Feature correlations are measured via Information Gain that determines the degree of association between features. The Information Gain (IG) of feature \( X \) to the class \( Y \) can be expressed as:

\[
IG(X,Y) = H(X) - H(X \mid Y),
\]

\[
H(X) = -\sum_i P(x_i) \log_2(P(x_i)),
\]

\[
H(X \mid Y) = -\sum_j P(y_j) \sum_i P(x_i \mid y_j) \log_2(P(x_i \mid y_j)).
\]

CFS uses the Best First search to explore the search space. It evaluates the merit of a feature by estimating its predictive ability and the redundancy it introduces to the selected feature set. Specifically, CFS calculates feature-class and feature-feature correlations first and then selects a
subset of features using the Best First search with a certain stopping criterion. It selects the most relevant features and by the greatest extent avoids the re-introduction of redundancy. As other filter models, CFS does not need to reserve any training data for the subsequent evaluation. Besides, it works well on smaller data sets.

**SVM parameter selection**

There are two parameters to be determined in SVM model with RBF Kernel: C and γ. Generally, increasing C and γ would improve classification accuracy on the training set, but also tends to lead to over-fitting.

Given that the two parameters play an important role in the performance of SVMs (Tay and Cao, 2001), improper selection of these two parameters may cause over-fitting or under-fitting. Since there is few reference to the parameter value, to raise the generalization accuracy, we conduct a cross validation (CV) process to decide optimal parameters using LIBSVM (Chang and Lin, 2011) for the best prediction performance. Then the parameter pair is tuned within a small range around the given optimal value.

The procedure is as follows:

- Estimate the accuracy of each parameter combination in the specified range: \( \log_2 C \in \{8, 9, \ldots, 16\} \) and \( \log_2 \gamma \in \{-5, -4, \ldots, 1\} \) by conducting a 5-fold cross validation on the training set.
- Choose the parameter pair \((C, \gamma)\) that leads to the highest CV accuracy rate.
- Tune the parameter pair by conducting a 5-fold cross validation with different combinations of \((C-200, C-100, C+100, C+200)\) and \((\gamma-0.0125, \gamma-0.025, \gamma+0.0125, \gamma+0.025)\) on the training set.
• Use the best parameter pair to create the model.

RESULTS COMPARISON

Neural Networks are still a popular tool for classification and forecast because of their unique learning capabilities (Huang et al., 2004). They are shown to be reliable in various different applications in industry, business and science (Widrow et al., 1994). A classical Neural Network consists of three layers: input layer, output layer and the hidden layer between those two. Nodes or units in each layer are fully connected according to the structure: information is fed forward from the input layer through the hidden layer, then to the output layer. Thus Neural Network learning is also called connectionist learning due to its connection quality (Han and Kamber, 2006).

Neural Network works according to the following process. Units in the input layers are corresponding to the attributes of the training tuple. The input flow passes through the input layer. Then it is weighted according to the weight of connection and fed simultaneously to neurons in the hidden layer. Than the output flow could go into another hidden layer is there are more than one hidden layer, although usually one hidden layer is used. The outputs from the last hidden layer are input into the output layer and the prediction result is produced by the output layer. During the whole process, the weighted sum of the outputs from the last layer could be applied with a certain activation function (non-linear function) and this enables Neural Network to model the forecast as the non-linear combination of the inputs.

Neural Network is often criticized for the long training time, the poor interpretation and the inability to determine the optimal network topology or structure in an efficient way. Parameters could only be selected by the trial-and-error method. Advantages of Neural Network, however,
are the high tolerance to the noisy data and the ability to classify untrained patterns (Han and Kamber, 2006). It can be used even when knowledge about the relationships between attributes and classes is quite limited.

As in other works (Tay and Cao, 2001; Kim, 2003; Tay, 2003), in this study, we used a standard three-layer fully connected Neural Network to perform as the benchmark of SVM. The input layer units are financial features, such as indices and interest rate, etc, while the output units are the binary class labels indicating the moving directions: up or down.

Besides the comparison between SVM and Neural Network, the results of SVM models with a carefully selected subset of inputs and that with the whole set of financial variables are also compared to show the influence of feature selection.
**Results**

**FEATURE SELECTION**

We combined features that survive in two individual feature selection processes in order to avoid certain inherent bias or defect of the algorithms. The details of the selected subsets of features for SSE series are shown below (Table 5).

Table 5: Selected features and the final feature combination (for the abbreviation please see Table 3).

<table>
<thead>
<tr>
<th></th>
<th>Fisher Score*</th>
<th>CFS</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SSE 50</strong></td>
<td>NI, EC, ST, JC, 10-year IR</td>
<td>NI</td>
<td>NI, EC, ST, JC, 10-year IR</td>
</tr>
<tr>
<td><strong>SSE 180</strong></td>
<td>NI, EC, ST, UC, 10-year IR</td>
<td>UC, NI</td>
<td>NI, EC, ST, UC, 10-year IR</td>
</tr>
<tr>
<td><strong>SSE Composite</strong></td>
<td>NI, EC, ST, JC, 10-year IR</td>
<td>UC, NI</td>
<td>NI, EC, ST, UC, JC, 10-year IR</td>
</tr>
<tr>
<td><strong>SZSE 100</strong></td>
<td>NI, ST, UC, EC, 0.5-year IR</td>
<td>NI</td>
<td>NI, EC, ST, UC, 0.5-year IR</td>
</tr>
<tr>
<td><strong>SZSE Composite</strong></td>
<td>NI, ST, EC, JC, 0.5-year IR</td>
<td>NI, UC</td>
<td>NI, EC, ST, UC, JC, 0.5-year IR</td>
</tr>
<tr>
<td><strong>CSI 300</strong></td>
<td>NI, EC, ST, 10-year IR, JC</td>
<td>NI</td>
<td>NI, EC, ST, JC, 10-year IR</td>
</tr>
</tbody>
</table>

*: Features listed in Fisher Score column are sorted according to their weights. The first feature has the highest weight (is the most useful) for the particular feature selection algorithm. And only the five most important features are selected under Fisher Score model.
Thus, basically the direction of the stock indices movement $D_t^{SSE}$ could be forecasted with the following model:

$$D_t^{SSE_{50}} = F(X_{t-1}^{10\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{JC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

$$D_t^{SSE_{180}} = F(X_{t-1}^{10\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{UC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

$$D_t^{SSECo} = F(X_{t-1}^{10\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{UC}, X_{t-1}^{JC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

$$D_t^{SZSE_{100}} = F(X_{t-1}^{0.5\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{UC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

$$D_t^{SZSECo} = F(X_{t-1}^{0.5\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{UC}, X_{t-1}^{JC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

$$D_t^{CSI_{300}} = F(X_{t-1}^{10\text{-yearIR}}, X_{t-1}^{EC}, X_{t-1}^{JC}, X_{t-1}^{NI}, X_{t-1}^{ST})$$

where $D_t^{SSE}$ is +1 when the index value at $t$ is greater than that at $t-1$; -1, otherwise. The threshold of the erection standard for the binary label is flexible. 5%, for example, could be a threshold value for the classification, above which $D_t^{SSE}$ is +1. Thus the forecasting model could play a role as the filter when some financial portfolios need to be selected.

**PARAMETER SELECTION**

After determining the features that will be introduced into the model, we used a 5-fold cross validation to find the parameter pair with the highest cross validation accuracy. Then the value of the parameter pair is tuned. The refined results are listed in Table 6.
Table 6: Parameters selected for each index feature combination.

<table>
<thead>
<tr>
<th>Index</th>
<th>Features</th>
<th>Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>SSE 50</td>
<td>NI, EC, ST, JC, 10-year IR</td>
<td>1024</td>
</tr>
<tr>
<td>SSE 180</td>
<td>NI, EC, ST, UC, 10-year IR</td>
<td>2048</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>NI, EC, ST, UC, JC, 10-year IR</td>
<td>2048</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>NI, EC, ST, UC, 0.5-year IR</td>
<td>2048</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>NI, EC, ST, UC, JC, 0.5-year IR</td>
<td>4096</td>
</tr>
<tr>
<td>CSI 300</td>
<td>NI, EC, ST, JC, 10-year IR</td>
<td>2048</td>
</tr>
<tr>
<td>SSE 50</td>
<td>All</td>
<td>2048</td>
</tr>
<tr>
<td>SSE 180</td>
<td>All</td>
<td>4096</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>All</td>
<td>2048</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>All</td>
<td>4096</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>All</td>
<td>4096</td>
</tr>
<tr>
<td>CSI 300</td>
<td>All</td>
<td>1024</td>
</tr>
</tbody>
</table>

RESULTS COMPARISON

We used IBM SPSS® Modeler to perform the comparison between SVM and Neural Network as well as the comparison between SVM models with or without feature selection step. The prediction accuracy and coincidence matrix are shown in Table 7 and 8, respectively.
### Table 7: The prediction accuracy of SVM and Neural Network.

<table>
<thead>
<tr>
<th>Index</th>
<th>SVM (Selected Features)</th>
<th>SVM (All Features)</th>
<th>Neural Network (Selected Features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE 50</td>
<td>61.06%</td>
<td>60.47%</td>
<td>60.18%</td>
</tr>
<tr>
<td>SSE 180</td>
<td>60.47%</td>
<td>61.06%</td>
<td>55.46%</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>61.65%</td>
<td>62.24%</td>
<td>58.70%</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>59.74%</td>
<td>60.26%</td>
<td>56.32%</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>61.58%</td>
<td>61.32%</td>
<td>60.26%</td>
</tr>
<tr>
<td>CSI 300</td>
<td>61.85%</td>
<td>62.22%</td>
<td>61.48%</td>
</tr>
</tbody>
</table>

### Table 8: The coincidence matrix of models with selected features.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Neural Network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neg</td>
<td>Pos</td>
</tr>
<tr>
<td>SSE 50</td>
<td>Neg</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>43</td>
</tr>
<tr>
<td>SSE 180</td>
<td>Neg</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>36</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>Neg</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>27</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>Neg</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>20</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>Neg</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>25</td>
</tr>
<tr>
<td>CSI 300</td>
<td>Neg</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Pos</td>
<td>13</td>
</tr>
</tbody>
</table>

Note: “Neg” means that the moving direction is -1 (down), while “Pos” means the opposite. The “Neg” and “Pos” in the rows represent actual observed conditions, and those in the columns represent predicted ones. The cell at the cross indicates the number of records for each combination of predicted and actual conditions.
And according to Table 8, we can calculate the sensitivity (true positive rate) and specificity (true negative rate) (Table 9) based on the following formula:

\[
sensitivity = \frac{true\_pos}{actual\_pos}
\]
\[
specificity = \frac{true\_neg}{actual\_neg}
\]

The actual “Pos” records for SSE 50, SSE 180, SSE Composite, SZSE 100, SZSE Composite and CSI 100 are 157, 163, 166, 187, 208 and 139, respectively; and the ratios of “Pos” records to the whole testing data are 46.31%, 48.08%, 48.97%, 49.21%, 54.74% and 51.29%, respectively. Since these ratios are nearly 50%, representing the number of +1 and -1 labeled tuples are almost the same, we didn’t need to balance the data structure for both training and testing. There is no potential for the bias to predict more +1 than -1 or the other way around.

Table 9: Sensitivity and specificity of models with selected features.

<table>
<thead>
<tr>
<th>Index</th>
<th>Sensitivity</th>
<th></th>
<th>Specificity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM</td>
<td>Neural Network</td>
<td>SVM</td>
<td>Neural Network</td>
</tr>
<tr>
<td>SSE 50</td>
<td>72.61%</td>
<td>74.52%</td>
<td>51.10%</td>
<td>47.80%</td>
</tr>
<tr>
<td>SSE 180</td>
<td>77.91%</td>
<td>83.44%</td>
<td>44.3%</td>
<td>29.55%</td>
</tr>
<tr>
<td>SSE Composite</td>
<td>83.73%</td>
<td>77.71%</td>
<td>40.46%</td>
<td>40.46%</td>
</tr>
<tr>
<td>SZSE 100</td>
<td>89.30%</td>
<td>83.42%</td>
<td>31.09%</td>
<td>30.05%</td>
</tr>
<tr>
<td>SZSE Composite</td>
<td>87.98%</td>
<td>83.17%</td>
<td>26.65%</td>
<td>32.56%</td>
</tr>
<tr>
<td>CSI 300</td>
<td>90.65%</td>
<td>85.61%</td>
<td>31.30%</td>
<td>35.88%</td>
</tr>
</tbody>
</table>

From Table 7 we can see that all accuracy of SVM is around 60% with the highest accuracy 61.85%, which comes from the model built for CSI 300 Index; and the lowest accuracy 59.74%, which comes from the model built for SZSE 100 Index. This is basically consistent with our
expected accuracy. Compared with the SVM model using all features, except SSE 50 and SZSE Composite (bolded in the table), whose accuracy is improved probably due to the removal of irrelevant or noisy features, the accuracy of the SVM model performing feature selection generally drops, but only less than 0.6%. And generally SVM outperforms Neural Network in prediction accuracy, which is in consistent with previous researches (Tay and Cao, 2001; Kim, 2003; Tay, 2003), but not all in a significant way: the greatest discrepancy between the results of these two models is 5.01% in predicting SSE 180 Index movement; while the smallest one is 0.37% obtained in predicting CSI 300 Index movement. Specifically, when we zoomed in the accuracy, we can find that the outstanding performance of SVM against Neural Network is mainly reflected in higher value of specificity (Table 9) or in other words, a lower value of false-negative rate, but the low value of specificity show the powerlessness of both models in predicting -1 labeled tuples, indicating that results lean to the positive prediction. Additionally, the sensitivity values of Neural Network in SSE 50 and SSE 180 as well as the specificity in SZSE Composite and CSI 300 are higher than those of SVM, suggesting that SVM could not, at least in this case, win Neural Network on both sides significantly: either with higher sensitivity or higher specificity.

Besides the traditional charts, *IBM SPSS® Modeler* allows us to evaluate each model by comparing gains between SVM and Neural Network (Fig. 2), which are defined as the proportion of total hits that occurs in each percentile:

\[
\frac{\text{number of hits in percentile}}{\text{total number of hits}} \times 100\%
\]

Here “hits” represent the situation that a tuple with label +1 is successfully classified. In the figures, “Bestline” indicates the perfect confidence where all the tuples with label +1 are
successfully classified at first (hits = 100%); while “Baseline” indicates the perfectly random distribution of hits among the sample where each percentile contains the equal number of hits. Between them lie the gains lines of SVM and Neural Network.

In Fig. 2A and 2F, SVM and Neural Network curves twist together, showing almost the same performance. This is in accordance with the similar prediction accuracy in SSE 50 Index and CSI 300 Index. In Fig. 2B to 2E, however, SVM outperforms Neural Network especially on the right part of the figures, suggesting a better performance of SVM beyond Neural Network in predicting the latest index movement.
Fig. 2: The gain charts for models with selected features. A to F are curves for SSE 50 Index, SSE 180 Index, SSE Composite Index, SZSE 100 Index, SZSE Composite Index and CSI 300 Index, respectively.
Conclusions

For the first time an SVM model is built based on the macroeconomic fundamental indicators to forecast the daily movement direction of the popular indices in Chinese stock markets. Neural Network is also used as the benchmark to the SVM model.

Several conclusions could be drawn from the experiment results obtained above.

1. Feature selection is a reliable step in the process of building a forecasting model. Although previous studies claimed that the set of financial variables they identified captured the most relevant information for intrinsic value of the stock index, among 16 available features, only several showed the high correlation with the class and are introduced into the next step of the experiment (Table 5). And data in Table 7 suggests that the involvement of feature selection would not sacrifice too much (actually less than 0.6%) or even improves the accuracy. The results indicate that models using the small set of financial variables that have been carefully selected achieved comparable or even better results compared to those using a larger or the whole set of financial variables. Besides, the results also show that these features cannot be consistently used, because different stock markets have their own “atmospheres”, where input attributes may have different weights in influencing the movement of the market. The key variable in one market may become useless in another market under different conditions. This would be even significant between developed and developing countries. For example, in our case, S&P 500 Index was excluded from all three models built for forecasting the movement direction of SSE series indices, while it plays an important role in predicting the NIKKEI 225 movement in Japanese market (Huang et al., 2005).
2. Our results partly conform to prior research results. The results show that fundamental models based on publicly available information built by SVM could provide some help to make profits from the prediction of the index movement. However, the results do not support the high prediction accuracy, say, more than 70%, in other studies regarding developed stock market, such as the prediction on the weekly movement direction of NIKKIE 225 (Huang et al., 2005). Although the daily prediction in our study is more complicated than the weekly one, in our experiment, macroeconomic fundamental indicators based on historical trends to some extent appear to lack predicative power, a finding that is similar to, but not exactly, the semi-strong form of the Efficient Market Hypothesis. The semi-strong form of the Efficient Market Hypothesis holds that current security prices fully reflect all public available information. Thus investors could not earn abnormal profits through the fundamental analysis. If so, the prediction accuracy of our results should be in line with the result of Random Walk model, which is 50%. But around 60% accuracy (Table 7) means it’s a better result that could lead to some abnormal profits. Typically, developed markets are semi-strong form efficient, while there is evidence of semi-strong form inefficiency in some emerging markets. This is proven by our results: Chinese stock market, albeit prosperous, is still under-developed.

3. Our results corroborate previous reports. SVM enjoys some advantages over Neural Network in prediction accuracy (Table 7). This is mainly embodied in better prediction results of SVM on the latest input tuples (Fig. 2B to 2E). But for SSE 50 Index and CSI 300 Index, SVM does not give much more information than Neural Network. However, both models show poor performances on the prediction of -1 tuples (low value of specificity) (Table 9) because they tend to give positive predictions to those tuples which should be -1.
Discussion

Features selected by the filter model

One shining point of this experiment is the performance of the feature selection. Several previous studies did not apply feature selections, but simply used traditional indicators when building the forecasting models. However, our results demonstrate that not all classical input attributes fit the new environment (Chinese stock market). By using two feature selection models we obtained comparable or even better results than those obtained using all the features (Table 7). The computational time reduction resulting from fewer input features is not significant in our experiment since the scale of input data is not very large. However, many financial models that may employ SVM approach because of its high accuracy may involve hundreds of potential input features and several decades of historical data. In those cases, feature selection could be useful in reducing the computing time, because generally SVM consumes plenty of time in training step when the data is huge.

In addition, some interesting phenomena are observed as a result in the feature selection step when we interpret the results.

Firstly, S&P 500 Index- generally used economic condition indicator of the US- is not included in the forecasting models at all, indicating that regardless of the micro-level (SSE 50 Index) or the macro-level (SSE Composite Index, SZSE Composite Index and CSI 300 Index) of the economy, the economic influence of US on China is less than that on other export-oriented countries, like Japan (Huang et al., 2005), probably because US is not the biggest export country of China. But exchange rate of USD against CNY does influence the Chinese economy greatly. Lower exchange rate would raise the cost for products and slow down the economic growth.
Secondly, although interest rate is included in all models, the time period is different between indices of Shanghai Stock Exchange and indices of Shenzhen Stock Exchange: 10-year Interest Rate is used by models for SSE Indices; while 0.5-year Interest Rate is used by models for SZSE Indices. This interesting phenomenon is consistent with the fact that major stocks in Shanghai Stock Exchange are weighted shares issued by relatively large companies with large capitalization, which care more about the mid- or long-term interest rate; while small-cap companies predominate at Shenzhen Stock Exchange and they may be more concerned with the short-term interest rate.

Thirdly, NIKKEI 225 Index, EURO STOXX 50 Index and exchange rate of EUR against CNY also appear in every model. European Union is the biggest trading partner of China, so it is not surprising that both the index and exchange rate of EU are selected into the model. However, Japan is listed as the third biggest partner, behind the US. Why is the Japanese economic index, NIKKEI 225 Index, selected instead of S&P 500 Index? One possible explanation is that China and Japan are Asian countries, which share similar cultural and economic climates, leading to more privities, connections and common in fluctuation of the stock markets between them that we cannot perceive and estimate according to the rank list.

One problem associated with our feature selection algorithm, namely filter model, is that although feature selection is traditionally independent of establishing the learning model, separating these two steps might result in a loss of information relevant to classification and thus lower prediction accuracy.
**Prediction Accuracy**

Besides the algorithm deficiency, the “abnormally” low prediction accuracy could be partially explained by the following reasons.

Firstly, data is not complete. A considerable part of daily data was deleted in order to make sure that on each date every attribute has its data. So the model was trained on only about half of the original data.

Secondly, prediction accuracy is significantly affected by the time window selection. According to a study related to S&P 500 Index (R. Rosillo, D. de la Fuente, and J. A. L. Brugos, unpublished data), the accuracy varies from less than 60% to more than 90% when different time periods were chosen. A relatively large time window would bring in more inconsistency, while a small one would lead to poor generalization. Albeit difficult to find, a properly selected time window would improve the results by leaps and bounds.

Thirdly, several macroeconomic attributes that previously proved to be important for establishing the model, such as GDP, CPI and so on, are not included due to the lack of daily data. CPI or inflation rate is an important indicator because of its close relationship with real interest rate. Investors generally estimate the monetary policy through the level of interest rate, since the manipulation by the government on interest rate is an effective way to regulate and control the cash flow into or out of the stock market. Higher interest rate would attract money from the stock market to saving or deposit. And there is a feedback loop between interest rate and the real economy (Fig. 3). The fluctuation of CPI or inflation rate (Fig. 4) would contort the interest rate and make it less powerful in forecast. Besides CPI, Purchasing Managers’ Index (PMI) could have raised the prediction accuracy if it was involved in the model building because it is a macroeconomic precursor indicator of financial activities reflecting purchasing managers’
acquisition on goods and services, and widely used as the basis of national macro-economy adjustment especially for the export-oriented economic entity like China. But, also there is only monthly data of PMI, so we cannot employ it.

![Fig. 3: The feedback loop between interest rate and real economy.](image-url)
Fourthly, and the most importantly, I think the fundamental discrepancy between developed stock markets (e.g. US) and developing stock markets (e.g. China) leads to the significant difference of the prediction power and the reduction of the model generalization. It is reflected in the following aspects:

(1) Unlike the developed markets, Chinese stock market is not sufficiently sound in institution, operation and supervision. Players on the ground acts more like speculators than investors. Therefore, the model or the thought of structuring the model that performs well in developed markets, such as the US, Japan, German and so on, would not deliver a reliable answer in emerging and undeveloped markets, such as BRICs (Brazil, Russia, India and China).

(2) As a country just stepped up from the panned economic system, China still emphases the important function of micro-control. For example, in order to prevent the stock market bubbles and “hard landing” of the economy, in May 2007, Chinese government suddenly raised the stamp-tax. All indices in China plunged, and SSE Composite Index was

Fig. 4: Chinese inflation rate between Jan 2003 and Jan 2012 (Source: http://www.tradingeconomics.com).
chopped at the midpoint. Although their influence is huge on the stock market, these micro-control measures are temporary events that could not be abstracted as indicators. Another example is that the interest rates in China have not been marketized yet, so the actual contribution of interest rates to the model would be discounted.

(3) Compared to developed markets, Chinese stock market is young. SSE series have extremely short history and may vibrate strongly. Thus in the early stage of indices development, the instability and un-expectation compound the difficulty in forecasting the moving direction. Additionally, unlike the case in the US, the correlation between the Chinese stock market and Chinese economy is relatively loose, because the cyclical fluctuation of state-owned businesses, which holds sizable proportion of the top companies, has greater relationship with the change of industrial structure and industrial competition pattern, rather than the cyclical fluctuation of macro-economy.

(4) Compared to stock markets in the US, where institutional investors are majority, in China, individual investors account for most of the security holders. This boosts the volatility of the market movement and makes the prediction model perform poorly.

**FUNDAMENTAL ANALYSIS VS. TECHNICAL ANALYSIS**

Fundamental analysis and technical analysis are two mainstream financial analysis methods. As introduced before, fundamental analysis attempts to find the intrinsic value of the assets. Basically fundamental analysis is often employed in value investment: one of the most popular investment strategies. When we built the financial models to predict the target, fundamental indicators tend to give more explicit explanations to the results. Besides, as to the prediction accuracy, the model with fundamental indicators performs better than that with technical ones.
Two studies related to the movement prediction of several blue chips in Chinese stock market could provide the support for this claim (Yang et al., 2002; Zhang et al., 2006).

**IS SVM really a good choice?**

Though machine learning has been in a “hotspot” recently, critics have raised doubts about applications of the machine learning techniques for financial prediction. One criticism is the potential bias of the small data set when used for the out-of-sample testing, which is unlikely to be representative of the full range of market behaviors owning to the complexity of the computation. Another criticism is the method for results evaluation. The traditional way, prediction accuracy rate of the classification engine is not enough to reflect the performance of a trading strategy associated with the true market. Therefore, the results published in the reports can be misleading and may lead to potential losses.

These criticisms are not new. Some of the criticisms are similar to those targeted at a broader concept: data mining, which is a process of investigating data until a statistically significant relation is found. By feature selection, data preparation and, more importantly, by the experience or intuition to financial variables, one can always find some patterns from the enormous data, use them as reference for decision and make profits through trading. Choosing the time frame is also important: same model in different time frames may have different performance which will result in gains or losses. The market is extremely complex and changes with dizzying abruptness. There is no model that could answer to all the changes. Thus, the most important factor for prediction is the manipulators, who select data models, features and make decision according to the model. Good models only add the strength to an individual person, just like adding wings to a
tiger. Beyond all doubt, SVM is proven to be one of the most efficient and reliable models for classification and prediction. A soldier will always choose the most powerful weapon. Why not?
Reference


Vitae

Zhongyuan Wei, who was born on 16 Nov, 1984 in Shanxi province, P. R. China, is a master student of the class of 2010 in Lehigh University, who pursues Master of Sciences degree of Management Science in Department of Industrial and Systems Engineering. He obtained Bachelor of Economics and Bachelor of Science in Biological Science in Peking University in 2007. Then he obtained Master of Science in Biochemistry and Molecular Biology in Chinese Academy of Sciences in 2010.