Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 869 – 874

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management,

ITQM 2014

Forecasting Direction of China Security Index 300 Movement with Least Squares Support Vector Machine

Shuai Wang[a,*],Wei Shang[b]

[a]*NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*
[b]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

Due to the complexity of financial market, it is a challenging task to forecast the direction of stock index movement. An accurate prediction of stock index movement may not only provide reference value for the investors to make effective strategy, but also for policy maker to monitor stock market, especially in the emerging market, such as China. In this paper, we investigate the predictability of Least Square Support Vector Machine (LSSVM) by predicting the daily movement direction of China Security Index 300 (CSI 300). For comparing purpose, another artificial intelligence (AI) model, Probabilistic Neural Network (PNN) and two Discriminant Analysis models are performed. Ten technical indicators are selected as input variables of the models. Experimental results reveal that LSSVM method is very promising for directional forecasting for that it outperforms PNN, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) in both training accuracy and testing accuracy.

Keywords: Directional prediction; Least squares support vector machine; Probabilistic Neural Network;

* Corresponding author. Tel.: 86-010-62610334.
E-mail address: wshuai@amss.ac.cn

doi:10.1016/j.procs.2014.05.338

1. Introduction

The financial market movements prediction is regarded as one of the most challenging tasks of time series prediction since the financial market is complicated, dynamic, evolutionary and nonlinear[1]. In addition, the affected factors in financial market include political events, general economic conditions, investors' expectations and psychology, and other financial market movements[2]. It's meaningful to study the stock index movement of emerging markets such as that of China. China security Index 300 future is the only financial future in China. As the underlying index, CSI 300 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 integral performance of China's Shanghai and Shenzhen stock markets[3]. Therefore, an accurate prediction of CSI 300 movement may not only provide reference value for the investors to make effective strategy, but also for policy maker to monitor stock market.

Considering the previous studies in forecasting stock index direction movement, it can be classified into two categories: statistical models and artificial intelligence (AI) models. For examples, Discriminant Analysis (including Linear Discriminant Analysis, Quadratic Discriminant Analysis)[4,5], logit and probit binary models[4], the generalized methods of moments (GMM) with Kalman filter[6], random walk[6] and case based reasoning (CBR)[7]. AI models include Support Vector Machine (SVM) [2, 5,7-9] and Artificial Neural Network (ANN) [2,4-7,10,11]. Especially ANNs used include Probabilistic Neural Network (PNN) [4,6], back propagation neural network (BP)[5,7] ,radial basis function neural network [10]and Bayesian regularized artificial neural network[11].To sum up, the experiment results of previous researches show that AI models has better performance.

Particularly, as a modified version of SVM, Least Square Support Vector Machine (LSSVM) was proposed by Suykens and Vandewalle[12] in 1999. It retains principle of the structural risk minimization (SRM) and has important improvement of calculating speed with traditional SVMs. It is because of changing inequality constraints into equations and takes a squared loss function. Therefore, LSSVM solves a system of equations instead of a quadratic programming problem. Due to these advantages of LSSVM, we employ it as the classification model.

This study aims to explore the predictability of Least Square Support Vector Machine (LSSVM).The rest of the article is organized as follows. In section 2, we introduce the basic theory of LSSVM for classification. The experimental study is illustrated in section 3. Section 4 contains the concluding remarks.

2. Theory of LSSVM for classification

Support vector machine (SVM) was firstly proposed by Vapnik in 1995 based on the principle of structural risk minimization (SRM), and it have been applied in many fields of classification and regression[13]. It has been proved to possess its excellent capabilities even for small sample, by minimizing an upper bound of the generalization error. However, SVM training is a time consuming process specially when analyzing huge dataset. For this purpose, least squares support vector machine (LSSVM) is proposed to overcome these shortcomings[8].In recent years, LSSVM has been successfully used for modeling economic time series[14, 15]. In the following part, there will be brief description of LSSVM for classification.

Considering binary class problem with the dataset $G=\{(x_i,y_i),i=1,2...l\}$, the input attributes vector is $X=(x_1,...x_n)$ and the output class labels are $y_i\in\{-1,1\}$.The basic concept of LSSVM is using linear model to implement nonlinear class decision boundaries by mapping the original input data X into a high-dimensional feature space via a nonlinear mapping function φ(•). Usually, the LSSVM classification function can be formulated as follows,

$$f(x)=Sign\left(w^T\varphi(x)+b\right) \tag{1}$$

where φ(x) is called the nonlinear function mapping from input space X into a high-dimensional feature space. Coefficients w and b are obtained by minimizing the upper bound of generalization error. Accordingly, Eq. (1) can be got by solving the following optimization problem:

$$\begin{aligned} &\min \quad \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{l} \xi_i^2 \\ &s.t. \quad y_i = w^T \varphi(x_i) + b + \xi_i, (i = 1, 2, ..., l) \end{aligned} \tag{2}$$

where ξ_i are the error variables and γ is the penalty parameter. By using Lagrangian function and the Karush-Kuhn-Tucker (KKT) conditions for optimality of Eq. (2), we can get the final classification solution of the primal problem:

$$f(x) = \mathrm{Sign}\left(\sum_{i=1}^{l} w_i K(x, x_i) + b \right) \tag{3}$$

In Eq. (3), K (•) is the kernel function which can simplify the use of a mapping. In this investigation, Gaussian RBF kernel function $K(x, x_i) = \exp\left(\|x - x_i\| / 2\sigma^2 \right)$ with a width of σ is used. More details about LSSVM can refer to[12, 16].

3. Experiment Study

In this section, the research data in this study is first described. Then, experimental results and corresponding analysis and explanations are reported.

3.1. Data Descriptions

In this study, ten technical indicators are used as input variables to predict the direction change of daily China Security Index 300[2, 7]. Table 1 summarizes the selected attributes and their formulas. The summary statistics for each indicator is presented in Table 2.The directions of daily change of China Security Index 300 are categorized as "0" and "1". "0" means that China Security Index 300 at time t is lower than that at time t-1 , while China Security Index 300 at time t is higher than that at time t-1 , the direction is "1".

Table 1. The technical indicators and their formulas.

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)	$\frac{H_t - C_{t-1}}{H_t - L_t}$
CCI (Commodity Channel Index)	$M_t - SM_t / 0.015 D_t$ where $M_t = (H_t + L_t + C_t)$, $SM_t = \left(\sum_{i=1}^{n} M_{t-i+1}\right) \Big/ n$, and $D_t = \left(\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert\right) \Big/ n$

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.

Table 2. Summary statistics of the indicators.

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

The data set covers the period from April 27, 2005 to February 15, 2012, with a total of 1653 observations. In these China Security Index 300 data, the former 80% of the data set (1322 observations) is taken as the training dataset which is used to determine the specifications of the models and parameters. The rest set of the data (331 observations) is chosen as testing dataset to evaluate the performances among various forecasting models. The number of "decrease" and "increase" for each year is given in Table 3.The numbers of decreasing direction is 734, with 44.4% of all samples.

The original data are scaled into the range of [-1, 1] to ensure that the larger value input attributes do not overwhelm smaller value inputs.

Table 3. The number of cases in the China Security Index 300 data set.

	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

3.2. Experimental Results

In this study, LSSVM is implemented via LSSVMlab1.8 toolbox of MATLAB software package. The kernel type selection of SVMs is usually based on application-domain knowledge and may reflect distribution of input values[17]. Due to the good performance of Gaussian kernels under smoothness assumptions[18], the Gaussian radial basis function is used as the kernel function of LSSVM. As the penalty parameter γ and kernel parameter σ should be predetermined and play important roles in the performance of SVM[18], 10-fold cross-validation with grid search method is used to tune the parameters.

For comparing purpose, Probabilistic Neural Network (PNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) are applied as the benchmark approaches to LSSVM.

Probabilistic Neural Network (PNN) was proposed by Specht in 1990, and it built on the Bayesian strategy of classification. A complete description of PNN and its mathematical background can refer to[19].The MATLAB Neural Network 5.0 toolbox was used to construct PNNs and test their classification accuracy.

Discriminant analysis is a statistical technique to study the differences between two or more groups of objects with respect to several input (independent) variables. In this study, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are employed by using MATLAB Statistic toolbox.

The classification performance is evaluated by hit ratio of training data and testing data respectively. Accordingly, the best results of different models are reported in Table 4.

Table 4. The performance Comparison of different models.

Evaluation indicator	LSSVM	PNN	QDA	LDA
Trianing accuracy	92.97	92.89	86.87	88.18
Testing accuracy	89.12	80.97	87.92	87.31

From Table 4, The LSSVM performs best in all these direction forecasting methods in terms of training data and testing data, which shows that LSSVM has stronger predictability. 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. Comparing the two Discriminant analysis methods, 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.

For further examination about whether LSSVM significantly outperforms the other models, the McNemar test is performed. The test is a nonparametric test in the analysis of pairs of matched samples (Y_{1i}, Y_{2i}), and is useful with before-after measurement of the same subjects [20]. Specifically, it is 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. Table 5 shows the result of the McNemar test to compare the prediction performance of the testing data.

Table 5. McNemar values (p-values) for comparison of performance.

	PNN	QDA	LDA
LSSVM	0.679(0.410)	4.654(0.031)	10.321 (0.001)
PNN		0.327(0.568)	2.326(0.127)

From table 5, it can be found that LSSVM outperforms LDA and QDA model at 1% and 5% significant level respectively. However, LSSVM does not significantly outperform PNN. In addition, the results of table 5 also demonstrate that PNN and two Discriminant analysis (QDA and LDA) do not significantly outperform each other.

4. Conclusion

It is important for the development of trading strategies in stock market to predict the direction of movements of stock index. Successful prediction is helpful for financial traders to decide whether to buy or sell. Due to the complexity of financial market, this study attempted to predict the direction of movement of China Security Index 300 with Least Square Support Vector Machine. Probabilistic Neural Network (PNN) and two Discriminant analysis models (QDA and LDA) are performed to predict the direction changes of CSI 300 on the daily data from 2005 to 2012. According to the empirical study, we find that LSSVM is superior to other direction movement predicting models in terms of training and testing data. This indicates that LSSVM is a promising method for financial direction movement prediction which may be a result of structural risk minimization principle and its improvement for SVM. In addition, PNN performs better than QDA and LDA in training data but not the same in testing data, which may because of that neural networks are more vulnerable to the over-fitting problem.

Acknowledgements

This research is supported by Beijing Municipal Commission of Education (Key Project of Science and Technology Plan, No. KZ201411232036) and the National Natural Science Foundation of China (No. 71171186, No. 91224006).

References

1. Abu-Mostafa Y S, Atiya A F. Introduction to financial forecasting. *Applied Intelligence*, 1996; **6**(3): 205-213.
2. Kara Y, Acar Boyacioglu M, Baykan Ö K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Application*s, 2011; **38**(5): 5311-5319.
3. Wei Y, Wang Y, Huang D. A copula–multifractal volatility hedging model for CSI 300 index futures. *Physica A: Statistical Mechanics and its Applications*, 2011; **390**(23): 4260-4272.
4. Leung M T, Daouk H, Chen A S. Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, 2000; **16**(2): 173-190.
5. Huang W, Nakamori Y, Wang S Y. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 2005; **32**(10): 2513-2522.
6. Chen A S, Leung M T, Daouk H. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, 2003; **30**(6): 901-923.
7. Kim K. Financial time series forecasting using support vector machines. *Neurocomputing*, 2003; **55**(1): 307-319.
8. Son Y, Noh D, Lee J. Forecasting trends of high-frequency KOSPI200 index data using learning classifiers. *Expert Systems with Applications*, 2012; **39**(14): 11607-11615.
9. Zhiqiang G, Huaiqing W, Quan L. Financial time series forecasting using LPP and SVM optimized by PSO. *Soft Computing*, 2013; **17**(5): 805-818.
10. Shen W, Guo X, Wu C, et al. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. *Knowledge-Based Systems*, 2011; **24**(3): 378-385.
11. Ticknor J L. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, 2013; **40**(14): 5501-5506.
12. Suykens J A K, Vandewalle J. Least squares support vector machine classifiers. *Neural processing letters*, 1999; **9**(3): 293-300.
13. Vapnik V. *The nature of statistical learning theory*. Springer; 2000.
14. Tang L, Yu L, Wang S, et al. A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. *Applied Energy*, 2012; **93**: 432-443.
15. Wang S, Yu L, Tang L, et al. A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China. *Energy*, 2011; **36** (11): 6542-6554.
16. Van Gestel T, De Brabanter J, De Moor B, et al. *Least squares support vector machines*. Singapore: World Scientific; 2002.
17. Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, 2004; **17**(1): 113-126.
18. Tay F E H, Cao L. Application of support vector machines in financial time series forecasting. *Omega*, 2001; **29**(4): 309-317.
19. Specht D F. Probabilistic neural networks. *Neural networks*, 1990; **3**(1): 109-118.
20. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947; **12**(2): 153-157.