Journal of Electronics and Information Science (2021) 6: 62-66
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2021.61010
ISSN 2371-9524

Vespa mandarina diffusion model based on AdaBoost and CNN

Zerong Wang, Yiran Wang

Leicester international Institute, Dalian University of Technology, Dalian 116000, China

Keywords: AdaBoost, CNN, Bagging, SVM

Abstract: In this paper, we build two different models at the same time. The first model is a classification model based on Adaboost, and the second model is an image recognition model based on CNN, which takes the pictures in 2021mcm _ problem _ files as input and classifies them. According to the classification results of the two models, we can find the accuracy of the Adaboost model is 0.57 and the accuracy of the CNN model is 0.971, which is found on the test data.

1. Introduction

Vespa mandarina is the world's largest wasp species and is considered to be a predator of bees. A small amount of wasp will destroy the entire European bee colony in a short time. In September 2019, the Vespa mandarina colony was found and killed on Vancouver Island, British Columbia; subsequently, the Washington State Department of Agriculture confirmed that a specimen of the dead wasp was found in Washington [1].

We need to build two different models at the same time. The first model is a classification model based on Adaboost, and the second model is an image recognition model based on CNN, which takes the pictures in 2021mcm _ problem _ files as input and classifies them.

2. Model Establishment and Solution

2.1 Introduction of Bagging model

Bagging is a technique to reduce generalization error by combining several models. The main idea is to train several different models separately, and then let all the models vote on the output of test samples. This is an example of a conventional strategy in machine learning, which is called model averaging. Model averaging is a very powerful and reliable method to reduce generalization error [2].

2.2 Introduction of KNN model

The idea of KNN is very simple and intuitive: if most of the k most similar samples in feature space belong to a certain category, then the sample also belongs to this category. In the classification decision, this method only determines the category of the samples to be classified according to the category of one or several nearest samples [3].

2.3 Introduction of CNN model

Convolutional neural networks (CNN) is a kind of Feedforward Neural Networks with convolution computation and deep structure, which is one of the representative algorithms of deep learning. Convolutional neural network has the ability of representation learning, and can classify the input information according to its hierarchical structure, so it is also called "Shift-Invariant Artificial Neural Networks (SIANN)" [4].

2.4 Model solution

After data cleaning, including removing blank values and standardizing, we keep 2080 groups, in which the category that has been identified as Asian bumblebee is set as 1, the category that has not been identified as Asian bumblebee is set as 0, and the unidentified individual is directly removed.

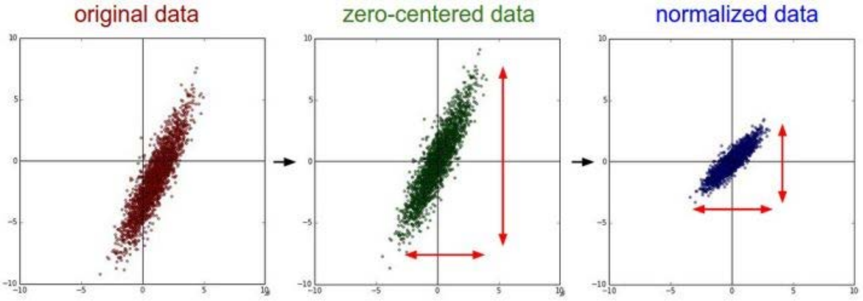


Figure 1: Standardizing

We then select longitude, latitude and time as feature input models, and take whether it is Asian bumblebee as output result. First of all, we can observe the three-dimensional distribution of data. We take 1867 groups of data as training sets and the remaining 213 groups as test sets for training. The specific distribution is as follows:

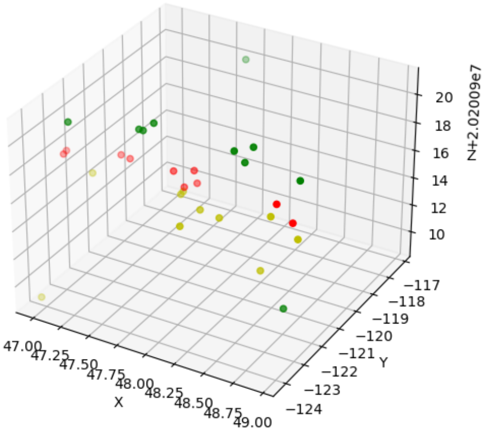


Figure 2: Training set distribution

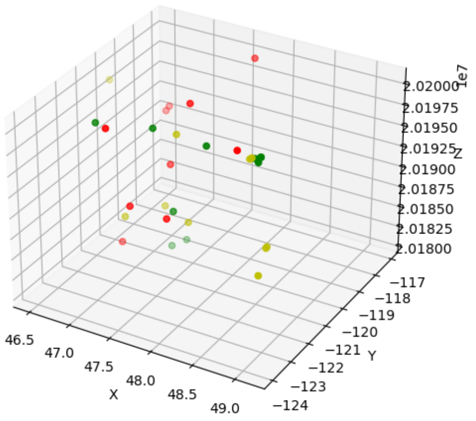
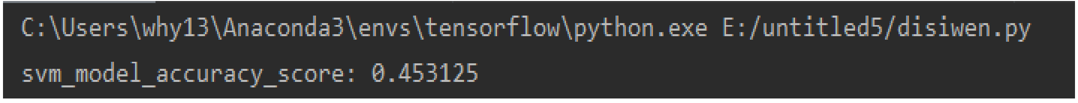


Figure 3: Test set distribution

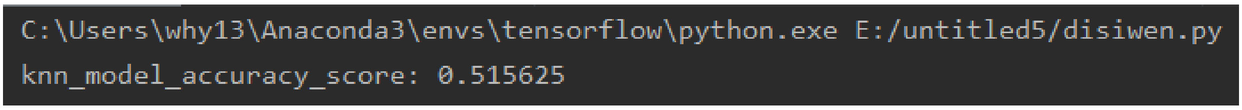
We use python software to run SVM model separately:



```
C:\Users\why13\Anaconda3\envs\tensorflow\python.exe E:/untitled5/disiwen.py
svm_model_accuracy_score: 0.453125
```

Figure 4: SVM accuracy

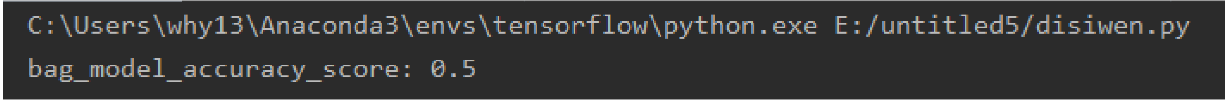
The accuracy rate is 0.453125, which is not up to 0.5, and the performance is not good. Then we use KNN model:



```
C:\Users\why13\Anaconda3\envs\tensorflow\python.exe E:/untitled5/disiwen.py
knn_model_accuracy_score: 0.515625
```

Figure 5: KNN accuracy

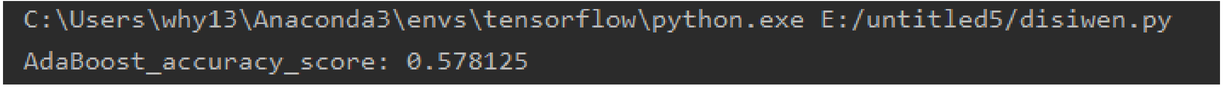
The accuracy rate is 0.515625, which is greatly improved compared with the previous SVM model. We then tested the Bagging model:



```
C:\Users\why13\Anaconda3\envs\tensorflow\python.exe E:/untitled5/disiwen.py
bag_model_accuracy_score: 0.5
```

Figure 6: Bagging accuracy

It is found that the correct rate is 0.5. Finally, we tested the AdaBoost model:

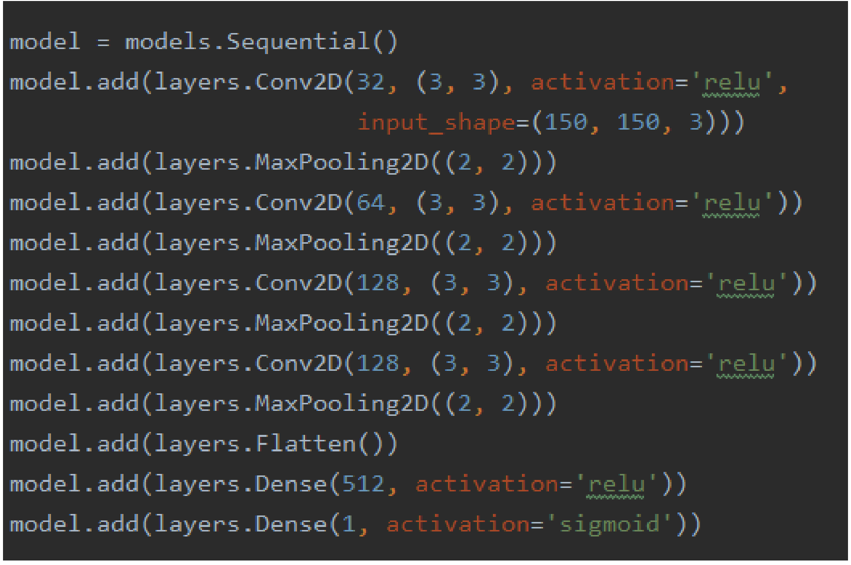


```
C:\Users\why13\Anaconda3\envs\tensorflow\python.exe E:/untitled5/disiwen.py
AdaBoost_accuracy_score: 0.578125
```

Figure 7: Adaboost accuracy

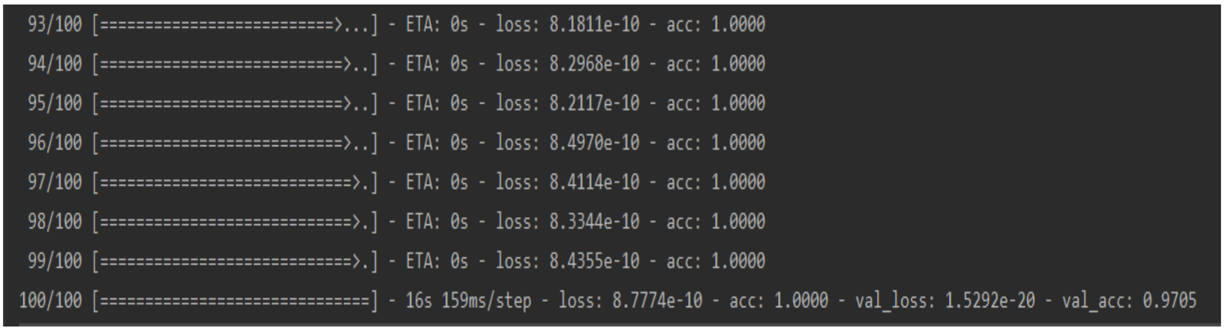
It is found that the correct rate is 0.578125 after many experiments. Finally, we choose Adaboost model as our model, which can be used to identify the probability of mistaking Asian bumblebee.

Secondly, according to the pictures in 2021 MCM _ problem _ files as input, we train a CNN model to recognize the Asian bumblebee pictures. The specific convolution layer of CNN model is as follows. We run 30 epoches to ensure the relative accuracy of the results.



```
model = models.Sequential()
model.add(layers.Conv2D(32, (3, 3), activation='relu',
                        input_shape=(150, 150, 3)))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D(64, (3, 3), activation='relu'))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D(128, (3, 3), activation='relu'))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Conv2D(128, (3, 3), activation='relu'))
model.add(layers.MaxPooling2D((2, 2)))
model.add(layers.Flatten())
model.add(layers.Dense(512, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
```

Figure 8: CNN structure



```
 93/100 [=========================>...] - ETA: 0s - loss: 8.1811e-10 - acc: 1.0000
 94/100 [==========================>..] - ETA: 0s - loss: 8.2968e-10 - acc: 1.0000
 95/100 [==========================>..] - ETA: 0s - loss: 8.2117e-10 - acc: 1.0000
 96/100 [==========================>..] - ETA: 0s - loss: 8.4970e-10 - acc: 1.0000
 97/100 [===========================>.] - ETA: 0s - loss: 8.4114e-10 - acc: 1.0000
 98/100 [===========================>.] - ETA: 0s - loss: 8.3344e-10 - acc: 1.0000
 99/100 [===========================>.] - ETA: 0s - loss: 8.4355e-10 - acc: 1.0000
100/100 [=============================] - 16s 159ms/step - loss: 8.7774e-10 - acc: 1.0000 - val_loss: 1.5292e-20 - val_acc: 0.9705
```

Figure 9: CNN accuracy

We can find that the accuracy of CNN model is 1.0000 in the training set and 0.9705 in the test set, which shows that the accuracy of the model is good. Finally, we will measure the probability of wrong classification according to the common prediction results of the two models.

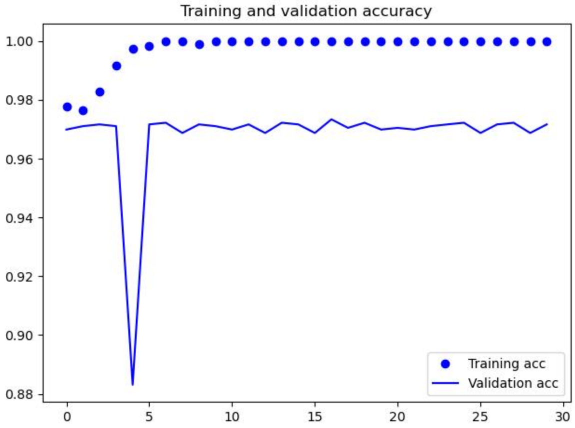


Figure 10: CNN accuracy

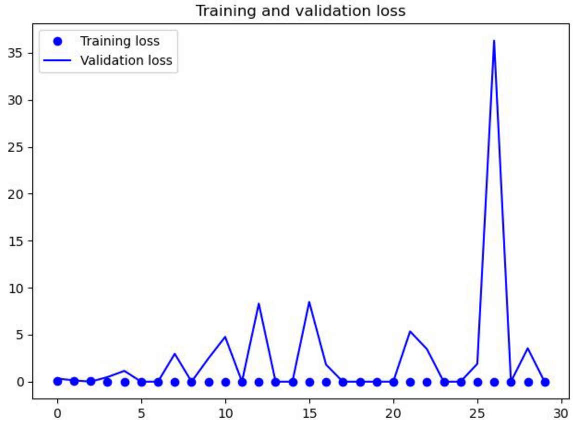


Figure 11: CNN loss

3. Model evaluation and promotion Strengths

On the basis of certain data, through the collection of Vespa Mandarina data, we can predict by computer, and get a set of models that meet the corresponding characteristics. After verification, it is enough to ensure its accuracy. On the other hands, we use more than one analysis method in the mathematical model, and we confirm each other to ensure that it meets the corresponding practical requirements of the given data in multiple dimensions. However, the data used in the mathematical model is still insufficient, and it is unable to select enough data for computer training, resulting in the lack of accuracy and precision of the model.

In this paper, we use python software to process data to determine the threshold and find outliers, which is still applicable to other mathematical problems and general models. According to the data, we have established a Adaboost model, and used big data to ensure its adaptability, which is universal and more suitable for the requirements of modern life [5].

References

[1] First Asian Giant Hornet Trapped in Washington. Western Farm Press, 2020.

[2] Xiaoming Xu. SVM parameter optimization and its application in classification [D]. Dalian Maritime University, 2014

[3] Federico Magliani,Andrea Prati. LSH kNN graph for diffusion on image retrieval [J]. Information Retrieval Journal,

2021 (prepublish).

[4] Feiyan Zhou, Linpeng Jin, Jun Dong. a review of convolutional neural networks [J]. Acta Sinica Sinica, 2017, 40 (06): 1229-1251

[5] Li Yuan, Geng zewei. Fault detection based on K-means clustering and local outlier algorithm [J]. Chemical automation and instrumentation, 2019, 46 (10): 816-821