ISSN: 2664-9640

Another Voice about the Sparrow Wasp

Hongyu Chen[1], Siyuan Wu[2], Haiyang Shang[2]

[1] College of Computer Science and Technology, Nanjing Tech University, Nanjing, Jiangsu, 211816, China

[2] College of Mechanical and Power Engineering, Nanjing Tech University, Nanjing, Jiangsu, 211816, China

Abstract

We build an ARIMA (0, 1, 1) model and an ARIMA (3, 1, 3) model for longitude and latitude. Taking the prediction accuracy as the radius, the prediction accuracy range is established, which plays an important role in determining the priority of the report. On this basis, we build a text analysis model to evaluate the impact of textual information on the probability of misjudgment. Furthermore, we make full use of image information. The gray-level co-occurrence matrix and gray-level analysis are established as the core algorithms, and a machine learning model based on the support vector machine (SVM) model is established on this basis. Finally, we set the weights based on the prediction accuracy of the SVM classification model, integrated the text analysis model with the SVM classification model, and obtained a misclassification probability evaluation model, which comprehensively explained the reported misjudgment probability.

Keywords

ARIMA; support vector machine (SVM); model integration.

1. Introduction

The “Vespa mandarinia” invasion, a sudden and aggressive phenomenon, has a potentially serious impact on local bees and has caused people’s anxiety. Therefore, Washington State has set up a hotline and website for people to report sightings of these wasps. However, because the life cycle and appearance of this wasp are similar to many other wasps [1]. How to identify wild wasps has become a problem. Many misjudgments occurred. As a result, how to prioritize the use of its limited resources for further follow-up investigations has become a major issue.

In this paper, we build two models: ARIMA model and machine learning supporting SVM model. Model 1 was used to predict the distribution location of bees, and Model 2 was used to determine the probability of misjudgment by public reports. Here, we explain the use of the model and give you practical strategies for solving the problem.

The first is strategy one, the strategy for resource allocation. First, you use Model 1 to predict the distribution location of the bee colony on that day, and determine the prediction accuracy range on the map by the model accuracy, and mark the reports falling in these ranges with high priority. Second, you use Model 2 to predict the false conviction rate for the high-priority reports and rank them according to the false conviction rate. Finally, you can determine your inspection strategy based on the sorting results, which is very helpful when you have limited resources.

The second strategy is the strategy for updating the model. Due to the small sample when we build the model, the prediction accuracy is limited to some extent, so you need to update the model constantly to better meet your needs. Finally, strategy three, which is a strategy for determining whether an insect pest is eradicated. First, you should check the top 50% reports of the day by applying Model 1 and Model 2. Second, you should review the locations that have

ISSN: 2664-9640

been treated. From May to December, we recommended the frequency of review at a 3-day interval three months before the review, and from January to April, we recommended the frequency at a 15- day interval three months before the review. If no new confirmed cases are found in the above two steps, we can preliminarily assume that the pest has been eliminated. Finally, you should revisit the site after a year, and if you do not find any new confirmed cases, you can be confident that the pest has been eradicated.

2. The Spread Prediction Model

We use the "Positive ID" set and the information artificially judged to be "Positive" in the "Unprocessed" set as the data set to establish the ARIMA model. The frequency is shown in the Fig.1.

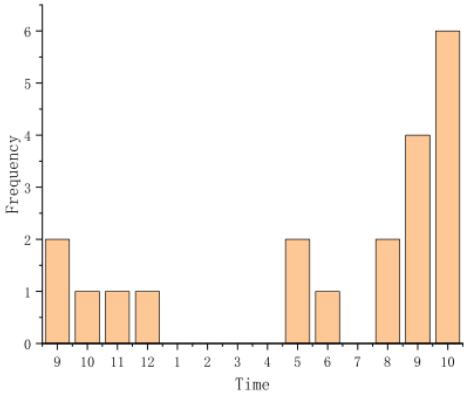

Figure 1 Data frequency distribution

Therefore, we choose the data from May to October 2020 as the data for establishing the ARIMA model. Since the amount of data is small and not in equal time interval, we carry out linear difference on the data to obtain the distribution data in equal time interval.

For ARIMA (p,d,q) model ,the general situation[2] is as following:

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-q} \quad (1)$$

where d represents the difference order of the data.

Now let's determine the parameters of the model. We tried several different model fitting, taking the results of significance t-test of each parameter of the model, the AIC information criterion of the model and the sum of squares of residuals as the judgment criteria. The smaller the P value is, the higher the model accuracy is. The smaller the AIC value is, the higher the model accuracy is. The smaller the sum of squares of residual error, the higher the model accuracy. The significance test, the AIC information criterion and the sum of squares of residuals are shown in table 1.

Table 1 Test of model parameters

	Ar(1)	Ar(2)	Ma(3)	Ma(2)	Ma(3)	AIC	Sum squared resid
ARIMA(3,1,0)	0	0				-9.49997	0.000671
ARIMA(3,1,3)	0	0	0.0001	0	0.0001	-9.57587	0.000599
ARIMA(3,1,1)	0	0.0107		0.0019		-9.53134	0.000642
ARIMA(3,1,2)	0	0.2656		0	0.1083	-9.52932	0.000635
ARIMA(2,1,3)	0.586	0				-9.57554	0.000611
ARIMA(1,1,3)	0					-9.58571	0.000612

ISSN: 2664-9640

Based on the above analysis, we establish the ARIMA (3,1,3) model to describe the longitude sequence. We used the same steps to analyze the longitude sequence and established the ARIMA (0,1,1) model to describe the latitude sequence.

Static prediction [3] refers to rolling forward prediction, that is, for each prediction, replace the predicted value with the real value, add it to the estimated interval, and then make forward prediction (use the actual value of the lag dependent variable instead of the predicted value to calculate the result of one-step-ahead prediction).The prediction results of the model are shown in the Fig.2 and Fig.3.

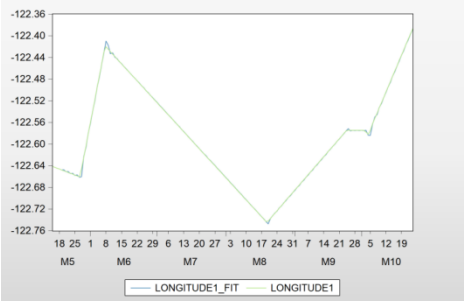

Figure 2 ARIMA(0,1,1) prediction

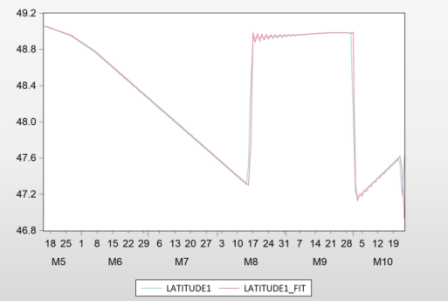

Figure 3 ARIMA(3,1,3) prediction

Among the parameters of model prediction results, "Mean Absolute Error" was selected as the index to evaluate the accuracy of the model. According to the relative scale of the sequence value and the error value, the error is converted, and the more intuitive error index is obtained. The results are shown in table 2.

Table 2 Model accuracy

		Mean Absolute Error	Accuracy
ARIMA(0,1,1)	longitude	0.000866	100m
ARIMA(3,1,1)	latitude	0.028947	300m

3. The Misclassification Probability Assessment Model

In this model, in order to better explain the data provided by the public report, we build text analysis model and support vector machine (SVM) classification model based on text information and image information, respectively. We then fused the models to establish a misclassification probability assessment model to ensure that the information reported by the public is fully utilized. Our work is shown in Fig.4.

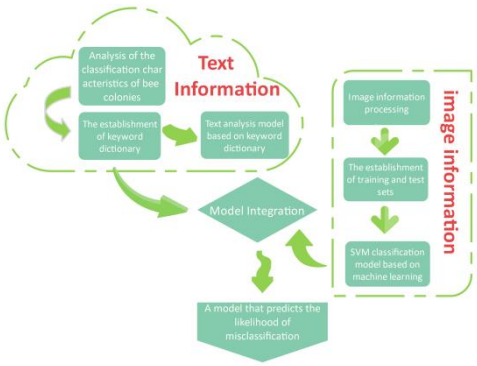

Figure 4 Overview of this model

Through traversal, we get the error probability of all the text words in the comments. We're going to sum it up and give it the value of a new index F, which reflects the likelihood that a new review is an erroneous observation.

ISSN: 2664-9640

$$F = \sum_{i=1}^{t} \sum_{j=1}^{n_i} p_{ij} \quad (2)$$

where $p_{ij} = \beta(w_j, v) \times h$ is the error probability corresponding to the qualified words.

Image feature extraction is the basic work of image recognition and classification, content-based image retrieval, image data mining and so on, among which the texture feature of image is of great significance to describe the image content.In this part, we will build a machine learning model based on support vector classifier to extract image information. The network model is as follows in Fig .5.

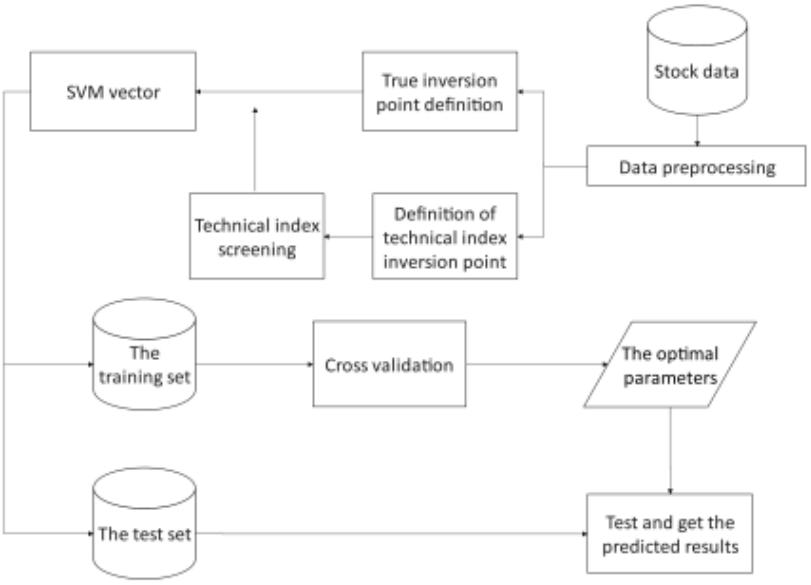

Figure 5 Overview of network structure

Based on the principle of structural risk minimization, support vector machine classification algorithm [4] minimizes the actual risk of machine learning by selecting the appropriate function subset and the discriminant function in the subset. When the sample is linearly separable, the optimal classification hyperplane of the sample should be found. If the sample is linearly unseparable, then relax variables are added. By using nonlinear mapping, the sample from the low-dimensional space is mapped to the high-dimensional attribute space, so as to make the sample linearly separable. The basic principles of the model are illustrated in Fig 6.

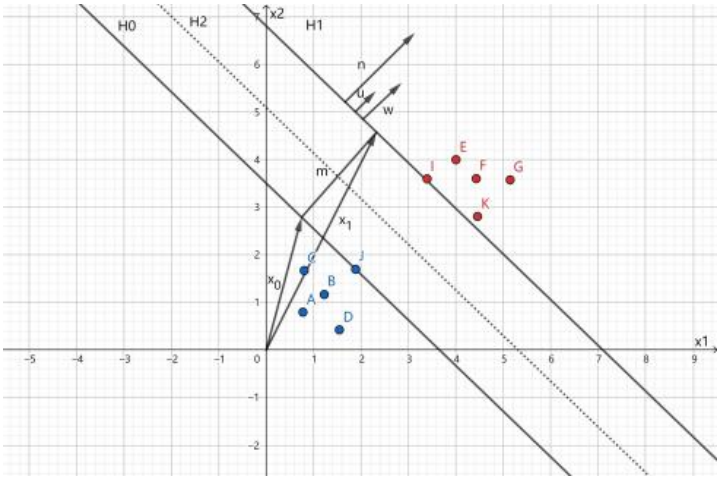

Figure 6 the basic principle of the SVM

Through the training of the samples of the training set, we obtained the SVM model as shown in Fig. 7. And based on the gray level co-occurrence matrix, we obtained the gray level features of the image, as shown in Fig 8.

ISSN: 2664-9640

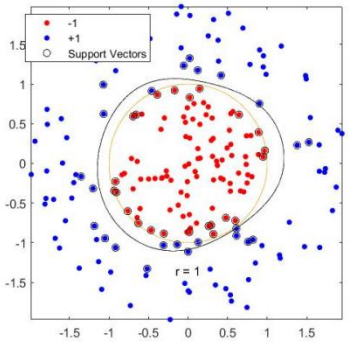

Figure 7 the SVM model after training

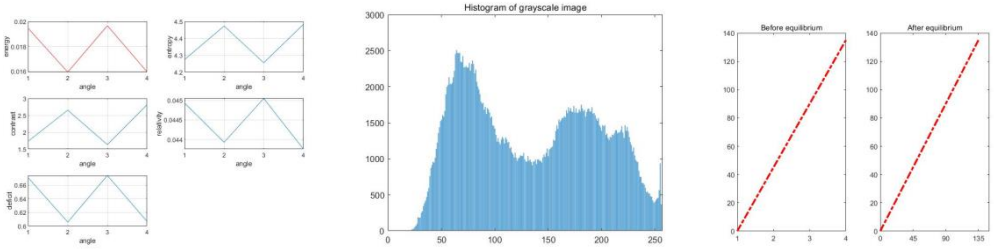

Figure 8 Gray level analysis results

Here, the confusion matrix of dichotomy prediction is established to evaluate the accuracy of the model. The confusion matrix is shown in table 3. At the same time, we define the correct judgment rate α, the error judgment rate γ according to the confusion matrix. After the prediction of the sample of the prediction set, we get that α = 0.5.

Table 3 The confusion matrix

Actual category	Predicted category	
	Positive	Negative
True	TP	FN
False	FP	TN

$$\alpha = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\gamma = \frac{FN+FN}{TP+TN+FP+FN} \tag{4}$$

We take the error judgment rate γ as the weight of SVM classification model, and the correct judgment rate α as the weight of text analysis model, and we establish the error classification probability M, which can be expressed as:

$$M = F\alpha + \theta\gamma \tag{5}$$

4. Model Analysis and Comprehensive Application

We use M as an indicator of the likelihood of the report becoming a positive goal. We rank public reports by M. The larger the value of M is, the more likely the report is to be “Negative ID”, while the smaller the value of M is, the more likely the report is to be “Positive ID”. In particular, when the value of M is equal to 0, the probability of the report being “Positive ID” is very high, and the relevant departments need to mobilize resources in time for confirmation.

We establish the ARIMA models of longitude sequence and latitude sequence respectively, and obtained the prediction accuracy of the models. We can apply the prediction to the distribution position of bee colony, and the prediction result is the longitude value and latitude value of the

ISSN: 2664-9640

predicted point. Taking the prediction point as the center and the prediction accuracy as the range, we establish the prediction accuracy range. Under the existing data set, the accuracy of the model in terms of longitude is 100m and in terms of latitude is 300m. The schematic diagram of prediction accuracy range is shown in Fig.9.

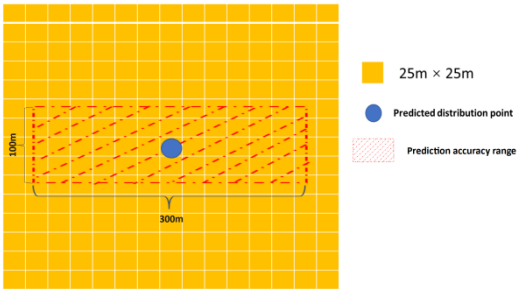

Figure 9 The schematic diagram of prediction accuracy range

In order to improve the quality of the data, we combined the ARIMA model to screen the reported data, and sorted out the data that appeared in the range of prediction accuracy every day. Then, we conduct error analysis on the error classification probability evaluation model. We limited the range of accuracy change to 20% and adjusted the number of SVM model training sets.

We found that when the sample fluctuation of the training set reached 35.7%, the precision change of the SVM model was 21.4%, which exceeded the range of precision change, and the model needed to be updated. Here, we use the weight in model fusion to convert the error analysis results of SVM model to the error analysis of text analysis model.

We conducted sensitivity analysis on the SVM classification model with the following steps: Firstly, we quantified the model, and meanwhile conducted gray analysis on the “Positive ID” and “Unprocessed” samples, and obtained the mean and standard deviation of their five attributes, such as moment of inertia, inverse difference and entropy. Then we apply the limit mean value theorem twice to get the standard error L for each index. We compare the standard error L with the errors obtained by the SVM model, and make judgments in multiple precisions to obtain multiple arrays based on multiple precisions. Based on the weights and values of various attributes, the score for each “Unprocessed” report was obtained and a score matrix, S_6, was established, which was again compared with SVM-based errors to obtain the final error and sensitivity.

5. Conclusion

In this paper, in order to fully explain the data provided by the public report, we set up a time series model using time and latitude and longitude data, a text analysis model using text information, and a SVM classification model using image information, and make a reasonable fusion of the models to make the model more explanatory.

In view of the limited resources of government agencies, through the comprehensive analysis and application of several models, we give practical strategies to help the government use the limited resources to solve the problem to the maximum extent.

References

[1] Arca, M., F. Mougel, T. Guillemaud, S. Dupas, Q. Rome, A. Perrard, F. Muller, A. Fossoud, C. Capdevielle-Dulac, M. Torres-Leguizamon, X. X. Chen, J. L. Tan, C. Jung, C. Villemant, G. Arnold, and J. F. Silvain.

ISSN: 2664-9640

2015. Reconstructing the invasion and the demographic history of the yellow-legged hornet, Vespa velutina, in Europe. Biological Invasions 17(8):2357-2371.
[2] Pal, Dr. Avishek, and Prakash, Dr. PKS. Practical Time Series Analysis. Birmingham: Packt Publishing, Limited, 2017. ProQuest Ebook Central.
[3] Wikipedia. Branchpredictor, 2021. https : //en.wikipedia.org/wiki/Branchpredictor.
[4] Zhang Shuya, Zhao Yiming, Li Junli. Image classification algorithm and implementation based on SVM [J]. Computer Engineering and Applications, 2007, 43(25):40-40.