A New Approach of Stock Price Trend Prediction Based on Logistic Regression Model

Jibing Gong
Computer Department of Yanshan University
Qinhuangdao, China
e-mail: gongjibing@163.com

Shengtao Sun
Computer Department of Yanshan University
Qinhuangdao, China
e-mail: xysst@ysu.edu.cn

***Abstract*—In our economic society, future stock price trend is very hot focus that the investors concern about. Challenges still exist in stock price prediction model regarding significant time-effectiveness of prediction, the complexity of methods and selection of feature index variables. In this paper, we present a new approach based on Logistic Regression to predict stock price trend of next month according to current month. Characteristics of our method include: (1) Feature Index Variables are easy to both understand for the private investor and obtain from daily stock trading information. (2) the prediction procedure includes unique and crucial operation of selecting optimizing prediction parameters. (3) significant time-effectiveness and strong purposefulness enable users predict stock price trend of next month just through considering current monthly financial data instead of needing a long term procedure of analyzing and collecting financial data.**

Shenzhen Development stock A (SDSA) from RESSET Financial Research Database is chosen as a study case. The SDSA's daily integrated data of three years from 2005 to 2007 is used to train and test our model. Our experiments show that prediction accuracies reach as high as at least 83%. In contrast to other methods, e.g. RBF-ANN prediction model, our model is lower in complexity and better accuracy in prediction.

Keywords- Stock Price Trend Prediction; Logistics Regression Model; Regression Coefficients

I. INTRODUCTION

Stock price trend prediction has always been one of the hottest topics in research. However, challenges still exist in stock price trend prediction regarding financial index selection for prediction model and low complexity in forecasting process.

In this paper, we present a new approach based on Logistic Regression to predict stock price trend of next month according to current month. In our proposed approach, Logistic Regression Prediction Model is constructed through a serial of procedure including choosing samples, training our model, and selecting optimizing regression coefficients. The innovative Feature Index Variables are introduced into our proposed model. They are easy to both understand for the private investor and obtain from daily stock trading information. Unique and crucial process of selecting optimizing regression parameters is included for obtaining better prediction accuracy.

This paper makes the following contributions to address those above challenges: (1) we propose a new approach based on Logistic Regression Model for predicting stock price trend; (2) we introduce innovative feature index variables into our prediction model. They are not only easy to both understand for the private investor and obtain from daily stock trading information, but also make our approach perform better than other method; (3) our proposed approach includes special optimization process to select optimizing regression parameters. This makes our prediction model have better accuracy.

The remainder of this paper is organized as follows. In section 2, we propose a new approach based on Logistic Regression Model to predict stock price trend, which includes four procedures and one Logistic Regression Model for stock price trend prediction (SPTP-LRM). Section 3 gives the experimental details and prediction results. Section 4 contrasts our proposed approach with the existing RBF-ANN model for stock price trend prediction, and analyzes comparison effects and results. The related work is presented in section 5. Finally, conclusions and future work are drawn in section 6.

II. OUR NEW APPROACH FOR STOCK PRICE TREND PREDICTION

A. Logistic Regression Model for Stock Price Trend Prediction (SPTP-LRM)

Logistic Regression is specially fit for those dependent variables for binomial or multinomial classification. In stock price prediction, the trend of next month's price may be categorized into two classes: '1' or '0'. '1' indicates average price of the next month is higher or equal to one of the current month, and '0' indicate a downtrend. We use previous stock prices, such as highest price, lowest price and so on, as financial index to enhance the predictability of the monthly stock price trend.

Logistic Regression Model is represented by

$$p = \frac{\exp(c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k)}{1 + \exp(c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k)} \quad (2\text{-}1)$$

To predict stock price trend of average price of the *i*-th month, such stock price is considered to have downtrend if the Logistic Regression value p_i is close to 0 (or is equal to 0). Otherwise, such stock price is considered to have uptrend

978-0-7695-3687-3/09 $25.00 © 2009 IEEE
DOI 10.1109/NISS.2009.267

if the Logistic Regression value p_i is close to 1(or is equal to 1). Moreover, the more far from 0 p_i is, the lesser probability the stock prices have downtrend, and vice versa. Natural logarithm operations are performed on the formula obtained through the likelihood function of joint density function with *n* samples, and then its Log Likelihood Function is [1]:

$$\ln LF = \sum_{i=1}^{n}\left[y_i\left(c_0 + \sum_{k=1}^{m} x_{ki} \right) - \ln\left(1+\exp\left(c_0 + \sum_{k=1}^{m} x_{ki} \right)\right)\right] \quad (2\text{-}2)$$

where x_{ki} represents the *k*-th feature index variable of the *i*-th month, $k = 0,1,2,\cdots,m$ and $i = 0,1,2,\cdots,12$. For estimating all parameters $c_j (j = 0,1,2,\cdots m)$ and obtaining the minimal value of $\ln LF$, we perform partial differentiation operation to all parameters c_j on formula (2-2), and then set its value as 0. Finally, we can obtain Likelihood Equations as follows [1]:

$$\frac{\partial \ln LF}{\partial c_0} = \sum_{i=1}^{n}\left[y_i - \frac{\exp\left(c_0 + \sum_{k=1}^{m} x_{ki} \right)}{1+\exp\left(c_0 + \sum_{k=1}^{m} x_{ki} \right)}\right] = 0 \quad (2\text{-}3)$$

$$\frac{\partial \ln LF}{\partial c_k} = \sum_{i=1}^{n}\left[y_i - \frac{\exp\left(c_0 + \sum_{k=1}^{m} x_{ki} \right)}{1+\exp\left(c_0 + \sum_{k=1}^{m} x_{ki} \right)}\right] \cdot x_{ki} = 0$$

$$k = 0,1,2,\cdots,m \quad (2\text{-}4)$$

We can have *m+1* equations through jointing formulas (2-3) and (2-4), and obtain solutions $\hat{c}_j$ of all above parameters $c_j (j = 0,1,2,\cdots,m)$. In another word, $\hat{c}_j$ just is the parameter values which we want to estimate.

Specially, we may have about 20 equations through jointing formulas (2-3) and (2-4) because there are about 20 trading days in a month. For example, we may obtain $c_{10}, c_{11}, \cdots, c_{1m}$ through training the samples of the 1st moth in 2005 year, and then we can continue to compute the rest of estimation value in turn. Finally, 12 groups of estimated value will be obtained. One of methods achieving $\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_m$ is to perform averaging operation through formula (2-5).

$$\hat{c}_0 = \frac{1}{k}\sum_{i=1}^{k} c_{i0} \quad , \quad \hat{c}_1 = \frac{1}{k}\sum_{i=1}^{k} c_{i1} \quad , \cdots, \quad \hat{c}_m = \frac{1}{k}\sum_{i=1}^{k} c_{im} \quad (2\text{-}5)$$

where *k* is the number of month including different group of samples, and *m* is the number of feature index variables. Here, set *k*=12 and *m*=9. $\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_m$ are those estimation parameters that we want to obtain for constructing our Logistic Regression Model. Another method is to selecting optimizing $\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_m$ from above 13 groups including 12 groups corresponding to 12 months and 1 group obtained through formula (2-5).

To predict stock price trend, the financial index chose should be irrelevant and sufficient. As a result, we choose Previous close price, Opening price, Highest price, Lowest price, Closing price, Composite weight price, Daily Turnover, Amount traded, and Number traded into consideration. The name list of feature index variables is displayed in Table I.

TABLE I. THE NAME LIST OF FEATURE INDEX VARIABLES

FIV	VN	FIV	VN
x_1	Previous close price	x_6	Composite weight price
x_2	Opening price	x_7	Daily Turnover
x_3	Highest price	x_8	Amount traded
x_4	Lowest price	x_9	number traded
x_5	Closing price		

FIV: Feature Index Variables, VN: Variables Name

B. Introduction of our proposed approach for predicting stock price trend

The basic idea of our proposed approach is embodied during the following processes.

1) Choosing samples. There are so many kinds of financial data in RESSET Financial Research Database [2], e.g. Stock data, Foreign Exchange data, Fund data, Futures data and so on. For predicting monthly stock price trend, we choose Stock Integrated Index data as training and testing samples. The form and style of the data is showed in Table II.

These columns in Table Ⅱ are just a part of all fields in stock integrated index data. These data are real and the kernels of stock data. Our samples data involve three years from 2005 to 2007. The number of samples reaches about 800.

Figure 1 displays stock price trend through analyzing stock integrated index data of Shenzhen Development Stock A (SDSA) from year 2005 to 2007, respectively. From Figure 1, rising and dropping information of stock prices showed by trend curves originates from practical stock integrated index data. So, it is very reasonable and effective that we chose stock integrated index data to predict stock price trend.

2) Training SPTP-LRM Model. The stock integrated index data of full year 2005 is used in training process. The iteration operations are repeated until the values of regression coefficients are generated. For example, we input the data of December of year 2005 into our SPTP-LRM Model with SPSS, and then obtain the results like the values in Table Ⅲ. In practical training process, not all data of year 2005 is useful and effective. These data of second month,

TABLE II. THE FORM AND STYLE OF TRAINING AND TESTING SAMPLES DATA

股票代码 _Stkcd	股票名称 _Lstknm	日期 _Date	前收盘价 _PrevClPr	开盘价 _Oppr	最高价 _Hipr	最低价 _Lopr	收盘价 _Clpr	复权价 _AdjClpr	成交量 _Trdvol	成交金额 _Trdsum	成交笔数 _Trades
000001	深发展 A	2005-01-04	6.59	6.59	6.59	6.46	6.52	4.54	1760832	11465603	1803
000001	深发展 A	2005-01-05	6.52	6.52	6.55	6.35	6.46	4.5	3222144	20718559	3054
000001	深发展 A	2005-01-06	6.46	6.5	6.59	6.45	6.52	4.54	2666413	17333840	2738
000001	深发展 A	2005-01-07	6.52	6.58	6.6	6.46	6.51	4.54	1886151	12302853	1660
000001	深发展 A	2005-01-10	6.51	6.51	6.59	6.37	6.59	4.59	2632055	17111498	2149
…	…	…	…	…	…	…	…	…	…	…	…

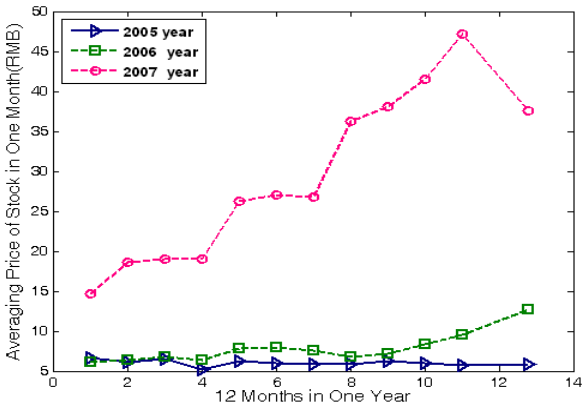

Figure 1. Price Trend of SDSA from year 2005 to 2007

fifth month and eighth month is invalid since they have linear dependence. Therefore, we cannot use these data to perform Logistic Regression. In addition, tenth month only includes seven records while SPSS need at least nine parameters to run. So, further analysis cannot be done for this reason. Finally, eight groups of prediction parameters were obtained to construct our model named as Model1, Model3, Model4, Model6, Model7, Mode9, Model11, and Model12, respectively.

TABLE III. REGRESSION COEFFICIENTS OF DECEMBER IN YEAR 2005 THROUGH THE TRAINING PROCESS

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 2[a]	x_{k1}	-359.682	5.815E5	.000	1	1.000	.000
	x_{k2}	-1.266E3	5.885E5	.000	1	.998	.000
	x_{k3}	-588.331	3.199E5	.000	1	.999	.000
	x_{k4}	2.104E3	7.501E5	.000	1	.998	.
	x_{k5}	-3.053E3	4.123E6	.000	1	.999	.000
	x_{k6}	4.572E3	5.632E6	.000	1	.999	.
	x_{k7}	.000	.201	.000	1	.999	1.000
	x_{k8}	.000	.032	.000	1	.999	1.000
	x_{k9}	.008	31.521	.000	1	1.000	1.008
	Constant	-29.047	1.513E6	.000	1	1.000	.000

In addition, we gained another two groups of technical parameters through averaging operation and observation. Trend curves described by those above 10 groups of regression coefficients are compared in Figure 2 and 3.

3) Selecting optimizing regression coefficients. We used financial data of year 2006 to determinate finial regression coefficients from 9 candidate groups of ones. Here, the nine candidate groups of regression coefficients include the eight groups of ones from training process and one group of ones from taking average operation (as showed formula 2-5). We offered some results of analyses on the SPSS and adjust the thresholds to 0.5 for better price trend prediction. Let $p_i = 1$ when $p_i \geq 0.5$, and let $p_i = 0$ when $p_i < 0.5$. The digital 1 represented uptrend of stock price in next overall month according to average price of the current month, and the digital 0 represented downtrend. Figure 2 and Figure 3 contrast all nine models with different groups of regression coefficients in order to select optimizing Feature Index Variables.

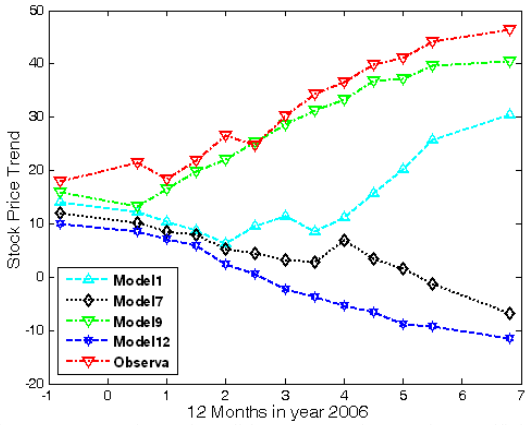

Figure 2. Comparison of candidate groups of regression coefficients (a)

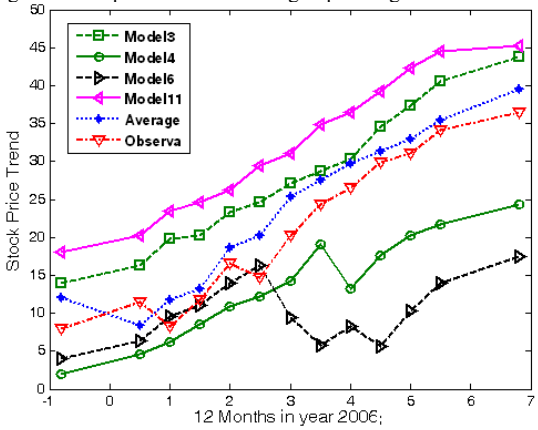

Figure 3. Comparison of candidate groups of regression coefficients (b)

From both Figure 2 and Figure 3, we may have conclusions that the "Average Model" mostly match up observation result. That is, the prediction result of the "Average Model" is best satisfying. Therefore, it was selected as optimizing regression coefficients in subsequent testing process.

4) Building and testing our model. According to selected optimizing regression coefficients, we build our logistic regression model as $P = \frac{e^s}{1+e^s}$ where $s = -480.198 - 556x_1 + 5.937x_2 - 38.654x_3 + 294.907x_4 + 772.777x_5 - 570.215x_6 + 0x_7 + 0x_8 + 0.008x_9$.

To test our SPTP-LRM model, we predicted monthly stock price trend of full year 2007.

The input of our model is averaging value of all feature index variables on current month, and the output is probability p_i representing SDSA price trend, $p_i \in [0,1]$. The greater probability p_i become, the greater possibility of stock price rising trend has, and vice versa.

III. EXPERIMENTS AND RESULTS

In our experiments, daily integrated data of Shenzhen Development stock A (SDSA) was used as training and testing samples, and obtained from RESSET Financial Research Database [2]. We applied the daily integrated data of full year 2005 as training data, the data of full year 2006 as selecting data, the data of full year 2007 as testing data. During the training process, we obtained 8 groups of regression coefficients. We got another group of regression coefficients through averaging the same kind of parameter from 8 groups (see also formula 2-5). For example, the value of parameter c_0 was achieved through the equation $c_0 = \frac{1}{n}\sum_{i=1}^{n} c_{n0}$, n=12. For more satisfying prediction results, we compared the nine groups of regression coefficients with observation price trend. Finally, from Figure 2 and 3, we selected "Average Model" as optimizing regression coefficients of our proposed STPT-LRM model.

We obtained stock price trend of every month in year 2007 with our STPT-LRM model through running SPSS. Table Ⅳ shows experiments data in testing our model. Columns from third to tenth are average value of every Feature Index Variables of 12 months. The last column displays the prediction results.

For clearly illustrate prediction effectiveness of our STPT-LRM model, we compared its prediction results with observation price trend in Figure 4. From Figure 4, there are two prediction errors, signed by dashed rectangle, which mean two inaccurate forecasting in 12 prediction points. So, we evaluated the prediction accuracy of our model is about 83.3%.

TABLE IV. EXPERIMENTS DATA IN TESTING OUR MODEL

Stock	Date	X1_Prev	X2_Oppr	X3_Hipr	X4_Lopr	X5_Clpr	X6_AdjClpr	X7_Trdvol	X8_Trdsum	Trend
SDSA	2007-01	16.26	16.44	16.88	16.05	16.49	11.5	69207082	1016722868	1
SDSA	2007-02	19.03	19.01	19.52	18.61	19.03	13.26	66285498	1222286994	1
SDSA	2007-03	18.75	18.71	19.1	18.31	18.74	13.06	22927483	418404551	0
SDSA	2007-04	21.59	21.66	22.27	21.32	21.96	15.3	31360278	609052911	1
SDSA	2007-05	27.69	27.84	28.57	27.12	27.99	19.51	25374071	689417015	1
SDSA	2007-06	32.45	32.86	33.73	31.28	32.28	27.76	130252684	4182345222	0
SDSA	2007-07	29.68	29.79	30.83	29.26	30.1	23.08	32376147	875614738	1
SDSA	2007-08	38.07	38	39.01	37.17	38.14	29.25	29143344	1042756522	1
SDSA	2007-09	36.93	36.96	37.76	36.33	37.03	28.4	19369706	741200572	1
SDSA	2007-10	39.4	40.3	41.46	39.87	40.81	31.3	27622941	1153482332	1
SDSA	2007-11	40.43	40.28	41.14	39.18	39.88	30.59	22331877	1072296819	1
SDSA	2007-12	37.76	37.82	38.57	37.24	37.89	29.06	33351253	1281146398	1

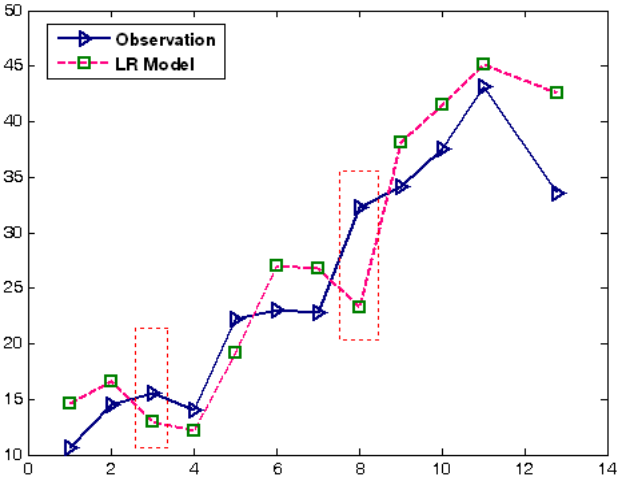

Figure 4. stock price trend curve of our proposed STPT-LRM model

IV. CONTRASTS AND ANALYSES

To fully illustrate good performance of our model, we compared our STPT-LRM model with Artificial Neural Network base Radial Basis Function (RBF-ANN) model. RBF-ANN model [3,4] was used to predict monthly SDSA average price trend of full year 2007.

Table Ⅴ shows weight parameters of RBF-ANN obtained with SPSS, and schematic diagram of ANN was showed in Figure 5. Experiments results and comparison with observation price trend were showed in Figure 6. For clearly displaying uptrend or downtrend of stock average price, we increase value in Y axis in order that all curves didn't overlap.

From Figure 6, we found two errors of prediction in 12 prediction points, and so the prediction accuracy of our proposed approach can reach at least 83.3%. Familiarly, from Figure 6, the prediction accuracy of ANN method can also reach at least 83.3%.

The same prediction accuracies state that our SPTP-LRM model is feasible and practical. In contrast to RBF-ANN model, our STPT-LRM model is lower in complexity, higher in efficiency, and easier to understand and realize. RBF-ANN model has some disadvantages, such as less efficiency and higher complexity, but its prediction accuracy will be higher than our model through longer term and repeated training on it.

Through contrasting with RBF-ANN model, our proposed approach has better prediction accuracy and is low

in complexity. The reasons include the following aspects: (1) to great extent, prediction ability of Logistic Regression is not sensitive to regression coefficients variance. Therefore, prediction accuracies of our proposed Model with different regression coefficient have stability. (2) selecting process was adopted by us to obtain optimizing regression coefficients through examining stock integrated index data of full year 2006. (3) in contrast to other logistic regression models for predicting stock price trend [5], feature index variables are more reasonable and comprehensible, and all of them are the kernel of the factors effecting stock price.

TABLE V. WEIGHT PARAMETERS OF RBF-ANN

		Predicted					
		Hidden Layer[a]				Output Layer	
Predictor		H(1)	H(2)	H(3)	H(4)	[Trend=0]	[Trend=1]
Input Layer	X1_PrevClPr	-.805	.975	.128	-.951		
	X2_Oppr	-.833	.974	.141	-.949		
	X3_Hipr	-1.291	.791	.619	-.927		
	X4_Lopr	-1.464	.830	.361	-.698		
	X5_Clpr	-1.778	.494	.904	-.605		
	X6_AdjClpr	-1.751	.510	.882	-.618		
	X7_Trdvol	-.030	-.044	.908	-.662		
	X8_Trdsum	-.088	-.022	.925	-.678		
	X9_Trades	.445	-.177	.912	-.695		
Hidden Unit Width		1.167	.850	1.228	.663		
Hidden Layer	H(1)					.839	.161
	H(2)					1.098	-.098
	H(3)					.158	.842
	H(4)					-.355	1.355

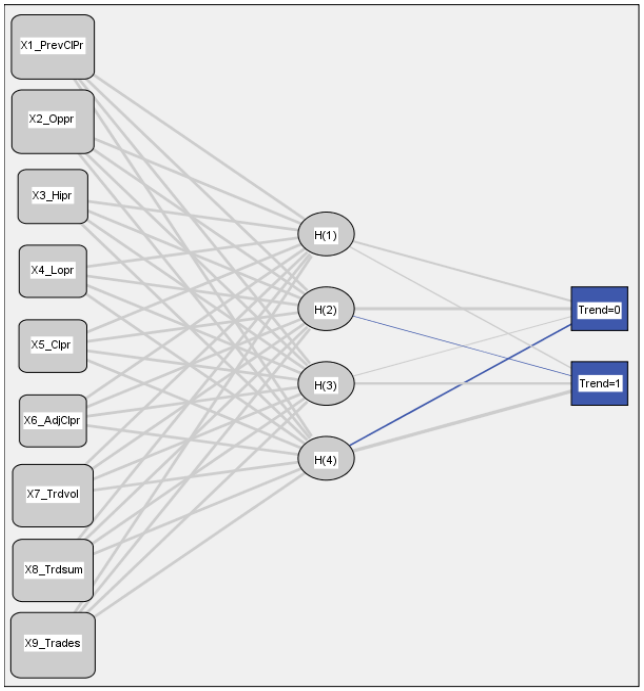

Figure 5. Schematic diagram of RBF-ANN

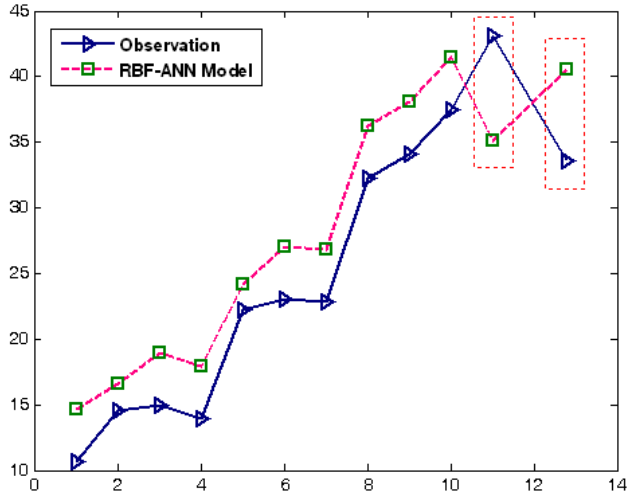

Figure 6. Stock price trend curve of RBF-ANN model

Through testing and contrasting experiments, we may have the following useful conclusions: (1) Logistic Regression Model is high effective and efficient in predicting stock price trend. (2) it is very important to select appropriate financial index in prediction model. (3) it is necessary to perform optimization process through generating different candidate groups of regression coefficients and selecting optimizing one group.

V. RELATED WORKS

Logistic regression model has its particular advantages in stock market prediction. In 2000, David West [6] built up 5 kinds of Neural Network Models and 5 kinds of statistical classification models. This model had the highest precision of discrimination among these 10 models. Sulin Pang used Logistic Regression Model to predict the tendency of stock price in 2004 [7-9]. In 2007 year, Zheng Mei and Miao Jia [10] applied Logit model to predict the stock tendency of Shanghai stock market and made some analysis and comparison with ARMA model.

Artificial Neural Network based on Radial Basis Function (RBF-ANN) is a kind of effective forward neural network, and fit for financial prediction system with nonlinear Time Series [3]. Yuean Yin constructed RBF-ANN model for predicting stock price trend through considering Index of Shanghai Stock Exchange as research objects. Prediction results were better, and showed that RBF-ANN has excellent learning capacity and generalization ability. Lihua Yue et al. applied RBF-ANN to predict stock price of short-term, and also obtain better prediction results.

VI. CONCLUSION AND FUTURE WORKS

The series of experiments reported in this paper showed that it is possible to use Logistic Regression as an approach effectively to predict future trend of stock price. Feature Index Variables are innovative and crucial to our proposed approach. They are not only easy to both understand for the private investor and obtain from daily stock trading information, but also make our approach perform better than or equal to other methods, e.g. RBF-ANN Prediction Model under the same conditions of Feature Index Variables. The

procedure of selecting optimizing group of regression coefficients improves prediction accuracy to as high as 83.3%. Our proposed approach is low in complexity and easy to understand or realize. Duo to use current monthly trading information, significant time-effectiveness and strong purposefulness enable users predict stock price trend of next month just through considering current monthly financial data instead of needing a long term procedure of analyzing and collecting financial data. However, one defect still exists in our approach considering that some feature index variables fail because of very small value (approximate to zero) of this parameter.

Future works will include the implementation of our proposed method instead of using SPSS to perform prediction. This is very important to further investigations of Logistic Regression Model for stock price trend. Moreover, since there exist the situation that some feature index variables fail sometimes, we will study how to select effective variables to improve prediction accuracy of stock price trend.

REFERENCES

[1] Sunlin Pang. An Application of Logistic Regression Model in Credit Risk Analysis [J]. Mathematics in Practice and Theory, 2006, 36(9), pp. 129-137.

[2] http://www1.resset.cn/product/index.jsp.

[3] Zhu Yun, Wang Xingyu. An Applicaiton of RBF Neural Network to Stock Market Trend Prediction [J]. Journal of East China University of Science and Technology, 2002, 28(5), pp. 547-550.

[4] Le liHua, Wen Rongsheng, Zhu hui. Stock Market Forecast Based on RBF Neural Network and MATLAB Realization [J]. SCI-TECH Information Development & Economy, 2008, 18(30), pp. 151-152.

[5] Yuzheng Zhai, Arthur Hsu, Saman K Halgamuge. Combining News and Technical Indicators in Daily Stock Price Trends Prediction [C]. Lecture Notes In Computer Science, Vol. 4493 archive. Proceedings of the 4th international symposium on Neural Networks: Advances in Neural Networks, Part III, 2007, pp. 1087 – 1096.

[6] West D. Neural network credit scoring models [J]. Computer & Operations Research, 2000, 27, pp.11131-1152.

[7] Sulin Pang, Jie Xiao. Analysis and forecasting to stock pricing using Logistic regression model [C]. Intelligent and Complex Systems, DCDIS Proceedings 2, 2004, pp. 221-226.

[8] Sunlin Pang. An Application of Logistic Model in Stock Forecasting [C]. 8th International Conference on Control, Automation, Robotics and Vision Kunming, China, 2004, p. 1491-1496.

[9] Pang Sunlin, Deng Feiqi, Wang Yanming. A Comparison of Forecasting Models of the Volatility in Shenzhen Stock Market [J]. Acta Mathematica Scientia, 2007, 27B(1),pp. 125-136.

[10] Zheng Mei, Miao Jia. The application of Logit Model to predicting Shanghai stock market [J]. Statistics & Decision, 2007, 3, pp. 102-104.