ISSN 2616-5775 Vol. 4, Issue 3: 77-82, DOI: 10.25236/AJCIS.2021.040312

Confirming the Buzz about Hornets

Junman Yu, Chenkang Wei, Yue Fang, Rui Kong

School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710600, China

Abstract: *With the occurrence of many Vespa mandarinia sightings, Washington has established various channels to encourage people to report Vespa mandarinia. This paper uses the Time Series to predict the latitude and longitude, and verify the prediction through Neural Network and Linear Regression Fitting the accuracy of the result. The relative error of bringing the prediction data into the Neural Network training model is 8.73%. Using SVM model in this paper, unverified date training and analysis were carried out on the original data, and the analysis results showed that the accuracy was 89.95% and the recall rate was 93%, demonstrating the good matching degree of the model. Finally using a Decision Tree Classification Model calculate the total weight corresponding to each unverified data through PYTHON programming.*

Keywords: *Neural Network, SVM Model, Decision Tree*

1. Introduction

Since the mandarinia wasp footprint was found on Vancouver Island in British Columbia, Canada in 2019 and its nest was quickly destroyed, there have been many confirmed pest sightings and many false sightings in neighboring Washington State. The local people's fear of the wild wasp is related to its characteristics and habits. The mandarinia wasp is the largest wasp species in the world. Its body length can reach 5 to 6 cm, and its head is more than 4 times larger than a normal bee. The presence of wild wasps has caused anxiety in the Washington state government. The local helpline and a website have been established for people to report sightings of these wasps.

2. BP Artificial Neural Network Model

2.1 Establishment of BP Artificial Neural Network Model

The outstanding characteristics and advantages of artificial neural network are mainly shown in the following three aspects:

First, it has a self-learning function. For example, in the realization of image recognition, we only need to input a variety of image templates and the corresponding recognition result set into the neural network in turn, and the network can recognize similar images with the help of self-learning function. Therefore, self-learning function is very important for image prediction. In the future, the artificial neural network computer will provide a variety of forecasts with broad application prospects, such as economic, market and profit forecasts.

Second, it has associative storage function, which can be realized through feedback network.

Third, it can quickly find the optimal solution. Finding the optimal solution of a complex problem usually requires large-scale computation. Using the feedback neural network specially designed for the problem and the high-performance computing power of computer, the optimal solution will be found faster.

So in this question, we build an artificial neural network model to predict latitude under the premise of known longitude, and compare it with the latitude and longitude predicted by the time series model to verify whether the previous model is reasonable and accurate.

First, we divide 14 groups of positive and 2609 negatives into two parts, one for training and one for testing prediction results. In order to make the prediction effect of the model better and the error smaller, 70% of the positive and negative in the sample are taken out, that is, 10 groups of positive and 1826 groups of negative are used for training data. Use detective, longitude, latitude,

ISSN 2616-5775 Vol. 4, Issue 3: 77-82, DOI: 10.25236/AJCIS.2021.040312

and image as input nodes, and labelstatus as output nodes. Through 10 repetitions of training and learning, positive determines the middle layer of the positive neural network. Then use the model to predict the remaining unverified and unprocedd. Through 4 sets of input indicators, the corresponding labelstatus can be predicted. This process is implemented in python.

2.2 Result Analysis

The data trained through the neural network is fitted and compared with the original data in the table, as shown in the following table:

Table 1: Comparison table of original latitude and neural network prediction latitude

ID	latitude	Forecast latitude	ID	latitude	Forecast latitude
1	-122.7009	-122.05558	6	-121.84395	-122.396
2	-122.6613	-122.20143	7	-121.9599	-121.616
3	-122.3544	-122.19667	8	-122.14892	-122.6051
4	-120.8265	-121.28395	9	-122.13832	-122.4621
5	-123.1641	-123.19184	10	-122.34219	-122.5785

It can be seen from the above table that the original latitude is similar to the latitude predicted by the neural network, and the performance of the generated neural network is shown in the figure:

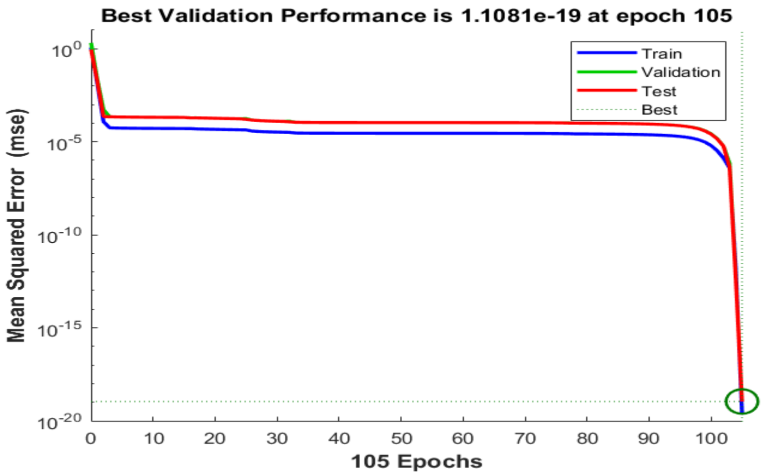

Figure 1: Performance graph of neural network

The abscissa of the figure is the number of training sessions, and the ordinate is the error rate. Among them, the blue Train line is the training error, the green validation line is the verification error, and the red Test line is the test error. It can be clearly seen from the picture that the error rate of the training model does not change much before 100 times. When the number of training times reaches 105 times, the error does not decrease in 6 consecutive tests. Stop training to avoid excessive learning. Prove that the predicted data is highly accurate and the original data is credible.

3. Machine Learning Model

3.1 Image Recognition Technology

The computer image recognition includes the following two main steps: image feature extraction and image classification prediction. Firstly, the input image is preprocessed into a form suitable for feature extraction and image features are extracted. Then the feature images are classified and predicted as shown in the figure

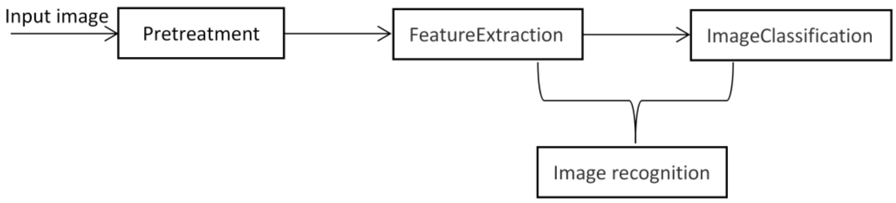

Figure 2: Image recognition process

The image preprocessing module can better extract image features, enhance the target image information and eliminate interference.

The traditional image recognition technology is not suitable for processing a large number of high-resolution images in the era of higher and higher image resolution. It can avoid the complex process of image feature extraction and reconstruction.

3.2 Machine Learning Selection of SVM

SVM (SupportVectorMachine) is a class of generalized linear classifiers that binary classification of data in a supervised learning method. Its decision boundary is the maximum margin hyperplane that is solved by the learning sample. The basic model is defined in the feature space with the largest interval linear model. SVM mainly deals with classification situations, and its significant advantages include:

- The mathematical theory is rigorous and explanatory, which simplifies the usual classification and regression problems;
- Be able to find key samples (ie: support vectors);
- After using nuclear technology, it can deal with regression task and nonlinear classification;

A small number of support vectors determine the final decision function, and the computational complexity depends on the number of support vectors rather than the dimension of the sample space, which avoids the "dimension disaster" in a sense.

In the case of only using the provided data set files and (possibly) provided image files to create, analyze and discuss the model that predicts the possibility of misclassification, the data in the data set need to be effectively classified first, so this article uses the SVM model.

SVM can be divided into the following three forms according to different data:

- Linear support vector machine, also called soft interval support vector machine, when the data is approximately linearly separable, by introducing a relaxation factor, the soft interval is maximized to learn a linear separable model.
- Linearly separable support vector machines, also called hard-interval support vector machines, process data that are linearly separable, and learn a linearly separable model by maximizing the hard interval.
- Non-linear support vector machine, when the data is linearly inseparable, after the data is mapped to a high-dimensional space by introducing a kernel function, a non-linear support vector machine is learned.

Since the features have a certain linear relationship, but they are not completely linear, the first form is selected, namely linear support vector machines.

3.3 Linear Support Vector Machine

When the data is approximately linearly separable, that is to say, there are noise points in the data, we introduce a relaxation factor to make the function interval plus the relaxation factor ξ be greater than or equal to 1, so the constraint condition becomes:

$$y_i(wx_{\mathrm{i}} + b) \geq -\xi_i + 1$$

For each relaxation factor $\boldsymbol{\xi_i}$, a cost needs to be paid, so the objective function, that is, the cost function becomes:

ISSN 2616-5775 Vol. 4, Issue 3: 77-82, DOI: 10.25236/AJCIS.2021.040312

$$\boldsymbol{C}\sum_{i=1}^{N}\boldsymbol{\xi_i}+\frac{\mathbf{1}}{\mathbf{2}}\parallel\boldsymbol{w}\parallel^{\mathbf{2}}$$

Among them, C>0 is called the penalty parameter. It is a hyperparameter that requires us to manually adjust the parameters. The larger the C value, the greater the penalty for misclassification, the narrower the interval width of the support vector machine, and the smaller the C value, the more the penalty for misclassification. Smaller, the wider the interval width of the support vector machine. The geometric meaning of ξ means that the distance between the misclassified data point and the correct classification side is the geometric distance.

Because the data is approximately separable and there are many noise points, when calculating the cost function, these misclassification points should be included in the cost function. The function interval of these misclassified points is $\boldsymbol{y}(\boldsymbol{wx}+\boldsymbol{b}) <= -\mathbf{1}$, so the cost function can be written as an aggregate function with a 0/1 loss function, that is, when the function interval of the data points minus 1 is less than 0 (Misclassification point), the cost function needs to be calculated. If the function interval of the data point minus 1 is greater than 0 (correct classification point), there is no need to calculate the cost function:

$$\min_{w,b}\frac{\mathbf{1}}{\mathbf{2}}\parallel\boldsymbol{w}\parallel^{\mathbf{2}}+\boldsymbol{C}\sum_{i=1}^{m}\boldsymbol{\tau_{0/1}}(\boldsymbol{y_i}(\boldsymbol{w^T x_i}+\boldsymbol{b})-\mathbf{1})$$

But the mathematical properties of the 0/1 loss function are not good, it is non-convex and discontinuous. So generally use his replacement loss function "hinge loss max(0,1-z)" to replace it, then the cost function, which is the objective function, becomes:

$$\begin{aligned}&\min_{w,b,} && \frac{1}{2}\parallel w\parallel^2+C\sum_{i=1}^{m}\xi_i\\ &\text{s.t.} && y_i(wx_i+b)\geq 1-\xi_i, i=1,2,\dots,N\\ & && \xi_i\geq 0, \mathrm{i}=1,2,\dots,N\end{aligned}$$

Therefore, the loss function of SVM can also be regarded as a hinge loss function with L2 regular term. The process of solving w and b is consistent with the linearly separable method, which can be achieved by introducing Lagrange multipliers.

3.4 Results and Analysis

Input the detectiondate, image, longitude, and latitude to get the corresponding analysis result (positive, negative, unverified).

Perform training analysis on unprocessed and unverified in the original data, and get the output result.

Combined with the above analysis, we choose precision rate and recall rate to describe. The accuracy of prediction results is high. It shows how many of the predicted positive samples are really positive samples. There are two possibilities of prediction, one is to predict the positive class TP, the other is to predict the negative class FP, namely

$$\mathbf{P}=\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$$

The recall rate is specific to the original sample and represents the proportion of positive samples correctly predicted in the sample.. There are also two possibilities to predict, one is to predict the positive class as TP, the other is to predict the negative class as FP, namely

$$\mathbf{R}=\frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$$

After calculation, the accuracy rate is 89.95%, and the recall rate is 93%. The data obtained by the model is real and highly reliable.

ISSN 2616-5775 Vol. 4, Issue 3: 77-82, DOI: 10.25236/AJCIS.2021.040312

4. Decision Tree Model

4.1 Establishment of decision tree model

Using training data to train a decision tree, the main idea is as follows, a total of 8 steps, the focus is on recursion:

- Custom information entropy calculation function, used to calculate the information entropy of the data set
- Custom data division function, used to divide the data set according to the specified value of the specified feature
- The self-data set of step2 is used as the function input to step1, and the information entropy H(Di) of the data set divided by a specified value (A=ai) of a specified feature can be calculated, and a specified feature according to a specified feature can be calculated at the same time The sample probability of the data set divided by the value (A=ai)
- Traverse each value of the feature, calculate the information entropy H(Di)of the data set divided under each value and multiply the sample probability $|\boldsymbol{Di}|/|\boldsymbol{D}|$, and then sum to obtain the empirical conditional entropy $\boldsymbol{H}(\boldsymbol{D} \mid \boldsymbol{A})$ of feature A to data set D.
- Calculate the information gain $\boldsymbol{g}(\boldsymbol{D},\boldsymbol{A}) = \boldsymbol{H}(\boldsymbol{D}) - \boldsymbol{H}(\boldsymbol{D} \mid \boldsymbol{A})$ of feature A to the data set
- By analogy, the information gain of each feature to the data set is calculated, and the feature with the largest information gain is taken as the best division feature to obtain the tree T1
- Continue step 3-6 for each node of T1, select the feature with the largest information gain, continue to divide the data, and get a new decision tree
- Until the information gain is less than the threshold, or there is no feature to be divided, or all instances under each branch have the same classification, the decision tree is completed

4.2 Result Analysis

Using time, longitude, and latitude as input training model, the final weight result is obtained:

1) The weight of latitude is 0.58740758;

2) The weight of longitude is 0.13526841;

3) The weight of month is 0.27732401

From this weight, it can be seen that latitude and month have a higher impact on unverified, and longitude has the least impact on unverified.

According to the analysis data, the final weight value of unverified is calculated (full score is 1), sorted in descending order, and the report closer to 1 will be investigated first.

4.3 Accuracy verification

The ROC curve is usually used to judge the pros and cons of a binary classifier. The ROC curve is called the receiver operating characteristic curve, and it can be used to determine the best critical point for the selection of positive indicators for screening tests. Any point on the curve represents a pair of sensitivity and specificity corresponding to the cut-off value of a certain positive result of a certain screening test. Usually the point closest to the upper left corner of the ROC curve is set as the best critical point.

The abscissa of the ROC curve is false positive rate (FPR), that is, the proportion of negative samples that are judged as positive, that is, the false acceptance rate; the ordinate is true positive rate (TPR), the positive samples are judged as positive samples. 1-TPR is the false rejection rate.

This curve reflects the sensitivity of the model. The closer it is to the upper left corner, the higher the sensitivity. The point in the upper left corner, (0,1), means FPR=0, TPR=1, indicating that the false acceptance rate is close to 0. The false rejection rate is close to 0, that is, the proportion of negative samples that are judged as positive is 0, and the negative samples are all judged as negative and the judgment is correct; the proportion of positive samples that are judged

ISSN 2616-5775 Vol. 4, Issue 3: 77-82, DOI: 10.25236/AJCIS.2021.040312

as positive is 1, and the positive the class samples are all judged correct. In summary, this is a perfect classifier, it classifies all samples correctly. The accuracy of this model is high.

5. Advantages and Generalization

Advantages:

①The calculation and solution of the model adopts professional mathematics software with high credibility;

②The various prediction models established are closely related to the actual situation, and the problem is solved in combination with the actual situation. A large number of relevant materials are consulted. Make the model have good versatility and generalization;

③In order to facilitate the description, this article draws a lot of statistical graphs to help understand, and the observability is strong.

④Neural network can map any complex nonlinear relationship, has strong robustness, memory ability, nonlinear mapping ability and strong self-learning ability, and has a broad application market.

Generalization:

This article has established models such as time series model, neural network training model, SVM model, decision tree classification model, scientifically and objectively to help the government interpret the data covered by public reports and provide strategies for prioritizing public reports. The various models established are consistent with the actual situation, the problem is solved in combination with the actual situation, and a large amount of relevant information is consulted, which can make the model have good versatility. It is of reference significance for the analysis of other invasive species in the local area, and provides analysis ideas for the analysis of the invasion of alien species worldwide.

References

[1] Washington state officials locate Asian giant hornet nest [J]. Farm Industry News, 2020.
[2] First Asian giant hornet trapped in Washington [J]. Southeast Farm Press, 2020.
[3] Xia Bing, Research on Evaluation Methods Based on Mathematical Models [M]. Peking University Press, Heilongjiang University Press, 2017: 69-80.
[4] Yang Haimin, Pan Zhisong, Bai Wei. A review of time series forecasting methods [J].Computer Science, 2019, 460128-35.