ISSN 2616-5805 Vol. 4, Issue 4: 22-26, DOI: 10.25236/AJMS.2023.040404

Application of BP neural network model based on whale algorithm optimization in text word separation

Hangyu Zeng

School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China

Abstract: *Wordle is a New York Times crossword puzzle that has become very popular recently. In this paper, Wordle is a New York Times crossword puzzle that has become very popular recently. In this paper, a SIR infectious disease model was developed to explain the reasons for the variation in the number of daily reported results and to predict the number of future reports. In this paper, a multi-output BP neural network model optimized by a whale optimization algorithm is built to predict the proportion of reports of the term EERIE on 1 March 2023 as (1.67, 3.47, 18.75, 39.20, 11.67, 13.63, 11.61). The uncertainty in the model arises mainly from the inadequate sample size and possible differences between the reported and true values. For the confidence in the predictions, the mean absolute error (MAPE) was used to measure the confidence in the predictions and the model had a confidence in the predictions of approximately 0.812.In this paper, the number of letter repetitions and the usage rate of the letters that make up a word are considered to be the two most important dimensions that affect the difficulty of a word, so this paper uses the K-Means algorithm to classify words according to these two dimensions and obtain five categories. Then, the data with labels were substituted into the decision tree model to obtain the specific classification criteria, and it was found that the main factor affecting the classification results was the rate of letter usage of the constituent words, with an error rate of 3.28% for 10-fold cross-validation. Finally, we used this model to classify the difficulty of the word EERIE, and the result was that EERIE belonged to the most difficult category of words. In this paper, we found four characteristics of the dataset: 1. the proportion of players choosing high difficulty gradually increases; 2. the difficulty of the game is stable over time; 3. the common use rate of words is stable; and 4. the proportion of unused words leads to a larger proportion of failures. To sum up, this paper is based on the requirements given in the title, comprehensive analysis of Wordle software and the relevant content of the New York Times data report, model construction is rigorous and accurate, with strong rational and practical significance.*

Keywords: *SIR Infectious Disease Model, Whale Optimization Algorithm, BP neural network, K-means Clus- Tering Algorithm*

1. Introduction

WOA-BP (Whale Optimization Algorithm-Back Propagation) is a deep learning method based on the Whale Optimization Algorithm and BP neural networks [1-3]. The initial weights and thresholds of the traditional BP neural network model during training are generated by rans'sdom numbers, which may have an impact on the structure of the network after training. At the same time, the sample size of the dataset in this paper is small, and the use of standard BP neural networks is prone to overfitting, resulting in poor prediction ability for new samples [4]. Therefore, this paper considers the use of the whale optimisation algorithm to optimise the initial weights and thresholds of the BP neural network, which reduces the data volume requirements of traditional modelling algorithms through its heuristic search and parallel computing features, resulting in a more stable WOA-BP neural network model [5-6].

We summarise the main steps of the WOA-optimised BP neural network as follows, as shown in Figure 1.

ISSN 2616-5805 Vol. 4, Issue 4: 22-26, DOI: 10.25236/AJMS.2023.040404

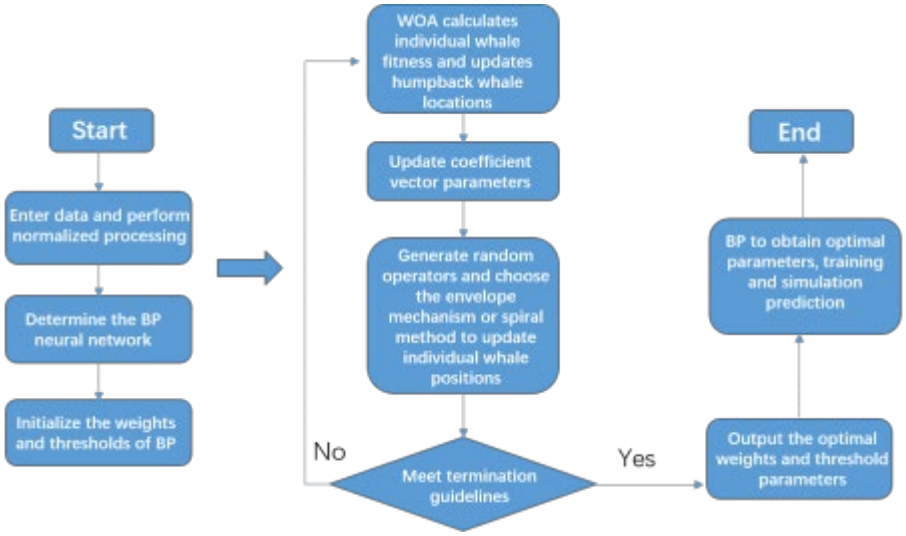

Figure 1: The operation process of WOA optimized BP neural network

This paper requires the prediction of a distribution of reported outcomes based on a given future date and word. Since the reported outcomes include a total of seven possibilities ranging from one to six attempts and failure to solve the puzzle, a multiple-output regression model is required to predict all seven cases in the distribution simultaneously. Therefore, this paper constructs a multiple-input, multiple-output prediction model based on the WOA-BP neural network algorithm to predict the distribution of reported outcomes for guessing a word at a future date by using the four word attributes identified in Problem 1 (presence of repeated letters, generality of letters contained in the word, frequency of use of the word, and number of vowel letters) and a specific time of day. Also, to measure the optimisation effect of WOA on the traditional BP-neural network [7], we compared the prediction results of the standard BP-neural network and the WOA-BP-neural network, respectively.

From the comparison plots of the predicted and true values of the BP-neural networks before and after the WOA optimisation, the fitted values of the WOA-optimised BP-neural networks were generally closer to the true values than the standard BP-neural networks when different numbers of attempts were estimated, i.e. the prediction deviations in Figure 2 and Figure 3 below were closer to 0. Therefore, for this data, the choice of the whale optimisation algorithm for neural network optimisation was feasible [8].

From the regression fit, the R2 for the training, validation and test sets were all above 0.75, giving an overall fit of 0.87. The model fits well and can be used to predict the distribution of reported outcomes for future dates.

Applying the model to predict the percentage of reported outcomes for the term "EERIE" on 1 March 2023 gave the following results: (1.67, 3.47, 18.75, 39.20, 11.67, 13.63, 11.61) for (1, 2, 3, 4, 5, 6, X).

In addition, we assessed the uncertainty and credibility of the model. In terms of uncertainty, the uncertainty of the model mainly arises from arbitrary uncertainty and cognitive uncertainty. The former mainly comes from data observation noise; the sample data used in this paper are the scores reported by users on Weibo every day, and the possible discrepancy between the reported and the real situation will lead to the cognitive uncertainty of the model; the latter is the uncertainty represented by the model parameters, which is mainly caused by the shortage of training data; the dataset used in this location only contains 341 sample data after cleaning, which is likely to be affected by the small sample size affecting the model parameters and causing cognitive uncertainty. Taken together, improving the accuracy of the reported data (or using data automatically collected by the game platform) and increasing the sample data size could effectively improve the uncertainty of the model and increase the prediction accuracy [9-10].

In terms of the credibility of the model, we assessed the model by the mean absolute percentage error (MAPE), which eliminates the effect of magnitude and takes the values [0, 1]; the smaller the value, the better the accuracy of the prediction model and the higher the credibility.

ISSN 2616-5805 Vol. 4, Issue 4: 22-26, DOI: 10.25236/AJMS.2023.040404

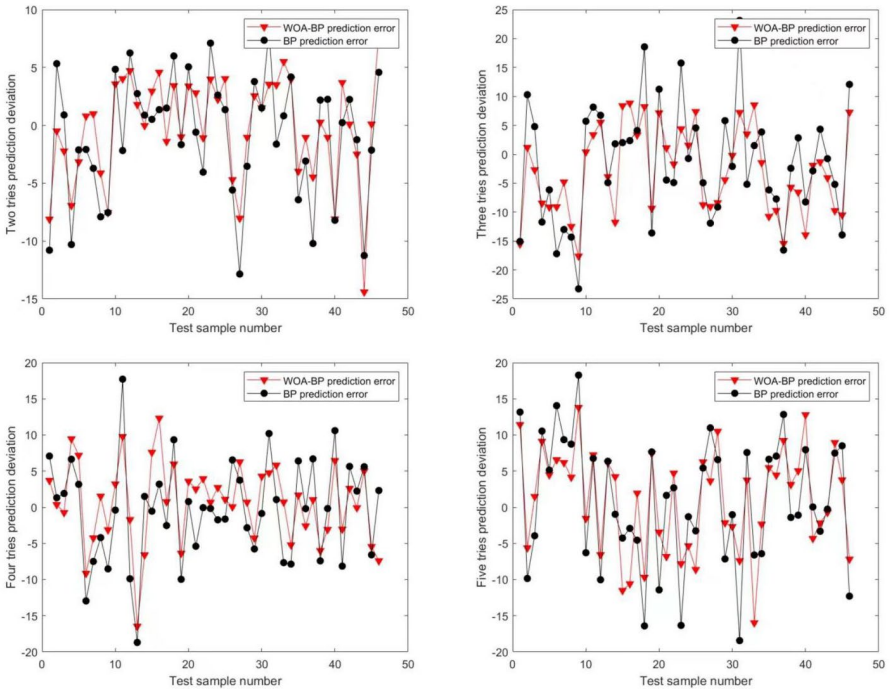

Figure 2: Comparison of predicted and true value errors of BP neural network before and after WOA optimization

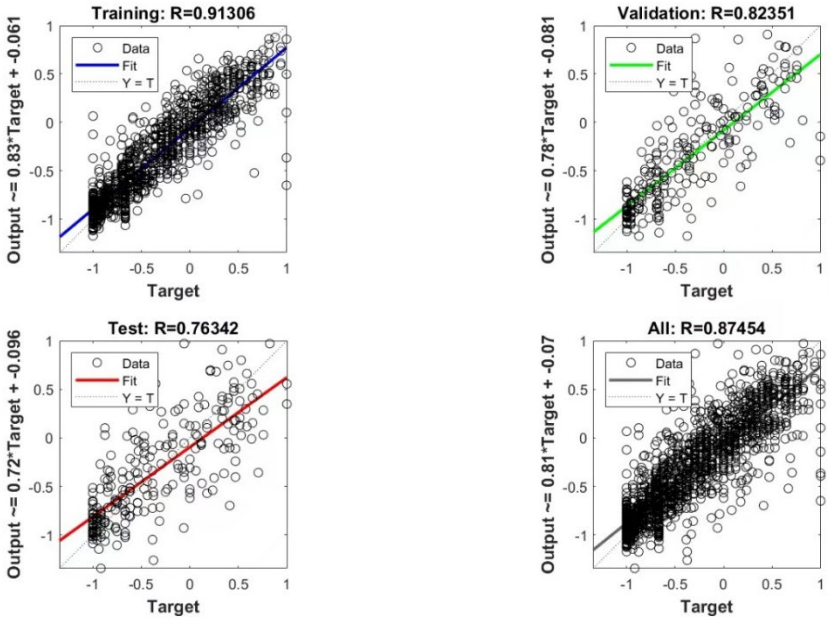

Figure 3: Comparison of predicted and true value errors of BP neural network before and after WOA optimization

2. The basic fundamental of K-Means

According to the principle of K-Means, as the number of selected centers k increases, the division of the sample becomes more refined, the degree of aggregation of each cluster increases, and the sum of squared errors (SSE) gradually becomes smaller. Therefore, when k is smaller than the optimal number of centers, the degree of aggregation increases as k increases, and the sum of squared errors decreases more rapidly. When k reaches the optimal number of centers, increasing k leads to a rapid decrease in the degree of aggregation and a slower decrease in the sum of squared errors, and then gradually flattens out as k continues to increase, and we can determine the optimal k value through the inflection point. The specific algorithm is shown below:

ISSN 2616-5805 Vol. 4, Issue 4: 22-26, DOI: 10.25236/AJMS.2023.040404

1) The sample is $x^1 \cdots x^m, x^{(i)} \in R$, the set is denoted as X

2) Randomly select k cluster centroids as , $\mu_1, \mu_2, \cdots, \mu_k \in R$, the set is denoted as U

3) Compute the clusters to which each sample belongs:

$$c^{(i)} = \arg\min_j \| x^{(i)} - \mu_j \|^2$$

Where $c^{(i)}$ indicates the cluster number of $x^{(i)}$ in the range [1, k] .min and the subsequent paradigms indicate that for each given sample x_i, the centroid u_j with the smallest Euclidean distance is selected from k cluster centroids. Arg is the subscript number of the result into u_j assigned to c_i.

4) After each cluster is divided, a new cluster centroid is computed for each cluster:

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

The equation $1\{c^{(i)} = j\}$ means if $c^{(i)} = j$ is 1, otherwise it is 0.

5) Repeat the above steps until the clustering centroids remain unchanged or stop when the expected number of iterations is reached.

3. Results

3.1 Data simulation

This paper requires a classification of words according to their difficulty. We can determine that the number of letter repetitions in a word and the commonness of the letters that make up the word correlate most strongly with the difficulty of guessing the word correctly. Therefore, the question uses these two indicators to characterise the difficulty of the words and, based on this, the words are clustered using the K-Means algorithm. The results of K-Means clustering are shown in Figure 4 and Figure 5 below.

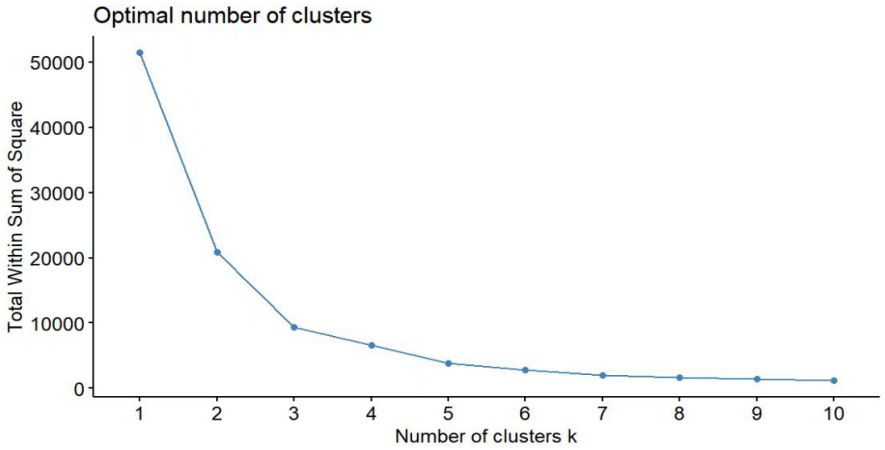

Figure 4: Time Density, Latitude and Longitude of People's Reports of Vespa Mandarinia

ISSN 2616-5805 Vol. 4, Issue 4: 22-26, DOI: 10.25236/AJMS.2023.040404

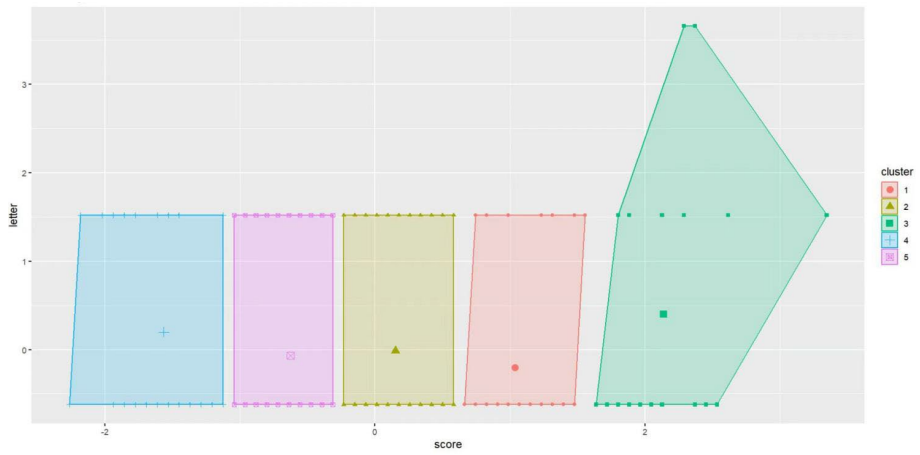

Figure 5: Time Density, Latitude and Longitude of People's Reports of Vespa Mandarinia

3.2 Analysis of experimental results

Combining the meaning of our two indicators and the correlation obtained in question 1, we can tell that the higher the number of letter repetitions and the less frequently the word uses letters, the more difficult the word is. Based on this, we can determine that category 4 is the easiest, category 5 is the second easiest, category 2 is medium, category 1 is harder and category 3 is difficult.

4. Conclusions

This paper begins by building a multi-output BP neural network model optimised by the Whale Optimisation algorithm with better prediction accuracy. After that, this paper considers the number of letter repetitions and the usage rate of the letters that make up a word as the two most important dimensions that affect the difficulty of a word, so this paper uses the K-Means algorithm to classify words according to these two dimensions and obtain five categories. The method is capable of clever text segmentation and can be generalised for future applications.

References

[1] Pang C, Ma WG, Li Cha-Wei, Jiang Y, Liao Cheng-Wang, Chen GQ. Study of microseismic source localization using a novel population intelligence optimization algorithm [J]. Geodesy and Geodynamics, 2023, 43(07):708-714.
[2] Zhang P, Wu Hsien-Teng, Lu Shengxin, Jia Chao, Wu Weiqiang. Optimal design of wind turbine grounding network based on whale algorithm [J]. High Voltage Electrical, 2023, 59(06):128-136+153.
[3] Zhai Huimin, Zhang Xinshi, Cheng Qixian, Wang Maorong, Ma Xueyao. Research on ecological evaluation of water resources in urban areas based on principal component analysis and BP neural network [J]. Journal of Hunan Institute of Technology (Natural Science Edition), 2023, 36(02):38-45.
[4] Xiong Haojie, Wei Yi. Research on multibeam sonar elevation data prediction based on improved CNN-BP [J]. Computer Science, 2023, 50(S1):575-578.
[5] Raziani S , Ahmadian S , Jalali S M J ,et al.An Efficient Hybrid Model Based on Modified Whale Optimization Algorithm and Multilayer Perceptron Neural Network for Medical Classification Problems[J].Journal of Bionic Engineering, 2022, 19(5):1504-1521.DOI:10.1007/s42235-022-00216-x.
[6] Kang Z , Qu Z .Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou[C]//2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA).IEEE, 2017.DOI:10.1109/CIAPP.2017.8167199.
[7] Li Yuntao. Short-term PV power prediction by LSTM model based on Bootstrap algorithm and whale optimization algorithm [J]. Information Technology and Informatization, 2023, (05):188-191.
[8] Shi M, Jing X. Improved whale optimization algorithm for inversion of leaf biochemical parameters [J]. Geospatial Information, 2023, 21(05):1-4.
[9] Lu ZW, Xie YY. Multi-objective engineering optimization of motor structures by Pareto whale algorithm [J]. Agricultural Equipment and Vehicle Engineering, 2023, 61(05):146-149.
[10] Xiao P, Wu KQ, Ding MIF. A multi-inverse composite whale optimization algorithm based on circular search mechanism [J]. Microelectronics and Computers, 2023, (05):1-11.