Indian Journal of Science and Technology, *Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/78722, January 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Tourism Arrival Forecasting using Genetic Algorithm based Neural Network

Edi Noersasongko, Fenty Tristanti Julfia, Abdul Syukur, Purwanto*, Ricardus Anggi Pramunendar and Catur Supriyanto

Dian Nuswantoro University, Semarang, Indonesia; edi-nur@dosen.dinus.ac.id, fentytristanti@gmail.com, abah. syukur01@gmail.com, purwanto@dsn.dinus.ac.id, ricardus.anggi@research.dinus.ac.id, caturs@research.dinus.ac.id

Abstract

Background: Tourism industry is very important for a country. Many tourists travel into a country will help and improve its economic growth. **Methods**: Many researchers used Backpropagation Neural Network (BPNN) for predicting tourist arrivals in a country. As the result, BPNN is proven to give good results, but the accuracy is still less than optimal. This study uses series dataset from the arrival of foreign tourists in the district of Central Java's: Magelang, Solo, and Wonosobo from 1991 to 2013. We compared the performance of BPNN, K-Nearest Neighbor (KNN) and Multiple Linier Regression (MLR). Genetic Algorithm is used to optimize the parameters of BPNN, such as learning rate, training cycle, and momentum. The performance is measured by Root Mean Square Error (RMSE). **Findings:** BPNN produces small error of prediction compare to KNN and MLR. KNN performed the worst when used to predict. **Improvements:** Genetic algorithm proved to be able to optimize the parameters of BPNN. GA is able to minimize the error of the prediction of BPNN.

Keywords: Data Mining Algorithm, Forecasting, Genetic Algorithm, Neural Network, Tourism Arrival

1. Introduction

The number of foreign tourists is able to increase the country's foreign exchange. It is caused by the high exchange foreign currency against the local currency. So one way to increase the value of a currency is to increase the number of foreign tourists. With the knowledge of the number of tourists will come, then the government can prepare strategic steps to build the tourism industry. Many of the benefits if the government has an accurate prediction system. Development of transport infrastructure and tourist sites can be the priority. As for investors, they can invest in the tourist locations[1].

Tourist arrivals are changing every month to make local government of Central Java province difficulty in predicting the level of tourist arrivals. With the forecast for tourist arrivals in Central Java of the three cities that are often visited by foreign tourists could be used by the government of Central Java province to avoid a decrease in the number of foreign tourists, but it can also be used by the providers of tourism facilities such as hotels and travel agency.

Neural networks have been also implemented for tourism forecasting. Authors in[2] used Neural Network, Single Exponential Smoothing (SES), Double Exponential Smoothing, and Triple Exponential Smoothing forecast revenue for tourism accommodation in South Africa showed that of the four methods used better methods of Neural Network evidenced by acquisition of Neural Network for MAE 49,52 and 6,15% to MAPE. Author in[3] used BPNN to forecast tourist arrivals in Mount Huangshan in China for monthly data, BPNN method is able to predict well. Authors in[4] used BPNN, Self-Regression, Gray System, and Exponential Smoothing to forecast tourist arrivals in Beijing obtained BPNN method is superior to other methods used for comparison.

A problem of BPNN is there are several parameter need to be adjusted manually, such as momentum, learning rate, and training cycle. Therefore, some researchers often proposed optimization algorithm such as Genetic Algorithm (GA) to adjust the parameters automatically.

This paper is organized as follows. Section 2 describes literature review of our study. Section 3 explains the

*Author for correspondence

methodology. Section 4 presents the experiment results and analysis. Section 5 provides a conclusion.

2. Theoretical Background

2.1 Time Series Data

Time series data is a data consists of one numeric variable sequence according to the time. Time series data are typically used for forecasting or prediction in data mining. The notation of time series data is x_t and modeled as x_{t-n},... x_{t-2}, x_{t-1},x_t . The goal of forecasting is to predict the future data values by modeling the existing time series data[5].

2.2 Multiple Linier Regression

Multiple Linier Regression (MLR) is a statistical method used to model the relationship between the independent variables and dependent variable. MLR is widely used to predict the future values in finance[6], stock[7], temperature prediction[8], etc.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_m X_m \quad (1)$$

Where Y is the future value, α is alpha value, X_1,..., X_m are past values, and β_1, …, and β_m are beta value for X_1, …, and X_m variables, respectively.

2.3 K-Nearest Neighbor

KNN is one of the simplest supervised learning algorithm in machine learning techniques. KNN may use as classification or prediction depend on the type of label data. In our case, KNN is used as prediction since we have numeric label to be predicted. KNN starts by initialing the number of k. In prediction, the number of k could be the even or odd number, different from classification that use the odd number. Next, the distance between the testing and training data is computed. The prediction is determined by compute the average k nearest distance neighbor of the test data. There are several distance measures could be used. In our experiment, we implement Euclidean Distance as shown in Equation 2.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

Where d is the distance value, x is the testing data and y is the training data.

2.4 Backpropagation Neural Network

Back propagation Neural Network (BPNN) is a supervised learning algorithm. It is widely applied in various engineering research fields. BPNN consists of training and testing phase. This model at least have three layers which are input, hidden and output layer as shown in Figure 1.

The input layer receives multivariate type data from the external sources. Then the input data is processed into the hidden layer and output layer. In an experiment, the model of BPNN may have more than two hidden layer. It is varies as well as the number of neuron in the hidden layer. BPNN needs initialize some parameter such as the number of layer, the value of bias, learning rate, momentum training cycle and error function. BPNN distribute the input value into the hidden and output layer through Equation (1) and (2). The error of prediction is computed by Equation (3). When the error of prediction is obtained, BPNN moves to backward to update the weight with Equation (6) and (7).

$$z = b_0 + \sum_{i=0}^{n} x_i v_{ij} \quad (3)$$

$$y = b_i + \sum_{i=0}^{n} x_i w_{ij} \quad (4)$$

$$\delta_y = (t_k - z_k) f'(y) \quad (5)$$

$$\Delta v = \sum_{j=1}^{m} \delta_y w_{ij} \quad (6)$$

$$\Delta w = \alpha \delta_y z \quad (7)$$

$$w_{new} = w_{old} + \Delta w \quad (8)$$

$$v_{new} = v_{old} + \Delta v \quad (9)$$

where x, z and y are the input, hidden and output value, respectively; b is the bias value; v and w are the weight of input-hidden and hidden-output value; δ is the error of the prediction.

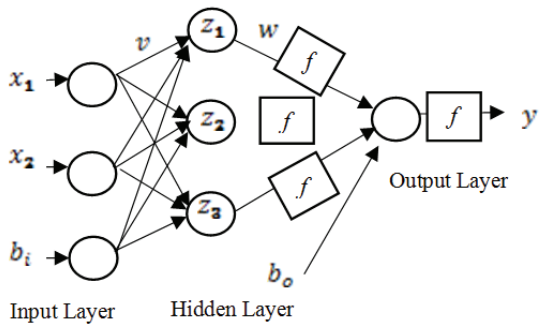

Figure 1. Single hidden layer BPNN.

2.5 Genetic Algorithm

Genetic Algorithm (GA) is an adaptive method and often to solve the optimization problem. GA consists of several stage, namely evaluation, selection, crossover, and mutation[9]. GA starts from creating alternate solutions (population). Each solution is represented as an individual or a chromosome. GA was created by John Holland in the 1960s and 1970s, GA uses analogy naturally based on natural selection. There are five main components in Genetic Algorithm:

- **The encoding technique**. How the chromosomes in the gene encodes, where for the gene is a part of the chromosome. Gene can be expressed in the form of a string of bits, for the example 100011.
- **Generating initial population**. Generate individual random process or through some particular method.
- **Selection**. Select individuals that will be used in the process of crossbreeding and mutation. Given this selection is used to find potential good parent. Early stage in the selection itself is looking for fitness values, the higher the fitness value that existed at the individual more likely to be chosen and used at a later stage.
- **Crossover**. Stage involves two parents to form a new individual. At this stage produces a new point in the search space that will be ready for testing.
- **Mutation**. Replace genes that are lost as a result of the selection process to the genes that did not recur in the initialization of the population.

3. Methodology

3.1 Data Collection

This study used the data from the Department of Tourism and Culture in the Province of Central Java, Central Bureau of Statistics in Central Java. The data was collected in every month from 1991 to 2013. Foreign tourists come to the three cities Magelang, Surakarta, and Wonosobo. Datasets are separated into training data (1991-2009) and testing data (2010-2013).

3.2 Design of the Experiment

In our experiment, we compared the prediction techniques between KNN, MLR, BPNN and BPNN with GA optimization. The prediction is evaluated by using Root Mean Square Error (RMSE). We use 7 neuron in the input layer, 7 neuron in the single hidden layer and 1 neuron in the output layer. First, we evaluate the training cycle of BPNN. Then, the training cycle that has the minimum RMSE is implemented to evaluate the learning rate. Finally, the best learning rate that has the minimum RMSE is implemented to evaluate the momentum of BPNN. The best model of BPNN after manually adjustment then is compared to BPNN with GA optimization. We used 10-fold validation to separate the proportion of training and testing dataset. The number of 10 means that 90% of the data is used to train the model and 10% is used to test the model.

3.3 Experiment Result

As explained before, we evaluate the parameter of BPNN one by one. Figure 2 shows the RMSE of training cycle evaluation. Based on our experiment, large number of training cycle attempt to reduce the error of forecasting. The RMSE of Magelang District far exceeded that of the other two districts, since it has many tourists come to Magelang. In the first ten training cycle, the RMSE of Magelang district decreased quite considerably. Compared to Magelang, the RMSE of Surakarta nearly equal to Wonosobo. Figure 3 shows the performance evaluation of learning rate. The value of learning rate starts from 0.1. The RMSE of Magelang increased significantly to over 10,000. Overwhile, Surakarta and Wonosobo rose steadily to reach around 1,300. It seems that minimum error of prediction was produced by the small of learning rate. Figure 4 shows the performance of momentum. The small momentum also produces the small error of prediction.

After reached the best parameter of BPNN above, we perform the evaluation of BPNN with GA optimization. In Figure 5 shows the error arrival tourism prediction in

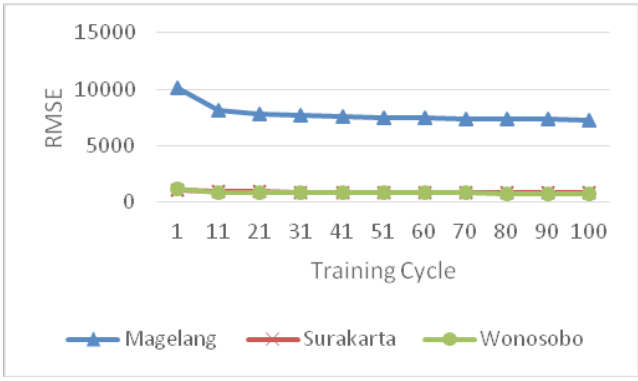

Figure 2. Training cycle adjustment.

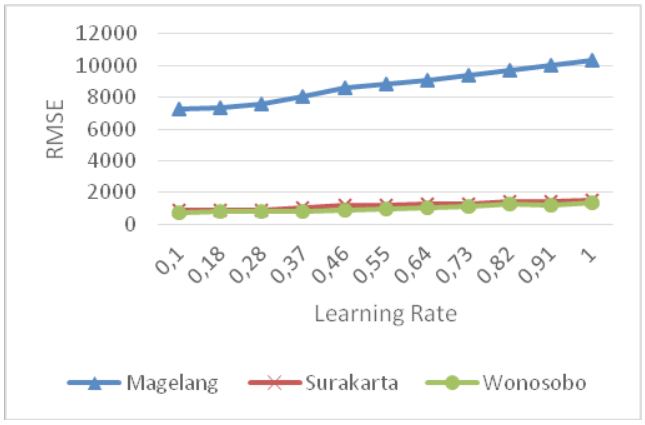

Figure 3. Learning rate adjustment.

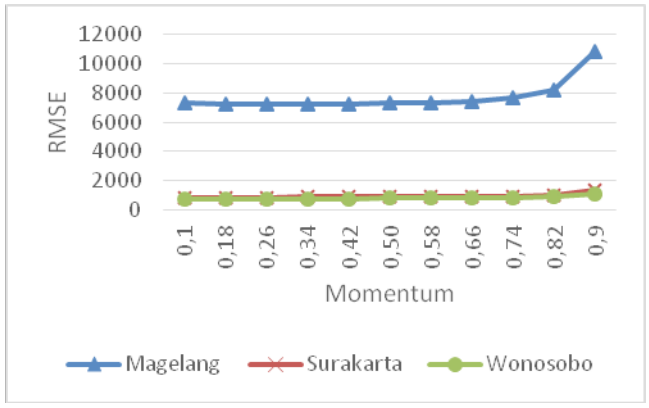

Figure 4. Momentum adjustment.

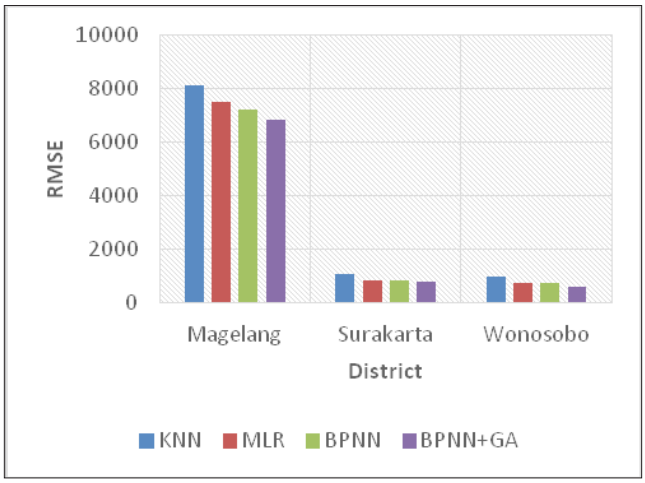

Figure 5. Comparison of BPNN and BPNN+GA.

three different districts in Central Java. BPNN with GA optimization is compared to KNN, MLR and BPNN. From there, it is clear that BPNN with GA optimization is outperforms other three algorithms. In all datasets, the RMSE of KNN is higher than the others. BPNN with GA optimization shows lower RMSE value than the others. Overall, the performance of BPNN with GA optimization is satisfactory. Figure 6, 7, and 8 show the comparison of prediction value against the actual value in Magelang, Wonosobo and Surakarta, respectively. In these figures, the gap between actual and prediction value is not too high.

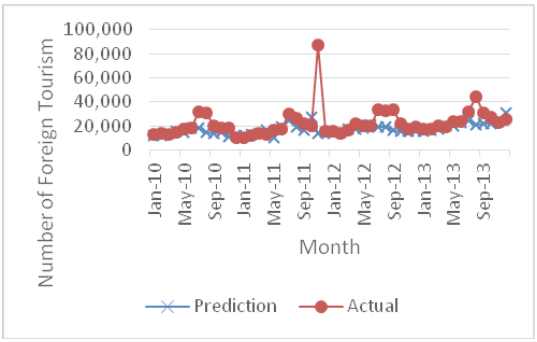

Figure 6. The line graph of actual and prediction of the number of tourism in Magelang District.

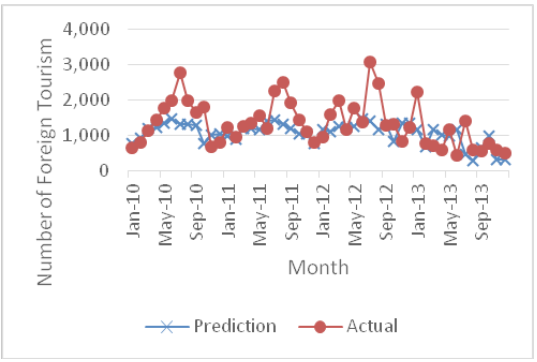

Figure 7. The line graph of actual and prediction of the number of tourism in Wonosobo District.

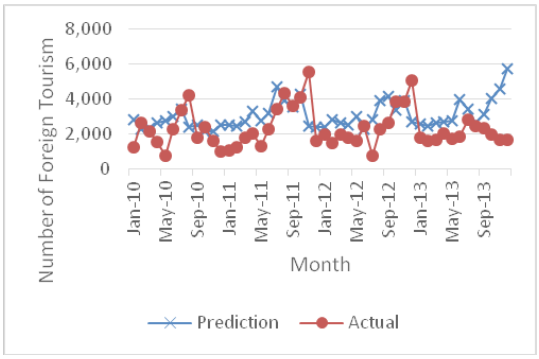

Figure 8. The line graph of actual and prediction of the number of tourism in Surakarta District.

4. Conclusion

Genetic Algorithm is a method capable of providing good RMSE results by optimizing the parameters of BPNN. By optimizing the parameters BPNN using Genetic Algorithms, RMSE obtained better results than just using BPNN without optimization. As for predictions of tourist arrivals in the three cities like Magelang, Surakarta, and Wonosobo from 2010 until 2013 showed a similar pattern in the chart that compared with actual data obtained from Statistics Tourism Central Java. It also can add attributes that affect tourist arrivals for a visit to Central Java, for example such as the number of tourist attraction, visitors of hotels or lodging, the number of tourists who come through the airport, and so forth.

5. References

1. Palmer A, Montano JJ, Sese A. Designing an artificial neural network for forecasting tourism time series. Tourism Management. 2006; 27(5):781–90.
2. Sigauke C, Darikwa T, Masemola M. Prediction of South Africa's tourism hotel accommodation monthly income: Challenges in an environment characterised by a world recession and a world cup. Mediterranean Journal of Social Sciences. 2014; 5(20):460–5.
3. Liu L-J. Tourist traffic prediction method based on the RBF neural network. Journal of Chemical and Pharmaceutical Research. 2014; 6(3):1121–5.
4. Lubo Z, Rui K. The forecast and analysis of Beijing inbound tourism demand. Proceedings of 2010 International Symposium on Tourism Resources and Management; 2010.
5. Anwar S, Watanabe K. Performance comparison of multiple linear regression and artificial neural networks in predicting depositor return of islamic bank. 2010 International Conference on E-business, Management and Economics; 2011.
6. Desai VS, Bharati R. A comparison of linear regression and neural network methods for predicting excess returns on large stocks. Annals of Operations Research. 1998; 78:127–63.
7. Singh S, Bhambri P, Gill J. Time series based temperature prediction using back propagation with genetic algorithm technique. International Journal of Computer Science Issues. 2011; 8(5).
8. Plummer EA. Time series forecasting with feed- forward neural networks: Guidelines and limitations [MS thesis]. Laramie, USA: Department of Computer Science, The Graduate School of The University of Wyoming; 2000.
9. Indhumathi S, Venkatesan D. Improving coverage deployment for dynamic nodes using genetic algorithm in wireless sensor networks. Indian Journal of Science and Technology. 2015 Jul; 8(16).