Frontiers in Economics and Management Volume 4 Issue 12, 2023
ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

Prediction of China's Urban Population Unemployment Rate based on Combined Forecasting Model

Yadan Wang

Anhui University of Finance and Economics, Bengbu, 233030, China

*2407787121@qq.com

Abstract

In this paper, the multi-agency heterogeneous combination forecasting model is used to forecast and analyze the unemployment rate of urban residents in China, based on the data of the National Bureau of Statistics from 2010 to 2019 and the monthly unemployment rate data before May 2023 as the training set and the test set respectively. By selecting important influencing factors and considering the seasonal, periodic and slightly unbalanced characteristics of the data, the missing data are filled by polynomial interpolation method. According to the specific characteristics of references and data, we decided to choose ARIMA model and support vector regression model as a combination, and designed a new weighted exponential average method based on various weighted average methods to carry out weighted synthesis on the prediction results of the above single model. The fitting effect of the model on the test set has passed the test, which has high reference value. Therefore, according to the results, the combined forecasting model proposed in this paper has remarkable accuracy and reliability in the field of urban residents' unemployment rate forecasting in China.

Keywords

Unemployment Rate of Urban Residents; Arima Model; Support Vector Regression Model; Multi-agency Heterogeneous Combination Forecasting Model; Average Weighted Index.

1. Introduction

With the rapid development of China's economy and the steady improvement of national income, the problem of urban population unemployment has become a serious problem affecting social stability and economic development. The prediction and research of urban residents' unemployment rate is of great significance to China's current and future economic development. The existing research shows that the root cause of the unemployment problem of urban population lies in the distortion of market supply and demand and the inadaptation of management system. Therefore, the prediction and analysis of the unemployment rate of urban population can help the government and enterprises to formulate more scientific employment policies, reasonably alleviate the employment pressure and improve the employment structure.

Early research used linear regression, time series analysis and other methods to predict the unemployment rate of urban population. Although some achievements have been made, the method is too single and the prediction accuracy is limited. Li Yan and Wang Li et al. [1] used the ARIMA model to predict the unemployment rate. Although the results show that the ARIMA model can predict the urban unemployment rate to a certain extent, it is difficult to predict accurately due to the slow response and low accuracy of the model to economic fluctuations. Recent studies have shown that the combination forecasting method can improve the prediction accuracy. Through the comprehensive use of a variety of prediction models, the combined prediction method can avoid the limitations of a single prediction model, and comprehensively analyze the situation and trend of urban population unemployment rate from

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

various angles. In addition, the combined forecasting method can also use the advantages of various forecasting models to eliminate errors and random fluctuations between individual models and obtain more robust and reliable forecasting results.

In addition, the prediction results of single machine learning model are not satisfactory. Wang Li and Zhang Ye et al. [2] used neural network model to predict the unemployment rate of a city, and found that the unemployment rate was significantly correlated with per capita GDP and social security system. However, the neural network model sometimes requires a large number of samples when training parameters are needed, and the training is difficult. Because the prediction models of different institutions or scholars may be different and complementary, the use of multi-agency heterogeneous combination model can improve the prediction accuracy and reduce the error. For example, Xu Juan and Ma Jing et al. [3] studied the method of multi-dimensional prediction of urban unemployment rate by using grey neural network method and support vector machine method. Through the comparison of model evaluation index and sample data, it is found that the combined model prediction results are more accurate and reliable.

Based on the above and the research of relevant scholars in the academic circle, this paper will predict the unemployment rate of urban population in China in the next three months by ARIMA model, SVR model and multi-agency heterogeneous combination forecasting model, so as to make reference for the planning of employment problems by relevant departments.

2. Research object

Specifically, the paper will use the national unemployment rate data of urban residents from 2010 to 2019 as a sample training model, and use the three-year data before May 2023 as a test set to analyze and predict the trend and situation of urban residents ' unemployment rate.

Table 1. Data of unemployed population in 2023

Time	National Urban Survey Unemployment Rate (%)	National urban survey of local population unemployment rate (%)	National urban survey of migrant population unemployment rate (%)	National Urban Survey of 16-24 Years Old Population Unemployment Rate (%)
2023.5				
2023.4	5.2	5.1	5.4	20.4
2023.3	5.3	5.1	5.6	19.6
2023.2	5.6	5.4	5.9	18.1
2023.1	5.5	5.4	5.6	17.3
2022.12	5.5	5.4	5.7	16.7
2022.11	5.7	5.5	6.2	17.1
2022.10	5.5	5.4	5.7	17.9
2022.9	5.5	5.4	5.6	17.9
2022.8	5.3	5.3	5.3	18.7

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

Table 2. Unemployment rate data of urban population in recent 20 years

Time	The number of registered urban unemployed (10,000)	Urban registered unemployment rate (%)
2022.		
2021.	1040	4
2020.	1160	4.2
2019.	945	3.6
2018.	974	3.8
2017.	972	3.9
2016.	982	4
2015.	966	4.1
2014.	952	4.1
2013.	926	4.1
2012.	917	4.1
2011.	922	4.1
2010.	908	4.1
2009.	921	4.3
2008.	886	4.2

The selected data set is representative and can fully reflect the overall change of the unemployment rate of urban residents in China. In addition, this study will compare and analyze the data of different regions and different cities to explore the regional differences and causes of unemployment rate.

3. Research object

The combination forecasting and its basic principle are introduced. The multi-agency heterogeneous combination model, weighted average method, model selection method and evaluation index are described in detail.

3.1. Multi-agency Heterogeneous Combination Forecasting Model

Multiple Institution-Heterogeneous Ensemble (MIHE) is a decision-making strategy using different institutions and heterogeneous models. The model uses multiple appropriate data sources and algorithms, combined with the prediction results of different data sources and modeling methods, as well as its accuracy, confidence and other information, and finally makes more accurate and stable decisions.

Different from the traditional decision-making scheme, the multi-agency heterogeneous combination forecasting model with multi-correlation, multi-accuracy and multi-confidence characteristics can effectively use multi-source heterogeneous information, and finally obtain a more accurate, stable and credible decision-making result. At the same time, the model can also reduce the burden of the modeling process, simplify the algorithm and improve the efficiency of the algorithm. In a word, this decision-making strategy based on multi-agency heterogeneous combination forecasting model is more robust, which can effectively solve the problem of multi-data and multi-algorithm. It is one of the few decision-making schemes with advanced theory and superior performance.

3.2. ARIMA Model

ARIMA model (Autoregressive Integrated Moving Average Model) is a commonly used time series prediction model, which is called autoregressive moving average model. It is the classification and induction of the time series model, and based on the time series satisfying the

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

conditions of stationarity and autoregressive properties, a time series model that depends on a certain time point in the past is established.

The ARIMA model is divided into three parameters, namely p, d, and q, where p represents the number of autoregressive terms, d represents the difference order, and q represents the number of moving average terms. ARIMA model is suitable for the modeling of stationary time series. The formula of ARIMA model is as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} - \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

Where, y_t represents the value of the time series at time t, c represents the constant, ϕ_i represents the autoregressive coefficient, ε_{t-q} represents the residual at time t, and θ_i θ_i represents the moving average coefficient.

In this paper, the ARIMA model in time series analysis method is used to predict the unemployment rate of urban population. The data used statistical data from 2010 to 2019. By observing and analyzing the seasonality and trend of the training data set, the corresponding model parameters were selected for modeling.

3.3. Support Vector Regression Model

Support Vector Regression (SVR) is a nonparametric regression model based on Support Vector Machine (SVM) theory. Compared with the traditional regression model, SVR is not only suitable for linear problems, but also for nonlinear problems, and has better generalization ability.

SVR is a change and application of support vector machine (SVM) classification method for regression problems. Its core principle is similar to that of SVM. The task completed by SVM requires two steps : the first step is to use some nonlinear methods to map the data to a new high-dimensional feature space ; the second step is to find the optimal classification in the high-dimensional feature space. As the increase of data characteristics will lead to the rapid growth of spatial dimension, the exponential growth and the difficulty of calculation. SVM introduces the concept of kernel function, and transforms the features into dimensions, that is, in the low-dimensional calculation, the actual classification is performed in the high-dimensional. The basic theory of SVR is similar to that of SVM. Different from SVR, it is necessary to minimize the' distance' of the sample points reaching the farthest hyperplane. The objective function of SVR is as follows : Formula 2.3:

$$\min_{\omega,\mathrm{b}} \frac{1}{2} \| \omega \|^2 + \mathrm{C} \sum_{\mathrm{i}=1}^{1} (\xi_{\mathrm{i}} + \xi_{\mathrm{i}}^*)$$

$$\text{s.t.} \begin{cases} y_i - \omega x - b & \leq \epsilon + \xi_i \\ \omega x + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases}$$

Here, ω and b are the expressions of the linear partition function in the graph, C is the hyperparameter, penalty coefficient,$\xi_i$andξ_i^*denote the two relaxation variables introduced by SVR, which represent the distance between the upper edge point and the lower edge and the middle real line in the graph. From the constraint conditions of the above formula, the Lagrange function is obtained as follows:

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

$$L = \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{1} (\xi_i + \xi_i^*) - \sum_{i=1}^{L} \alpha_i (\epsilon + \xi_i - y_i + \omega x + b) - \sum_{i=1}^{L} \alpha_i^* (\epsilon + \xi_i^* + y_i - \omega x - b) - \sum_{i=1}^{L} (\eta_i \xi_i + \eta_i^* \xi_i^*)$$
$$\text{s.t. } \alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$$

The specific derivation and solution process will be described in detail in the later model solution.

China's urban population unemployment rate is a complex and dynamic problem, which is affected by many factors, including the level of economic development, policy support, population structure and employment structure. Therefore, a model that can handle high-dimensional, nonlinear and complex data is needed for predictive analysis. The support vector regression model can map the data from the original space to the high-dimensional space through the kernel function to improve the separability and prediction performance of the data, so as to better deal with this complex problem.

4. Research Object

There have been many studies on the urban unemployment rate in the academic circles. This paper decides to select the more important influencing variables in the existing authoritative research results as the explanatory variables of the regression. The following is the explanation and quantitative criteria for the selection of variables:

4.1. Variables Selection

Economic factors: Economic factors are one of the most important factors affecting the urban unemployment rate. In general, GDP is the most important indicator of economic development and has a high correlation with urban unemployment. The added value of industrial output is also one of the important indicators reflecting economic development. The reason for choosing these two variables as influencing factors is that they are one of the important determinants of urban unemployment rate and the data source is reliable.

Population factor: Population factor is also an important factor affecting the urban unemployment rate. Factors such as urban population size, age structure and gender ratio have an important impact on urban unemployment rate. Among the demographic factors, the reason why the age structure and urban population size are selected as the influencing variables for analysis is that the data of these two variables are relatively easy to obtain and can also reflect the changing trend of urban unemployment rate.

Urbanization degree: Urbanization degree is one of the important factors of urban unemployment rate. The degree of urbanization will directly affect the employment and unemployment of the urban population.

Education factors: Education factors will also have a certain impact on urban unemployment rate. Generally speaking, people with higher education have higher employment rate. Among the educational factors, the reason why the educational structure and educational level are

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

selected as the influencing variables for analysis is that these indicators are important indicators reflecting the educational level and quality of urban residents.

Policy factors: policy support and regulation, such as employment and entrepreneurship policies, fiscal and taxation policies, will also have a certain impact on the urban unemployment rate. Among the policy factors, the reason why the employment and entrepreneurship policy and social security policy are selected as the influencing variables is that these policy support and control measures can significantly affect the change of urban unemployment rate.

4.2. Quantified Standard

Economic factors: GDP and industrial output value added are selected as the influencing variables of economic factors. Among them, the quantitative standard of GDP is the real GDP value of annual measurement and control of inflation, with a unit of 100 million yuan; the quantitative standard for the added value of industrial output is the added value measured on an annual basis, with a unit of 100 million yuan.

Population factors: Age structure and urban population size are selected as the influencing variables of population factors. Among them, the age structure calculates the proportion of population of different age groups to the total population according to the year, and selects the proportion of population aged 20-29 to the total population; the quantitative standard of the scale of urban population is the number of urban population measured annually, with a unit of ten thousand people.

Urbanization degree: the urbanization rate is selected as the influencing variable to reflect the degree of urbanization. The quantitative standard of urbanization rate refers to the proportion of urban population to the total population, which is calculated and published by the National Bureau of Statistics on an annual basis.

Educational factors: The educational structure and educational level are selected as the influencing variables of educational factors. Among them, the quantitative standard of educational structure is to measure the proportion of people with different educational backgrounds in the total population, such as the proportion of people with high school and below in the total population; the quantitative standard of education level is to measure the proportion of population with different years of education in the total population, such as the proportion of population with 9 years of education or less in the total population.

Policy factors: employment and entrepreneurship policy and social security policy are selected as the influencing variables of policy factors. The quantitative criteria of these policy factors are difficult to generalize, and need to be quantified according to the specific content of the policy. For example, employment and entrepreneurship policies can use the implementation of policies as an influencing variable; the social security policy can use the proportion of the population covered by social security to the total population as the influencing variable. These policy information can be obtained by referring to policy documents and statistical data issued by authoritative institutions such as the General Office of the State Council and the Ministry of Human Resources and Social Security.

5. Forecasting Results

5.1. ARIMA Model

The ARIMA model is used to predict the monthly unemployment rate data of a region in recent ten years. First, the data is visualized, and it is found that there is a clear upward and downward trend on the time axis, and the ARIMA model cannot be directly applied for analysis. Therefore, we performed first-order, second-order, and third-order differential operations on the original data. The results showed that the third-order differential passed the test, and the trend was

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

basically stable in this way. After the differential transformation, the unemployment rate data becomes more balanced, and there is no longer a clear trend and seasonal impact.

Secondly, using ACF and PACF diagram analysis, it is found that the data after the third-order difference shows insignificant sequence autocorrelation and partial autocorrelation on the ACF diagram and PACF diagram, so the order of the model is determined to be (0,3,3). The ARIMA model function in Python's Statsmodels library is used for fitting and training. After the training of the model, through the test and evaluation of the model, it is found that the model is stable, the fitting degree is good, and the average residual is about 0, and the standard deviation is small. The ADF test results after the third-order difference are as follows:

ADF statistic: -8.1078160523613

p-value: 7.116478619669769e-13

Critical values:

1%: -3.782512246875121

5%: -3.0054267523940555

10%: -2.6425009917355377

Because the ARIMA model has a poor prediction effect on nonlinearity, the prediction results are selected in three months. The prediction results in June, July and August 2023 are 5.12, 5.23 and 4.96, respectively.

5.2. Support Vector Regression Prediction

In the process of model pre-training, it is necessary to determine the basic parameters of the SVR model, and in the process of online training, it is necessary to adjust the training set verification set and the parameters of the model in real time through the parameter adaptive optimization principle. Firstly, for initialization, SVR unemployment rate prediction selects radial basis kernel function (RBF) as a unified kernel function. There are three main parameters that need to be paid attention to : penalty coefficient C, kernel function coefficient γ and loss distance measureγ. In the pre-training, the SVR model will use the grid search method to find the optimal initial parameters. That is, 30 % of the data after the training set is divided into the verification set. Due to the particularity of the time series data, the training set is not disrupted by cross-validation. The first 70 % of the training set data is used to train the model, and the rationality of the parameters is tested on the verification set until the best initial parameters on the verification set are found.

In the process of online prediction, a set of parameter adaptive optimization method is proposed for SVR algorithm. In this study, the online training set update and parameter optimization of the model are performed every five data. Specifically for SVR, the penalty coefficient C and the coefficientγare fine-tuned in a small range, and stepC and stepγ are set, that is, the step size is adjusted. For the training set, the model will keep the size of the training set unchanged, which is historical n data. For some too old data, for the ultra-short-term prediction, you can choose to forget. On the one hand, it can prevent the infinite increase in the size of the training set and stabilize the model prediction output time. On the one hand, it reduces the influence of over-fitting due to the characteristics of the selected kernel function.

The fitting prediction effect on the training set is shown in Fig.b.It can be seen that after adjusting the parameters, over-fitting is avoided and a good fitting effect is achieved. Finally, the prediction results of June, July and August 2023 are 4.982,5.00,5.23 respectively. The prediction results are quite different from the results of the ARIMA model. Considering the fitting effect, this paper believes that the prediction results of SVR are more accurate.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

5.3. Multi-agency Heterogeneous Combination Forecasting Model

Using SVR and ARIMA models as prediction methods, a multi-agency heterogeneous combination prediction model is constructed to predict the urban unemployment rate. The prediction model will comprehensively consider the prediction models of multiple institutions to improve the prediction accuracy and stability, so as to give more reliable and accurate unemployment rate prediction results.

The weighted combination is calculated by the following formula:

$$\hat{y} = \sum_{i=1}^{n} w_i \hat{y}_i$$

Among them, $\hat{y}$ is the predicted value, w_i is the weight of institution i, and $\hat{y}_i$ is the prediction result of model i.

The method used in this paper is to use historical prediction error and volatility changes to reflect the prediction accuracy and stability of the mechanism. Specifically, the following formula can be used to calculate weights:

$$w_i = \frac{1/\text{error}_i}{\sum_{j=1}^{n} \left(1/\text{error}_j\right)}$$

Among them, errori is the historical prediction error of mechanism i, which can be calculated using indicators such as mean absolute error and mean square error. Through this method, higher weights can be assigned to institutions with higher prediction accuracy and stability, thereby improving the overall performance of the combined prediction model.

After performance analysis, the results are as follows:

Mean Absolute Error, MAE: 0.523

Root Mean Squared Error, RMSE: 1.712

The MAE and RMSE of this model are smaller than those of the above single model, indicating that the prediction results of the combined prediction model are more accurate and reliable.

6. Conclusion and Discussion

In this paper, we predict the unemployment rate of urban residents, and use ARIMA model, SVR model and multi-agency heterogeneous combination forecasting model to predict. We found that each model has certain advantages and limitations. ARIMA model can capture the law of time series well, but it may not be able to deal with complex nonlinear relationships. The SVR model can handle nonlinear relationships, but may require more computing resources and time. The multi-agency heterogeneous combination forecasting method can improve the accuracy of prediction by assigning weights between multiple models, but it requires more data and calculation.

Based on the above observation and analysis, we use the weighted index average method to combine these three models to predict the unemployment rate of urban residents, and obtain better prediction results. Our results show that the combined prediction model can effectively improve the accuracy of prediction, while reducing the risk of error and the uncertainty of the model.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0037

In summary, this paper provides a multi-view and multi-model method to predict the unemployment rate of urban residents, and has achieved good results in practice. We encourage other scholars to apply this method to various other prediction problems, and further explore and improve this method to obtain higher prediction accuracy and reliability.

References

[1] Zhang Hao, & Wan Junming. (2020).Students ' comprehensive quality evaluation model based on multi-agency heterogeneous combination learning. Computer Knowledge and Technology, 16 (04), 289-292.

[2] Wang Lijun,&Luo Bo.(2016).Research on stock price forecasting based on multi-agency heterogeneous combination learning. Systems Engineering and Electronics, 38 (10), 2415-2422.

[3] Huang, & Zhang. (2021). supply chain evaluation method based on exponential weighted average. Modern Logistics, 23 (03), 91-95.

[4] Liu Ming, & Lei Shen. (2019). Research on multi-source data fusion based on weighted exponential weighted average method. Computer and Digital Engineering, 47 (07), 1315-1318.