Research on Forecast Model of Elderly Pension Expenses Based on Data Mining Algorithm

Bin Zhu1,*, Li Zhang1,Yonggang Zhang1
1Gongqing College of Nanchang University, Jiujiang, jiangxi 332020,China

*Corresponding author email:zhubinax@ncupyh.edu.cn

Abstract. With the intensification of the aging of our country's population, research on how to better protect the living standards of the elderly, reduce the burden of the working population to support the elderly, and ensure the orderly and healthy development of the society, is currently the focus of attention. In order to prevent aging from bringing further negative impacts on our society or economy, this paper uses data mining algorithms as the research basis and uses time series models to make effective predictions. Therefore, this article first decomposes these data by analyzing the characteristics of elderly consumption data. The decomposed sequence has short correlation characteristics, and the prediction accuracy is significantly higher than that of the traditional time series model. However, this kind of time series model can only predict some offline data. In order to make it can be used for dynamic data prediction, this paper uses RBFNN online prediction algorithm. The RBF neural network is partially improved, and the parameters are calculated by improving the later SGD algorithm. The experimental conclusion shows that the prediction accuracy and efficiency of this algorithm are obviously more accurate than the neighbor clustering online training algorithm, and it achieves more effective online prediction. It can be concluded that the role of the elderly social security budget in economic management can fully reflect the income and expenditure of the elderly social security funds, improve the efficiency of fund use, and establish a unified national social security budget, thereby reducing the impact of aging on society Negative impact.

Key words: Data Mining, Pension Expenses, Online Prediction Algorithm, Rbf Neural Network

1. Introduction
The influence of an elderly society on a society is manifold. From an economic point of view, as the proportion of the elderly population rises and the proportion of the labor force declines, the demographic dividend that my country once had will disappear, which will cause my country to have to start transformation and upgrading from the original extensive economic growth model[1-2]. From the perspective of fiscal expenditure, in order to better provide the elderly with a high-quality elderly life, the government's fiscal expenditure structure will change. In order to vigorously support the elderly care industry, the government will increase fiscal expenditure in this area and adjust the
country’s development. Strategy \[3\]. At the end of the 1990s, the mechanism of combining social poolings and personal account pension funds with each other needed part of the younger population structure and fund management system to maintain. At present, our country has entered an old-age society, and social pension insurance expenditures have increased while incomes have decreased. Over the long term, accumulation has caused my country’s social security accounts to make ends meet, resulting in empty accounts \[4\]. If things go on like this, social old-age insurance cannot support the task of providing pensions for the elderly in our country.

Due to the rapid development of many new technologies, such as computers and the Internet of Everything, and various data interleaving, the era of big data has also arrived. Big data presents the characteristics of scale, diversity, and high speed. The data stream is often high-speed real-time data stream, so it needs fast and continuous real-time processing to reflect the true value of big data \[5-6\]. Under the influence of big data, the cost of living has also entered the era of big data, producing a steady stream of data that needs to be processed \[7\]. Big data technology also brings spring to consumers. Through the establishment of various consumer big data centers, various data can be mined to generate more profit for the efficiency of capital use. Consumption forecasting can estimate future consumption and make decisions in advance. There are many models and algorithms used for prediction, but each model has its own advantages and disadvantages. For different data sets and prediction application scenarios, the models and algorithms used are different \[8\].

Based on big data technology, this paper studies the flow prediction algorithm for the big data of the elderly, introduces the development and research status of the elderly consumption data, the flow computing platform and the elderly consumption flow prediction algorithm at home and abroad, and carries out the traditional prediction algorithm. After analysis and comparison, the offline prediction algorithm based on the LMD model and the online prediction algorithm of the RBF neural network are proposed. Finally, the parallelization scheme of the online prediction algorithm is implemented, which verifies that the algorithm realizes real-time traffic prediction in logistics big data applications. The possibility of the big data real-time prediction algorithm is finally summarized and prospected \[9-10\].

2. Algorithm Optimization

2.1. Time series forecasting model

Time series forecasting refers to the estimation of future phenomena according to the historical change law of time series to predict the future. In order to better use time series forecasting, a mathematical model is usually established to fit historical time series to make forecasts. After so long of development and growth, the series model can be divided into: stationary time series model, non-stationary time series model and multivariate time series model. Here are several commonly used time series models:

(1) ARMA and ARIMA models

The structure of the ARMA model is as follows:

\[ Y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} + \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_q x_{t-q} + \mu_t \]

\[ \beta_p \neq 0, \alpha_q \neq 0 \]

\[ E(\mu_t) = 0, Var(e_t) = 0, E(e_t e_s) = 0, s \neq t \]

\[ E(Y_t | e_t) = 0, \forall s < t \]  \hspace{1cm} (2.1)

Introducing the backshift operator B, that \( Bx^t = x^{t-1} \) is \( B^k x(n) = x(n-k) \), the ARMA model can be simply written as

\[ \Phi(B)x_t = \Theta(B)e_t \]  \hspace{1cm} (2.2)

Where:

\[ \Phi(B) = 1 - \beta_1 B - \cdots - \beta_p B^p \]  \hspace{1cm} is a polynomial of order p autoregressive coefficients.
\[ \Theta(B) = 1 - \alpha_1 B - \cdots - \alpha_q B^q \]
is the q-order moving average coefficient polynomial.

The ARIMA model structure is as follows:

\[
\begin{cases}
\Phi(B) \nabla^d x_t = \theta(B) \epsilon_t \\
E(\epsilon_t) = 0, Var(\epsilon_t) = \sigma_\epsilon^2 \\
E(\epsilon_t \epsilon_s) = 0, s \neq t \\
E x_t \epsilon_t = 0, \forall s < t
\end{cases}
\]  

(2.3)

Where:
\[ \nabla^d = (1 - B)^d ; \]
\[ \Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \]
is the autoregressive coefficient polynomial of the stable reversible ARMA model;
\[ \Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \]
is the moving smoothing coefficient polynomial of stable reversible ARMA model.

(2) ARCH and GARCH models

The ARMA model or ARIMA is often used in time series models, but the ARMA model is based on linear modeling and assumes that the sequence is homoscedastic. When the residual sequence has heteroscedasticity, the fitting effect is not ideal, so the prediction accuracy for nonlinear time series is not very high. The method of homogeneity of variance is relatively weak in operability and can only have partial value in theory. In order to facilitate the estimation of the heteroscedasticity function, we use the autoregressive conditional heteroscedasticity (ARCH) model.

Autoregressive conditional heteroscedasticity (ARCH) model:
\[
\begin{cases}
x_t = f(t, x_{t-1}, x_{t-2}, \ldots) + \epsilon_t \\
\epsilon_t = \sqrt{h_t} \\
\eta_t = \omega + \sum_{j=1}^q \lambda_j \epsilon_{t-j}^2
\end{cases}
\]  

(2.4)

where, \( f(t, x_{t-1}, x_{t-2}, \ldots) \) is \( \{x_t\} \) regression function; this model is abbreviated as ARCH(p,q).

But the ARCH model can only be used in the process of short-term autocorrelation of heteroscedasticity functions. In order to solve the long-term autocorrelation of the heteroscedasticity function of the residual series, the generalized autoregressive conditional heteroscedasticity (GARCH) model is proposed, which is a generalization of the ARCH model, which can better fit time series with heteroscedasticity.

The GARCH model structure is as follows:
\[
\begin{cases}
x_t = f(t, x_{t-1}, x_{t-2}, \ldots) + \epsilon_t \\
\epsilon_t = \sqrt{h_t} \\
\eta_t = \omega + \sum_{j=1}^p \eta_j \epsilon_{t-j}^2 + \sum_{j=1}^q \lambda_j \sqrt{h_{t-j}} \epsilon_{t-j}^2
\end{cases}
\]  

(2.5)

Where \( f(t, x_{t-1}, x_{t-2}, \ldots) \) is \( \{x_t\} \) the regression function of \( \{\} \); this model is abbreviated as GARCH(p,q).

2.2. RBF neural network

RBF neural network can be used for function approximation and pattern classification. Compared with other artificial neural networks, RBF has the characteristics of simple structure, fast learning speed and strong approximation ability. Generally speaking, the RBF neural network structure is similar to the
forward network structure, and its topology is shown in Figure 2.1.

![Figure 2.1. RBF topology diagram](image)

According to the topology of the RBF neural network, we can divide it into two stages: the first stage is the mapping from the input layer to the hidden layer; the second stage is the mapping from the hidden layer to the output layer. Under normal circumstances, the Gaussian function is used as the excitation function, and the parameters of the Gaussian function are determined by the data center and width. Each hidden layer unit is a pattern recognition of training samples, and a network model with better performance can be realized through effective learning algorithms and error correction. Therefore, the hidden layer can be regarded as a new space, and the input layer can be mapped to this space. The excitation function is the pattern recognition of the training samples. The data center represents the clustering center of the training samples. The closer the distance, the greater the similarity. Moreover, the nonlinear transformation unit is only sensitive to data with relatively large similarity. The larger the distance, the smaller the output value.

3. Establishment of early warning system

3.1. Prediction algorithm model

(1) LMD can transform long-term related pension expenses into short-term related pension expenses series, so that short-term correlation models can be established to predict the short-term related expenses series. Based on this, a prediction algorithm for pension expenses based on LMD and GARCH is proposed.

Preprocess the old pension expenses $S(t)$ to get the time series of pension expenses $x(t)$. First determine the time interval $m$ (m can be determined according to the predicted scale), and divide the original flow data duration $t$ evenly according to the time interval $m$. Divide into $\mu = \lfloor t / m \rfloor$ time intervals: $[t_0, t_1], [t_1, t_2], ..., [t_{\mu-1}, t_{\mu}]$, and then sum the flow in each time interval to get the flow time series $x(t)$:

$$x_1(t), x_2(t), ..., x_{\mu}(t).$$

(2) Decompose $x(t)$ with LMD, and get several PF components $PF_1, PF_2...PF_n$ and the remainder $u(t)$. According to the LMD definition and decomposition process, the first PF component of the original signal can be obtained by multiplying the envelope signal $a_1(t)$ and the pure FM signal $s_{1,\mu}(t)$, that is,

$$PF_1(t) = a_1(t)s_{1,\mu}(t).$$

separates the first PF component from the original signal $x(t)$, and you will get a For the new signal $\mu(t)$, use it as the original data and repeat the loop $k$ times until $u_k(t)$ is a monotonic function. As follows:
According to formula 3.1, \( x(t) \) can be decomposed into the sum of several PF components PF1, PF2...PFN and the remainder u(t).

(3) Use GARCH model for modeling.

(4) Use the AIC criterion to determine the order of GARCH(r,s). The AIC criterion can be used to determine the order of the GARCH model. After the order is determined, the maximum likelihood function estimation method and iterative technique are used to calculate the estimated values of all unknown parameters.

(5) Use GARCH model to predict. After the parameters are confirmed, the recursive algorithm is used to predict the PF components PF1, PF2...PFN and the margin u(t) respectively to obtain PF1, PF2,..., PFN and u(t), and then add them to obtain the final predicted value \( x(t) \).

(6) Sum the predicted components to get the final predicted cost \( x(t) = PF1 + PF2 + ... + PFN + u(t) \).

4. Comparative analysis of results

System application
For \( x(t) \), the first 1000 data are used as training data for parameter estimation and model establishment; the last 100 data are used as prediction data to detect the difference between the predicted value and the actual value. According to the prediction algorithm based on LMD and GARCH proposed in the previous section, the original data has been tested for self-similarity and LMD decomposition and other steps, as long as the PF components PF1, PF2...PFN and the remainder u after LMD decomposition are required. (t) Use GARCH model for modeling and prediction respectively. Before using the GARCH model for modeling, first check whether the residual sequence of each PF component and the residual u(t) has heteroscedasticity. The essence of heteroscedasticity test is to perform heteroscedasticity correlation test. This paper mainly uses LM heteroscedasticity test method to test the heteroscedasticity of each PF component and the margin u(t). Table 4.1 shows the results \( PF1(t) \) of the heteroscedasticity test.

| Table 4.1. Test of heteroscedasticity |
|--------------------------------------|
| Q   | 1    | 2    | 3    | 4    | 5    |
|-----|------|------|------|------|------|
| Pr>Q| <.0001| <.0001| <.0001| <.0001| <.0001|
| LM  | 53.23| 55.75| 57.34| 58.03| 57.97|
| P>LM| <.0001| <.0001| <.0001| <.0001| <.0001|

According to Table 4.1, until the lag of 5 periods, whether it is Q test or LM multiplier test, it shows that the regression residual item of 1 has conditional heteroscedasticity. Using the same method to test other components and the margin u(t), the residual terms of PF1, PF2...PFN and the margin u(t) all have heteroscedasticity, indicating that the GARCH model can be used for modeling and prediction.
Figure 4.1. 2012-2020 my country's basic pension social average wage replacement rate

As can be seen from Figure 4.1 above, the overall replacement rate of my country's pensions has dropped from about 71% to 41% at present in the past few years. We analyze the reasons for the decline in the replacement rate of pensions: for the older elderly, their adjusted pension amount cannot keep up with the increase in social wages, which will inevitably lead to a decline in the replacement rate; for the younger elderly, their pensions It is the sum of the basic pension and the transitional pension, and the transitional pension is calculated based on the working years before July 1992. Over time, the newly retired slightly younger elderly will inevitably be slightly more than those who retire early. Younger old people receive fewer transitional pensions, which also leads to a drop in the replacement rate.

We use an algorithm model to establish a pension income and expenditure model, and calculate the total income and total expenditure of the pension from 2020 to 2030. Figure 4.2 for specific data. The forecast results show that there will be a gap in the social pooled pension for employees of urban enterprises in my country in 2026, and the gap will continue to grow over time. In 2030, the gap will reach 156.323 billion yuan.

Figure 4.2. The pension gap between 2020 and 2030

Pension reform also needs to consider the implementation impact of other policies, such as the population policy. This article assumes that the urban total fertility rate is maintained at 1.292. With the deepening of my country's population policy reform, the full liberalization of "second-child" fertility is just around the corner. People's concept of fertility will change. At that time, my country's paying population will increase and population policies will change. It will be more and more effective in alleviating the pension gap, and we need to do more in-depth research. In addition, we propose to transfer state-owned shares into the social security fund, how to manage these shares in the future, and what kind of incentive mechanism should be established to maintain the continuous profitability of
state-owned shares to fill the pension fund. These are also issues that we need to study in depth. This article will not do an in-depth investigation.

5. Result

In the context of the widespread application of big data technology, the forecasting of pension expenses is more time-effective and comprehensive, and it also brings a lot of convenience to the pension work. And due to the continuous expansion and development of the current pension insurance fund scale, its working mode is also facing changes. Big data technology is also capable of covering massive amounts of data while strengthening the exploration and research of correlation relationships, which can be highly applied and fit in the endowment insurance fund work, improve its work efficiency and quality, make the endowment insurance fund forecast work more standardized, and better promote the full coverage of auditing. It will also become an inevitable trend to apply big data technology to the prediction of pension insurance funds. Due to the limited scope of my participation in practice, the collection of data on pension insurance funds is not complete, which leads to the lack of depth of research in this article. At the same time, coupled with my inadequate understanding of big data information technology, this article has less research on information technology related to big data technology. In addition, this article still has many deficiencies in the research and analysis of big data of pension insurance funds. It is necessary to further expand the thinking in the follow-up research, introduce more methods, and combine the in-depth study of relevant knowledge to make up for the research flaws of this article.

Acknowledgments
Research Projects of Humanities and Social Sciences in Colleges and Universities of Jiangxi Province, Exploring the Home-based Care Service Model of Medical Care Integration Based on Pharmaceutical Care, Item number: SH17202.

References
[1] Zhang L. Design of a sports culture data fusion system based on a data mining algorithm[J]. Personal and Ubiquitous Computing, 2020, 24(1):75-86.
[2] Li J. Optimal design of transportation distance in logistics supply chain model based on data mining algorithm[J]. Cluster Computing, 2019, 22(1):1-10.
[3] Long Y. Research on art innovation teaching platform based on data mining algorithm[J]. Cluster Computing, 2018(2):1-7.
[4] Wendong X, Yuanfeng L, Deli C. Design of public bicycle scheduling model based on data mining algorithm[J]. 2018, 10(1):68-76.
[5] Shezi B, Jafra N, Sartorius B, et al. Developing a predictive model for fine particulate matter concentrations in low socio-economic households in Durban, South Africa[J]. Indoor Air, 2018, 28(2):228-237.
[6] Fábio Muradas Girardi a, B L M D S , C C D F. A predictive model to distinguish malignant and benign thyroid nodules based on age, gender and ultrasonographic features[J]. Brazilian Journal of Otorhinolaryngology, 2019, 85(1):24-31.
[7] Wiefeke N, Nicholson N. The retirement security of employee owners in ESOPs in S Corporations[J]. Journal of Participation & Employee Ownership, 2019, 2(3):212-221.
[8] Zhang Y, Zhang C, Zhao Y, et al. Wind speed prediction with RBF neural network based on PCA and ICA[J]. Elektrotechnicky Casopis/Journal of Electrical Engineering, 2018, 69(2):148-155.
[9] Zhang J W, Ye S B, Liu H, et al. Filtering out Antenna Effects From GPR Data by an RBF Neural Network[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(9):1378-1382.
[10] Song E, Wang Y, Ding S, et al. An application of RBF neural network theory in diesel engine control[J]. Harbin Gongcheng Daxue Xuebao/Journal of Harbin Engineering University, 2018, 39(5):908-914.