中国科学院数学与系统科学研究院 | 国家数学与交叉科学中心

# FORECASTING DIRECTION OF CHINA SECURITY INDEX 300 MOVEMENT WITH LEAST SQUARES SUPPORT VECTOR MACHINE

**Shuai Wang ,Wei Shang**

Academy of Mathematics and Systems Sciences
Chinese Academy of Sciences

***Beijing, China , June 2014***

# Contents

# Background & Motivation

A challenging task to forecast the direction of stock index movement.

Due to the complexity of the financial market
& its various affected factors

**Features of Financial Market**

- Complicated
- Dynamic
- Evolutionary
- Nonlinear

**Affected Factors of Financial Market**

- Political events
- Economic fundamentals
- Investors' sentiment
- Other markets' movements

# Background & Motivation

**An accurate prediction of stock index movement**

## Investors


**provide reference value for the investors to make effective strategy**

## Policy Makers

**Also for policy maker to monitor stock market**

# CSI 300 Index

The first equity index launched by the **two** exchanges (Shanghai and Shenzhen) together.

**CSI 300 Index (000300.SS)** - Shanghai ★ Follow

**2,147.28** ↓8.70 (0.40%) 3:04AM EDT

| | | | |
|---|---|---|---|
| Prev Close: | 2,155.98 | Day's Range: | 2,145.75 - 2,160.03 |
| Open: | 2,154.40 | 52wk Range: | 2,086.97 - 2,644.36 |

Quotes delayed, except where indicated otherwise. Currency in CNY.

The underlying index of China security Index 300 future ---the only financial future in China.

- replicate the performance of **300** stocks traded in the Shanghai and Shenzhen stock exchanges.
- Covers about **one seventh** of all stocks listed on China's stock markets and about **60%** of the markets' value.


**It is able to reflect the price fluctuation and performance of China's Shanghai and Shenzhen stock markets**

# Details of CSI 300 Index

## The ten largest companies

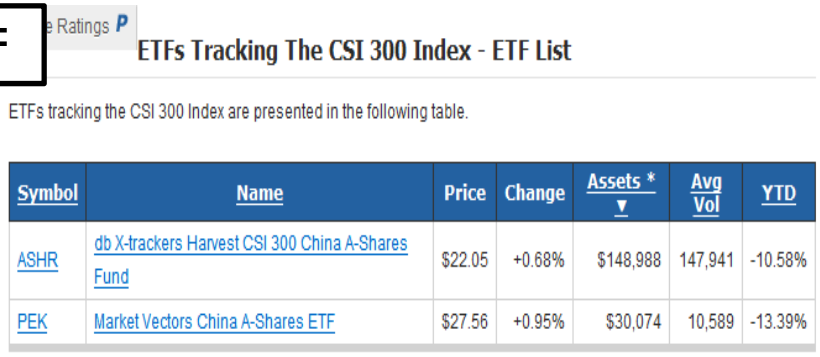
Ping An Insurance Group Co of China Ltd 3.92%
Citic Securities Co Ltd 3.64%
China Merchants Bank Co Ltd 2.98%
China Petroleum & Chemical Group 2.89%
Bank of Communications Co Ltd 2.60%
Baoshan Iron & Steel Co Ltd 2.49%
China Yangtze Power Co Ltd 2.39%
China Minsheng Banking Corp Ltd 2.24%
Shanghai Pudong Development Bank 2.23%
China Vanke Co Ltd 1.93%

## .The sector weightings

Finance 36.38%
Industry 15.93%
Basic Materials 13.55%
Energy 9.75%
Utilities 7.53%
Consumer Goods 7.01%
Capital 4.90%
Information Technology 2.11%
Telecommunications 1.50%
Health 1.42%

since April 8, 2005.
Its value is normalized relative to a base of 1000 on December 31, 2004.

ETF




| Symbol | Name | Price | Change | Assets * | Avg Vol | YTD |
|---|---|---|---|---|---|---|
| ASHR | db X-trackers Harvest CSI 300 China A-Shares Fund | $22.05 | +0.68% | $148,988 | 147,941 | -10.58% |
| PEK | Market Vectors China A-Shares ETF | $27.56 | +0.95% | $30,074 | 10,589 | -13.39% |

# Contents

# *Classification*

- ✓ Predicts categorical class labels(discrete or nominal)
- ✓ Classifies records (constructs a model) based on the training set including the class Labels and classifying attributes and then uses the rules(model) to classify new records

## A two-step process

### Model construction

Describe a set of predetermined classes

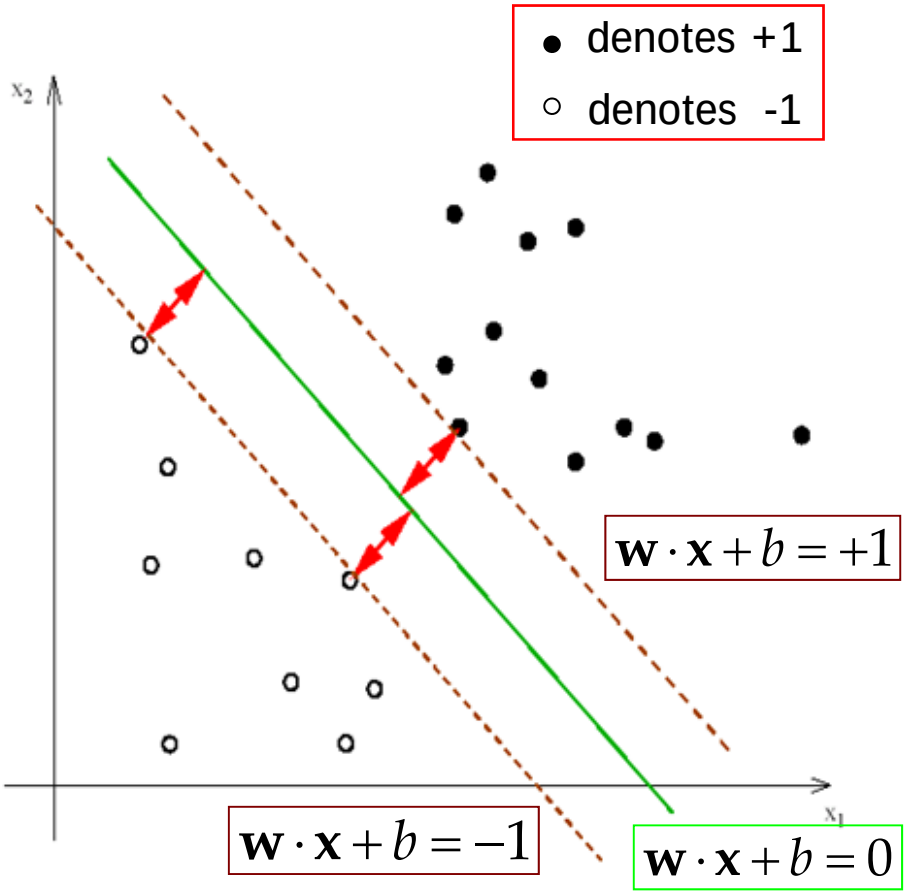
- ➢ Each sample is assumed to belong to a predefined class, as determined by the class label attribute
- ➢ The set of samples used for model construction is **training set**.
- ➢ The model is represented as classification rules, decision tree, or mathematical formulae.

### Model usage

Classify future or unknown objects

- ➢ Estimate accuracy of the model
  - The known label of test sample is compared with the classified result from the model.
  - Accuracy rate is the percentage of **testing set** samples that are correctly classified by the model.
  - Test set is independent of training set, otherwise over-fitting will occur
- ➢ If the accuracy is acceptable, use the model to **classify data** samples whose class labels are not known.

# *SVC Mathematically*




Given a set of linearly separable training examples,

$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$

Learning is to solve the following constrained minimization problem,

$$\text{Minimize: } \frac{\mathbf{w}\cdot\mathbf{w}}{2} \quad \left(\text{margin} = \frac{2}{\sqrt{\mathbf{w}\cdot\mathbf{w}}}\right)$$

$$\text{Subject to: } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1 \quad \text{for } y_i = +1$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1 \quad \text{for } y_i = -1$$

# LSSVC

SVC : a high computational complexity specially when computing large-scale QP problem

➢LSSVC takes equality constraints instead of inequality constraints in SVC.
➢ A squared loss function is taken for error variable in LSSVC

$$\min \quad \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{l} \xi_i^2$$

$$s.t. \quad y_i = w^T \varphi(x_i) + b + \xi_i, (i = 1, 2, ..., l)$$

where $\xi_i$ are the error variables and $\gamma$ is the penalty parameter

The final classification solution

$$f(x) = \text{Sign}\left( \sum_{i=1}^{l} w_i K(x, x_i) + b \right)$$

K (•) is the kernel function which can simplify the use of a mapping.

Gaussian RBF kernel function

$$K(x, x_i) = \exp\left( \|x - x_i\| / 2\sigma^2 \right)$$

# Benchmark methods

## AI: PNN

Probabilistic Neural Network (PNN) was proposed by Specht in 1990, and it built on the Bayesian strategy of classification.

## Discriminant analysis

- Discriminant analysis is a statistical technique to study the differences between two or more groups of objects with respect to several input (independent) variables.
- Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are employed

# Data Descriptions

Data range : April 27, 2005 to February 15, 2012, with a total of 1653 observations.

| X: Indicator name |
|---|
| MA10 (Simple 10-day moving average) |
| WMA10 (Weighted 10-day moving average) |
| MTM (Momentmum) |
| Stochastic K % |
| Stochastic D % |
| RSI (Relative Strength Index) |
| MACD (Moving average convergence divergence) |
| WR (Larry William's R %) |
| A/D Oscillator (Accumulation/Distribution) |
| CCI (Commodity Channel Index) |

- **Training dataset:** the former 80% of the data set (1322 observations to determine the specifications of the models and parameters.
- **Testing dataset:** the rest set of the data (331 observations) to evaluate the performances among various forecasting models.

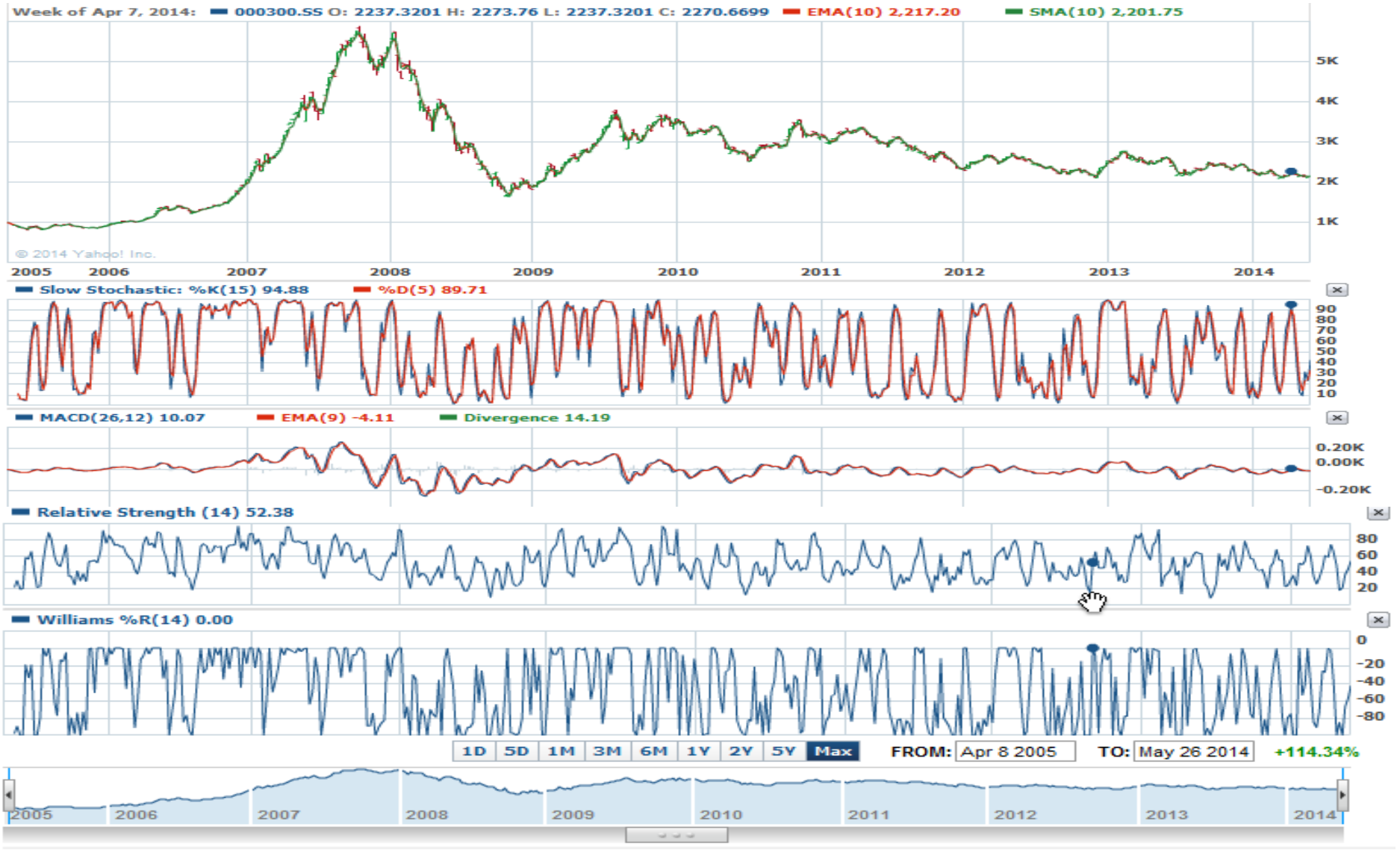
- Class one: Y=0. China Security Index 300 at time *t* is lower than that at time *t*-1
- Class two : Y=1. China Security Index 300 at time t is higher than that at time t-1

# Formula of Indicators

| Indicator name | Formula |
|---|---|
| MA10 (Simple 10-day moving average) | $\frac{C_t + C_{t-1} + ... + C_{t-9}}{10}$ |
| WMA10 (Weighted 10-day moving average) | $\frac{n \times C_t + (n-1) \times C_{t-1} + ... + C_{t-9}}{(n + (n-1) + ... + 1)}$ |
| MTM (Momentmum) | $C_t$ - $C_{t-n}$ |
| Stochastic $K$ % | $\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$ where $LL_t$ and $HH_t$ mean the lowest low and highest high in the last t days, respectively |
| Stochastic $D$ % | $\left(\sum_{i=0}^{n-1} \%K_{t-i}\right) / n$ |
| RSI (Relative Strength Index) | $100 - \frac{100}{1 + \left(\left(\sum_{i=0}^{n-1} Up_{t-i}\right)/n\right) / \left(\left(\sum_{i=0}^{n-1} Dw_{t-i}\right)/n\right)}$ where $Up_t$ means upward change and $Dw_t$ means downward change at time $t$. |
| MACD (Moving average convergence divergence) | $2 \times (DIFF - DEA)$, where $DIFF = EMA(C_t,12) - EMA(C_t,26)$, $DEA = EMA(DIFF,9)$, and $EMA(X,n) = \left(2 \times X + (n-1) \times EMA(X,n\text{-}1)\right) / (n+1)$ |
| WR (Larry William's R %) | $\frac{H_n - C_t}{H_n - L_n} \times 100$ |
| A/D Oscillator (Accumulation/Distribution) | $\frac{H_t - C_{t\text{-}1}}{H_t - L_t}$ |
| CCI (Commodity Channel Index) | $M_t - SM_t / 0.015 D_t$ where $M_t = (H_t + L_t + C_t)$, $SM_t = \left(\sum_{i=1}^{n} M_{t-i+1}\right)/n$, and $D_t = \left(\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert\right)/n$ |

Note: $C_t$ is the closing price at time $t$, $L_t$ is the low price at time $t$, $H_t$ is the high price at time $t$.

# Indicators


Week of Apr 7, 2014: 000300.SS O: 2237.3201 H: 2273.76 L: 2237.3201 C: 2270.6699 EMA(10) 2,217.20 SMA(10) 2,201.75
5K
4K
3K
2K
1K
© 2014 Yahoo! Inc.
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Slow Stochastic: %K(15) 94.88 %D(5) 89.71
90 80 70 60 50 40 30 20 10
MACD(26,12) 10.07 EMA(9) -4.11 Divergence 14.19
0.20K 0.00K -0.20K
Relative Strength (14) 52.38
80 60 40 20
Williams %R(14) 0.00
0 -20 -40 -60 -80
1D 5D 1M 3M 6M 1Y 2Y 5Y Max FROM: Apr 8 2005 TO: May 26 2014 +114.34%
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

# Summary statistics

| Indicator name | Max | Min | Mean | Standard deviation |
|---|---|---|---|---|
| MA10 | 5726.471 | 839.746 | 2699.383 | 1181.275 |
| WMA10 | 5765.633 | 837.377 | 2700.802 | 1180.632 |
| MTM | 896.980 | -1076.050 | 11.177 | 230.996 |
| *K* % | 99.100 | 4.353 | 57.956 | 27.473 |
| *D* % | 97.723 | 6.928 | 57.880 | 25.055 |
| RSI | 97.361 | 5.215 | 53.606 | 21.060 |
| MACD | 185.662 | -186.016 | 0.163 | 43.577 |
| WR | 100.000 | 0.000 | 41.957 | 33.485 |
| A/D Oscillator | 658.684 | -129.784 | 49.296 | 47.018 |
| CCI | 292.600 | -373.868 | 13.333 | 110.922 |

| | Year | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | |
| Decrease | 81 | 85 | 82 | 137 | 86 | 121 | 129 | 13 | 734 |
| % | 48.21 | 35.27 | 33.88 | 55.69 | 35.25 | 50.00 | 52.87 | 50.00 | 44.40 |
| Increase | 87 | 156 | 160 | 109 | 158 | 121 | 115 | 13 | 919 |
| % | 51.79 | 64.73 | 66.12 | 44.31 | 64.75 | 50.00 | 47.13 | 50.00 | 55.60 |
| Total | 168 | 241 | 242 | 246 | 244 | 242 | 244 | 26 | 1653 |

# Contents

# Empirical Results

➢The LSSVC performs best in all these direction forecasting methods in terms of training data and testing data.
➢The other artificial intelligence (AI) model, PNN performs better than Discriminant analysis in terms of training data, but has inferior performance in testing data. It may because of the neural networks are vulnerable to the over-fitting problem.
➢QDA performs better than LDA in terms of testing data, despite of inferior prediction performance of training data. The main reason may be that LDA assumes equal covariance in

all of the classes, which is not consistent with the properties of input variables.

| Evaluation indicator | LSSVC | PNN | QDA | LDA |
|---|---|---|---|---|
| Training accuracy | 92.97 | 92.89 | 86.87 | 88.18 |
| Testing accuracy | 89.12 | 80.97 | 87.92 | 87.31 |

# McNemar Test

**McNemar Test**:

✓one degree of freedom chi-square test which is applied to 2 × 2 contingency tables with a dichotomous variable, to determine whether the row and column marginal frequencies are equal.

✓The null hypothesis assumes that the total rows are equal to the sum of columns in the contingency table.

Comparison

McNemar values (p-values) for comparison of performance.

| | PNN | QDA | LDA |
|---|---|---|---|
| LSSVC | 0.679(0.410) | 4.654(0.031) | 10.321 (0.001) |
| PNN | | 0.327(0.568) | 2.326(0.127) |

- LSSVC outperforms LDA and QDA model at 1% and 5% significant level respectively.
- However, LSSVM does not significantly outperform PNN.
- PNN and two Discriminant analysis (QDA and LDA) do not significantly outperform each other.

# Contents

# Summary

**Main Works**

- Applied LSSVC to predict the movement of CSI 300 index.
- Compared the performance with PNN and two Discriminant analysis

**Results of Empirical Study**

- LSSVC performs best in all these direction forecasting methods in terms of training data and testing data.
- PNN performs better than Discriminant analysis in terms of training data, but has inferior performance in testing data.

**Main Conclusion**

- LSSVC is a promising method to forecast the direction of stock index.