Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2108549

Based on the prediction model of wasp occurrence

Summary

This paper mainly studies the development of wasp population prediction and the identification and prediction of wasp. To predict the development of the wasp population, it is necessary to use the data and information of the relevant reports. After a comprehensive analysis of the relevant reports, how the population will develop, and how to use the relevant information to improve the recognition rate when identifying the wasp. Based on these problems, a mathematical model was established to provide a method for the forecast and identification of the wasp development, and a solution to the report that led to the priority investigation was most likely to be a positive sighting.

Aiming at problem 1, predicting the development trend of the wasp population based on a given positive report and geographic location is essentially a time series analysis of the data.

Aiming at the second question, this paper uses neural network to establish a wasp recognition model of neural network model, divides the test set and the verification set according to the positive and negative data sets, and predicts the unprocessed data based on the positive and negative data.

In response to question three, the corresponding probability is obtained through the softmax mapping using the data batch processing technology, and the report that the priority investigation is most likely to be a positive sighting.

In response to question four, take appropriate frequency to update according to the analyzed characteristics and use batch processing technology to complete the update of the model

For question 5, use the predictions in public reports to analyze the development of wasp populations

The time series model is used to predict the future development of the wasp population, and the results of model predictions such as image recognition show that the wasp has been eliminated in the state.

Keywords: LSTM; Resnet; Cellular Automata

Contents

1 Introduction

In recent years, the topic of the Asian giant hornets destroying nests has continued to appear. The occurrence of such incidents not only endangers human life and health, but also destroys the production of honey and endangers the growth and reproduction of crops. This impact has attracted widespread attention. At the same time, with the rapid development of machine learning and deep learning technology, the Asian giant hornets, as one of the main pest occurrence prediction methods, has attracted much attention.

1.1 Problem Background

The emergence of the Asian giant hornet caused a huge disaster to people in most areas. However, humans are not the prey of these Asian giant hornets. Because the hornets toxin can cause severe clinical symptoms such as hemolysis, blood clotting disorders, and allergic reactions, the specific molecular mechanism of the Asian giant hornet toxin causing these symptoms is not particularly clear. In clinical treatment, there is a lack of diagnostic markers and quick and effective treatments for the Asian giant hornet poisoning after bite. It is often impossible to receive timely treatment or even treatment, which causes thousands of deaths every year. These wasps pose a greater threat to the hive of bees, and they can kill all the bees in a hive in a short time. These attacks not only reduce honey production but also interfere with the normal pollination of plants by bees and threaten billions of dollars in crops. Therefore, research on pest control is particularly important. Although Washington State has established a helpline and a website for people to report on the sightings of the Asian giant hornet, it is worrying how to use these reports to prioritize the allocation of limited public resources. Public reports are often due to deviations in format and presentation. , Authenticity, timeliness， and other reasons reduce its availability.

1.2 Restatement of the Problem

Considering the background information and constraints identified in the problem statement, we need to address the following issues:

1. Build a model that can predict the development of pests over time.
2. According to the analysis of the model, the report of how the investigation is most likely to be a positive.
3. Discuss how often and how to update the model over time to adapt to more new reports.
4. Analyze and discuss the survival of pests.

1.3 Literature Review

Since the 1980s, the application of computer vision technology in agricultural production has been undertaken at home and abroad. Due to the slow development of computer hardware resources at that time, it was not suitable for large-scale high-performance computing. At that time, image recognition in agriculture could not meet the computing requirements, and image recognition could only be conducted in restricted environments. For example, linearize the collected insects into black-and-white images in advance, and then recognize them according to

the morphological characteristics of the collected insects. With the increase in computing power of computing resources, the theory of machine learning has been extensively used, and the rapid development of deep learning has also enabled computer vision to better help solve some agricultural production problems.

1.4 Our Work

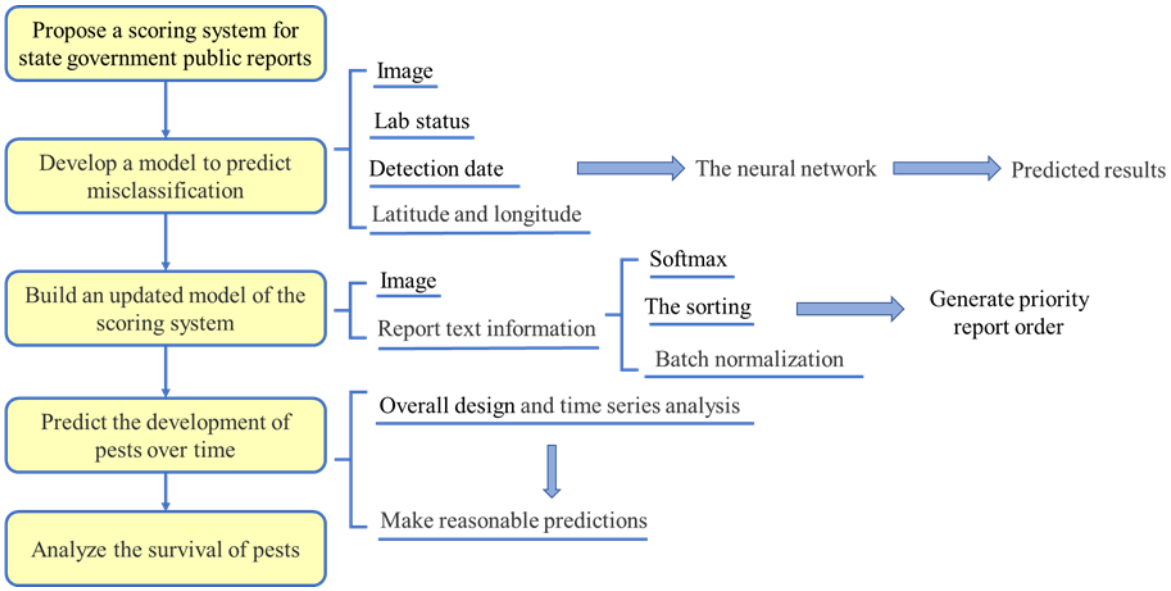

Figure 1: Model overview

The figure above shows the main workflow of our model. Establishing models, data processing, and general methods of models.

2 Assumptions and Justifications

1. We assume that the wasp in the model has a finite fixed survival time. The wasp can only survive the survival time, and will die if it exceeds the survival time.

To simulate normal existence wasp populations breeding cycle by setting the demise of the survival time of wasps, hornets in order to achieve reproduction

2. We assume that the river in the model cannot build nests. In some areas, bees will appear with a low probability. Wasps can survive by preying on bees to expand the scale of reproduction. When resources are abundant, the possibility of nesting increases, within 30*30 square kilometers Rivers, the probability of nesting increases, human pest control, there are more than 3 nests within 4*4 square kilometers, and the honeycombs in the neighborhood of von Neumann will be cleared.

According to the facts of the real world, simulated impossible events are stipulated, and the real situation and the living habits of pests are simulated to achieve better results.

3. We assume that the hive bee species in the model is a single bee species that can complete the process of self-reproduction.

By simplifying the relationship within the bee colony population, reducing factors that have little effect on the model in reality, reducing the difficulty of model simulation, and achieving better computer simulation results.

4. We assume that adjacent regions are associated with two-dimensional von Neumann type adjacent.

The real world is a three-dimensional world. The model is simplified and simplified to two-dimensional. The real world is surrounded by radiation, which is not conducive to the establishment of the model and computer simulation. The use of two-dimensional von Neumann adjacent reduces the difficulty of model.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description
t_i	Detection time i
Lon_i	Longitude of submit report i
Lat_i	Latitude of submit report i
$Note_i$	Witness report message i
C_i	Lab Comment i
$\triangle t_i$	Time interval i
φC_i	Expression analysis of lab comment i
Ls_1	Lab status 1
Ls_2	Lab status 2
Ls_3	Lab status 3
Ls_4	Lab status 4
φN_i	Analysis of the expression of the sighting report i

Von Neumann type adjacent two-dimensional: the one-dimensional space is divided into nine areas, intermediate areas as the reference system, only the neighborhood of the intermediate region area which four vertically and horizontally, not adjacent to the other regions.

Self-reproduction: In the hypothetical vespa society, there is no distinction between wasps and no need to reproduce through heterosexual mating. Any two wasps can reproduce their offspring and maintain exponential growth under sufficient resources.

4 Time series analysis model based on LSTM

Time series forecasting analysis is to use the characteristics of an event in the past period of time to predict the characteristics of the event in the future. The time series model is dependent on the sequence of events, and the results of the input model after changing the sequence of values of the same size are different. The most commonly used and most powerful tool for

time series models is recurrent neural network (RNN). Compared with the independent characteristics of the calculation results of ordinary neural networks, each hidden layer calculation result of RNN is related to the current input and the previous hidden layer result. Through this method, the calculation result of RNN has the characteristic of memorizing the previous results. The LSTM (Long Short-Term Memory) model is a variant of RNN that can deal with the limitations of the RNN model.

Our model uses the main idea of LSTM-RNN, which is to find the characteristics of an event in the past period of time to predict the characteristics of the event in the next period.

4.1 Data Description

In order to clarify the spread of the wasps over time in the reported sightings, and to determine the prediction of the possible spread of the wasps in the future, the latitude and longitude of the 14 Positive IDs in the sighting reports were counted and the Latitude range of the samples of the wasps that were verified as genuine In [48.7775,49.1494], Longitude range: in [-123.9431,-122.4186]. So it turns out that the Asian Hornet is only within a small range compared to the reported ones. We decided to first use LSTM-RNN for time series prediction analysis on a small sample of positive indicators in the witness report.

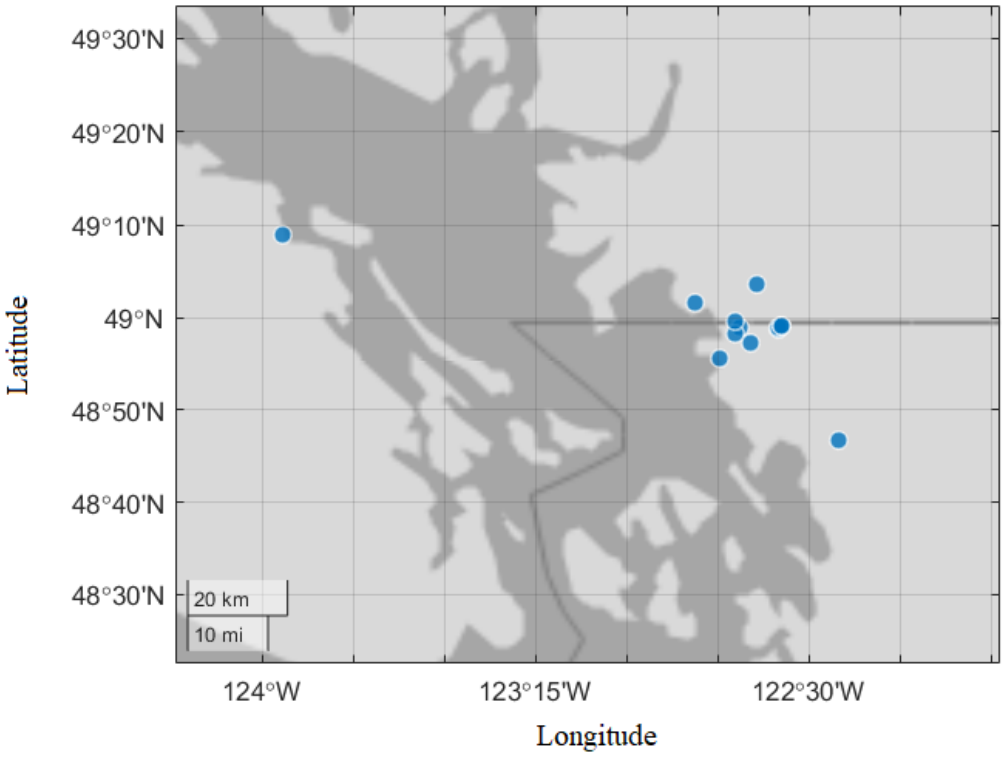

Figure 3 PostiveID distribute on maps

Extract the detection date, longitude and latitude of the 14 positive data, create a 3*14 matrix, and make the difference between the two adjacent elements of the detection date column to obtain the time interval sequence.

4.2 LSTM-RNN

Time series forecasting analysis is to use the characteristics of an event in the past period of time to predict the characteristics of the event in the future. The time series model is dependent on the sequence of events, and the results of the input model after changing the sequence of values of the same size are different. The most commonly used and most powerful tool for time series models is recurrent neural network (RNN). Compared with the independent characteristics of the calculation results of ordinary neural networks, each hidden layer calculation result of RNN is related to the current input and the previous hidden layer result. Through this method, the calculation result of RNN has the characteristic of memorizing the previous results. The LSTM (Long Short-Term Memory) model is a variant of RNN that can deal with the limitations of the RNN model.

Input gate i_t : Control the information of the current word into the memory unit. When understanding a sentence, the current word may or may not be important to the meaning of the whole sentence. The purpose of the input gate is to judge the importance of the current word to the overall situation. When the switch is turned on, the network will not consider the current input.

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

Forgetting door: Control the information from the memory unit at the previous moment into the memory unit. When understanding a sentence, the current word may continue to describe the meaning of the above, or it may start to describe new content from the current word, which has nothing to do with the above. Contrary to the input gate, it does not judge the importance of the current word, but judges the importance of the memory unit at the previous moment in calculating the current memory unit. When the switch is turned on, the network will not consider the memory unit at the previous moment.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

Output gate: The purpose of the output gate is to generate hidden units from memory cells. Not all the information in is related to hidden units, and may contain a lot of useless information. Therefore, the function of is to determine which parts of the are useful and which parts are useless.

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$

Memory unit: It combines the information of the current word and the memory unit of the previous moment. This is very similar to the residual approximation idea in ResNet. Through the "short-circuit connection" from to to, the gradient has to be effectively backpropagated. When is in the closed state, the gradient of can be directly transmitted along the lowest short-

circuit line to, without being affected by the parameter W. This is the key to LSTM's ability to effectively alleviate the disappearance of the gradient.

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

We do a time series analysis based on the geographic location changes within the time interval, and import the preprocessed data into the model. Before that, we need to define our LSTM-RNN model as follows:

We do a time series analysis based on the geographic location changes within the time interval, and import the preprocessed data into the model. Before that, we need to define our LSTM-RNN model as follows:

Define the network: We will build an LSTM neural network with 1 input time step and 1 input feature in the visible layer, 10 storage units in the LSTM hidden layer, and 1 in the fully connected output layer Neuron with linear (default) activation function

Compile the network: We will use an efficient ADAM optimization algorithm with default configuration and mean square error loss function, because this is a regression problem.

Fit the network: We will adapt the network to 1,000 epochs and use batches equal to the number of patterns in the training set. We will also close all detailed output.

Evaluate the network: We will evaluate the network on the training data set. Usually we will evaluate the model on a test or validation set.

Make predictions: We will make predictions on the training input data. Similarly, we usually make predictions on data without knowing the correct answer.

4.3 The Solution of Time series analysis

Pseudocodes

Algorithm 1: LSTM-RNN with Adam optimization regression

Input: x_t

Output: y_t

The loaded single sequence x_i can be get to cut n slice $\sigma(x_i)$
Get slice to index matrix **σ(x)**
Initial sequence **σ(x)** model added layer: embedded layer **E**
Embedded matrix **E** passes through the LSTM layer $LSTM(\mathbf{E})$
    The matrix $LSTM(\mathbf{E})$ through dropout layer and dense layer $\nabla LSTM(\mathbf{E})$
    Increase the attractiveness of softmax get prediction y_t

end

Explanation

We first standardized 4400 sighting reports, deleted items with missing information, extracted the data set, and created an information matrix to store the positive indicators. According to the geographic location information analysis of the positive indicators, we found that the positive indicators did not have direct generalization. Ability to analyze the time series model of positive indicators.

The preprocessed data was analyzed through LSTM and the predictive analysis was established using autoregressive fitting and the following results were obtained

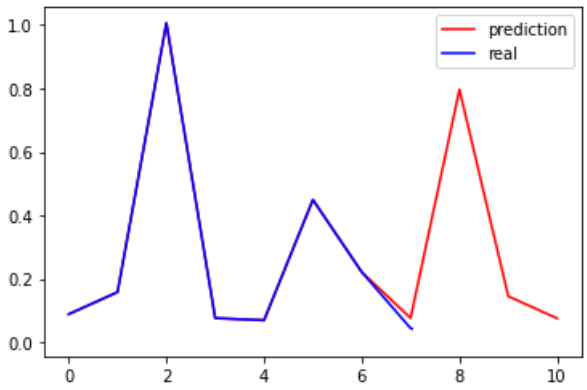

Figure 4 predication

From the results of model prediction and fitting, it can be known that the spread of wasp population over time is seasonal and conforms to the reproduction characteristics of wasp. The mean square error of MSE calculated according to the model is 20%, which is within a reasonable error range.

5 Cellular Automata Simulation Based on Positive Index

According to the study on the habit of wasp population and the data of the positive indicators in the sighting report, we can simulate the development of wasp population by using computer intelligent algorithms to get the prediction effect we need.

5.1 Cellular Automata

For the analysis of the characteristics of pests, we can make corresponding assumptions and eliminate unnecessary factors to achieve the effect of simplifying the model according to the biological characteristics of the pests, the characteristics of the environment, and the corresponding actions and policies. Therefore, we have made the following assumptions about cellular automata:

1.Vespa survives the survival time

2.No nesting on the river

3.The hive bee species is a single bee species

4.The neighboring regions are associated with von Neumann type neighbors

5.The possibility of nesting increases when living resources are abundant, and there is a river within 30*30 square kilometers, and the probability of nesting increases

6.Human pest control, there are more than 3 nests within 4*4 square kilometers, and the honeycombs in the neighborhood of von Neumann will be cleared.

The consideration of the seasonal factors to the wasp was added, and the simulation was carried out. In order to simplify the model, the von Neumann type adjacent

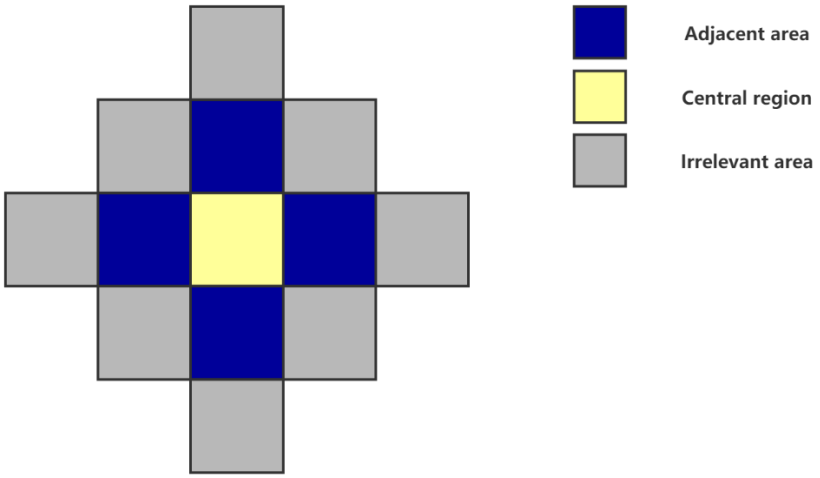

Figure 5 Von Neumann Cellular Automata

5.2 The Solution of Cellular Automata

According to the above rules, a cellular automata model is established to predict and the following results are obtained.

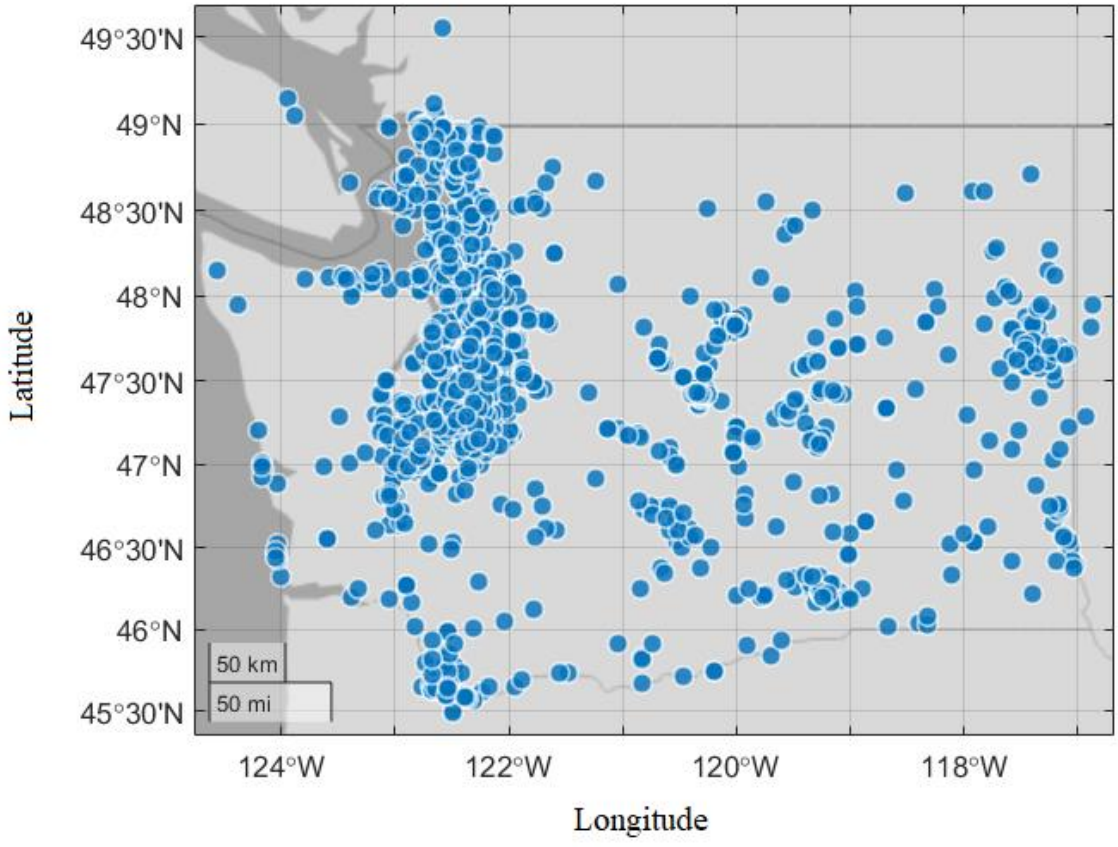

Figure 6 Sim Distribution

It can be found that the simulation results are consistent with the characteristics of the wasp population, which is more conducive to the reproduction of the wasp population in resource-rich areas

6 ANN based on feature collection

By analyzing the data set of 4400 sighting reports, it can be known that the report contains a large number of data that have not been evaluated by the laboratory and have large uncertain factors. According to the state government report, most reported sightings mistake other wasps for wasps , So we need to eliminate data that cannot be confirmed, and keep the data that has been judged as positive and negative in the report. We preprocess the witness report and select the judged indicators as our data set.

6.1 BPANN

Numericalize valuable text data, and process the remaining data according to the data processing model. After processing, several valuable parameters are obtained as input items to build a BP neural network.

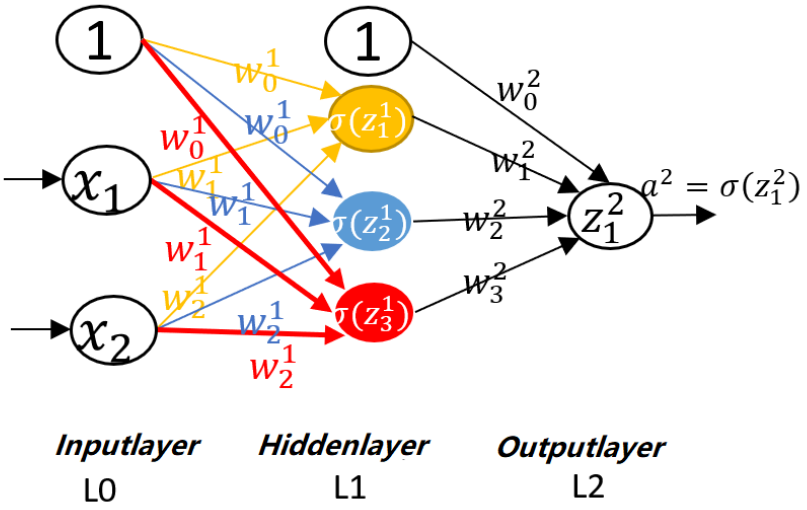

Figure 7 BPANN

We selected the negative and positive indicators that were successfully processed by the laboratory, and found that there were fewer positive indicators and a large number of negative indicators. We solved the WordAVGmodel model of the text returned by the laboratory to obtain the corresponding emotional index in the laboratory response as an addition to the model. Itemized, improve the utilization of the data set, use the latitude and longitude information and the date of detection as training data, divide the training set and test set samples according to the negative and positive indicators marked by the laboratory, establish a BP neural network model, and pass the preprocessed data into Train in the model and get the test effect in the test set.

According to the above model solution process, we know that more positive indicators

are missing in the training data, and the image set in the witness report is an unstandardized data set, so we choose to use the BeeAndWasps image data set from the Kaggle website as the training wasp image recognition The auxiliary picture of the algorithm, by choosing to use the Resnet neural network, the neural network model can learn deeper features. The following figure is a schematic diagram of the Resnet structure.

6.2 Vespa image recognition based on Resnet

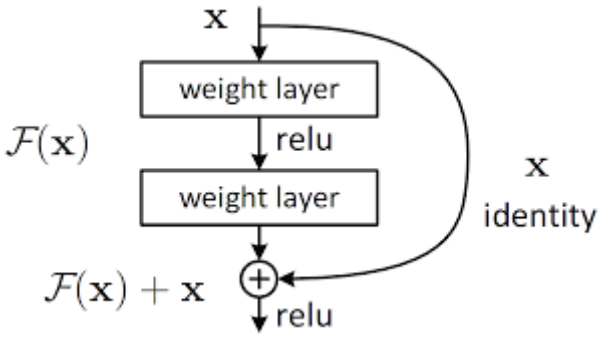

Figure 8 Resnet

6.3 The Solution of Neural Network

Randomly scramble the pre-processed data, divide the training set and test set, pass the training set into BPANN for training, use the test set to test the effect of the trained model, and use the trained model to analyze the unprocessed data in the laboratory. Prediction, pass the unprocessed data into the neural network model to get the predicted result

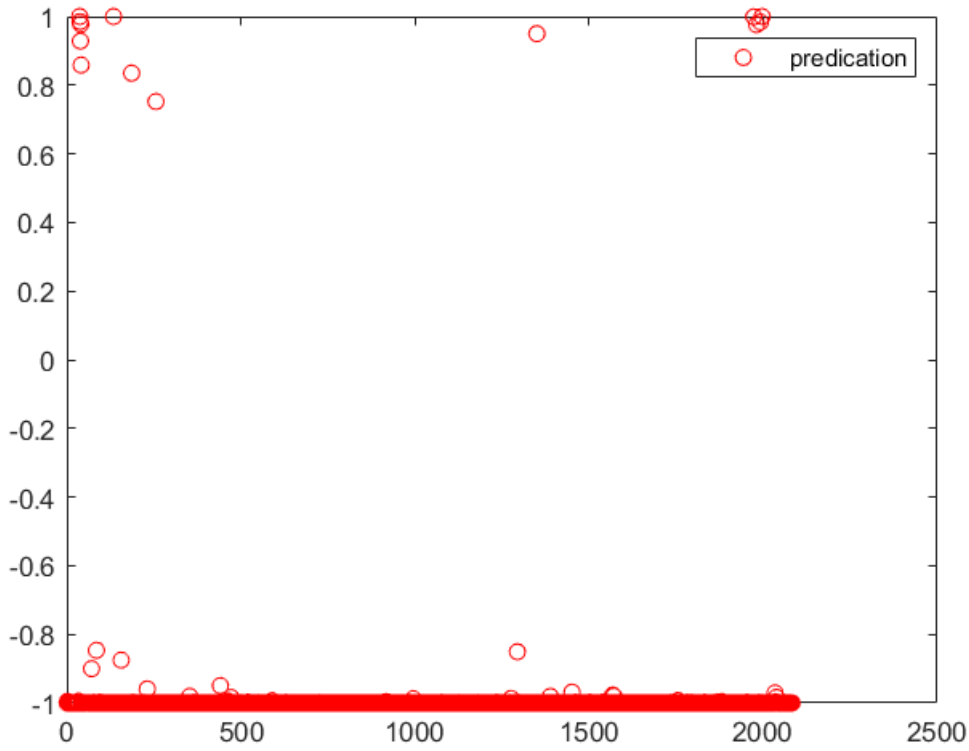

Figure 9 predicate of un-verify and un-process

Through the preliminary analysis of the data set and the analysis of the prediction results, it can be concluded that the model has a stronger ability to judge errors than correct judgments. According to the prediction results, conclusions can be drawn. There are a large number of reports in the report that misjudge local bees as wasps.

According to the data distribution after image processing, it can be seen that Washington State has eliminated this kind of pest.

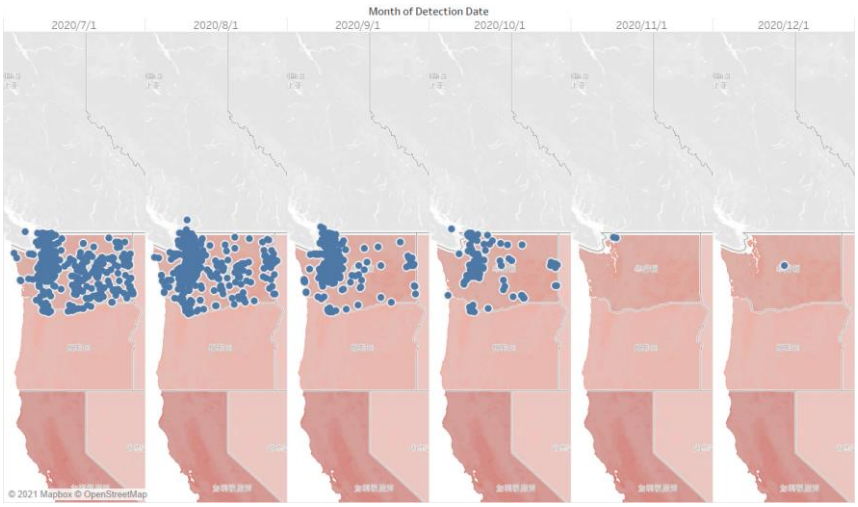

Figure 10 predicate of un-verify and un-process

7 Error Analysis

Through the performance of the model test set and the pre-processed data, it can be seen that the generalization ability of the model is poor. Although the calculated mse mean square error is 0.048193, the data set lacks many positive data, making the model unable to judge correctly Ability to get better training

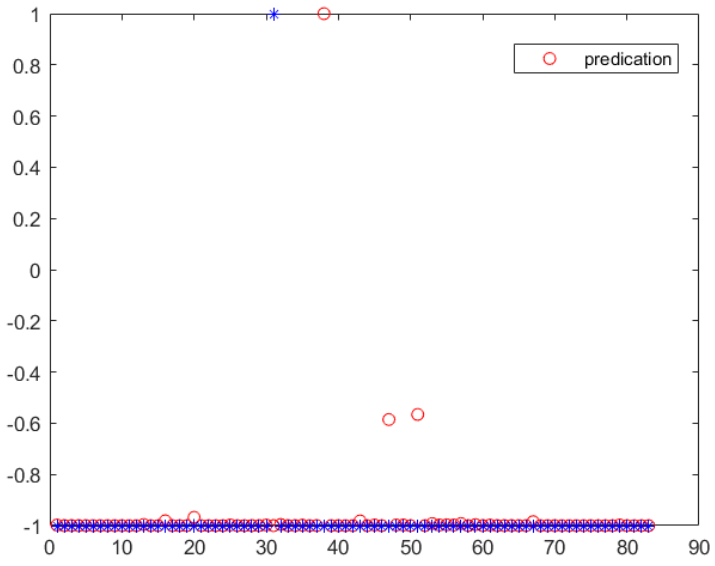

Figure 11 Testset performance

8 Model Evaluation and Further Discussion

The model should be added to batch data processing. As time develops, more data can be obtained. The batch data can be imported into the image recognition model, and the predicted results can be softmax mapped to obtain the probability of negative and positive batch data, and then sort the obtained probability vectors , Obtain the possibility sequence of the sighting report, and allocate public resources according to the sequence provided by this sequence. According to the above model, it can be known that the development of the wasp population has seasonal characteristics, and the update frequency should be adopted quarterly.

8.1 Strengths

The data set was objectively analyzed, a variety of models were selected to comprehensively analyze the problem, and the utilization rate of the data set was improved through natural language processing technology.

8.2 Weaknesses

The generalization ability of the model is weak.

8.3 Further Discussion

Standardized processing on the data set facilitates the extraction of more features to improve the accuracy of the model.

Conclusion

We conducted a comprehensive analysis of the data. First, the data was preprocessed, time series analysis was used, and the development trend of the wasp population was obtained. The neural network model was used to extract the features of the data set and the unprocessed data was reasonable Prediction. Through the establishment of an image recognition model, the use of the images in the witness report has been completed, and the credibility of the prediction has been improved. Finally, we propose a method to adapt the model to more data, add a batch data processing mechanism, and obtain result by using softmax mapping According to the positive probability of batch data, public resources can be better allocated.

References

[1] Li Hang. Statistical learning methods[M]. Tsinghua University Press, 2012.

[2] Lin Lefen, Li Yongxin, LIN, et al. Research on the Supply and Demand Matching of Commercial Bank Credit Products for Small and Medium-sized Enterprises[J]. Journal of Nanjing University (Philosophy, Humanities and Social Sciences), 2016.

[3] Li Beilei. Multi-adaptive least square curve fitting method and its application [D]. Yangtze University, 2014.

[4] Zhuo Jinwu. Application of MATLAB in Mathematical Modeling. Second Edition [M]. Beijing University of Aeronautics and Astronautics Press, 2014.

[5] Zhou Zhihua. Machine learning: Machine learning[M]. Tsinghua University Press,

2016.

[6] Xuan Shibin. Weighted sparse PCA algorithm and its application[J]. Journal of Chongqing University, 2014(04):49-54.

[7] Harrington Li Rui. Machine learning in action: Machine learning in action[M]. People's Posts and Telecommunications Press, 2013.

[8] Shao Fengjing. Principles and Algorithms of Data Mining[M]. Water Resources and Hydropower Press, 2003.