Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2103686

Analysis of confirming the real hornet and strategies

Summary

After the first detection of the Asian giant hornet, several confirmed sightings of the pest have occurred in neighboring Washington State, as well as a multitude of mistaken sightings. Although the number of reports is amazingly large, the ratio of mistaken is also high. With our exploitation of the data provided to us to describe the report submitted previously, we find the number positive reports is only 14 in all 4440 reports that is 0.3% report is real hornet. So, we analyze and build a model to predict and prioritize these reports for efficient usage of resource from government.

Firstly, we build a spatial-temporal model to predict the spread of the pests. After we clear the initial data to remove some invalid value, we focus on 14 positive report and get the range of Latitude and Longitude. Considering the range of nest is 30 km, we divide the whole area into pieces. By the referred growth-diffusion model for the insect movement, we use least square method to get the spatial-temporal function is $N(d,t) = 0.9 + 2e^{-10} * \mathrm{d}^{9.5} * \mathrm{t}^{3.24}$ which reflects the spread of pests.

Secondly, we use the SVM method to create a model that predicts the likelihood of a mistaken classification. We use the method called one-hot encoding to process the txt data and use image cropping to get the main part, then we extract the RGB value and ambiguous value of each picture. Using above characters and the detection date, Latitude, Longitude as input layer with the lab status as output layer, then we train a SVM discriminator to output whether the report is positive ID or not. For the imbalance data set, we propose a new way to test our model which just test in minority. Finally, after we take the picture into consideration, the accuracy of our model can reach surprisingly 100%.

Thirdly, we use PCA and k-means clustering to build our model to prioritize the predicted positive report. PCA is used to lower the dimension of variables from reports. Then by use the Silhouette coefficient to confirm the number of Clustering centers, we use k means to get centers and use the minimum distance as the score to prioritize the positive reports.

Fourthly, to confirm the interval of updating, we use all submit date from positive reports and get the delta-time interval $\Delta \mathrm{t}_1$ and find most interval is less than 40 days. Then we can choose 40 days to ensure that there will be more likely new positive reports to update our models. For updating, we should remove some ages-ago reports and add new reports to re-train the SVM

and get new centers.

Fifthly, we use all reports to calculate the delta submitting time Δt_2 for adjacent reports and find that the exponent distribution can preferably fit the data. Then we use the exponent distribution to fitting it and get the top 99% percent point 0.38 as the crucial point. In order to eliminate the chance, we set up a continuous month, daily report $R_i < 2.6$ copies to eliminate the pest.

Finally, we summarize above models and take the background of agriculture in American into consideration. Then we prepare a Memorandum for the Washington State Department of Agriculture which contains some proposed strategies.

Keywords: Grid method, insect growth-diffusion model, one-hot encoding, picture processing, SVM, PCA, k-means clustering, exponent distribution

Contents

1 Introduction

1.1 Problem Background

Pest control is related to agricultural development and ecological balance, which is a key concern of agricultural departments. The Asian giant wasp is highly aggressive. Not only does it have the characteristics of stinging humans, but also prey on other bees, posing a potential threat to the local bee population, so it is considered as an agricultural pest.

After the discovery and destruction of the Asian giant wasp on Canada's Vancouver Island in September 2019, there have also been several sightings of the pest in the neighboring state of Washington. In order to effectively control and destroy this threatening Asian giant wasp, Washington State encourages people to provide information about the wasp through hotlines and websites.

However, a significant number of sightings turn out to be misjudge. How to interpret the data provided by public reports and optimize the use of limited resources to follow up the survey? At the request of the Washington State Department of Agriculture, we will conduct a modeling analysis to propose solutions.

1.2 Our work

We do such things ...

- Preprocess the given data provided by people to discuss whether the transmission patterns can be determined.
- Based on the data set and image information to develop and analyze the misclassification model.
- Discuss why our model promote the likelihood of positive data of the prior investigation.
- According to the analysis above, clarify how to improve the model if more data is given as time goes by.
- Explain what confirms the pest has been removed from Washington State?

2 Preparation of the Models

2.1 Assumptions

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1 Notations

Symbol	Definition
N	the population of the Asian giant wasp
d	diffusion distance in one direction
D	diffusion coefficient
t	the spread time of pests
md	minimum distance to two clustering centers
Δt_1	time interval of adjacent positive report
Δt_2	time interval of adjacent report
R_i	number of daily reports

3 The Models

3.1 Model 1: Population Growth-Dispersion Model

3.1.1 Details about Model 1

In order to study the spatiotemporal variation of Asiatic giant wasp, a unified model of insect population growth and dispersal was established, which could satisfy both growth and dispersal laws.

By referring to the paper, we know that in theory the growth-diffusion model should have the following form:

$$\frac{\partial N}{\partial t} = D\frac{\partial^2 N}{\partial x^2} + \varphi(N, x, t) \tag{1}$$

Based on the theoretical model, taking the logarithmic diffusion model as the diffusion condition and Weibull growth process as the growth condition.

Logarithmic diffusion model:

$$\frac{\partial N}{\partial x} = a + bx \tag{2}$$

Weibull growth process:

$$N(a + bt^c) \tag{3}$$

The growth-diffusion model was obtained as followed:

$$N(x,t) = A + Bx^{\alpha}t^{\beta} \tag{4}$$

3.1.2 Solution and Results

Firstly, we find that data related to pest sightings were divided into Positive ID, Negative ID, Unverified and Unprocessed. After screening, we found 14 positive identities.

Then, we preprocessed the longitude, latitude, number and time of each of the 14 data.

For longitude and latitude, we use Subdivision mesh method. One wasp queen is known to nest over a range of about 30 kilometers. And we can calculate the distance per unit of latitude and longitude. We know that the latitude varies by one degree and the distance varies by 110km, so 0.4 degrees longitude is equal to 30 kilometers. The relationship between longitude and distance differs from latitude in that it is multiplied by the cosine of latitude (48°), so 0.25 degrees latitude is equal to 30 kilometers.

Then we start at (-123.9431,48.7775),30 kilometers are a unit, setting up a coordinate system. We get the coordinates of 14 data points.

For number, we count the number of each report as 1. If we find a report with the same time and place, we will add 1 on the original basis.

For time, we start with September, one unit per month. We get the simplified time of 14 data points.

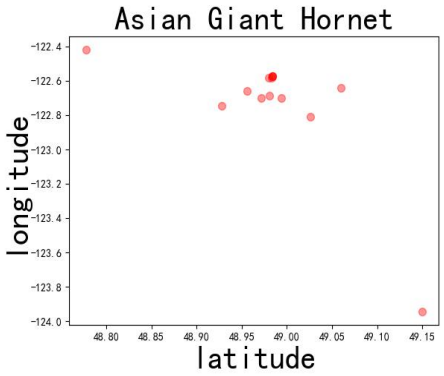

Figure 1 Visualize positive ID

Table 2 Preprocess positive ID

t	x	y	number	d
0	0	1	1	1.487576002
0	3	1	1	3.220638378
1	3	1	1	3.201321641
2	3	1	1	3.000319794
3	3	1	1	3.24039228
8	3	1	2	3.444554037
9	4	0	1	3.811220002
11	3	1	1	3.0547298
12	3	1	4	3.519294154
13	3	1	1	3.499849964

We made these 10 groups of data into a three-dimensional image, and it can be found that with the passage of time, the number of the Asian giant hornet population is increasing, and the population distribution is more and more diffuse.

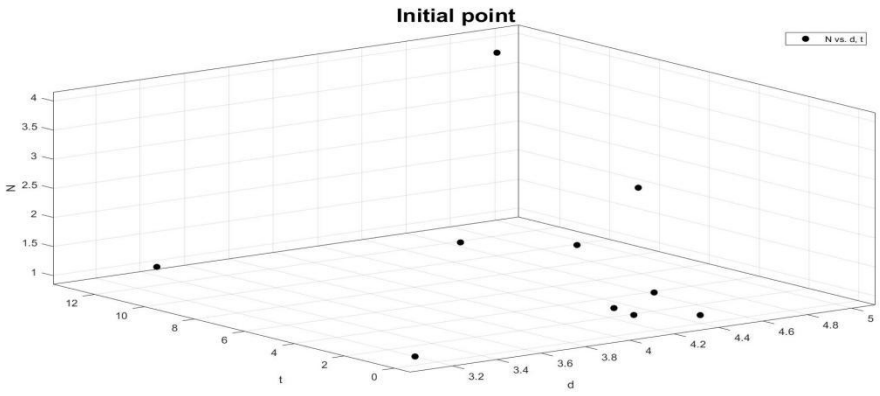

Figure 2 Initial points

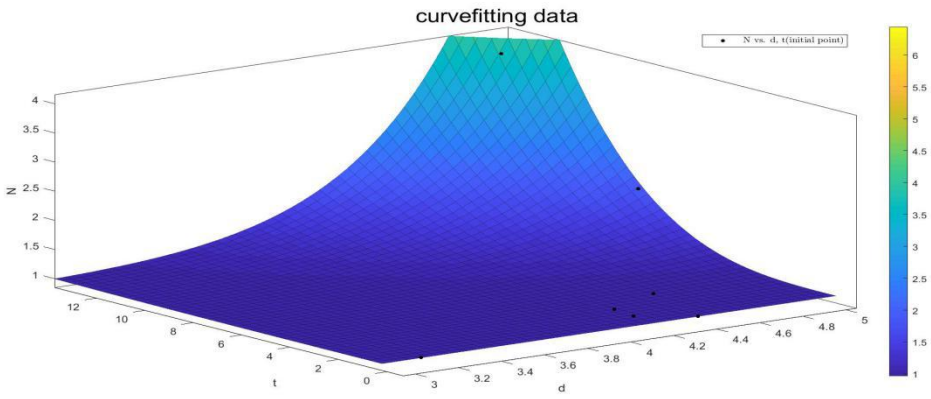

Figure 3 Curve-fitting data

After that, we substituted 14 sets of data into the insect population growth-dispersal model.

And we use fit computation to figure out the unknown. (Fit computation did not converge: Fitting stopped because the number of iterations or function evaluations exceeded the specified maximum.)

Fit found when optimization terminated:

General model:

$$f(x,y)=a+b\cdot x^{c}\cdot y^{d} \tag{5}$$

Coefficients (with 95% confidence bounds):

$$\begin{aligned} &a=0.9664\ \ (0.3041,1.629)\\ &b=2.088\cdot10^{-10}\ \ (-8.579\cdot10^{-9},8.996\cdot10^{-9})\\ &\alpha=9.576\ \ (-13.28,32.43)\\ &\beta=3.24\ \ (-1.262,7.742) \end{aligned} \tag{6}$$

Goodness of fit:

$$\begin{aligned} &\text{SSE}: 2.308 \\ &\text{R-square}: 0.7252 \\ &\text{Adjusted R-square}: 0.5878 \\ &\text{RMSE}: 0.6203 \end{aligned} \tag{7}$$

Finally, the spatial-temporal function is $N(d,t) = 0.9 + 2e^{-10} * d^{9.5} * t^{3.24}$ which reflect the spread of pest.

Also, the fitting error is calculated, and it is found that the geometric error of most data is very low.

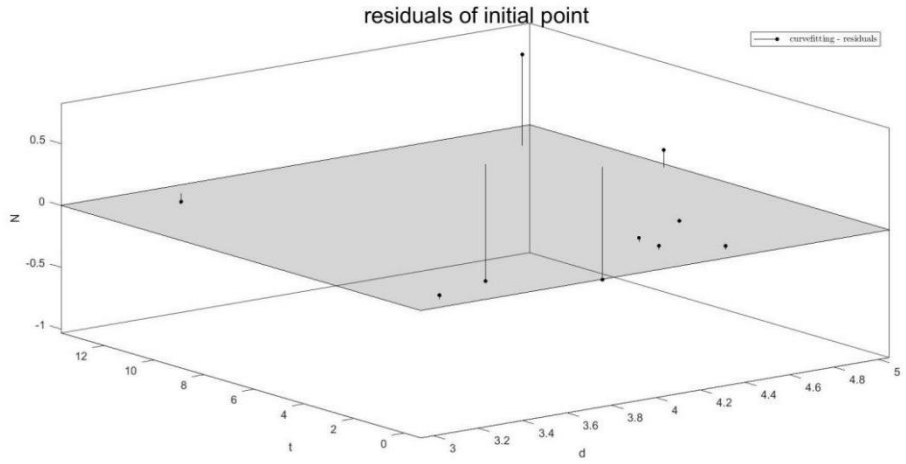

Figure 4 Residuals of initial point

3.2 Model 2: Recognition and Classification Model

3.2.1 Using Data Set Based on Text

(a) Data Cleaning

The data set file includes data like date, notes, lab comments and so on. But part of the data is invalid.

Data has timeliness, so we only select the data of "time" from 2019 to 2020 for research, and delete the data that is not in this time interval.

Table 3 Invalid Detection Date

Detection Date
<Null>
12/30/1899
12/30/1899
4/21/1600
12/30/1899
……

(a) Null and format mistake

Detection Date
2016/6/20
2018/7/1
2016/8/25
2018/7/20
2018/6/22
……

(b) ages ago

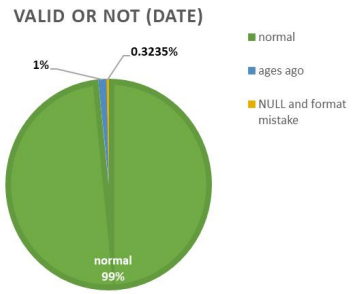

Figure 5 The proportion of valid value(date)

Notes are an eyewitness's description and assessment of a witnessed event, often a descriptive or emotional sentence. But some notes in the data set fail to fit this feature, even containing dates.

Table 4 Invalid Notes

Notes
10:45 AM
99115
44050
44065
(null)
……

Table 5 Valid Notes

Notes
One dead wasp seen in Blaine, and suspect flying nearby
Hornet specimen sent to WSU
We found about a dozen of these bees in our backfield last October while we were storing bee boxes, Sorry no pictures
……

Take Detection Date, Notes, Latitude and Longitude as the input and Lab Status as the output for training. The data set was divided into training set and prediction set to verify the possibility of prediction error, namely the accuracy of the model. Since classification prediction is involved, SVM is used to solve the problem.

The "date" attribute is represented by four parameters: “day”, “month”, “weekday” and “is weekend”.

Table 6 Denote date by four attributes

Notes	Lab Status	Latitude	Longitude	day	month	weekday	is weekend
You can call me on…	Negative ID	47.3074	-122.23041	20	5	3	0
While walking I noticed…	Negative ID	47.737396	-122.170982	14	6	7	1
WSDA submitted for…	Positive ID	48.98422	-122.574726	28	9	1	0
Working in the back..	Negative ID	47.15096	-122.082851	29	7	3	0
Wondering if it could…	Negative ID	48.846752	-122.609404	10	7	5	0

(is weekend: 0 denotes NO, 1 denotes Yes)

Text is a kind of sequential data. We use the method called one-hot encoding to process the txt data.

A sentence can be seen as a sequence of characters or words. We split each note into several words and count the high frequency words of them among positive ID.

It is worth noting that function words such as "of" and "to" are often used in daily expressions, but it is not necessary to include them in the word frequency statistics of texts, so the words we count are those content words with emotional colors.

Table 7 Top10 high-frequency words

hornet	at	one	an	if
seen	like	with	We	sure

Having got the value of each parameter, we normalize them, and then SVM is used for creation and training. The method used here is cross-validation. Choose a variety of equidistant C and G training to find the most suitable C and G. But the training was long so we choose to skip this step, input parameter directly.

$$cmd = '-t\ 2\ -c\ 42.2243\ -g\ 2.639' \quad (8)$$

(b) Data Process

After the data cleaning, 2,069 pieces of data left. For error analysis, 1500 items were selected as the training set and the remaining 569 items as the test set. The results are shown as followed:

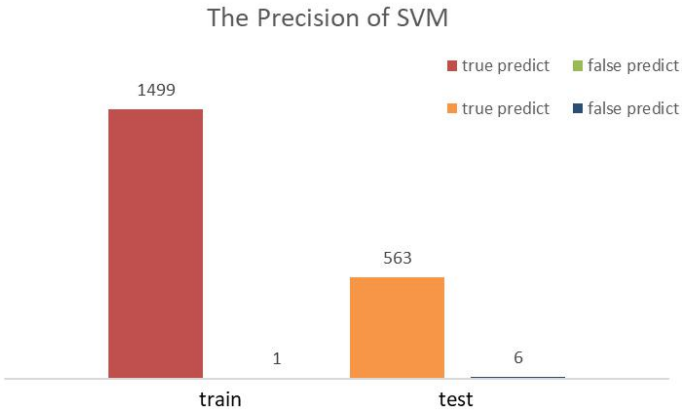

Figure 6 Error Analysis

The results of SVM simulation test are:

$$\begin{aligned} &\text{Accuracy} = 99.9333\%\ (=1499/1500)\ \ \text{(classification)} \\ &\text{Accuracy} = 98.9455\%\ (=563/569)\ \text{(classification)} \end{aligned} \quad (8)$$

Matlab gives a comparison of the predicted values of the test set in SVM:

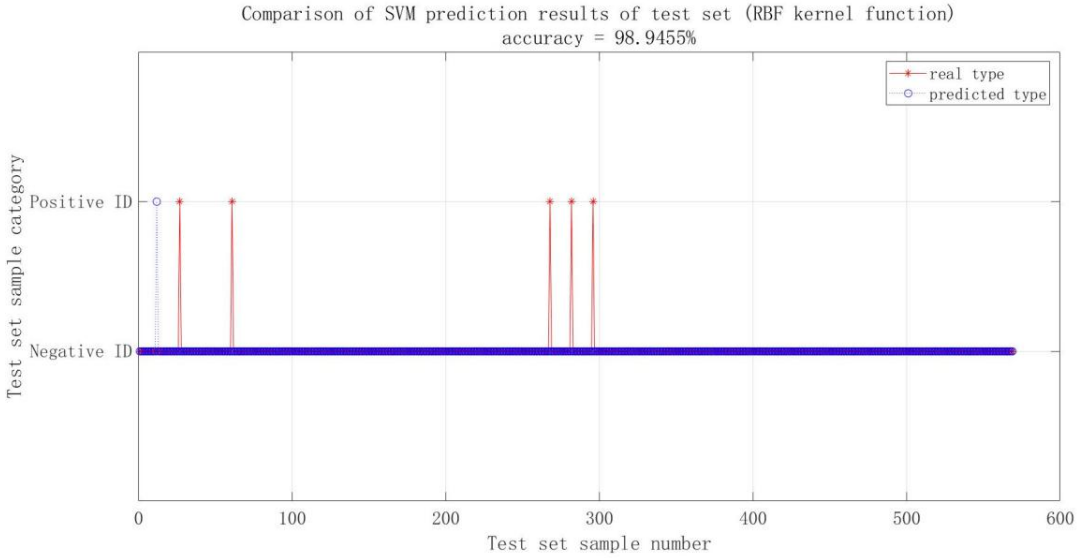

Figure 7 Comparison of real and predicted type (the overall sample)

Accuracy seems high. But our data are lopsided. The number of positive and negative ID varies greatly, and the data set is inbalanced. In order to make the two kinds of data equal in quantity, we usually copy or synthesize the samples of small number to make the quantity of them close to the majority. Now we extract the 14 positive IDs and test them.

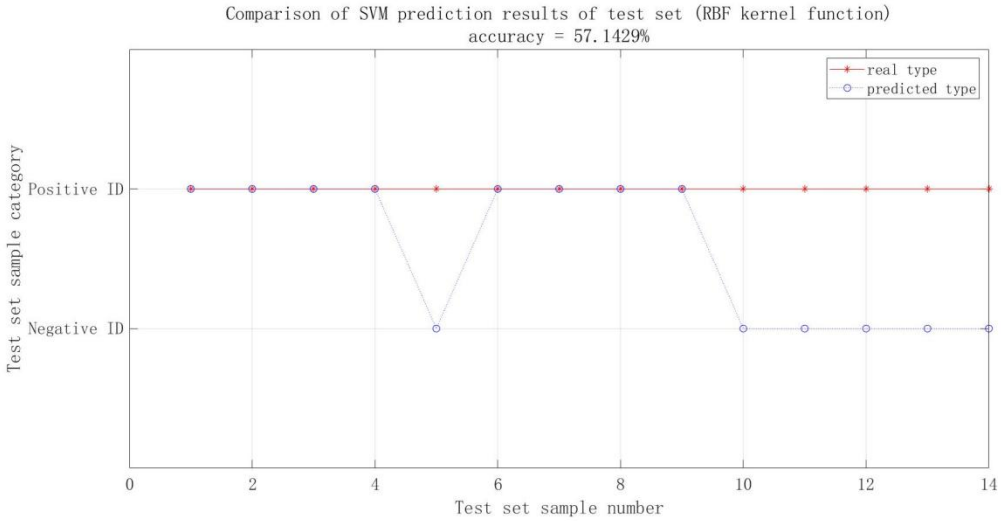

Figure 8 Comparison of real and predicted type (the positive sample)

3.2.2 Using Image Files Based on Picture

The computer cannot recognize a piece of text directly, so we count the frequency of word to process the data. Similarly, in order for the computer to read the information in the picture, we try to grasp the digital features of the pixels.

Each piece of data relates to one Global ID. During the text processing, the irregular data of time and notes in the DataSet have already been deleted. Then according to the table file Images_by_globalID, we delete those data whose corresponding file type is not image.

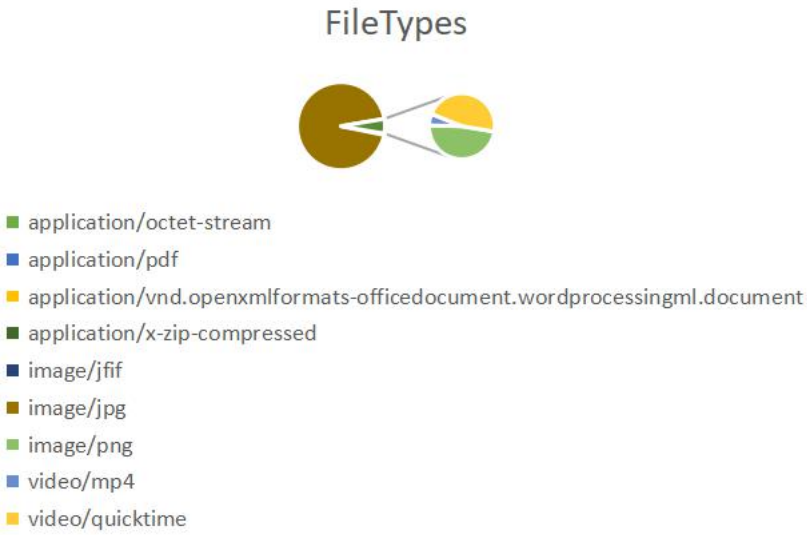

Figure 9 File's types

Due to the existence of the background, insect makes up only a small proportion of some picture. Cropping out the insect in the image, which will remove most of the irrelevant background, could help to improve the recognition rate. We use Python to follow the process below for image cropping.

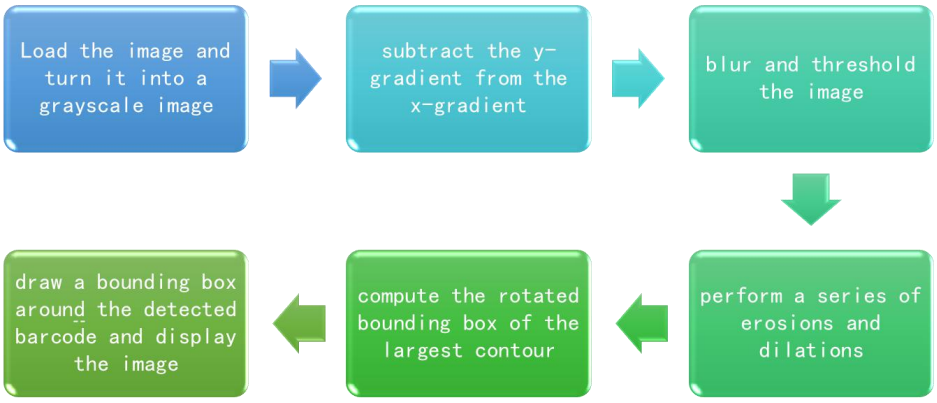

Figure 10 The flow chart of cropping

We can get something like this:

Figure 11 The original picture

Figure 12 The processed picture

Crop according to the green area and calculate the RGB value of the cropped image. Combined with the two indexes of RGB and “ambigious”, the image recognition and classification prediction are carried out.

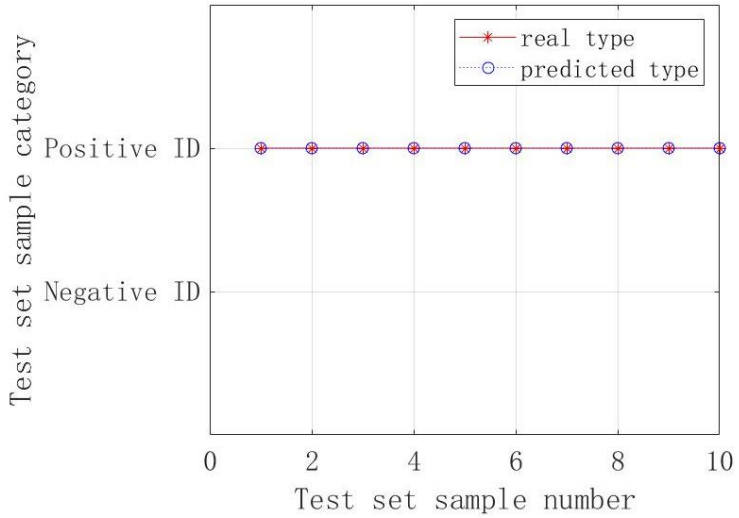

Figure 13 Comparison of SVM prediction results of test set (RBF kernels)

Dramatically, both train set (1914 samples) and test set (10 samples) have an accuracy of 100%.

3.3 Principal Components Analysis and K-means Clustering

By the previous model, we can predict the likelihood of a mistaken classification. But the resource of government is limited, we should build a model to prioritize the report predicted to be positive ID to make better use of the resource from government.

3.3.1 Model of Principal Components Analysis

Through model 2, every report has 10 variables. If we desire to calculate the priority of the report, we need to using PCA, to simplify too many variables and discard factors that have little influence on the report. we decided to reduce the variables to three.

We set up the matrix with 10 pieces of 20-dimensional data, 10 samples.

$$\mathrm{x} = (x_1^*, x_2^*, \ldots\ldots, x_{20}^*) \quad (9)$$

Firstly, we preprocessed the data. We arrange the data variables in columns, i.e., one sample per behavior. And then we take the average of the columns, the average of the features.

Then, we calculated the eigenvalues, corresponding eigenvectors of covariance matrix and covariance matrix from the variables.

After that, we arranged the eigenvectors into matrices according to their corresponding eigenvalues in rows from top to bottom, and take the first three rows to form matrix A (with 3*10 dimensions):

$$\mathrm{A} = \begin{bmatrix} [0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0], \\ [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0], \\ [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1] \end{bmatrix} \quad (10)$$

By observing the eigenvector, the first principal component is found to be related to longitude, so the first principal component is called longitude component.

The second principal component is related to the green color in the image, which is called the green channel component.

The third principal component is related to the blue color in the image and is called the blue channel component.

Then, we need to reduce the dimensions of the matrix A to three-dimensions.

$$Y = A^T X \quad (11)$$

And calculate the principal component score for each sample:

$$\begin{bmatrix} [-123.943134, 63.82378203, 44.13805117], \\ [-122.581335, 138.38446567, 109.75395567], \\ [-122.418612, 144.39688619, 120.20634567], \\ [-122.745016, 117.31779441, 124.15313847], \\ [-122.641648, 132.49061196, 124.34322941], \\ [-122.700941, 157.60111866, 139.99316194], \\ [-122.582465, 154.94646534, 146.31122309], \\ [-122.661037, 143.86489732, 146.82921424], \\ [-122.702242, 180.11293896, 188.44429822], \\ [-122.688503, 113.4677369, 88.62933482] \end{bmatrix} \quad (12)$$

In addition, we draw a three-dimensional scatter plot according to the principal component scores of 10 samples:

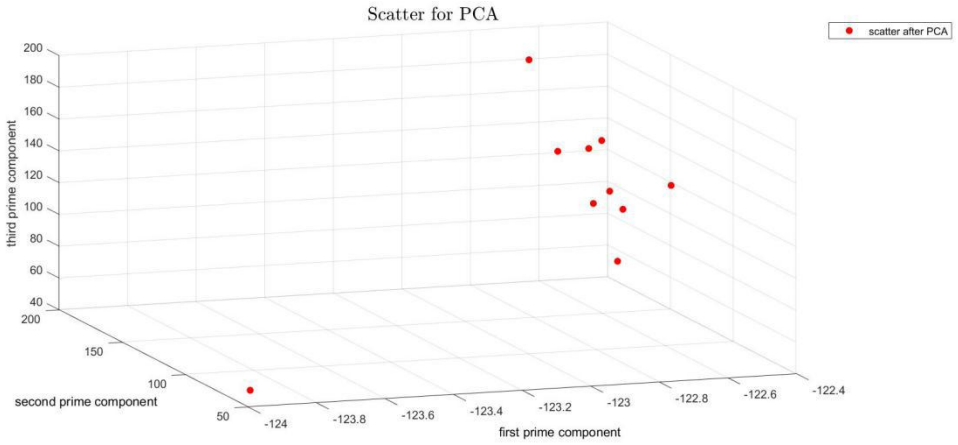

Figure 14 Scatter for PCA

3.3.2 Model of k-means Cluster Analysis

We observed the three-dimensional scatter plots of the samples drawn by principal component analysis, and found that most of the samples showed a centralized trend, indicating that the Asiatic Giant WASP is characterized by concentration.

In order to study the correlation between the 10 samples and the characteristics of Asiatic Giant WASP and judge the priority of the samples, cluster analysis was adopted to find the center point of the samples, which was the characteristics of Asiatic Giant WASP. The distance from the sample to the center point is used as a reference value to determine the priority.

First of all, we need to determine the number of clusters.

We used the Silhouette coefficient as the indicator of clustering.

Silhouette coefficient:

$$S = \frac{b - a}{\max(b - a)}$$

For a single sample,

a : the average distance from other samples in the same category

b : the average distance of the samples in the different categories closest to it (13)

The range of Silhouette coefficient is [-1,1]. The closer the distance between samples of the same category is and the farther the distance between samples of different categories is, the higher the score will be.

Generally speaking, the number of clusters is from 2 to 10, so we used Python to traverse the size of the contour coefficient of 10 samples when they were divided into groups from 2 to 10, and draw a line chart:

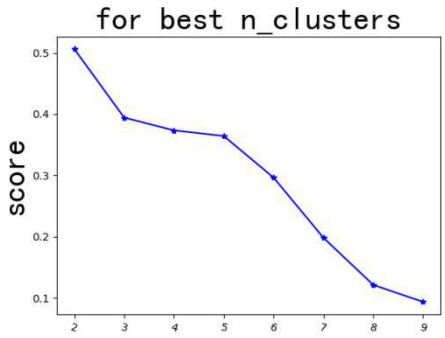

Figure 15 n_clusters

As can be seen from the line chart, when the samples are divided into two groups, the Silhouette coefficient is the largest, and the samples of the same category are the closest and the samples of different categories are the farthest. Therefore, we chose to divide the samples into two groups.

Then, we calculated the center of the two groups,

(-122.629162, 146.13939731,137.50432084)

(-123.3158185, 88.64575946,66.38369299)

After that, calculate the distance between each sample and the two center points, and take the minimum value (min distance md) as score:

(It's better if the score is less because the report is closer to the centers which can ensure it more likely report the true pest and need the attention of government.)

d1	d2
124.47830073323188	72.67828739473113
65.99589416311188	100.86413414014937
77.49755514001764	103.84956174885205
64.4958931617721	97.57593870416329

the first four reports' distances

3.3.3 Solution and Results

According to the distance from 10 samples to the center point, we conduct priority ordering. The smaller the distance, the higher the priority. The priority ordering of the 10 positive report given by this problem is as follows:

Global ID	priority
{5AC8034E-5B46-4294-85F0-5B13117EBEFE}	1
{A717D86F-23E9-4C8C-9F12-198A71113E93}	2
{0FAC3767-EAC4-477A-B5F0-24AF8A40BD09}	3
{1C6D0EAB-F68D-411D-974E-1233618854CC}	4
{FC6E894B-F6DF-4FDC-853A-D7372D253988}	5
{AD56E8D0-CC43-45B5-B042-94D1712322B9}	6
{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}	7
{BEAC832C-0783-414A-9354-C297F38570AD}	8
{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}	9
{7F3B6DB6-2ED4-4415-8DC2-3F03EC88F353}1	10

3.4 Model 4: Updating model

As time goes by, if there are other new reports, they should be added to the training set to re-train the model we built.

When a next Asian Giant Hornet is found, our sample data will change subsequently. So, we care about the occurring frequency of positive ID. Sort the 14 positive identities according to the submission time, and calculate the time difference between the adjacent IDs. Afterwards, use SPSS to analyze the data.

Table 8 Time intervals of the submission time of positive ID

0	**20**	**113**	**5**	**4**	**7**	**3**
70	33	5	0	0	11	

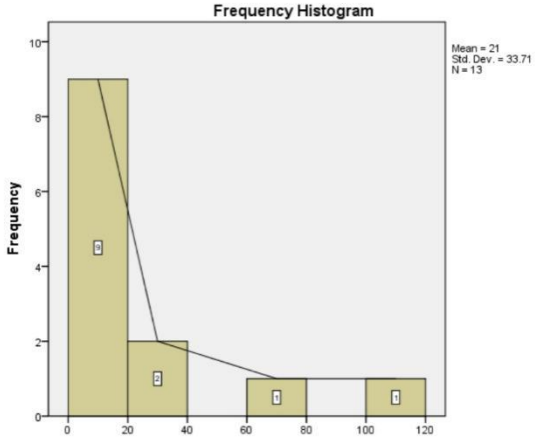

Figure 16 Frequency histogram of intervals

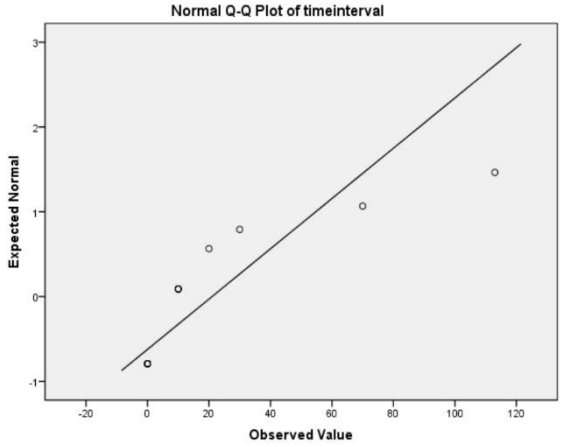

Figure 17 Normal Q-Q plot

Table 9 Analysis of data

Mean	*21.00*	*Std. Deviation*	*33.710*	
95% Confidence Interval for Mean	9.349	Minimum	0	
Lower Bound	.63	Maximum	113	
Upper Bound	41.37	Range	113	
5% Trimmed Mean	17.06	Interquartile Range	25	
Median	10.00	Skewness	2.174	.616
Variance	1136.333	Kurtosis	4.449	1.191

Considering skewness and kurtosis, the data do not fit the normal distribution very well. But we can see that most of the numbers are less than 40. Thus, we reckon updating the model every 40 days performs better.

3.5 Model 5: Whether the pests are eradicated

Since the model can be further updated only when the data is submitted, we analyzed the submission time when solving the fourth problem. Problem five is to determine whether the hornet is extinct, so we need to analyze the observation time.

First, visualize the data, and then calculate the interval.

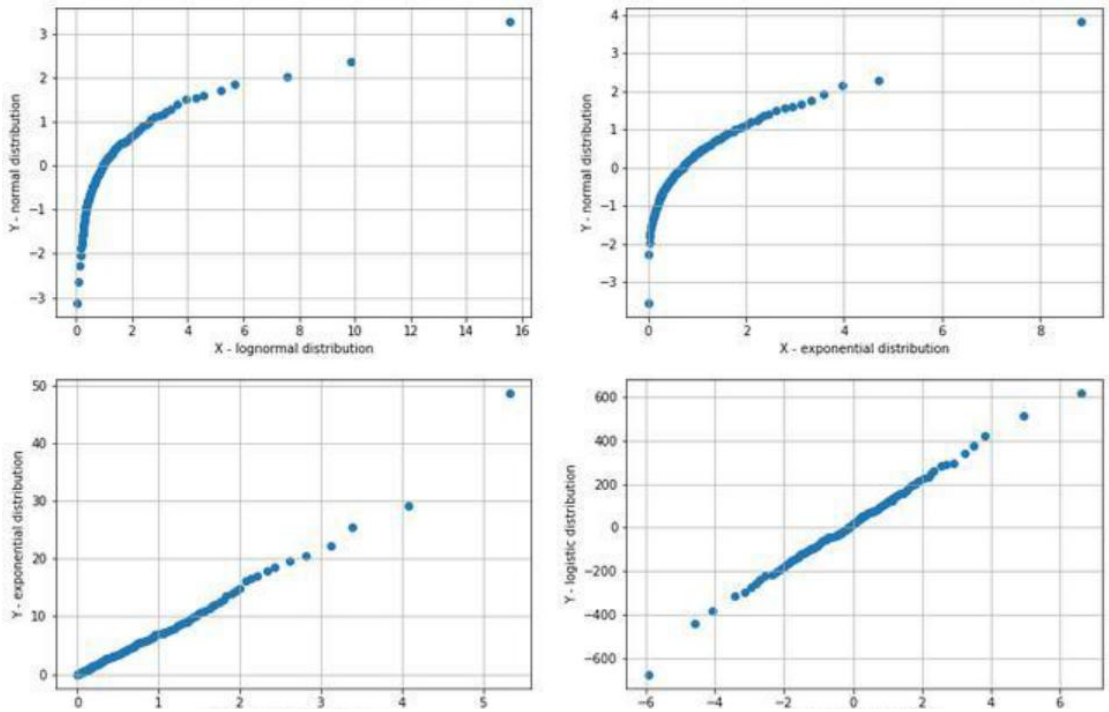

Figure 18 Some Q-Q plots

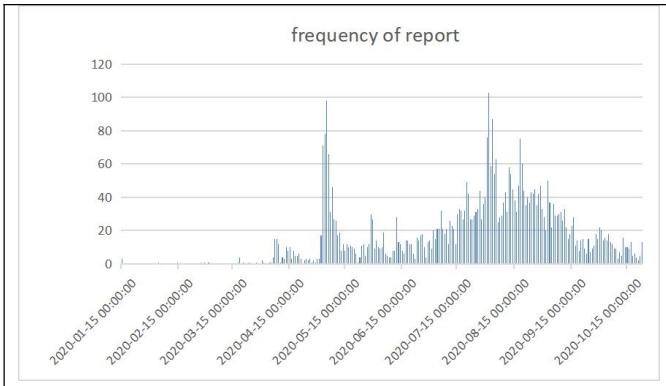

Figure 19 Frequency of report

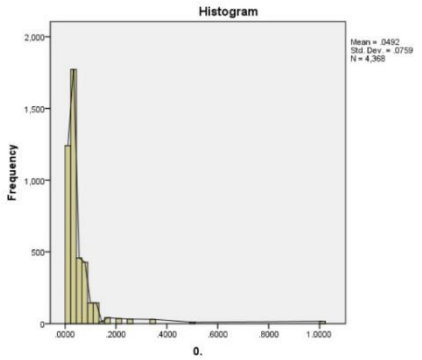

Figure 20 Detection intervals

Sig is the test value of significant difference, which is generally compared with 0.05 or 0.01. It indicates significant difference if sig is less than 0.05 or 0.01. In this model, sig is equal to 0.247, which means exponential distribution is adaptable.

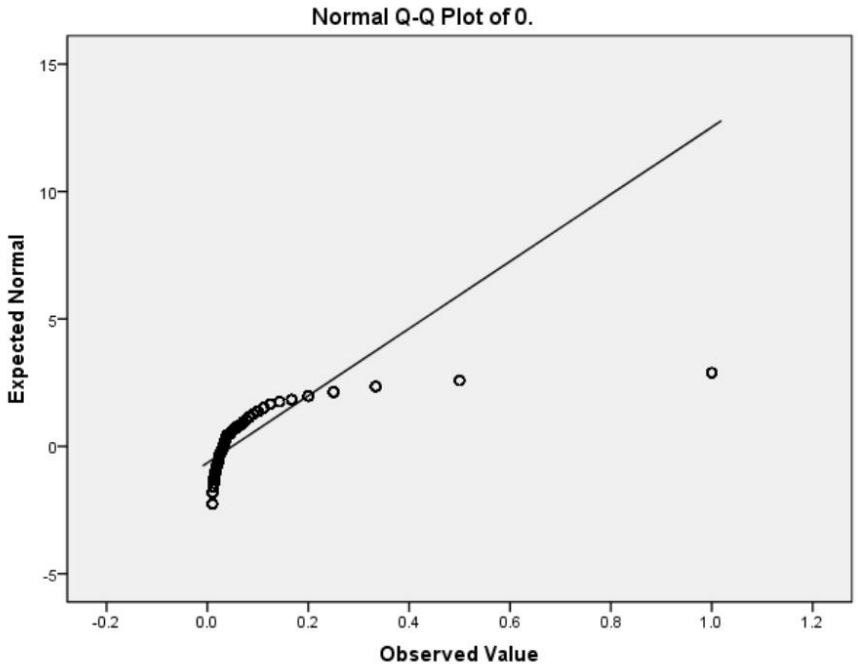

Figure 21 Normal Q-Q Plot

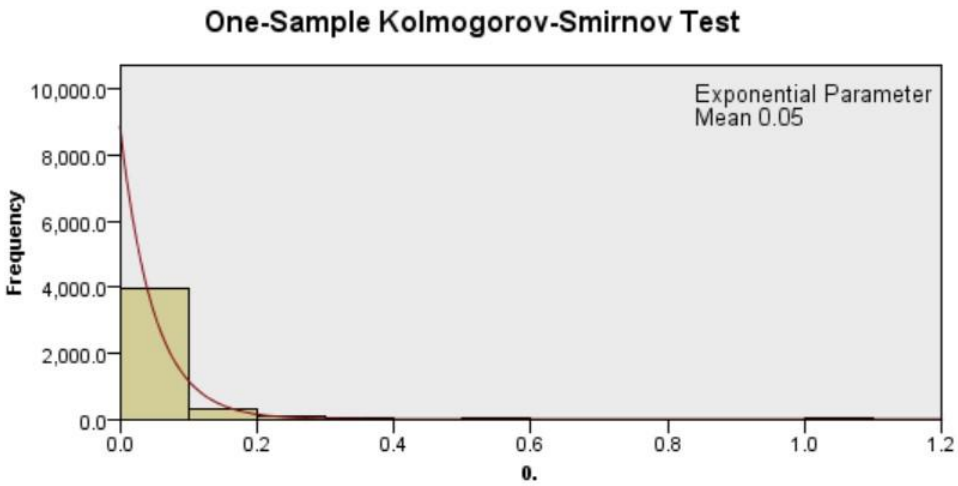

Figure 22 Kolmogorov-Smirnov Test

Use exponential distribution to fit it, and try to get the top 99% percent point.

$$1-\lambda e^{-\lambda t} < 0.99$$

$$t = 0.38 \qquad (14)$$

That means if the adjacent time of detection is more than 0.38 days, which equals that the reports obtained per day are less than 2.6, we have 99% possibility to think the hornet is distinct.

Considering contingency, we conclude that if witness less than 2.6 for a month, the pests were exterminated

Memorandum

To: Washington State Department of Agriculture
From: Team 2103686
Date: February 9th, 2021

Knowing the problems posed by the Asian giant, we established several customized models to explain the data provided by the public reporting and to select the public reporting that should be investigated in the highest priority.

Then we are writing to report our key results in these four days.

The background of American agriculture and the giant hornet

The giant hornet known as the "killer bee" has been discovered in the US state of Washington. Many beekeepers fear them because they are dangerous to humans and animals, especially the European honeybee, which is common in the area. In November 2019, a beekeeper in Washington state found that thousands of his bees were dying after their heads were torn off. And with colonies in the United States declining in recent years, many scientists say an attack by the giant hornet could wipe it out.

Due to the potential severe impact on local honeybee populations, the presence of the giant hornet can cause a good deal of anxiety. The State of Washington has created helplines and a website for people to report sightings of these hornets. Based on these reports from the public, the state must decide how to prioritize its limited resources to follow-up with additional investigation.

Introduction to the model

1. Population Growth-Dispersion Model: This model established a link between the location of the pests, the number of pests, and time. By Applying this model, we can predict the spread of this pest over time.

2. Recognition and Classification Model: This model can be used to identify images, categorize sightings, and determine whether it's the giant hornet.

3. Principal Components Analysis Model： Using this model, the variables for each sighting can be simplified to extract the main features of the giant hornet.

4. Cluster Analysis Model： This model is used to score confirmed sightings by citizens and to determine the priority of sightings for further investigation

5. Updating Model: This statistical model is used to calculate the effect of

the time interval on the mode.

Suggestion to Washington State Department of Agriculture

1. When encountering the giant hornet sightings provided by citizens, Recognition and Classification Model can be used to detect whether they are the giant hornet or not. If so, the score is calculated based on the priority score, which is then used to consider whether to use government resources for further investigation, since most of the sightings are not real giant hornet.

2. Over time, if there are other new reports, they should be added to the training set to retrain the model we have built. The model should be updated in a timely manner. We believe that updating the model every 40 days works best.

3. The government will continue to observe the pests after control until the time interval between sightings is greater than 2.6 days before proving that the pests have been eradicated

References

[1] Zhao Xinqiu, he Hailong, Yang Dongdong, Duan Siyu. Application of improved convolution neural network in image classification [J]. High tech communication, 2018,28 (z2): 930-936

[2] Yang Linnan. Research and application of occurrence prediction model of striped rice borer and Liriomyza sativae [D]. University of Electronic Science and technology, 2010

[3] Mu Wenxiu, Hong Lei, Wang Han. Application of intelligent insect classification recognition algorithm based on machine learning [J]. Digital technology and application, 2018,36 (11): 118-119

[4] Zhang Wenjun, Gu Dexiang. Two models for estimating the density of moving insects and the number of infectious diseases [J]. Acta biomathematica Sinica, 1995 (03): 60-64

[5] Lu Minyan. Research on distributed multi moving object detection based on digital recognition [D]. Shandong University, 2016

[6] Zhang Wenjun, Gu Dexiang. Study on a kind of spatiotemporal dynamic model of insect population [J]. Ecological science, 2001 (04): 1-7

[7] Zhou Guofa. Study on the growth diffusion model of insect population [J]. Acta biomathematica Sinica, 1996 (04): 17

Appendix: Program Codes

Here are the program codes we used in our research.

I. mesh method

```
pla=alld.loc[:,'Latitude']
plo=alld.loc[:,'Longitude']
#print(pla)
x=[]
y=[]
for i in range(14):
    x.append((plo[i]+123.9431)/0.4)
    y.append((pla[i]-48.7775)/0.25)
alld['x']=x
alld['y']=y
```

II. one hot encoding

```
text = list(data['Notes'])
tokens = tokenize(text)
print(tokens)
#print(tokens)
mydict=count_corpus(tokens)
f='dict.txt'
tmp=mydict.items()
d_order=sorted(tmp,key=lambda x:x[1],reverse=True)
print(d_order)
with open(f,'w',encoding="utf-8") as fo:
    for k,v in d_order:
        print("%s,%s" %(k,v),sep='',file=fo)
key=[]
for i in range(21,31,1):
    key.append(d_order[i][0])
print(key)
print(data._stat_axis.values.tolist() )
for i in key:
    data[i] = data.apply(lambda _: 0, axis=1)
for i in range(2069):
    for j in tokens[i]:
        if j in key:
            data.loc[i,j]+=1
class_mapping = {'Negative ID':0, 'Positive ID':1}
data['Lab Status'] = data['Lab Status'].map(class_mapping)
```

III. image cropping and get RGB and ambiguous

```
text = list(data['Notes'])
tokens = tokenize(text)
print(tokens)
#print(tokens)
mydict=count_corpus(tokens)
f='dict.txt'
tmp=mydict.items()
d_order=sorted(tmp,key=lambda x:x[1],reverse=True)
print(d_order)
with open(f,'w',encoding="utf-8") as fo:
    for k,v in d_order:
        print("%s,%s" %(k,v),sep='',file=fo)
key=[]
for i in range(21,31,1):
    key.append(d_order[i][0])
print(key)
print(data._stat_axis.values.tolist() )
for i in key:
    data[i] = data.apply(lambda _: 0, axis=1)
for i in range(2069):
    for j in tokens[i]:
        if j in key:
            data.loc[i,j]+=1
class_mapping = {'Negative ID':0, 'Positive ID':1}
data['Lab Status'] = data['Lab Status'].map(class_mapping)

```

IV. PCA

```
    dataCov = datasTad.cov()
    newData1 = np.array(dataCov)
# means filling
    mean=np.mean(newData1)
    if(np.isnan(mean)):
        numFeat = shape(newData1)[1]
        for i in range(numFeat):
            newData1[i]=0
    else:
        numFeat = shape(newData1)[1]
        for i in range(numFeat):
            if np.isnan(newData1[i]):
                newData1[i]=mean

    eigenValue, eigenVector = np.linalg.eig(newData1)
    sorceEigenValue = np.argsort(eigenValue)
    nPcaEigenVector = sorceEigenValue[-n:]
    pcaEigenVector = eigenVector[nPcaEigenVector]
    PCAX = np.dot(dataX , pcaEigenVector.T)
    return PCAX ,pcaEigenVector
```

V. K-means

```
def train_cluster(train_vecs, model_name = None , start_k = 2 , end_k = 10 ):
    打印("培训集群")
    分数= []
    型号= []
    对于 我 在 范围内(start_k, end_k):
        kmeans_model = KMeans(n_clusters = i, n_jobs = multiprocessing.cpu_count() , )
        kmeans_model.fit(train_vecs)
        分数= Silhouette_score(train_vecs, kmeans_model.labels_, metric = '欧几里得' )
        scores.append(score)    #保存每一个k值的score值，在这里用欧式距离
        print ('{}表示分数损失= {}'. format(i, 分数) )
        models.append(kmeans_model)
    plt.plot(范围(2 , 10 ) , 分数, 'B * - ' )
    #设置横纵坐标的名称以及对应字体格式
    font2 = { 'family' :  'SimHei' ,
              'weight' :  'normal' ,
              '大小' :  30 ,
              }
    plt.xlabel('n_clusters' , font2)
    plt.ylabel('score' , font2)
    plt.title('for best n_clusters' , font2)
    plt.xticks(rotation = -15 )    #设置x轴标签旋转角度
    plt.savefig('for best n.jpg' )
    plt.show()
    best_model =模型[scores.index(max(scores)) ]
    返回 best_model
```