Research on Asian hornet Invasion Prediction Based on LSTM and BP neural network

Chubing Chen, Nanyu Zheng and Minghua Huang

Foshan University, Foshan, Guangdong 528000, China

Keywords: LSTM model, BP neural network, genetic algorithm

Abstract: The Asian hornet is the largest wasp species in the world and is a greater potential threat to native bee populations, and the occurrence of invasive Asian hornets in Washington State poses a significant challenge to public safety and native species. First, we pre-processed the data for better analysis, including data quantification and balancing the data set using the SMOTE algorithm. And two models were built. One is an AGH propagation prediction model based on LSTM algorithm. The other is a prediction error classification probability model based on BP neural network. Second, we applied model 1 to predict the latitude and longitude of the new nest of Asian hornet as [48.88721431, -122.47042182], and the mean square error of the model was 0.0017433 and the distance from the previous nest was 29. 234 km, which was consistent with the nesting habit of the queen bee. Third, we used the four dimensions of Detection Date, Notes, Latitude, and Longitude as the input parameters of Model 2 for training, and obtained a goodness-of-fit of 0. 7862.The unprocessed dataset was predicted using model 2 to obtain the priority investigation report ranking of the Unprocessed dataset, where the sighting report with Global ID 26DF8E2 -DAOC- 4F87-A65A-2331 15BAFCCD was ranked the highest and should be investigated most priority.

1. Introduction

Biological invasion is one of the main issues affecting the biodiversity of colonies, and invasive species will have an adverse impact on the invasive area.In September 2019, nests of the Asian hornet (AGH) were discovered on Vancouver Island, British Columbia, Canada.(In September 2019, the nest of the Asian Hornet (AGH) was discovered on Vancouver Island, British Columbia, Canada.) The Asian hornet (AGH, Vespa mandarinia) is the largest known hornet in the world, measuring about 2.5 cm-4.5 cm in length [1], a small number of bumblebees can destroy European bee colonies in a short period of time, posing a major threat to the colony's biological species and the public, and a major public safety issue.Washington State has established a help line to provide an effective way for people to report sightings of the wasp.

2. Data Exploration

Since the amount of data provided in the question is large and not intuitive, we first analyze and visualize some data statistically in order to observe. In order to have a deeper understanding of the Asian Hornet movement, we visualize the latitude and longitude on all data sets, and calculate the latitude and longitude of 14 Positive IDs, and find that the Latitude range of the samples of these verified Asian Hornets: [48.7775,49.1494] , Longitude range: [-123.9431,-122.4186]. Therefore, it is proved that the distribution area of AGH is only in a smaller range than the reported colony distribution area.

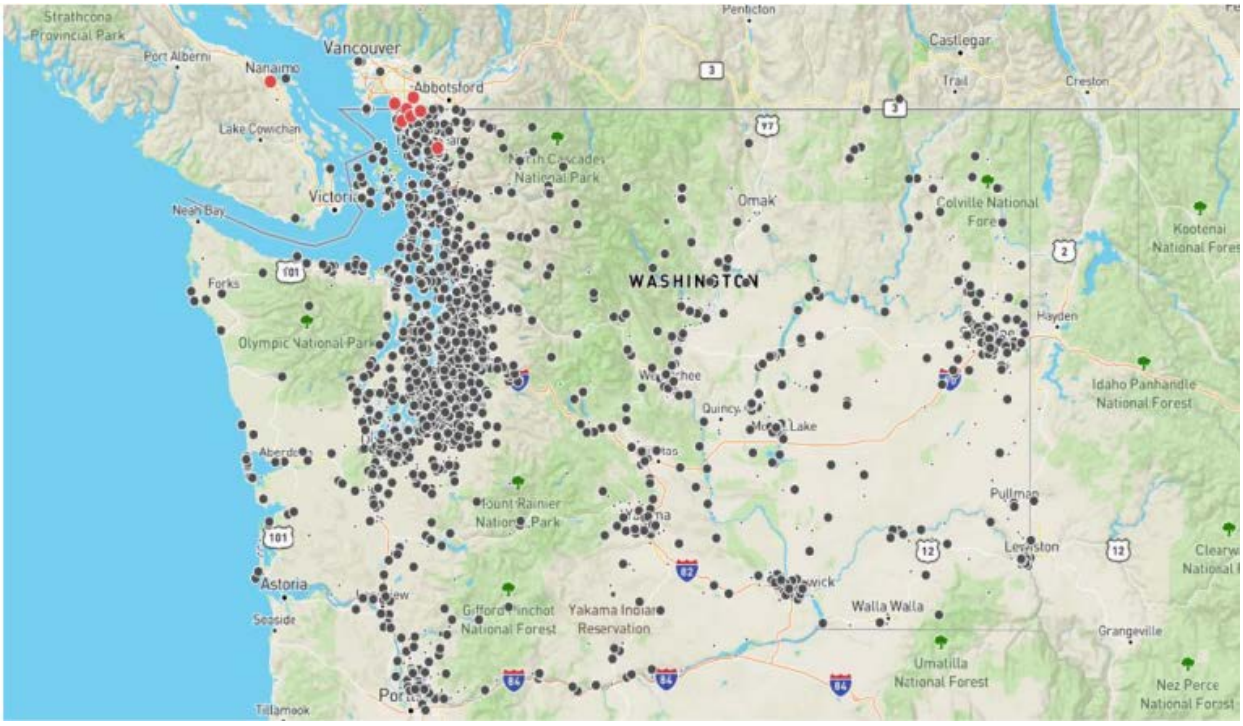

Figure 1. Population distribution map

3. LSTM model

Due to the data limitations of the data sets given in the title, there are only 14 data identified as positive sighting reports, which is a small amount of data, and our team believes that predicting the propagation of AGH in a certain period of time is mainly based on the latitude and longitude data of positive sighting reports. Through reviewing a large amount of literature and repeated attempts under existing conditions, our team decide to use the LSTM model to predict the propagation of AGH based on latitude and longitude data.

Firstly, the 14 available positive sighting reports are sorted positively according to the time of sighting, and secondly, the training and testing sets of the LSTM model are divided. By reviewing a large amount of literature, we use the latitude and longitude data of the sorted positive sighting reports as the data set, and in order to make the model training outstanding, a balance between the input layer of the LSTM model and the number of training samples is needed. Therefore, we divide the latitude and longitude datasets into equal steps and programmatically traverse the training set from 1 to 11 datasets to loop through the inputs and output the corresponding generated training set, and by comparing the corresponding accuracies, we can obtain that the model training effffect is optimal when the divided dataset is a set of 6.

Then, we use K-Means clustering method to cluster the existing 14 latitude and longitude data by the time information of the positive sighting report and the location distance of the sighting report, and we can get that the number of clusters is 8.The nesting distance of a new queen bee was estimated to be 30 km. We solved and compared the predicted latitude and longitude data with the known 8 latitude and longitude mean points by Euclidean distance respectively, and it can be concluded that the distance between the predicted location and the third category among them is about 29.234 km, which is the prediction result that best fits the actual situation, and there fore the predicted points are classified in this category, and thus the prediction of AGH in The main purpose of predicting the propagation of AGH in a certain period of time is achieved with high accuracy. The following figure shows the predicted propagation direction and geographical location of AGH.

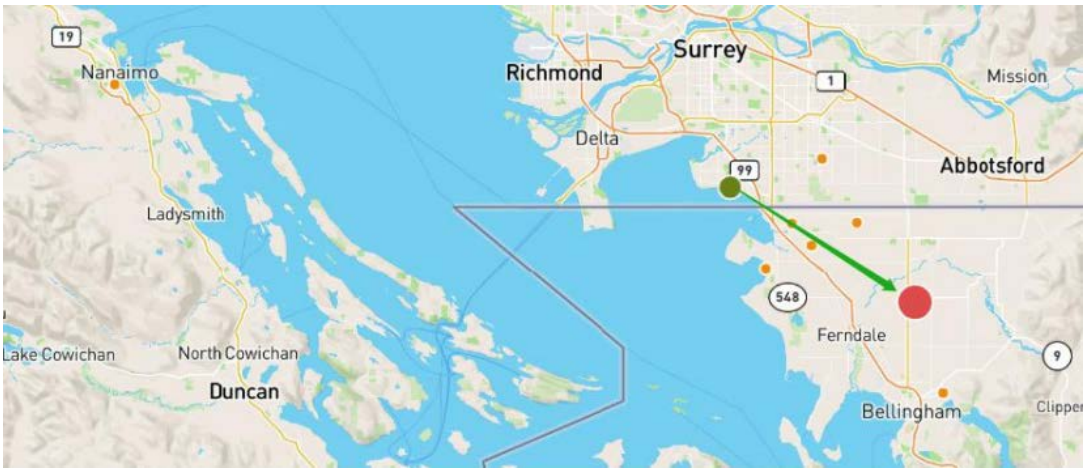

Figure 2. Prediction of AGH propagation direction and geographical location

The figure shows the geographical location of each category, where the red point represents the predicted point, the green point represents the third category, i.e., the original nesting point, and the orange point represents the remaining category, and the arrow shows the direction of AGH propagation. The formula for latitude and longitude versus distance is as follows:

$$\text{haversin}\,(\theta) = \sin^2(\theta/2) = (1 - \cos(\theta))/2$$

Among them, the haversin function is a function for any angle. The Haversine formula can be obtained:

$$\text{hav}(\Theta) = \text{hav}\,(\varphi_2 - \varphi_1) + \cos\,(\varphi_1)\cos\,(\varphi_2)\,\text{hav}\,(\lambda_2 - \lambda_1)$$

4. BP Neural Network model

4.1 Introduction of the model

The network topology generally includes a three-layer structure: an input layer, a hidden layer, and an output layer. The learning process of the BP neural network algorithm includes forward propagation of the signal and backward propagation of the error [2].

During the forward propagation, the data is input from the input layer, processed by the implicit layer, and passed to the output layer. If the output result is inconsistent with the expected result or has a large gap, the error is passed back as an adjustment signal, and then the weight matrix between neurons is adjusted to reduce the error, and the connection weights and node thresholds are continuously updated until the error of the output result is gradually reduced to an acceptable level or reaches the set number of iterations, and finally the output result is obtained.

4.2 Solution of the model

We need to predict the probability of misclassification, that is, the probability that the predicted report result is wrong, and common prediction models, such as SVM models, can only predict the result of 0,1, and cannot predict the result of wrong Probability information, so after consulting a large number of documents and actual operations, our team decided to use BP neural network to solve problem 2.After the data preprocessing performed above, four forms of data are available, the month-time information of the sighting AGH for the positive sighting report and the false sighting report, the data generated by 0-1 normalization for comments, the latitude and longitude data of the sighting location for the positive sighting report and the false sighting report, and the data generated by 0-1 normalization for the classification results of the positive sighting report and the false sighting report.

We use the first 3 data as the input layer data of the training set of BP neural network, and after a large number of attempts, we set the number of neurons in the hidden layer to 12, and use the normalized data of classification results of positive sighting reports and false sighting reports as the

output layer data of the training set, and set the number of iterations to 1000 for predicting the results of report classification. Finally, the output layer is used to generate the data with the probability that the predicted result is misclassified, as shown in Fig.

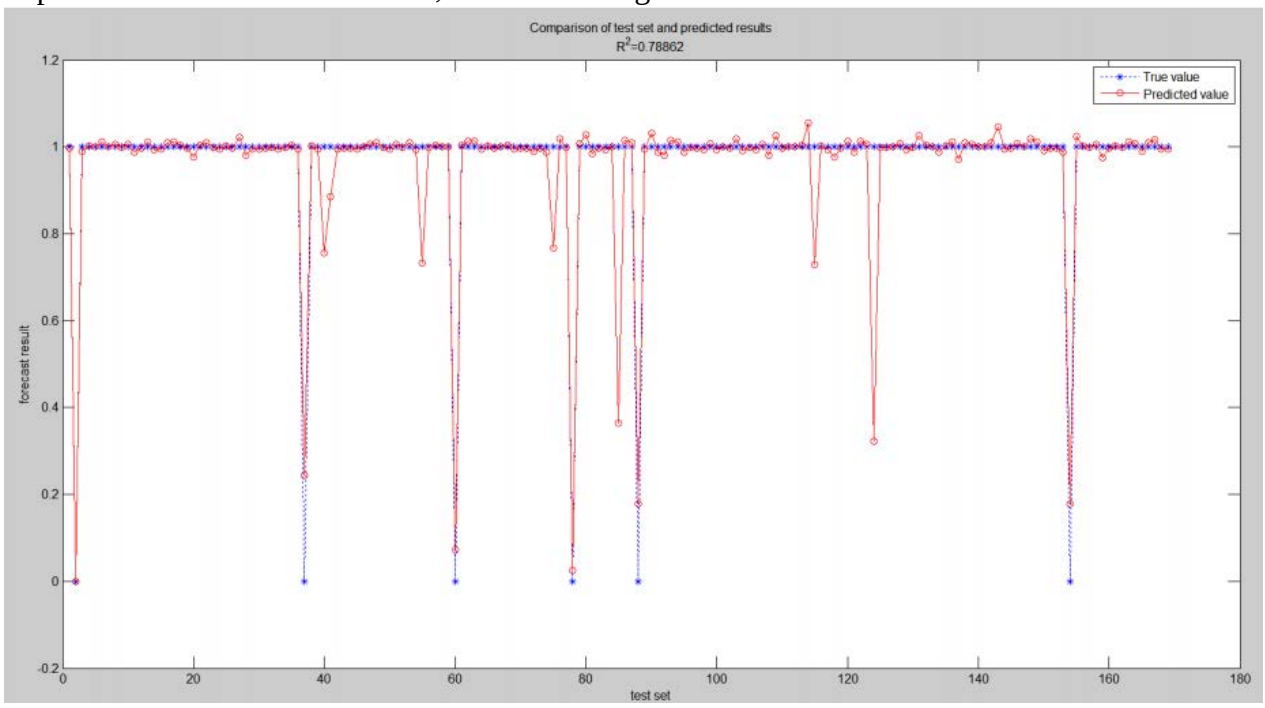

Figure 3. Prediction results

After normalization, we adjust the probability interval to 0-1, as shown in Figure 4.

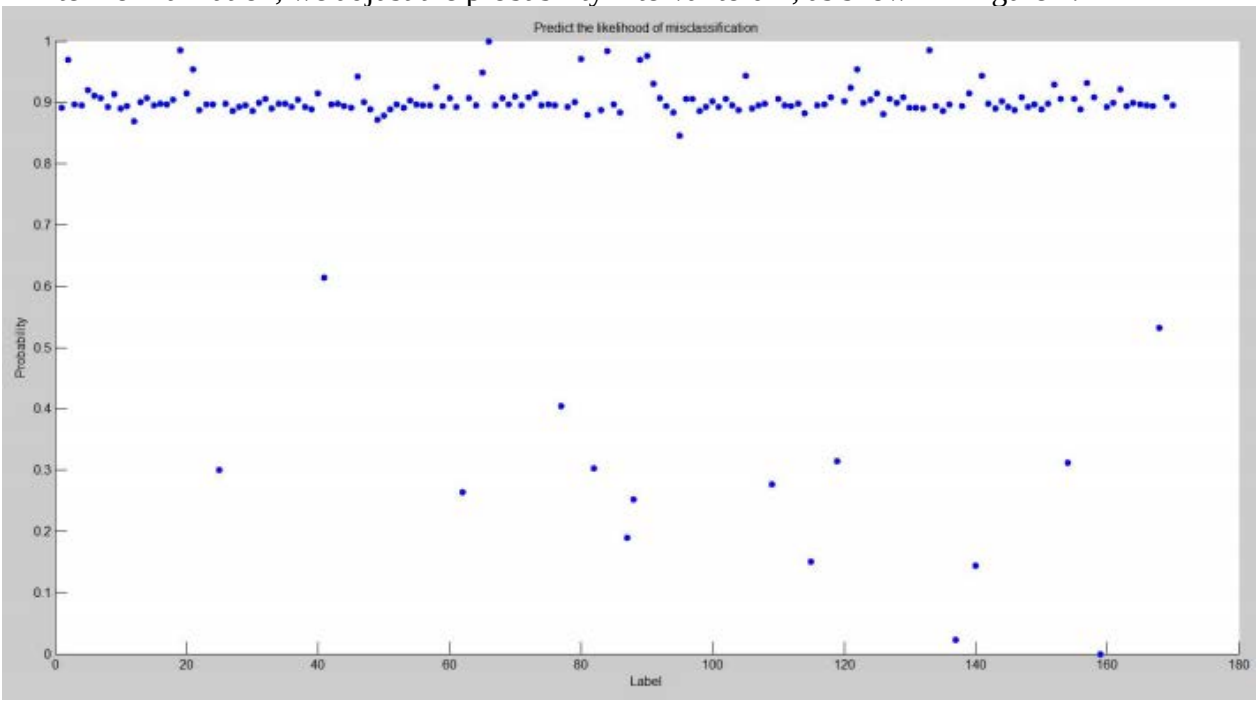

Figure 4. Probability Prediction distribution

We find that the vast majority of the reported predicted probabilities were in the range of 0.9-1, which is consistent with the actual situation of misclassification. The model's R2 reached 0.78862, with a high accuracy and better prediction results.Since most reported sightings mistake other wasps for Asian hornets, and the BP neural network model can more accurately determine that the sighting report is misclassfied, therefore,we can use this model for better resource dispatch by government agencies to speed up the pace of pest control.

4.3 Application of the model

Based on the BP neural network model established above, by analyzing the problem, we need to use the established model to generate the output probability of misclassification, through normalized data processing, the output probability can be ranked positively. when the output probability tends to 1, it means that the probability of misclassification is higher, On the contrary, when it tends to 0, it means that the probability of misclassification is lower, then the probability of correct classification is higher, therefore, the report priority survey ranking is derived through the probability ascending ranking, and the resources are allocated to the reports that need to be investigated first, so as to maximize the use of resources and reduce the cost of investigation in the case of limited resources.

Taking the unprocessed dataset in the Unprocessed dataset as an example for analysis, it is found through statistics that there are 15 unprocessed datasets.First, using the BP neural network model established above, the 15 unprocessed datasets are used as input for application analysis, we can obtain the probability distribution of the error classification generated by the unprocessed datasets, then we rank the generated error classification probabilities in ascending order and generate a priority survey ranking table as follows.

Table.1. Probability Rank

Global ID	Probability	Rank
26DDF8E2-DA0C-4F87-A65A-233115BAFCCD	0.1	1
153C4ACC-72AE-4D87-AA80-4C714417F8C6	0.956127869	2
46B8640F-7273-4517-AAD4-10D6CE6385F8	0.957484771	3
5BBFCFBA-27A6-46AB-9440-06FB025C2EEE	0.967230574	4
82D11527-5AA3-4982-A26C-0CAE8BBC40D7	0.967233433	5
23756338-4E29-4F92-ADE0-F0375321FB8B	0.968593854	6
3E50801D-9DBB-43DE-8D32-31CFA88C74D9	0.973296525	7
9BA7BDD9-01A5-4776-99B0-89FCE08CA53B	0.974591537	8
D94B286F-6AB6-4731-96BC-4B0F52188C21	0.984257644	9
DE321EDF-2949-415D-BD71-F4D7C9F81D81	0.985643438	10
665C417D-A34B-4B58-973A-569EB01C4769	0.990029073	11
1E2B3656-E2CD-4DA9-8CEF-FDE70664643B	0.990029073	12
72BE3A9B-2F2C-4051-B93E-C2C8875EA63D	0.994269381	13
7BD09981-CC6F-4FF8-8395-DABFF0D07A96	0.996936251	14
C248B633-A567-4C44-BC3B-FC4D7C74EFD1	1	15

By analyzing the ranking, we can obtain that the probability of misclassification for GobalID of 26DDF8E2-DA0C-4F87-A65A-233115BAFCCD is less than 0.5, which is the misclassification probability of this ID is the smallest.In other words, it is predicted to have the highest probability of being a positive classification report. Therefore, when the government agency has limited resources, the report with GlobalID of 26DDF8E2-DA0C-4F87-A65A- 233115BAFCCD can be considered as the highest priority report for resource allocation, and when the agency still has sufficient resources, the remaining datasets are considered to be prioritized for investigation in order of ranking, as a resource scheduling strategy for the government agency.

4.4 Promotion of the model

The invasion of AGH poses a serious threat to the survival of honey bee colonies in Washington State, and it is therefore imperative to eradicate this voracious predator and thus maintain the ecological balance of Washington State. Through statistical analysis of sighting reports in recent years, we have obtained the trend of changes in the number of reports.

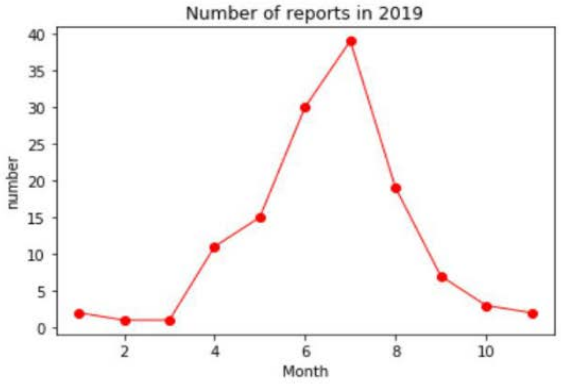

Figure 5. 2019 Reporting Frequency

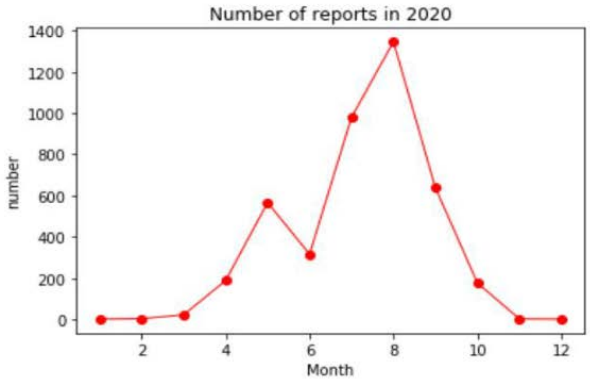

Figure 6. 2020 Reporting Frequency

4.5 Optimization of the model

Genetic Algorithm (GA) is a global search heuristic algorithm for solving optimization in the field of artificial intelligence in computer science, which is a kind of evolutionary algorithm and overcomes the shortcomings of traditional algorithms that tend to fall into local minima. It is appropriate for solving complex nonlinear problems and can be used for model optimization of BP neural networks. The basic idea of genetic algorithm based optimization of BP neural network algorithm[3] is to combine BP neural network algorithm with genetic algorithm, when the convergence speed of BP training network is slow, the threshold and weight of each hidden layer node of BP neural network is used as the input information of GA and they are encoded for generating chromosomes, and then the selection operator, crossover operator and variation operator of genetic algorithm[4] are used to generate new offspring as the initial values of BP algorithm, and finally continue to train the network using BP algorithm.

The prediction accuracy of the BP neural network model optimized based on genetic algorithm is improved from 0.78 to 0.82 compared with the model before optimization, with the following results.

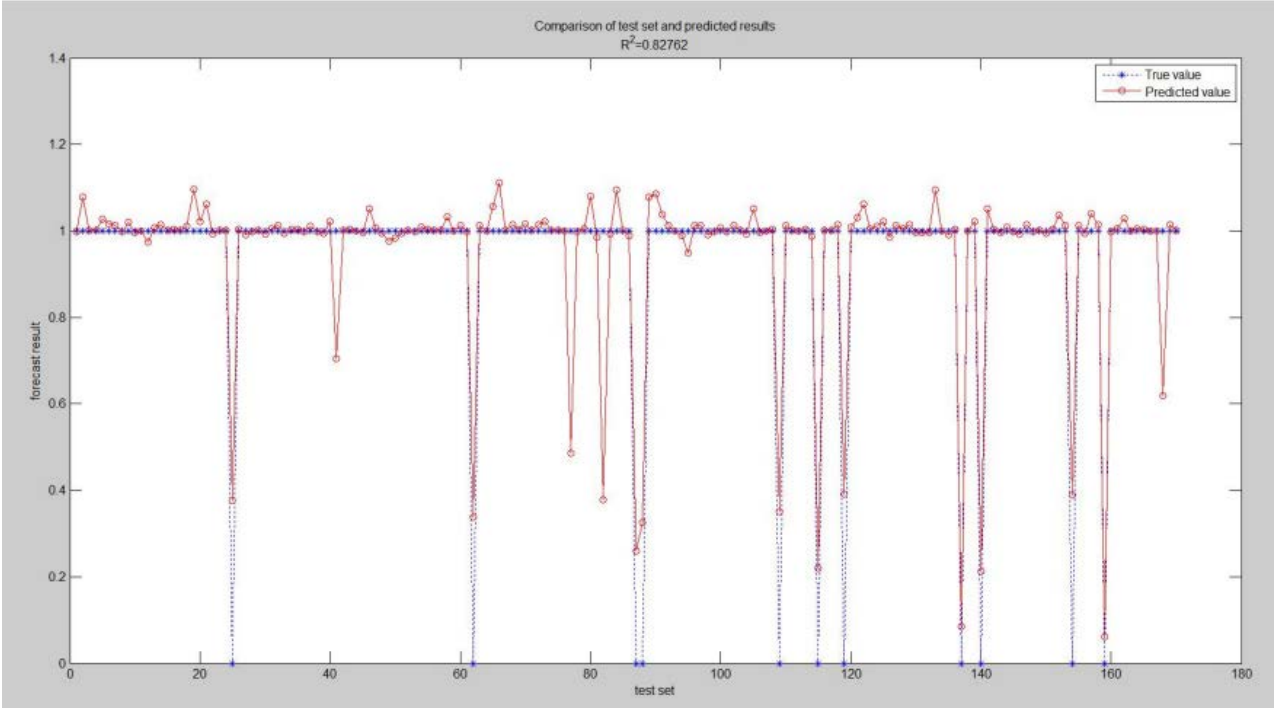

Figure 7. Comparison of test set and predicted results

By introducing the genetic algorithm, the training effect of the model is further improved, which enhances the prediction effect of the BP neural network model to a certain extent and makes the problem solution more reasonable and accurate.

5. Conclusion

To better facilitate the control of Asian hornets, we proposed two models to predict the propagation of AGH and the category of sighting reports, respectively, with good accuracy and robustness.

The LSTM-based AGH propagation model was considered for latitude and longitude data, and the geographic location of the next hive was predicted to be [48.88721431, -122.47042182] with a model mean square error of 0.0017433 and a distance of 29.234 km from the previous hive, which is known to be consistent with the actual situation when combined with the nesting habits of queen bees.

The BP neural network was used to construct a probabilistic model for misclassification of sighting reports, to predict the unprocessed dataset, and to rank the predicted results so as to prioritize the allocation of survey resources, and then we introduced a genetic algorithm to optimize the BP neural network. For the input of new samples, our team's model uses a combination of full and incremental updates, and the update frequency is chosen to be updated month by month.

By statistically analyzing the cyclical changes in the annual bee population growth, our team concluded that when the population size showed a stable or decreasing trend after March and the model judgment results all belonged to Negative ID, the population of AGH had been eliminated.

References

[1] Alaniz, Alberto J., Mario A. Carvajal, and Pablo M. Vergara. "Giants are coming? Predicting the potential spread and impacts of the giant Asian hornet (Vespa mandarinia, Hymenoptera: Vespidae) in the USA."Pest Management Science 77.1 (2021): 104-112.

[2] Li JM, He GQ. Airport visibility prediction based on BP neural network [J]. China Science and Technology Information, 2021(Z1): 39-41+44.

[3] Xie-Nan Ren. Optimization study of BP neural network based on genetic algorithm and MATLAB simulation [D]. Tianjin Normal University, 2014.

[4] Ben Wu Rui. Research on nuclear pipeline load identification and optimization based ongenetic algorithm and neural network [D]. Dalian University of Technology, 2020.