Environment, Resource and Ecology Journal (2021) 5: 37-45
Clausius Scientific Press, Canada
DOI: 10.23977/erej.2021.050106
ISSN 2616-3756

The Prediction and Classification of Vespa Mandarinia Based On LSTM and Decision Tree

Cheng Yi

Tongji University, Shanghai, 200082

Keywords: Vespa mandarinia, the Asian giant hornet, LSTM, CNN, rating, level, Decision Tree, classification

Abstract: As it know that Vespa mandarinia hunt bees and other natural creatures in large quantities, and its venom is very harmful to the human body. Once discovered in the United States, they attracted widespread attention from relevant departments and the public.In view of this situation, this article aims to solve these five problems: The first problem is to predict and analyze the spread of the Vespa mandarinia. The second problem is to establish a model based on the providing information to analyze whether the witnesses misclassified or not. The third problem is to carry out a quantitative analysis of the priority processing order of the report on the basis of the second model. The fourth problem is to use statistics to analyze the update time of the model. The last question is to judge whether the Vespa mandarinia is eradicated or not according to the model and providing data.

1. Introduction

Asian giant hornet is the world's largest hornet. The hornet is a predator of European bees and can invade and destroy their nests. In fact, a small number of hornet can destroy the entire European bee community in a short period of time. At the same time, recent studies have shown that they are also predators of many agricultural pests. Therefore, the correct and reasonable identification, research and prediction of Asian giant hornet are of great guiding significance for maintaining local ecological balance and ensuring the local agricultural security.

In order to avoid public anxiety in Asia, Washington has set up a helpline and a website for people to report sightings of the hornets. So our purpose is to establish a model to discuss, analyze and predict the existing data, to correctly explain the data provided by the public, and to take strategies to determine which reports are the most accurate.

2. Model of Spread over Time

2.1 Data Analysis

In the official data set, we extracted several more important data volumes:

- Detection Date: Record the order in which pests were discovered.
- Lab Status: Measures the accuracy of the submitted data.
- Latitude, Longitude: Record the geographic location where the pest was found.

By processing the latitude and longitude coordinates in the data set, we can get a heat map of the geographic location of the wasp witnessed by the masses (Figure 1), and we will find that the distribution of pests shows a trend of clustering.

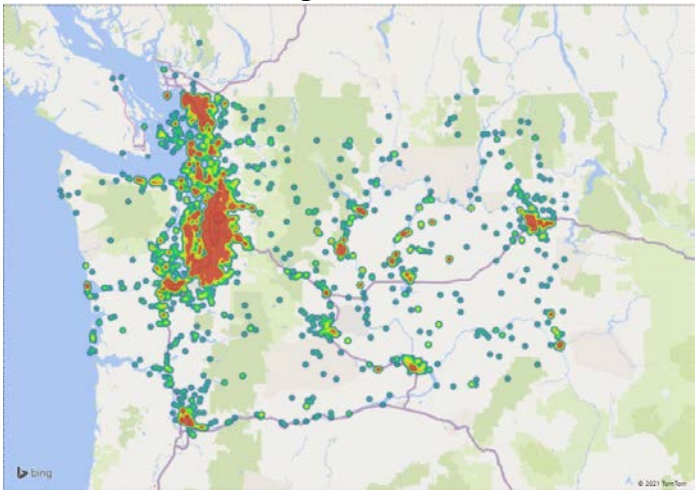


Figure 1. Wasp Distribution Heat Map

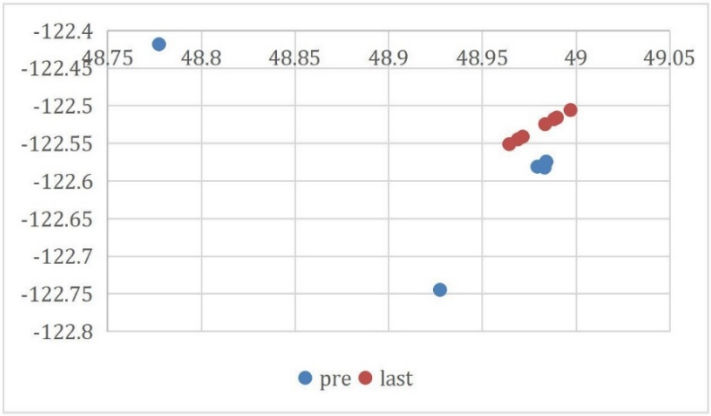
There are many factors that affect the spread of pests. Assuming that we do not consider the terrain in this problem, the impact of extreme weather conditions on the spread of wasps. The analysis elements for wasp propagation include: the flying speed of the wasp and the breeding range of the wasp. Given in the title: "A new queen has a range estimated at 30km for establishing her nest". Then we can use 30km as the tolerance range when evaluating the accuracy of the prediction results. If the value of latitude and longitude converted into distance is within 30km, we judge the prediction result as valid prediction data.

2.2 LSTM Model

The first is to process data.In the given data set file, extract the data items whose "laboratory status" is "positive ID", and sort them according to the order of detection date. Since the goal is to find out the propagation law of pests over time, we choose LSTM for time series prediction when constructing the model. The implementation code of LSTM is shown in the **appendix A**.

Among the 14 confirmation data, select the first 7 as the training set and conduct model training, and use the last 7 data as the test set to compare with the prediction results to judge the reliability of the prediction results

Table 1 Comparison of Actual Geographic Location and Predicted Geographic Location



Although the amount of data is relatively small, the trend is roughly correct based on the results reflected in the current amount of data. Since the error range was determined to be 30km, the original longitude and latitude coordinates were used as the center of the circle and 30km as the radius to draw the center map. If the predicted point is within the circle, it means that it is a correct prediction of the wasp's propagation path.

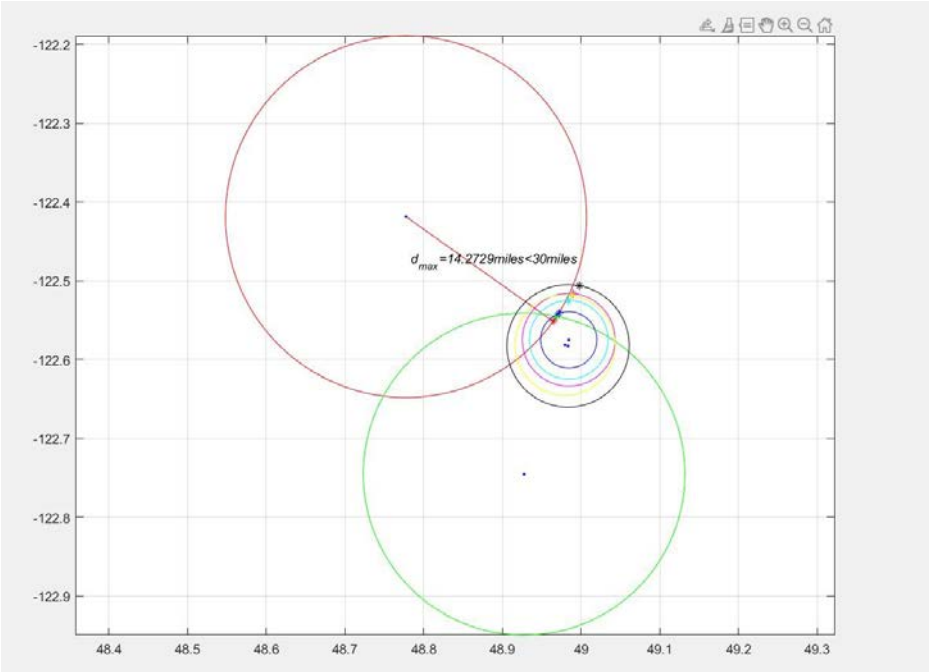


Figure 2. Wrongly Estimated Center Plot (Made by MATLAB)

3. Model for Judging Misclassification Based On Decisionmaking Tree

3.1 Analysis

Since the provided data set files include information such as videos, pictures, and some submitted comments, a model for predicting the possibility of misclassification should be established based on this information. Firstly, quantitative analysis is performed on the preprocessed excel file. Detection

date, latitude and longitude, notes and pictures or videos submitted by users can all be used as the basis for judging whether the classification is incorrect or not. However, because the dimensions of these four data are different, it is decided to quantify these four indicators and set a five-level indicator for each indicator.

Grade. Secondly, because the material contains a lot of image data, if manual judgment is made, the amount is too much and too time-consuming, so it is necessary to establish a model that can perform image recognition. Finally, based on the results of ratings and image recognition, a decision tree model that can judge whether the classification is wrong is established.

3.2 Decision-making Tree Model

3.2.1 CNN that recognizes pictures of the Asian giant hornet

First of all, there are many algorithms and models that can be used for image recognition. After looking up some information, I chose to use CNN (Convolutional Neural Network) to build a model to identify the images in the provided data set. The built CNN network model is shown in the following figure:

Layer(type)	Output Shape	Param #
conv2d_36(Conv2D	(None,148,148,32)	896
max_pooling2d_36(MaxPooling	(None,74,74,32)	0
conv2d_37(Conv2D)	(None,72,72,64)	18496
max_pooling2d_37(MaxPooling	(None,36,36,64)	0
conv2d_38(Conv2D)	(None,34,34,128)	73856
max_pooling2d_38(MaxPooling	(None,17,17,128)	0
conv2d_39(Conv2D)	(None,15,15,128)	147584
max_pooling2d_39(MaxPooling	(None,7,7,128)	0
flatten_9(Flatten)	(None,3272)	0
dense_15(Dense)	(None,512)	3211776
dense_16(Dense)	(None,2)	1026

Figure 3. CNN Network Model

Among them, Conv2D is a convolutional layer, which uses the convolution kernel to perform inner product operations on the pixels of the picture to extract image features; MaxPooling is a convolutional layer, which uses the maximum value to replace the selected area; Flatten is a smoothing layer that takes multi-dimensional input into Dense is the transition from the convolutional layer to the fully connected layer; Dense is the fully connected layer, which is classified according to feature combinations.

Next, there are two more tables provided to connect the data through GlobalID to classify the pictures. They are divided into four categories: Positive, Negative, Unverified and Unprocessed. Since the information provided contains not only pictures but also videos, take a screenshot of the moment when the Asian giant hornet appears in the video and put it into the data set.

Then read in the pictures in the data set and set a label for each picture. Since the size, number of channels, and format of the pictures provided are different, standardized processing is carried out

after reading in. Due to computer performance, the processing is 150*150 RGB three-channel pictures. Because this question is to discuss the possibility of misclassification, there are mainly two types of labels [1,0] and [0,1], the former means the Asian giant hornet, and the latter means not.

Finally, the data set is randomly divided into a training set and a verification set, and the training set is thrown into the built CNN model for training, and the images in the verification set are verified. The accuracy rate during the training process is 0.9672, the loss rate is 0.0574, the accuracy rate in the verification set is 0.9875, and the loss rate is 0.206. The line graph of the accuracy and loss rate drawn is shown in the following figure:

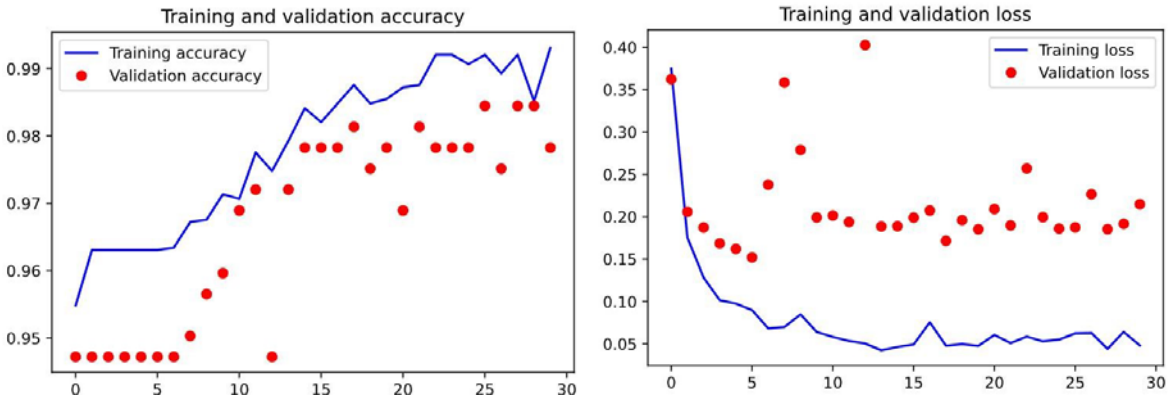


Figure 4. Line chart of accuracy and loss

3.2.2 Rating rules and results

- Detection date rating

In collating the data set, we found that the detection date corresponding to Positive ID is September to December in 2019, and May to June and August to October in 2020, which are in a relatively concentrated and narrow range. Taking into account the months when the Asian giant hornet is active, we believe that the closer the sighting time is to the sighting time of the positive ID, the higher the credibility of the corresponding sighting results, and the greater the probability that the sighting insect is the Asian giant hornet. Therefore, based on the hornet's breeding and active months, we have made a 5-level rating on the Detection date: the month where the positive ID is located is D5, and the sighting time adjacent to the month where the positive ID was discovered is rated D4, which means that for 2019 In August, the January, April, July and November of 2020 will be graded as D4, and so on, the grades will be assigned to October-November 2018, July 2019, February-March 2020 and December 2020. It is D3, the JanuarySeptember 2018 and December 2018 are rated as D2, and the time before 2018 and Null are rated as D1.

- Location rating

For the latitude and longitude of the sighting, we found that the Latitude range of the samples of the Asian giant hornet corresponding to the 14 positive IDs is [48.7775,49.1494], and the Longitude range is

[-123.9431,-122.4186], so it proved to be the corresponding to the sighting report of the Asian giant hornet .The latitude and longitude of is only within a small range relative to all reported latitudes and longitudes, so the closer the sighting location is to those sightings of positive ID, the greater the possibility of being judged as the Asian giant hornet. Due to the activity habits of the Asian giant

hornet, we can find that the activity range will not exceed 25 miles of the hive, so we will represent the longitude and latitude on the coordinate plane, and take the mean value of the longitude and latitude corresponding to the positive ID sighting location as the center of the circle, using 5 equal distances The latitude and longitude of all sighting reports are divided into 5 parts by the concentric circles of, starting from the center of the circle, the 5 parts from the inside to the outside are rated as L5, L4, L3, L2, L1.

- Notes rating

As a note for the data uploaded by witnesses, it is very likely that there is a description of the wasp seen by the public in the notes. Since the person submitting the report is most likely to be ordinary people, they do not have very professional tools, so they cannot be too precise in their description of the wasps they saw. Therefore, here we choose to read the description of this insect from Pennsylvania State University, extract the key words describing the main characteristics of it and submit it to the public within a certain error range Comment for evaluation.

For 4440 entry click comments, we have counted the most frequently occurring words among them. Except for some daily expressions, the most likely description words are: "very", "giant", "yellow", "2", etc. (Figure 7). According to the number of keywords appearing in the notes, the credibility is rated: a total of 15 keywords are counted, and n keywords appearing in the comments are tentatively rated as nlevel. Finally, due to the use of a model with a credit rating of 5, Summarize the originally divided levels into a five-level model in equal proportions, and finally divide Notes into five levels N5, N4, N3, N2, and N1.

l Picture rating

If the witness provides a picture or video, put the picture into the CNN convolutional neural network for recognition, and rank the final result. If the recognition result is the Asian giant hornet, the corresponding level is P5; if not, the corresponding level is P3; If the witness does not provide a picture, the corresponding rating is P0.

According to this rating principle, according to the weight calculation formula, the probability that each eyewitness on the five-point scale is correctly judged as the Asian giant hornet, the formula is as follows:

Possibility = DL*0.3+ LL*0.3+ NL*0.25+ PL*0.15

On this basis, we carried out a correlation analysis of the variables and got the data as shown in the figure below:

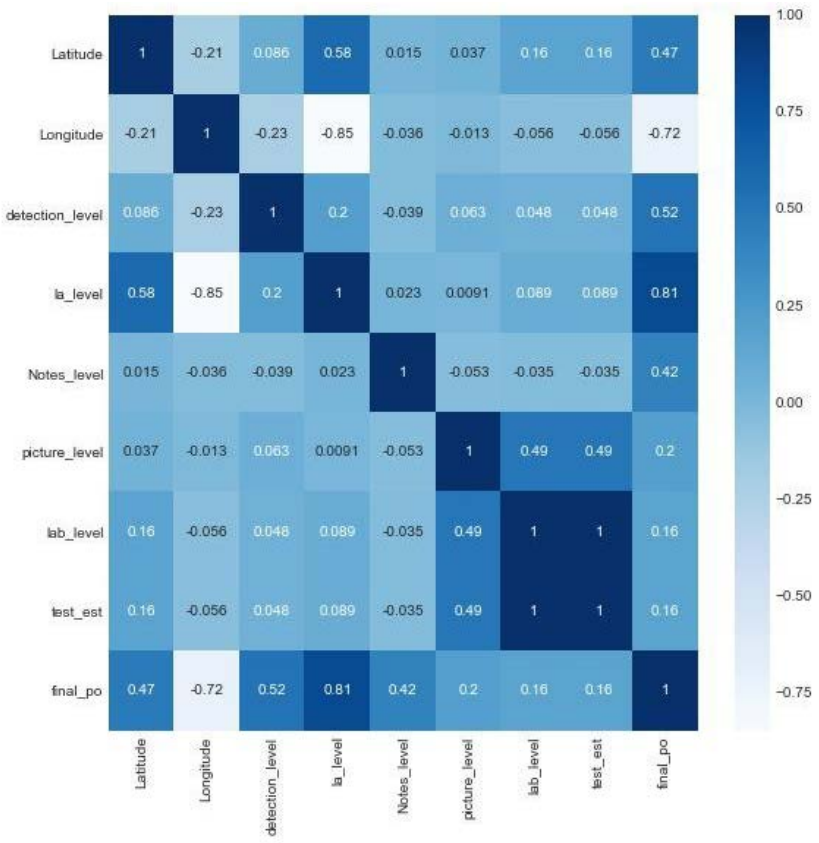


Figure 5. Data Correlation Analysis

From the correlation analysis between the variables, we can see that our selection and processing of data is more reasonable.

3.2.3 Construction of Decision Tree model

Since only 2083 of the 4440 pieces of data given have been officially confirmed as Asian giant hornet or not, it is unknown whether the remaining 2357 adjusted data is misclassified, so it is necessary to determine whether this part of the data may be misclassified. For sexual judgment, we use a decision tree model.

First, we analyzed the given data set and processed data, and selected several representative data as input to the decision tree: "Latitude", "Longitude", "detection_level", "la_level", "Notes_level", "picture_level", "final_po".

Then, we determined whether the desired result is the Asian giant hornet or not, "0" means the data is not the Asian giant hornet, and "1" means the date is the Asian giant hornet , and construct a classification decision tree model. The classification decision tree model is a tree structure for classifying instances. It consists of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. The internal node represents a feature and an attribute, while the leaf node represents a category. Put the confirmed data and results into the decision tree model for training. The decision tree will find the relationship between the put feature data and the final result, and each internal node will generate an ifthen conditional judgment. The constructed classification decision tree model is shown in the figure below:

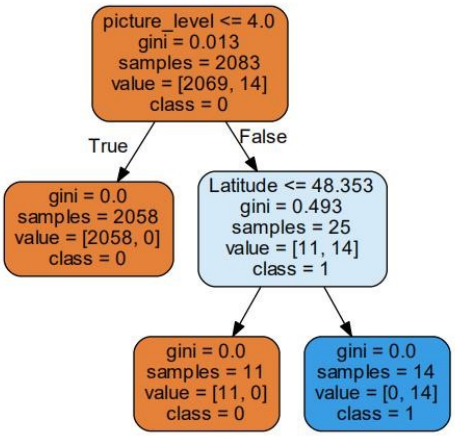


Figure 6. Classification Decision Tree Model in Problem Two

Finally, put the data that has not been officially confirmed into the decision tree for judgment, get the result, and perform statistical analysis. A total of 2,357 pieces of data were verified, of which 2,353 pieces of data were judged not to be the Asian giant hornet, and 4 pieces of data were judged to be the Asian giant hornet, so the false positive rate of witnesses reached 99.8%. The specific results are shown in the following table:

Result	Amount	Propotion
0	2353	99.83%
1	4	0.17%

Figure 7. Misjudgment Rate

3.3 Model Evaluation

Because this question needs to judge the possibility of the eyewitness's misclassification, the known results are put into the decision tree model for training, and the unconfirmed data is tested as the test data, and the false judgment rate of the witnesses is obtained However, due to the small amount of data and few witnesses who provided effective image data, when the CNN convolutional neural network was used for training, there were not enough samples to provide the model to determine the Asian giant hornet. Recognition is learning, so there is a higher error rate in recognizing the Asian giant hornet through pictures. In addition, when building a decision tree model to learn whether to misjudge or not, there are fewer feature data that can be referred to, so there will be a certain error in whether it is misjudgment.

4. Strengths and Weaknesses

Data preprocessing.When faced with big data problem, the data processing is very important. Through this step, we greatly improve the quality of the data. Thus, it is more efficient and convenient for us to solve the problem.Lack necessary data. The given positive ID data is too small, so the lack

of correct data during model judgment may lead to missed detection

References

[1] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell.Long-Term Recurrent Convolutional Networks for Visual Recognition and Description, 2016
[2] QIN Chuan. Image recognition based on convolutional neural network [J]. Electronic Technology and Software Engineering, 2020 (01): 98-99.
[3] AlexNet Recognition of Dog and Cat Data Set with TensorFlow (Cats vs. Dogs) https://blog.csdn.net/xiamencomingsoon/article/details/112263353?utm_medium=distribute.pc_relevant.n one-task-blog-baidujs_utm_term-14&spm=1001.2101.3001.4242