Analysis of Wasp Transmission through Prediction Classification Model and Support Vector Machine

Bao Wei¹, *, FangChao Liu¹, YuTing Dong²

¹Nanjing University of Science and Technology ZiJin College of Computer, Nanjing, Jiangsu, 210046
²Nanjing University of Science and Technology ZiJin College of intelligence manufacturing, Nanjing, Jiangsu, 210046

*Corresponding author: baowei2020@njust.edu.cn

Abstract. Due to the hazards caused by Vespa Mandarina, the public reports collected in Washington State were classified. Therefore, two models were established in this paper: The Grey Prediction Model and the Prediction Classification Model. We begin with the Grey Prediction Model to predict the problem of whether a pest can spread over a period of time. Based on the gray scale prediction theory and combining with experts’ identification of the sighting report as the data of Vespa Mandarina, GM (1,1) model has been established by using the gray scale prediction method.

Keywords: Prediction Classification Model, SVM.

1. Background
As the world's largest hornet, it preys on a wide range of bees, not only causing damage to bee populations, but also related to local agricultural production. The life cycle and habits of Vespa mandarina are similar to those of other wasps. They will pass the winter in the soil, and the next spring the new queen will nest within a radius of 30 kilometers.

Since the destruction of the island's hives, there have been so many sightings of the pests in neighboring Washington state that people have become concerned, so a website and helpline have been set up, and New York state has decided how to allocate resources based on the reports it has collected so it can investigate the situation further.

2. Our Work
We establish three models to solve the problems in our paper.

In order to predict the propagation of Vespa Mandarina in a period of time, and the number of data identified as Vespa Mandarina is 14, which belongs to medium and short term data, we first established a gray prediction model to predict its longitude and latitude.

Using the collected eyewitness accounts and expert appraisal result model is established in this paper, through the influence results of raw data for processing and quantitative analysis, the five factors as input, appraisal result as the output, so as to build prediction error classification model, through expert appraisal of the results and fail to identify the results of the division training set and test
set B, for training and testing of the model, so as to test and verify the accuracy of prediction error probability.

The prediction error classification model established in Question 2 was utilized to pre-process the text, longitude and latitude, time and pictures of the sighting report, and put them into the trained model. Finally, the possibility of the sighting report being Vespa Mandarinia was output, and the probability was sorted from high to low, so as to obtain the possible constructive sighting report.

According to the later collected eyewitness reports, the time when experts can later identify the correctness of the report results is used as a time point for data input to retrain the model, namely, the time point for updating the model.

By prediction error classification model, divided the expert appraisal and did not identify the sightings of substitution model, the choice may be the Asian hornet sightings, and then to predict and experts identified sightings, select correctly reported that month Vespa mandarinia on the number of times and collected monthly report, and its ratio was calculated by using time series analysis to forecast, when the future the proportion tends to zero, namely the pest has been eliminated.

3. Spread of Vespa mandarinia prediction
In this section, the Vespa mandarinia transmission, is the prediction of their latitude and longitude, through the analysis of the limited experimental data, and by using the theory of grey prediction model is established, the purpose is based on the analysis of these data, realize the prediction of unknown data. In this paper, we study the spread of harmful organisms In a period of time is a typical grey system, so you can use GM (1, 1) model to forecast.
3.1. Data Description
Considering the life cycle of Vespa mandarinia, the current seasons' profile dies out and the only individuals that survive are overworking Queens. Due to local winter about December of the year to March, this period of time without moving the hornets, spring, only to start the new queen bee colonies, start a new breeding, access to information, the development of worker bees breeding for three to six weeks, survival time for two to four weeks, so we can take a year is divided into the wasp species of a study of its life cycle.

Said the analysis shows that the data set Positive ID has determined the position of the bumblebee, other ID are not identified or has ruled out data, therefore, this article use EXCEL to 14 Positive ID of screening out the DataSet data set, and the sorted in chronological order, the spread of the discussion is that they bear on the forecast, so just Detection Date, Latitude and Longitude of the three indexes, the results shown in the following table.

| Detection Date | Latitude   | Longitude  |
|----------------|------------|------------|
| 2019/9/19      | 49.149394  | -123.94313 |
| 2019/9/30      | 48.993892  | -122.70224 |
| 2019/10/30     | 48.971949  | -122.70094 |
| 2019/11/13     | 49.025831  | -122.81065 |
| 2019/12/8      | 48.980994  | -122.6885  |
| 2020/5/15      | 49.060215  | -122.64165 |
| 2020/5/27      | 48.955587  | -122.66104 |
| 2020/6/7       | 48.777534  | -122.41861 |
| 2020/8/17      | 48.927519  | -122.74502 |
| 2020/9/21      | 48.984269  | -122.57481 |
| 2020/9/28      | 48.98422   | -122.57473 |
| 2020/9/29      | 48.984172  | -122.57472 |
| 2020/9/30      | 48.979497  | -122.58134 |
| 2020/10/1      | 48.983375  | -122.58247 |

It was visualized with Tableau software and observed that the 14 points were within a certain range, as shown in the figure below.

![Figure 2 Data point visualization](image)

3.2. Longitude and Latitude Model
The topic is required to study the spread of wasps in a period of time in the future. This paper selects the data of 9 periods after May 2020 in the table for prediction.
This paper, by using the gray prediction is through the scattered on the timeline sequence of discrete data as a set of continuous change, with the method of accumulative and b-b, unknown factor weakening of the grey system, strengthening the influence degree of the known factors, and finally build a continuous differential equations with time as the variables, through mathematical method to determine the parameters in the equation, so as to realize forecasting purposes.

We set up a GM (1,1) prediction model with the position of wasp, namely latitude and longitude, as the dependent variables respectively

First of all, we need in order to sum up the original data to weaken the randomness and volatility of the original data. Suppose our original data as following

\[ x^{(0)} = \left( x^{(0)}(1), x^{(0)}(2), \cdots x^{(0)}(n) \right) \]  

(1)

The sequence is generated by the accumulation of the original data adjacent to the mean value

\[ z^{(1)} = \left( z^{(1)}(2), z^{(1)}(3), \cdots z^{(1)}(n) \right) \]  

(2)

Then, according to the grey prediction theory, a first-order unitary differential equation GM (1,1) of the albino form of T is established for x

\[ \frac{dx^{(1)}(t)}{dt} + \alpha X^{(1)} = b \]  

(3)

Where, \( \alpha \), \( b \) are the coefficients needed to be solved by the model, \( \alpha \) is called the development coefficient, and \( b \) is called the gray action. In order to solve the problem conveniently, we combine the two parameters with solution into a parameter matrix, and regard the gray parameter matrix as the \( \beta \) of least square regression, so the solution of the gray parameter matrix can be transformed into a solution problem of the least square method to the \( \beta \) parameter matrix

\[ \hat{\beta} = \left( X^T X \right)^{-1} X^T Y \]  

(4)

By the formula can be very simple to calculate our gray parameter matrix, put the matrix into

\[ \hat{x}^{(1)}(m + 1) = x^{(0)}(1) - \frac{\hat{b}}{a} e^{- \frac{a}{m}} + \frac{\hat{b}}{a}, m = 1, 2, \cdots, n - 1 \]  

(5)

We then multiply the results, and we get the predicted value.

3.3. Analysis of the Result

Using Matlab to solve the model, to observe 2019 witnesses is found that the number of wasps, can be determined subject over a period of time to predict the number of times at about 4 times, and since then the hornets will enter the next winter, was seen again, and will be the birth of a new queen, consider to the operation of grey forecasting model can't be negative, negative longitude on behalf of
the scriptures, thus removing symbolic computation does not affect the results, final prediction results as shown in the figure below

**Figure 3** Latitude initial and analog data  **Figure 4** Longitude initial and analog data

Before the longitude and latitude prediction of wasps, we should carry out a quasi-exponential law test on them. Take the latitude as an example to test, the data with a smoothness ratio less than 0.5 accounted for 87.5%. Except for the first two periods, the data with a smoothness ratio less than 0.5 accounted for 100%, which shows that it meets the requirements of grey prediction

**Figure 5** Smooth degree of original data

Calculated respectively using model built in this paper, the traditional GM (1, 1), new information GM (1, 1) and metabolic GM (1, 1) model for the prediction error sum of squares of trial group were 0.0087294, 0.008729, 0.026482, because the new information GM (1, 1) model of minimum error sum of squares, so we chose to use the new information to estimate it.

Regarding the accuracy of prediction, when using GM(1,1) model for the prediction, we need to check its degree of fitting to the original data, and use residual test and grade-ratio deviation to calculate it.

Relative residuals:
\[
\varepsilon_r(k) = \left| \frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)} \right| \times 100\%, k = 2, 3, \ldots, n
\]  

(6)

Average residue:

\[
\bar{\varepsilon}_r = \frac{1}{n-1} \sum_{k=2}^{n} |\varepsilon_r(k)|
\]  

(7)

Stepwise Ratio :

\[
\sigma(k) = \frac{x^{(0)}(k)}{x^{(0)}(k-1)} \quad (k = 2, 3, \ldots, n)
\]  

(8)

Relative residuals:

\[
\eta(k) = \left| 1 - \frac{1 - 0.5\sigma}{1 + 0.5\sigma} \frac{1}{\sigma(k)} \right|
\]  

(9)

Average residuals:

\[
\bar{\eta} = \frac{1}{n-1} \sum_{k=2}^{n} \eta(k)
\]  

(10)

After calculation, the average relative residual sum of latitude predicted by the new information model is 0.00080757, and the average class specific deviation is 0.0014213. The results show that the model has a very good fitting degree to the original data, as shown in the figure below.

Figure 6 Accuracy test
It is mentioned in the data that the new queen bee is expected to nest within the range of 30 kilometers. In Excel, this paper calculates the longitude and latitude by using the function relation formula, and it can be found that the distance between two adjacent determined events meets the requirements. The final predicted results and the distance between them are shown in the table below:

| Latitude prediction data | Longitude prediction data | distance    |
|--------------------------|---------------------------|-------------|
| 49.0207                  | -122.5758                 | 4.17873398  |
| 49.0371                  | -122.5729                 | 1.83581416  |
| 49.0534                  | -122.5757                 | 1.82476105  |
| 49.0698                  | -122.567                  | 1.83665101  |
| 49.0862                  | -122.5641                 | 1.83579013  |
| 49.1027                  | -122.5611                 | 1.84767487  |
| 49.1191                  | -122.5582                 | 1.83577403  |

4. Report Accuracy Judgment Model
Collected using eyewitness accounts and expert appraisal result model is established in this paper, through the influence results of raw data for processing and quantitative analysis, the five factors as input, the appraisal result as the output, so as to build A model, through expert appraisal results and has not been able to identify the results are divided into training set and test set B, for training and testing of the model, so as to test and verify the accuracy of prediction error probability.

4.1. Main Factors and Data Processing of the Model
There are numerous factors that influence experts' identification of an Asian bumblebee. Delete the main data collected, they include text reports from eyewitnesses, pictures taken, the time of the finder, and the longitude and latitude of the location of the discovery.

![Figure 7 flow of Text processing](image_url)

First, this article uses Python to read the Chinese text of a DataSet table into a computationally convenient NumPy character matrix. Then, this paper attempts to transform the text data into the data form of word vector, so that the neural network can carry out further calculation.

This article uses Python's Tokenizer to segment text from its original statement form, dividing the statement into semantic chunks. Then we build a semantic dictionary based on the data composed of all the texts in the data set. In this paper, the most commonly used four thousand semantic blocks in the textual data part of the whole dataset are counted as dictionaries. The semantic blocks are then mapped in turn so that each semantic block is transformed into data. The matrix of text data is obtained.
Since the data size obtained from each text is not uniform at this time. This paper fills in the data obtained from each text with zero processing. Since about 99% of the text length in the data set used throughout this paper is less than 50, the length is unified and normalized to 50 in this paper. In other words, zero fill processing is carried out for the missing part of each piece of converted data whose length is less than 50. This results in a matrix with rows equal to the total number of reports and columns equal to 50.

Then, the data obtained in the previous step were sent as input data to the self-built LSTM Neural network for training. The model was divided into an Embedding layer to convert input data into word vectors, a LSTM layer to extract features, a Dropout layer to mitigate overfitting, and three layers of full connection. The activation function of the first two layers of the broad connection layer is the ReLU activation function, and the last layer is the Sigmoid activation function, which is used to output the predicted probability.

Its model structure is shown as follows:

LSTM\(^{(2)}\) introduces the core element Cell, which can be considered as multiple copies of the same neural network. The hidden layer of each moment in the LSTM structure contains multiple memory blocks, each block contains multiple memory cells, and each memory Cell contains one Cell and three gates.

In essence, LSTM neural network is a recursive neural network, and the data of this network at each moment will be transferred to the next moment, as shown in the figure below:

![Figure 8 recursive neural network](image)

SVM aims to find an optimal classification plane and minimize the error of all training samples from the optimal classification plane. After 6 times of training, the model constructed in this paper freezes the whole connection layer and puts its parameters into the SVR regression model for classification.

Through the above SVR method, the model finally outputs the regression results:

| regression analysis | sign |
|---------------------|------|
| 0.85801295          | 0    |
| 0.85161589          | 0    |
| 0.85598182          | 0    |
| 1.06516779          | 1    |
| 1.04057986          | 1    |
| 1.07075034          | 1    |
| 1.03707648          | 1    |

Due to limited space, this paper only presents the regression results of the first seven text data.
It can be seen that the regression result of data labeled 0 is significantly smaller than that of data labeled 1. Therefore, the regression result is used as the text probability index, and the lower the index, the higher the probability of Asian bumblebee.

4.2. Solution of the Model

This model can ignore the error of the real value in a certain upper and lower range, and its solution is characterized by the minimization of the function, which ensures the sparsity of the dual variable, the existence of the global minimum solution and the optimization of the reliable generalization bound. By using the idea of programming to minimize the error, and further introducing Lagrange function, the dual optimization problem of SVR\(^4\) can be obtained:

\[
\max_{\alpha, \alpha^*} \sum_{i=1}^{m} y_i (\hat{a}_i - \alpha_i) - \varepsilon (\hat{a}_i + \alpha_i) - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\hat{a}_i - \alpha_i) (\hat{a}_j - \alpha_j) \alpha_i^* x_i^T x_j
\]

s.t. \(\sum_{i=1}^{m} (\hat{a}_i - \alpha_i) = 0\)

\(0 \leq \alpha_i \leq C\)

\(0 \leq \alpha^*_i \leq C\)

It can control the penalty of misleading samples and obtain the optimal solution:

\[
f(x) = \sum_{i=1}^{m} (\hat{a}_i - \alpha_i) k(x, x_i) + b
\]

kernel function: \(k(x, x_i) = \phi(x) \phi(x_i)\)

5. Sensitivity Analysis

The following paper attempts to change the image index calculated by the model in the image part to observe the changes in the number of Asian bumblebee predicted by the model after the data change. In this paper, the image indexes of all samples were multiplied by a parameter \(\alpha\), and the remaining parameters were not changed.

The horizontal axis represents the number of Asian bumblebees in the test set, and the vertical axis represents the parameter \(\alpha\). The results are as follows:
It can be concluded from the above figure:

With the increase of $\alpha$, the number of Asian bumblebees was forecast to increase gradually. This paper believes that this is due to the increase of the image index, which indicates that the similarity between the original image and the Asian bumblebee is increasing.

As $\alpha$ increases, the sensitivity of the model to the predicted consequences increases until $\alpha$ is less than 2.7, and decreases slightly when $\alpha$ is between 2.7 and 3.

6. Model Evaluation and Further Discussion

6.1. Strengths

1. Prediction error classification model applies data processing methods such as image recognition and text data processing, which can easily train the model under different adaptive conditions

2. The result is a Washington State report on Asian bumblebee sightings that predicts probabilities, while addressing the difficulties encountered in processing model data that have been well addressed with a small number of innovations

3. This model is widely used, and the field of deep learning has been solved in a wide range at present. The model can be trained dynamically over time, which is more convenient

4. Gray scale prediction model and time series model, in the prediction of a small number of correct reports about bumblebees, the prediction of small samples plays a very good effect

5. Through data enhancement, we obtained millions of data with reasonable label structure, which enabled the model to overcome the difficulty of extremely unreasonable label structure of data set and solve the problem of too few training data

6.2. Weaknesses

1. In the model prediction proof of the fifth question, the effect of gray scale prediction is not so good, because gray scale prediction is not suitable for long-term prediction analysis

2. In the prediction misclassification model, in the sensitivity analysis, when the image probability index is all 0, there are still three samples that are considered to be Asian bumblebee, which has a high influence on the model

3. This paper used Platt scaling to achieve data conversion to probability, but this method could not get the true probability of the label in specific circumstances. The probability index obtained in this paper is proportional to the true probability, but it has little impact on the final result of the model.
References

[1] https://blog.csdn.net/savinger/article/details/89308831
[2] https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[3] (Zeng Shaohua. Research and Application of Support Vector Machine Regression Algorithm [D]. Chongqing University, 2006.).
[4] https://blog.csdn.net/viewcode/article/details/12840405
[5] https://zhuanlan.zhihu.com/p/41679153