Predicting China's Economic Running State Using Machine Learning

Jianxiang Luo [1], Yonggang Fu [2,a]

[1] *School of Science, Jimei University, Xiamen, Fujian, 361021, China*
[2] *Computer engineering college , Jimei University , Xiamen, Fujian, 361021, China*

Abstract. China's business index of macro-economic includes early warning index, coincidence index, leading index and lagging index, among which early warning index reflects the economic running state. However, obtaining these indexes is a complex and daunting task. To simplify the task, this article mainly explores how to use machine learning algorithms including multiple linear regression(MLR), support vector machine regression(SVM), random forest(RF), artificial neural network(ANN) and extreme learning machine(ELM) to accurately predict early warning index. Finally, it can be found that the warning index can be well predicted by above machine learning algorithms with coincidence index, leading index and lagging index to be variables, furthermore, extreme learning machine and random forest are superior to other methods.

1 Introduction

Business index of macro-economic composed of coincidence index, leading index, early warning index and lagging index reflects the strength of the national economic growth momentum. The coincidence index reflects the basic trend of the current economy, which is composed of industrial production, employment, social demand (investment, consumption, foreign trade), social income (national taxation, corporate profits and household income), etc. The leading index is used to predict the future trend of the economy. The lagging index is mainly used to confirm the peak and valley of economic cycle. The early warning index classifies the state of economic operation into five levels: "red light", "yellow light", "green light", "light blue light" and "blue light". "Red light" means overheated economy, "yellow light" means partially heated economy, "green light" means normal economic operation, "light blue light" means cold economy, and "blue light" means excessively cold economy.

Economic sentiment index is derived from the business climate survey, which is a statistical investigation system and compiled by conducting regular questionnaire survey on entrepreneurs according to their judgments and expectations to enterprise operation and macroeconomic condition. It reflects the status of production , operation of enterprises, economic operation situation to predict the future trend of the development of the economy. The process of getting indexes is complex and time-consuming [1-3], so it is of great significance to explore new ways to obtain indexes more easily.

Machine learning has been booming in recent years. An amazing feature of machine learning is its strong predictive ability. Machine learning has been widely used in various fields [5-16]. Support vector machine regression is originally proposed by Vapnik in 1995, and were widely used in all walks of life. Other researchers use it to predict the pressure drop during evaporation of R407C [17], toxicity of ionic liquids [18-19], partition coefficients [20], etc. Artificial neural network has been a research hotspot in the field of artificial intelligence since 1980s. Extreme learning machine is proposed by Huang et al ,and applied for classification in 2014 [21]. To some extent, ELM is a type of feedforward neural network. ELM can be used to achieve good generalization performance at extremely fast learning speed in different fields [22]. Simple linear regression, multiple linear regression, logical regression and random forest et al. are also widely used in various fields to making predictions.

In this paper, we explore the relationships among early warning index, coincidence index, leading index and lagging index based on machine learning methods. finally, it can be observed that early warning index which reflects the economic running state can be well predicted with coincidence index, leading index and lagging indexes to be variables. However, support vector machine regression and the random forest approach are superior to other machine learning methods. The main contribution of this study is to provide an easy way to obtain early warning index.

2 Model and Data

For the purpose of testing the predictive capability of different machine learning methods with coincidence index, lagging index and leading index to be inputs and early warning index to be output , we rely on the data

[a] Corresponding author: yonggangfu@jmu.edu.cn

set from CEIC. The data set consists of the monthly business index of macro-economic from 1991 to 2017, with a total of 324 data. The data need to be preprocessed before predicting, the data are normalized by equation(1):

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

$X_{\min}$ and $X_{\max}$ are the maximum and the minimum values corresponding to the indexes, respectively. X is the actual value of the index, and X_{norm} is the normalized value of the index . Furtherrmore, 75% of the data was used for training, and 25% of the data was used for testing.

In this paper, five machine learning methods were explored to predict the economic running state i.e. , multiple linear regression, support vector machine regression, random forest, artificial neural network and extreme learning machine. Mean square error (MSE) and squared correlation coefficient (R^2) are selected to evaluate the model performance. MSE is the mean square sum of the error of the corresponding points between the predicted data and the original data, and the closer it is to 0, the more successful the data prediction is. R^2 is to describe the goodness or badness of model fitting, and the closer it is to 1, the better the model fitting is. MSE and R^2 are given by equation(2) and equation(3):

$$MSE = \sum_{i=1}^{N} (y_i^{act} - y_i^{\exp})^2 / N \tag{2}$$

$$R^2 = \frac{\sum_{i=1}^{N} (y_i^{\exp} - \bar{y})^2 - \sum_{i=1}^{N} (y_i^{act} - y_i^{\exp})^2}{\sum_{i=1}^{N} (y_i^{\exp} - \bar{y})^2} \tag{3}$$

where $y^{\exp}$ is the predicted value (output), and y^{act} is the actual value. $\bar{y}$ is the average value of the actual values, and N is the number of data in the data set.

3 Methods

3.1 Multiple linear regression (MLR)

Multiple linear regression is a commonly used modeling method and a simple regression method. In this model, the main goal is to find the best fitting straight line with the given data. Linear regression is given by equation (2):

$$Y = b_0 + \sum_{j=1}^{n} b_j X_j \tag{2}$$

where b_j are the regression coefficients of the corresponding variable, n is the number of variables (inputs) in the model, the regression coefficients are mainly obtained by minimizing the error between predicted values and real values. If a training set with points (x_1, y_1), (x_2, y_2), … , (x_N, y_N) is given, the error (RSS) is given by equation (3) :

$$RSS(\beta) = \sum_{i=1}^{i=N} (y_i - b_0 - \sum_{j=1}^{j=n} x_{ij} b_j) \tag{3}$$

3.2 Support vector machine regression (SVM)

The main purpose of SVM is to obtain a optimal hyperplane. The support vector machine regression method reduces the constraint in error and no longer consider the residual in the training data set. Therefore, the goal of support vector machine (SVM) is to find the functions $f_{sv}(x) = \sum_{m=1}^{M} w_m \phi_m(x) + b$, where ϕ_m is a kernel function which can map data to high-dimensional space, and w_m the respective weights. SVM can take regression problem as an optimization problem as follows:

The objective function is given by $Min \frac{1}{2} |w|^2$

The constraint condition is given by the following inequalities(4):

$$\begin{cases} y_i - \sum_{m=1}^{M} w_m \phi_m(x) - b \le \varepsilon \\ \sum_{m=1}^{M} w_m \phi_m(x) + b - y_i \le \varepsilon \end{cases} \tag{4}$$

3.3 Random forest (RF)

Random forest is characteristiced by bagging and random feature selection. The RF approach selects a subset of characteristics to be split at each node during the tree formation, and each tree is built independently using the boot sample of the training data. The general algorithm for RF is as follows:

1. The bagging idea is used to randomly generate n sample subsets from the training data set.

2. Taking advantage of the idea of random subspace, randomly selecting f features, conducting node splitting, creating a single regression decision subtree from the sample subset, and repeating the same method

to build n trees, finally, each tree grows freely without cutting branches to form a forest.

3. Predicting the output of the new data set. The predicted values are the average of the predicted results of all decision trees.

The output of FP prediction can be expressed as equation(5):

$$f(x)=\frac{1}{n}\sum_{j=1}^{n} f_j(x) \tag{5}$$

where $f(x)$ is the predicted value, and $f_j(x)$ is the individual prediction of a tree for an input vector.

3.4 Artificial neural network (ANN)

Artificial neural network algorithm is based on the mathematical model inspired by behaviors of biological neurons. It abstracts the neural network of human brain from a certain point of view to establishes some simple model, and forms different networks according to different connection modes. The artificial neural network method consists the input layers, hidden layers and output layers, wherein hidden layer contains a given number of neurons that take input from the input layer and connect their outputs to the output layer. If the artificial neural network has more than one hidden layer, the outermost layer is connected between the innermost layer and the output. Each line connecting two neurons is associated with a given weight. In the hidden layer, the output of a neuron can be obtained from the following equation (6):

$$h_i = s\left(\sum_{i=1}^{N} v_i x_i + T_i^{hid}\right) \tag{6}$$

where $s(\ )$ is the transfer function, N is the number of inputs, v_i is the weight of layer i, x_i is the input value, and T_i^{hid} is the threshold term of hidden neurons.

3.5 Extreme learning machine (ELM)

ELM is a new algorithm for single hidden layer feedforward neural network. Compared with the slow training speed of the traditional feedforward neural network, a great chance to fall into local minimum point and sensitivity of learning rate selection and other shortcomings, the ELM method generates the weights between input layer and hidden layer and neuron threshold of hidden layer without adjustment in the process of training, and only need to set the number of neurons in hidden layer to obtain the optimal solution. ELM is Characterized by the proposition of using random independent nonlinear feature transformation. Inherently, with two important characteristics, interpolation capability and universal approximation capability. ELM is widely applied in various fields for extremely fast learning speed.

4 Results and discussion

The outcomes of multiple linear regression, support vector machine regression, random forest, artificial neural network and extreme learning machine are presented in **Figure1**, **Figure2**, **Figure3**, **Figure4** and **Figure5**, respectively. The data predicted from training and test set are highlighted in blue, which are shown respectively, and the coefficient of determination R^2 (calculated based on either the training or test sets) is indicated.

Figure 1 shows the prediction of the multiple linear regression, which presents that multiple linear regression provides an acceptable model for the prediction with the MSE and R^2 values of the training set and the test set to be 0.0047, 0.9014, 0.0061 and 0.9037, respectively. Besides regression coefficients are -0.01467 , -0.06193 and 1.1157.

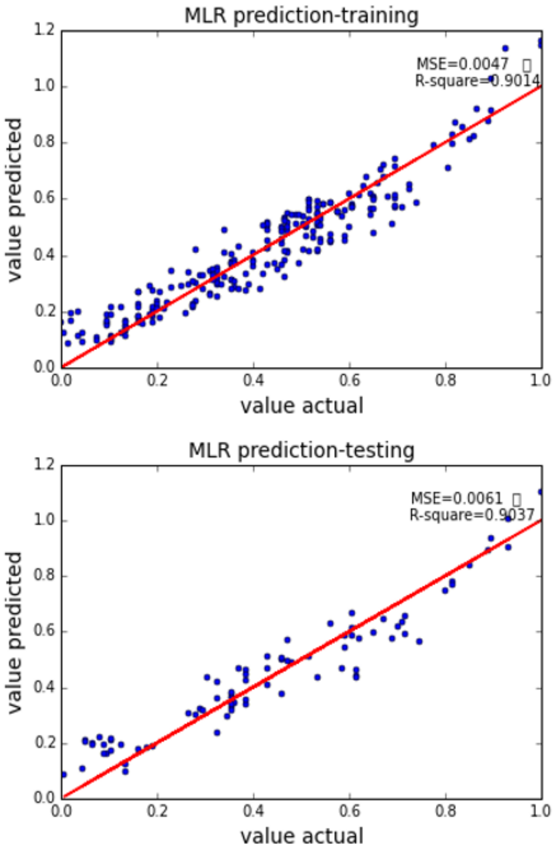

Figure 1. Predicted early warning indexes using MLR, compared to the actual values on the training set and test set

The SVM method can be used to train a model based on both linear and non-linear data. **Figure 2** shows the predicted values using the SVM method, as compared with the measured values. As we can see, the SVM method can also give a good prediction with the MSE and R^2 values of the training set and the test set to be 0.0035, 0.9259, 0.0057 and 0.9098, respectively. Furthermore, RBF function is selected to be kernel function.

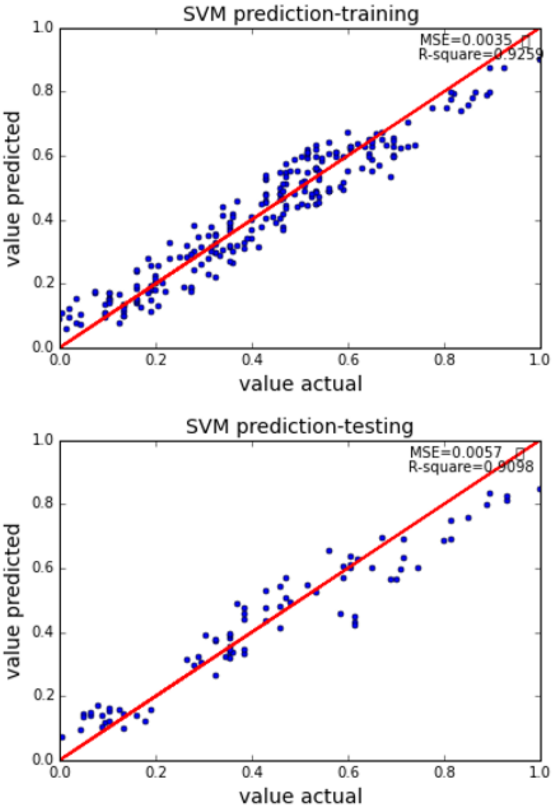

Figure 2. Predicted early warning indexes using SVM, compared to the actual values on the training set and test set

In general, the RF approach can be optimized by increasing the number of trees, but it can also ultimately lead to an overfitting situation. To find the best number of trees, we gradually increase the number of trees from 5 to 100. In this process, the R^2 and MSE values of the test set are computed in **Table 1**. It can be observed that the change of the number of trees has little effect on the prediction. MSE reaches the minimum value and R^2 reaches the maximum value when the number of trees increases to 20, so optimal number of trees is identified to be 20. **Figure3** shows the predicted values using the RF method, as compared with the measured values with the MSE and R^2 values of the training set and the test set to be 0.0070, 0.9855, 0.0044 and 0.9388, respectively. The prediction of the RF method on training set is vary excellent ,but it doesn't do the same in the test set. Maybe the overfitting situation ultimately happened.

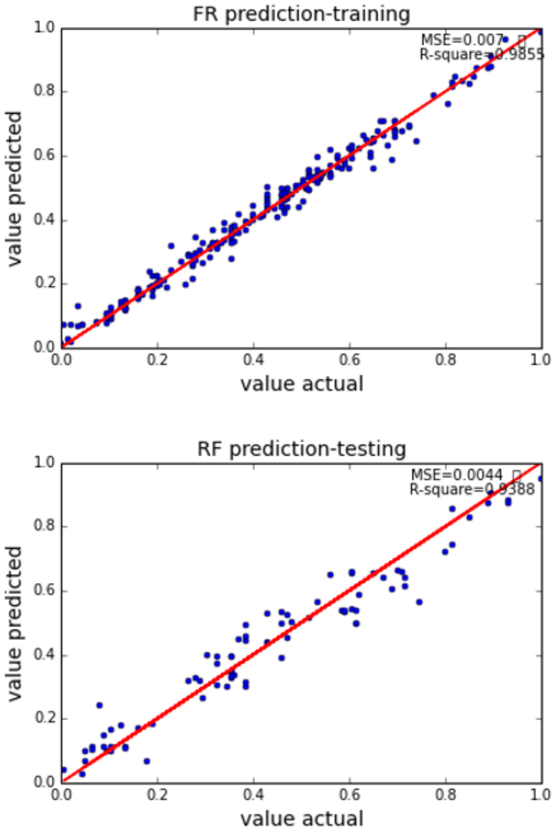

Figure 3. Predicted early warning indexes using RF, compared to the actual values on the training set and test set

Table 1. R^2 and MSE values of the test set with respect to the number of trees used in the random forest algorithm.

Number	5	10	15	20	25
MSE	0.0042	0.0041	0.0039	0.0039	0.0041
R^2	0.9333	0.9349	0.9382	0.9388	0.9354
Number	30	35	40	45	50
MSE	0.0049	0.0040	0.0049	0.0053	0.0050
R^2	0.9223	0.9223	0.9231	0.9159	0.9219
Number	55	60	65	70	75
MSE	0.0046	0.0051	0.0045	0.0046	0.0040
R^2	0.9275	0.9196	0.9291	0.0046	0.9364
Number	80	85	90	95	100
MSE	0.0039	0.0059	0.0045	0.0044	0.0042

R^2	0.9384	0.9078	0.9298	0.9313	0.9336

Artificial neural network method can improve the predictive ability by increasing the number of hidden layers and the number of neurons like the random forest method by increasing the number of trees, and it can also lead to overfitting. Likewise, we increase the number of neurons from 1 to 20 with 1 hidden layer to observe the corresponding changes. We find that the optimal number of neurons is 5, and the sigmiod function is chosen as the activation function. **Figure3** shows the predicted values using the ANN method, as compared with the measured values with the MSE and R^2 values of the training set and the test set to be 0.0041, 0.9118, 0.0061 and 0.9056.

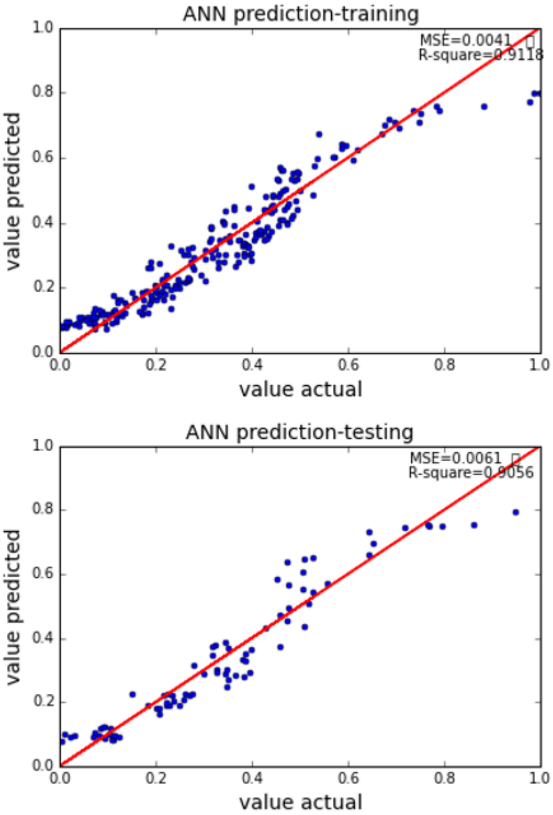

Figure 4. Predicted early warning indexes using NNT, compared to the actual values on the training set and test set

In the same way, changing the number of neuron to predict early warning indexes, the optimal number of neurons was finally determined to be 10, and radbas function was selected to be the activation function. The comparison between the predicted results using the ELM method and the real values is shown in **Figure5** with the MSE and R^2 values of the training set and the test set to be 0.0036, 0.9252, 0.0051 and 0.9211..

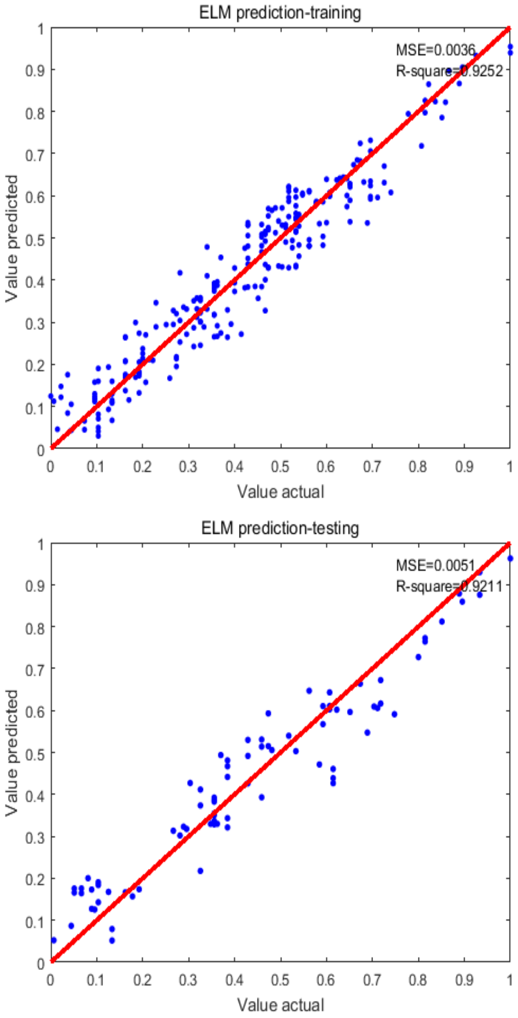

Figure 5. Predicted early warning indexes using ELM, compared to the actual values on the training set and test set

The comparison of outcomes of different machine learning methods is shown in **Table 2**, which suggests that all those machine learning methods can make a good prediction. As we can see RF and ELM both give an excellent prediction to the early warning index, and we also can find that RF is superior to the other machine learning methods.

Table 2. Comparisons of the statistical parameters by different method

Method	MLR	SVM	FR	ANN	ELM
MSE (training)	0.0047	0.0035	0.0070	0.0041	0.0036
R^2 (training)	0.9014	0.9259	0.9855	0.9118	0.9252
MSE (testing)	0.0061	0.0057	0.0044	0.0061	0.0051
R^2 (testing)	0.9037	0.9098	0.9388	0.9056	0.9211

5 Conclusions

In this study, to solve the problem that obtaining business index of macro-economic is a complex and daunting task, business index of macro-economic is explored based on 5 machine learning methods. It can be found that early warning index can be well predicted by the five machine learning methods and accurately predicted by the RF and ELM approach with the other indexes to be inputs, which means machine learning models provide a more convenient way to obtain earning warning index.

References

1. Baiqian Song, Journal of Guangxi Economic Management Cadre College. J. **15**, 22-25(2003)
2. Lu Han, Farm Economic Management. J. 34-36 (2016).
3. Xuewen Li, Compilation and research application of macro economic prosperity index in hunan province [D]. Hunan Agriculture, 2014
4. Garzón M B, Blazek R, Neteler M, et al., Ecological Modelling. J. **197**, 383-393(2006)
5. Ganapathi A, Kuno H, Dayal U, et al., Icde. J. 592-603 (2009)
6. Sharma N, Sharma P, Irwin D, et al., Predicting solar generation from weather forecasts using machine learning[C], 2012:528-533.
7. Cooper G F, Aliferis C F, Ambrosino R, et al., Artificial Intelligence in Medicine. J. **9**, 9(1997)
8. Rubinstein N D, Mayrose I, Pupko T. Molecular Immunology. J. **46**, 840-847 (2009)
9. Bhardwaj N, Langlois R E, Zhao G, et al., Nucleic Acids Research. J. **33**, 6486 (2005)
10. Han L, Cui J, Lin H, et al., Proteomics. J. **6**, 4023-4037 (2010)
11. Khandelwal A, Krasowski M D, Reschly E J, et al., Chemical Research in Toxicology. J. **21**, 1457-67(2008)
12. Long P. Predicting electricity distribution feeder failures using machine learning analysis[C]. AAAI Press, 2006:1705-1711.
13. Choi E, Safety and Health at Work . J. **8**, 371-377 (2017)
14. Dušan Cogoljević, Meysam Alizamir, Ivan Piljan, Tatjana Piljan, Katarina Prljić, Statistical Mechanics and its Applications . J. **495**, 211-214(2018)
15. Krishnan N M A, Mangalathu S, Smedskjaer M M, et al., Journal of Non-Crystalline Solids. J. **487**, 37-45(2018)
16. Xu Du, Jingyu Feng ,Shaoqing Lv, Telecommunications Science. J. **33**,66-75 (2017)
17. A. Khosravi, J.J.G. Pabon, R.N.N. Koury, L. Machado, Applied Thermal Engineering. J. **133**, 361-370(2018)
18. Zhao Y, Zhao J, Huang Y, et al., Journal of Hazardous Materials . J. **278**, 320-329(2014)
19. Cao L, Zhu P, Zhao Y, et al. Journal of Hazardous Materials. J. **352**, 17-26(2018).
20. Golmohammadi H, Dashtbozorgi Z, Molecular Informatics . J. **31**, 867-878(2012)
21. Chorowski J, Wang J, Zurada J M. Neurocomputing . J. **128**, 507-516(2014)
22. Huang G B, Siew C K. Extreme learning machine: RBF network case[C], 2004. Icarcv 2004. IEEE, 2012:1029-1036 Vol. 2.