Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2109268

Summary

With the progress of human civilization, human footprints spread all over the world. However, this progress has brought about the problem of species invasion. In the recent years, species invasion has become a serious problem all around the world. It has caused the destruction of the local ecological environment and has seriously affected human production and life. Currently, a species invasion event has occurred in Washington State. Bumblebees from East Asia invaded Washington State. The Asian Hornet has had a serious impact on the local ecology and life. 1. This pest will kill many of honey bees and destroy the ecological environment. 2. This kind of pests will attack humans and threaten human safety.

In order to predict and control the spread of these pests, we have established some mathematical models. Through these models, we will make some suggestions for pest prevention and control.

To begin with, we do some data preprocessing. The data visualization and data cleaning process shows two major problems of the data set. 1. The lack of useful data. 2. Too few features available. In order to solve these problems, we apply k-nearest neighbors algorithm to do data augmentation. The major challenge we facing in the model building process is the lack of data.

First, to answer the question whether the spread of this pest over time could be predicted, and with what level of precision. We build some models to predicts the rate and location of the spread of the pests.

The first model is developed to predict the rate of the spread of the pests, which is a birth-death process model. We later simplified this model into a nonhomogeneous Poisson process model. To estimate the parameter λ, we use the gradient descent algorithm. To overcome the problem of the lack of data. We apply bootstrap algorithm for a better estimation. And the Chi-square test shows a good fitness of the data. The result indicates that a single nest can product a new nest every season.

The second model is used to predict the location of the spread of the pests, which is a bivariate normal distribution model. We use the location of original nest as the center of this distribution $\boldsymbol{\mu}$. And then, we apply the maximum likelihood estimation to estimate the covariance matrix $\boldsymbol{\Sigma}$. To test the normality of the data, we apply the Shapiro-Wilk test. The result shows that there is a 95% probability that the new nest is within 10km of the original nest.

The evolving network model uses the previous two models and gives a great data visualization.

Second, we applying some machine learning algorithm classify whether the report is positive or not. The algorithm includes Support Vector Machine (SVM), random forest and the convolutional neural network. After processing, we design a way to fuse results as a whole and attain a comprehensive prediction.

Third, by the result of random forest algorithm in the second part, we could rank the importance of features to discuss how our classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings.

Fourth, we discuss some plausible incremental learning algorithms to update our machine learning model to achieve better prediction results when we're given additional new reports over time.

Fifth, the branch process model gives the evidence that the pest has been eradicated in Washington State. As long as the pest transmission rate in the season is less than 1, they will eventually become extinct.

Finally, we summarize the strengths and weakness of the model, and give some future work such as the use of oversampling method, EM algorithm and Bayesian inference to improve our model.

Pest Prediction and Identification Based on Random Process and Unbalanced Machine Learning

Contents

1 Introduction

1.1 Background

Species invasion has become a serious ecological and environmental problem worldwide. Rafferty gave the definition in her paper that "Invasive species, also called introduced species, alien species, or exotic species, any nonnative species that significantly modifies or disrupts the ecosystems it colonizes." (Rafferty, 2019, February 7).

Washington State was recently occupied by invasive species Vespa mandarinia also known as Aisan giant hornets. Researchers have given out the harm it may bring in their paper. "V. mandarinia are an invasion concern due to their ability to kill honey bees and affect humans." (Zhu, Gutierrez Illan, Looney, & Crowder, 2020). Our goal is to build mathematical models to control this pest.

1.2 Restatement of the Problem

There are totally 5 questions, including many facets of statistical modeling expertise.

1. A task of developing a model to predict the number and the location of the pests.
2. Pattern recognition problem which needs machine learning and some statistical models.
3. Use the result in the second problem and sort the importance of the features in the data set.
4. Develop an learning model which can update by the new data set.
5. Develop a new model based on the result of the first problem

1.3 Overview of Our Work

1. Develop random process and normal distribution model to predict the number and the location.
2. Apply Machine learning algorithms to classify the reports.
3. Use the result of random forest to sort the features.
4. Apply an increment learning algorithm based on the previous model.
5. Develop a branching process based on the previous results.

2 Assumptions and Symbol Explanation

2.1 Assumptions

In order to better optimize our model and increase the accuracy of the model, we need to make some reasonable assumptions about the model.

1) The model assumes that Asian giant hornets are more likely to establish colonies closer to the mother colony.

2) The model assumes that the Asian giant hornets can only start from one colony to establish the next colony, not multiple colonies can establish a colony at the same time.

3) The model does not consider the destruction of the Asian giant Hornets' colony.

4) The model assumes that there will be no more Asian giant hornets immigration.

5) The model assumes that the rate at which a single nest establishes a colony is a constant.

6) The model assumes that the number of detection equals that of nests of Asian giant hornets.

7) The model assumes that the location of detection is close to the nest.

2.2 Symbol Explanation

Symbol	Definition
$X_{n\times p}$	The data set with n rows and p columns
D	Euclidean distance
Pr	The symbol of probability
E	The symbol of expectation
μ	The death rate of a colony
λ	The birth rate of a colony
θ	The immigrant of Asian giant hornets
T	The random variable of time
X	The random variable of the number of colonies
m	The expect value of the number of colonies
N	The real value of the number of colonies
α	The confidence level
V	The nodes of a network
E	The edges of a network
$\text{loss}(x, class)$	The cross entropy loss
$\mathcal{D}$	The distribution of given image dataset

Table 1: The Definition of the Essential Mathematical Symbols

3 Data Preprocessing

3.1 Data Visualization and Data Cleaning

The data set given by the Washington State Department of Agriculture has 4440 rows and 8 columns which is a relatively huge data set. And therefore, it's important to do some data preprocessing before we begin to build the model.

3.1.1 The Distribution of Asian Giant Hornets

In order to begin the data analysis, one efficient step is to visualize the data set. From the data set, we draw the data on the latitude and longitude of the detection location from the data set. And with the detection time t and the lab status s, we get the maps of detection location. By Tableau, we draw two

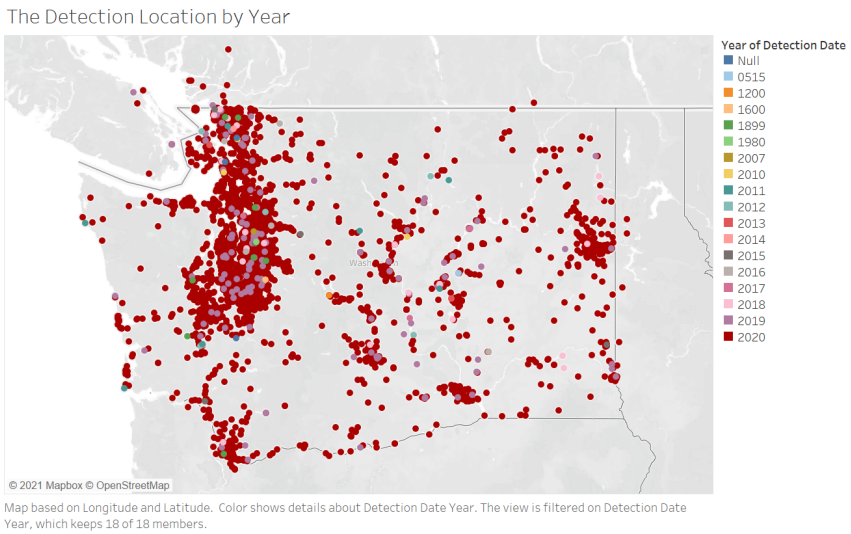

Figure 1: The Detection Map by Year

maps respectively (Figure 1, Figure 2), one is marked by the year of detection t, and the other is marked by the lab status s.

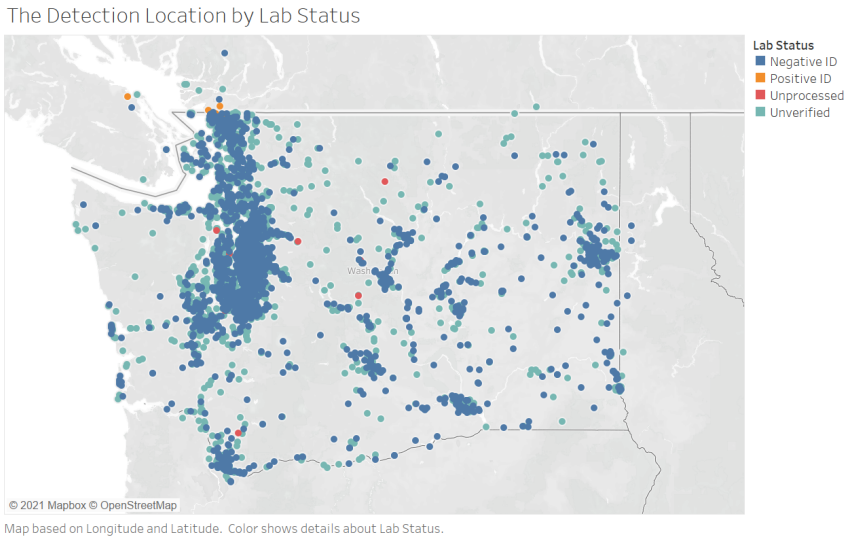

Figure 2: The Detection Map by Lab Status

According to the maps, there are two major problems of the data set.

1. There are some mistaken dates in the data set. Such as null and the year before 2007.
2. The proportion of samples which are labeled as positive IDs is too low in the data set.

And hence, the major problem in the model building step is the lack of useful data.

3.1.2 Data Cleaning and Data Selection

For the first problem, it's simple to extract the mistaken data and then remove it from the data set. After deleting all the null values which is also called the data cleaning process, we get a new data set which contains 4426 rows and 8 columns.

For the second problem, however, it's necessary to apply some data augmentation algorithms to get a balanced designed data set which means the data set contains enough positive ID and negative ID to the future statistics analysis and learning algorithms.

In order to decrease the influence of the unbalanced sample on the establishment of our model, we undersample some data from the original data set after data cleaning to increase the proportion of positive samples.

Figure 2 shows that the positive samples mainly fall on between latitude $48.9° \sim 49.1°$, longitude $-122.8° \sim -122.5°$. The figure of samples fall on this area is given as follows (Figure 3).

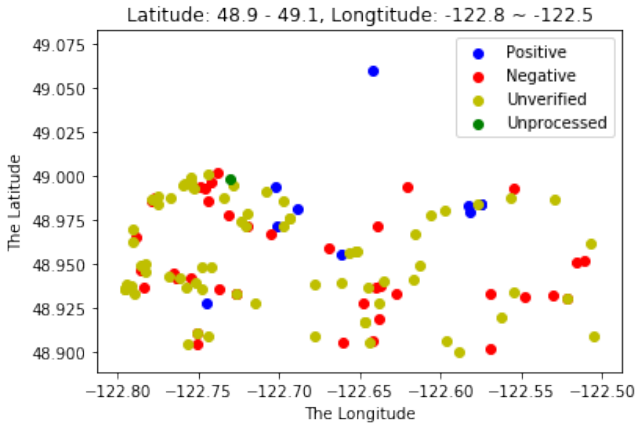

Figure 3: The Figure of Samples Fall between Latitude $48.9° \sim 49.1°$, Longitude $-122.8° \sim -122.5°$

By the 1-st hypothesis, we have the reason to aussme that there are more positive samples in this area in the unverified samples. This assumption is highly credible. From the description in the problem C, "A new queen has a range estimated at 30km for establishing her nest." Actually, according to the report of the PennState Extension (Skvarla, 2020, May 7), "Asian giant hornets only fly 0.5–1.25 miles (1–2 km) on average (and never more than 5 miles (8 km)) from the nest in search of food". Therefore, we have the reason to believe that Asian giant hornets are more likely to establish new colonies closer to the original colony. Where the positive samples gather, there are more hidden positive samples.

In the next subsection, we'll develop a algorithm to find these positive samples.

3.2 Data Augmentation Algorithms

According to figure 2, there are a lot of samples which are labeled as "unverified" and "unprocessed" in the data set. They are so called missing data. Our aim in the next subsection is to implement a algorithm to complete the data set.

3.2.1 K-Nearest Neighbors Algorithm

K-nearest neighbors (KNN) algorithm is a non-parametric classification method in statistics. It is widely used in predicting the missing values. We note that even though KNN algorithm is a non-parametric model. It's a semi-supervised learning algorithm. The algorithm is given as follows.

Algorithm 1 K-Nearest Neighbors Algorithm

Require: Data set $X_{n\times p}$ with rows $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$, Vector y with labeled elements $\{y_{(1)}, \ldots, y_{(k)}\}$ and unlabeled elements $\{y_{(k+1)}, \ldots, y_{(n)}\}$

Ensure: The labeled elements $\{y_{(k+1)}, \ldots, y_{(n)}\}$

1: **Distance:** $D_{ij} \leftarrow ||\boldsymbol{x_i} - \boldsymbol{x_j}||^2, 1 \le i, j \le n$

2: **Sort:** For each $i = (K+1), \ldots, (n)$, sort D_{ij}, $D_{i(1)} \le \cdots \le D_{i(n)}$

3: **Choose:** For each $i = (K+1), \ldots, (n)$, Choose k $y^{(1)}, \ldots, y^{(k)} \in y$ with k-nearest distance D_{ij}

4: **Assignment:** For each $i = (K+1), \ldots, (n)$, $y_i \leftarrow \frac{1}{k}\sum_{j=1}^{k} y^{(j)}$

Similar to K-means algorithm, KNN algorithm uses Euclidean distance as its metric. After calculating the distance between each observation, the algorithm needs to sort the distance. And then choose the k-nearest neighbors. And predict the unlabeled values with the information of the k-nearest neighbors. Here, we use the mean of the k-nearest neighbors.

However, as the 1-st hypothesis states, the Asian giant hornets tend to establish their new colonies near the original colony. When applying the KNN algorithm, we assign the $k = 1$ which means the label of the unverified samples is the same as the nearest labeled sample. The result is as follows (Figure 4).

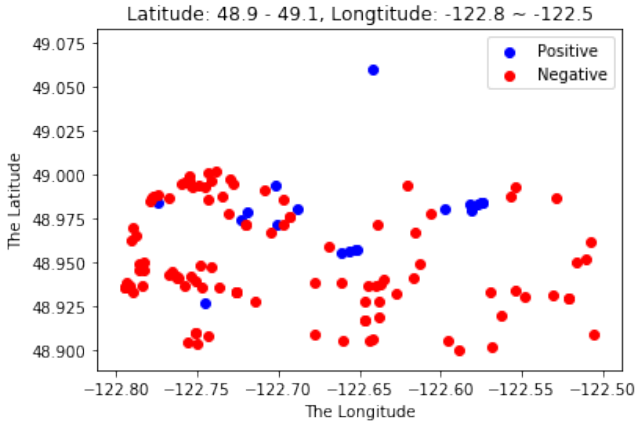

Figure 4: The Figure of Samples after Applying KNN Algorithm

The result of KNN algorithm is perfect. There are more positive samples in this area after applying KNN algorithm with high accuracy. Meanwhile, actually, the drawback of KNN algorithm is that it cannot deal with the problem of imbalanced designed data set. And hence the positive samples are still few in the area.

We use the result from KNN algorithm in the next section to predict the spread of the pest.

4 Immigration and Transmission Model

4.1 Birth-Death Processing

The Asian giant hornets establishing colony gives us such an intuition, we have reason to analogize this behavior to human immigrants. Therefore, we can consider a linear growth model with immigration which is a special case of Birth-Death Processing. By applying the model, we can get the number of the Asian giant hornets over time t.

Birth-Death processing is a continuous-time Markov chain which has the definition as follows.

Definition 1. *Suppose there is a continuous time countable states process* $\{X_t : t \geq 0\}$. *The process is defined as continuous-time a Markov chain if* $t \geq 0$ *and* $p > p_n > \cdots > p_0$, $i_0, i_1, \ldots, i_n, i, j \in S$,

$$Pr(X_{t+p} = j | X_p = i, X_{p_n} = i_n, \ldots, X_{p_0} = i_0) = Pr(X_{t+p} | X_p = i). \tag{1}$$

The definition indicates that the the probability of occurrence of this event is only related to the occurrence of the previous event, which is the same as our 2-nd hypothesis, so this assumption needs to be made to adopt continuous-time Markov chain. Such a assumption is creditable since the new nest is establish by only one queen from the original nest (Skvarla, 2020, May 7).

The linear growth model with immigration is stated as follow.

Each individual in the population gives birth at an exponential rate λ. Meanwhile, there is an exponential rate of increase θ of the population due to an external source such as immigration. Deaths are assumed to occur at an exponential rate μ for each individual. In this model:

$$\mu_n = n\mu, \quad n \geq 1 \qquad \lambda_n = n\lambda + \theta, \quad n \geq 0 \tag{2}$$

where n is the order of current state.

Let $X(t)$ denote the population size at time t. Suppose that $X(0) = i$ and let $m(t) = E(X(t))$. The model is aimed to find m(t).

As preparation, we consider two independent random variables $T_i \sim \text{expo}(\lambda_i), \quad i = 1, 2$.

Let $h \to 0$ (h = o(1)). Then, we have

$$Pr(T_1 \leq h) = 1 - e^{-\lambda_1 h} = \lambda_1 h + o(h) \tag{3}$$

by applying Taylor expansion $e^x = 1 + x + o(x)$. And further,

$$Pr(T_1 + T_2 \leq h) = \int_0^h Pr(T_2 \leq h - s)\lambda_1 e^{-\lambda_1 s} ds = o(h) \tag{4}$$

Back to the model. Note that if $X(t) = n$, then the state transition function is given by.

$$X(t+h) = \begin{cases} n+1 & w.p. \quad (n\lambda + \theta)h + o(h) \\ n-1 & w.p. \quad n\mu h + o(h) \\ n & w.p. \quad 1 - (\theta + (\lambda + \mu)n)h + o(h) \end{cases} \tag{5}$$

Thus,

$$E(X(t+h)|X(t)) = X(t) + (\theta + (\lambda - \mu)X(t))h + o(h) \tag{6}$$

Therefore, we have the following differential equation.

$$m'(t) = \theta + (\lambda - \mu)m(t) \tag{7}$$

Solve the differential equation, we get

$$m(t) = x_0 \exp((\lambda - \mu)t) + \frac{\theta}{\mu - \lambda}(1 - \exp(\lambda - \mu)t) \tag{8}$$

where x_0 is the population at time $t = 0$.

4.2 Nonhomogeneous Poisson Process

By the 3-rd and 4-th hypotheses, the parameters of the model $\mu = 0$ and $\theta = 0$ which means the model doesn't consider the destruction of the Asian giant Hornets' colony and there will be no more Asian giant hornets immigration in Washington State. Also, by 5-th hypothesis, the birth rate λ for an individual is an constant.

The state transition function becomes a state transition function of a nonhomogeneous Poisson process which has the following definition.

Definition 2. *The counting process $\{X(t), t \geq 0\}$ is said to be a nonhomogeneous Possion process with rate $\lambda > 0$ if the following axioms hold:*

1) $X(0) = 0$.

2) $\{X(t), t \geq 0\}$ has independent increments.

3) $X(t + h) - X(t) \sim Possion(\int_s^t \lambda(u)du)$.

Since it's not the point of our model, we omit the proof. The nonhomogeneous Poisson process has the following useful theorem.

Theorem 1. *Suppose $\{X(n), n \in \mathbb{N}\}$ is a Possion process with rate $\lambda(t) > 0$, then $X(n) \sim Poisson(\sum_n \lambda_n)$.*

Since data is always discrete in a data set. We can use the theorem to fit a nonhomogeneous Poisson process to predict the spread of the pest.

4.2.1 Gradient Descent Algorithm

To estimate the λ in the nonhomogeneous Poisson process model, we need to use the gradient descent algorithm.

We note that since $\mu = \theta = 0$ in the model. The expected population is given by $m(t) = x_0 e^{\lambda t}$.

In order to estimate the λ, we need to set a cost function $J(\lambda)$. And minimize the cost function to get the best estimation of λ.

$$J(\lambda) = \frac{1}{2n} \sum_{t=1}^{n} [m(t) - N(t)]^2 \tag{9}$$

where n is the number of the states and $N(t)$ is the real value of the population at time t. And the gradient is given by

$$\frac{d}{d\lambda} J(\lambda) = \frac{1}{n} \sum_{t=1}^{n} [e^{\lambda t} - N(t)] t e^{\lambda t} \tag{10}$$

after we substitute $m(t) = x_0 e^{\lambda t}$ into the cost function.

By applying gradient descent algorithm, we can get the extreme point of cost function which is the estimation of λ. The algorithm is stated as follows.

Algorithm 2 Gradient Descent Algorithm

Require: Initialized λ, learning rate α, cost function $J(\lambda)$

Ensure: Estimated λ

1: **repeat**
2: $temp \leftarrow \lambda - \alpha \frac{d}{d\lambda} J(\lambda)$
3: $\lambda \leftarrow temp$
4: **until** convergence

The advantage of the algorithm is that it's simple to calculate the gradient of the cost function. And thus, it's widely used in machine learning algorithms. The drawback of such an algorithm is that it is a convex optimization algorithm. Only when the function is convex can we get the true parameter estimates. However, the cost function we defined is not convex. Thus, we may get a estimated λ which is not best.

However, due to the lack of positive samples, the gradient descent algorithm is not enough for the parameter estimation. We need bootstrap method which we'll state later.

4.2.2 Bootstrap Method

Bootstrap Algorithm is an algorithm for obtaining confidence intervals (CIs) of parameter under the lack of samples which is appropriate for our data set.

Our goal is to get the point estimate $\hat{\lambda}$ and the confidence intervals (bootstrap interval) of λ. The essential steps are as follows.

1. Calculate the point estimate $\hat{\lambda}$ by gradient descent algorithm.
2. Generate a bootstrap sample $\boldsymbol{x^*} = (X_1^*, \ldots, X_n^*)^T$ with $\{X_i^*\}_{i=1}^n \overset{iid}{\sim} Poisson(\sum \lambda)$ and compute the corresponding bootstrap replication $\hat{\lambda}$ by gradient descent algorithm.
3. Independently repeating step 2 G times, we obtain G bootstrap replications $\{\hat{\lambda}^*(g)\}_{g=1}^G$.
4. A $100(1-\alpha)\%$ bootstrap CI for λ is $[\lambda_L^*, \lambda_U^*]$ where λ_L^*, λ_U^* are the $(\alpha/2)G$-th and the $(1-\alpha/2)G$-th order statistics of $\{\hat{\lambda}^*(g)\}_{g=1}^G$.

The advantage of bootstrap algorithm is it can overcome the problem of lack of sample which makes our result more accurate. However, it's computationally expensive since we need to repeat gradient descent algorithm.

4.2.3 Result and Chi-Square Test

In this part, we'll discuss how to fit the nonhomogeneous Poisson process by our data set. We note that the data set we used is processed by the KNN algorithm.

We choose season as the unit of time. The season starts from the summer of 2019, there was the first detection in Washington state. We have the time vector $t = (0, 1, 2, 3, 4, 5)$ and the the vector of number of detection by the season $d(t) = (1, 3, 4, 9, 10, 19)$.

By 6-th hypothesis, we view the number of detection $d(t) = N(t)$ where $N(t)$ is the nests of the Asian giant hornets and also the population of Asian giant hornets. The assumption is creditable since the recent research mentioned that if you find a Asian giant hornet. Usually, there will be a nest near it (Zhu et al., 2020). Since they can fly far away.

By setting $G = 10000$ and $\alpha = 0.05$, we get the 95% CI of λ is $[0.34417, 1.0099]$ and the point estimate is $\hat{\lambda} = 0.64381$.

This result indicates that the rate for a single nest produces new nests every season is about 0.64.

For the goodness of fit, we apply Chi-square test to the data set. We note that since $X(t+s) - X(s) \sim \text{Poisson}(\sum_n \lambda_n)$, then $X(n) - X(n-1) \sim \text{Poisson}(\lambda)$ where n is positive integer ($\lambda_n = n\lambda$). Therefore, we can get the frequency distribution of the corresponding Poisson distribution with $\lambda = \hat{\lambda}$ The data set shows the following frequency distribution (Table 2).

Number of Detection	0	1	2	3	4	5	6	7	8	9	Total
Frequency of Events	0	3	1	0	0	1	0	0	0	1	6

Table 2: The Frequency Distribution of the Number of Detection

The hypotheses are given as

$$H_0 : \text{The distribution is Poisson}(\hat{\lambda}) \quad v.s. \quad H_1 : \text{The distribution is not Poisson}(\hat{\lambda}).$$

The test statistics is given by

$$Q_{n0} = \sum_{j=0}^{m} \frac{(N_j - np_{j0})^2}{np_{j0}} \xrightarrow{L} \chi^2(m-q-1) \quad as \quad n \to \infty$$

where m is the sample size and q is the the number of parameters. If $Q_{n0} > \chi^2(\alpha, m-q-1)$, we reject the null hypothesis, where α is the confidence level.

After applying the test, we get the statistics $Q_{n0} = 14.0$. And at 95% confidence level, the p-value is $0.12235 > 0.05$. Thus, we cannot reject the null hypothesis. The distribution is Poisson distribution. The model fits the data well.

4.3 Bivariate Normal Distribution Model

In the previous subsection, we build a random process model to predict the number of the Asian giant hornets over time. However, it's not enough to predict the number of the pests. We need to build a model to predict the location of pests.

Since we assume that Asian giant hornets are always tend to establish new colonies near the old colony. It's naturally that we can consider the distance of the new colonies near the old colony as a

bivariate normal distribution which has the high probability density at the center of the distribution. And the center of the distribution is the location of the old colony.

The pdf of the bivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi}|\Sigma_j|^{-1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big). \tag{11}$$

By 7-th hypothesis, we aussume that the location of detection is the location of the nest. The location where they build a new nest follows a bivariate normal distribution centered on the original nest. The hypothesis is creditable since they cannot fly far away.

4.3.1 Maximum Likelihood Estimation

In this part, we only need to estimate the covariance matrix $\boldsymbol{\Sigma}$ the sample since the mean vector $\boldsymbol{\mu_0}$ is the location of the original colony which we've already known. Suppose we have m observations from a bivariate normal population, each of size 2: $\boldsymbol{X_1}, \ldots, \boldsymbol{X_m}$. The maximum likelihood estimate of $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m}(X_i - \boldsymbol{\mu_0})(X_i - \boldsymbol{\mu_0})^T. \tag{12}$$

4.3.2 Result and Shapiro–Wilk Test

By applying the maximum likelihood estimation on the postive samples, we get the covariance matrix as follows $\hat{\boldsymbol{\Sigma}} = [[0.00064672, 0.00039802], [0.00039802, 0.0042248]]$.

More generally, a $100(1-\alpha)\%$ confidence region (CR) for a bivariate normal distribution is as follows.

$$(\boldsymbol{X}-\mu)^T\Sigma^{-1}(\boldsymbol{X}-\mu) \leq \chi^2_{2,1-\alpha} \tag{13}$$

According to the three-sigma rule in statistics, we plot its three CRs: 68.27% CR, 95.45% CR, 99.73% CR separately (Figure 5).

The semi-minor and semi-major of the three CRs are $(a_1, b_1) = (0.025431°, 0.065000°)$, $(a_2, b_2) = (0.050861°, 0.13000°)$, $(a_3, b_3) = (0.076292°, 0.19500°)$. This means that their radius is approximately 5km, 10km, 15km.

The model gives us a result that, from the original colony, the Asian giant hornets has 68.27% probability to build their new colonies within 5km, 95.45% probability to build their new colonies within 10km, 99.73% probability to build their new colonies within 15km.

For the goodness of fit of the bivariate normal distribution model, we choose Shapiro–Wilk Test as our hypothesis testing. The test is a test of normality in frequentist statistics. The hypotheses are given as

$$H_0 : \text{The distribution is normal} \quad v.s. \quad H_1 : \text{The distribution is not normal .}$$

Due to the lack of the positive sample. After applying the test, the p-value is close to 0. Therefore, we reject the normal assumption. However, the model is still useful if we get more data in the future.

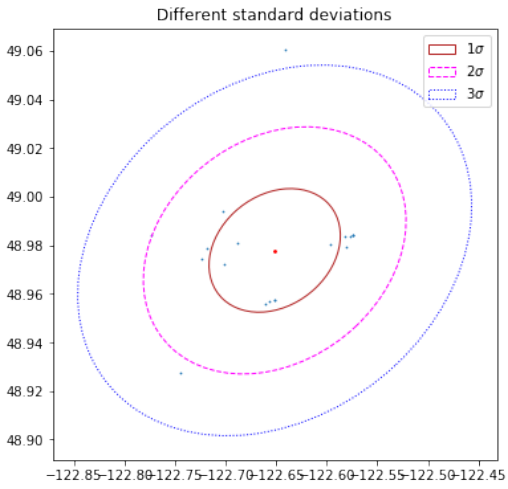

Figure 5: The Confidence Region of the Bivariate Normal Distribution

4.4 Evolving Network

In the section 4.2, we develop a model to predict the number of the pest over time. The rate at which a single nest produces new nests every quarter is about 0.64. In the section 4.3, we develop a model to predict the distance of the spread. The Asian giant hornets has 68.27% probability to build their new colonies within 5km, 95.45% probability to build their new colonies within 10km, 99.73% probability to build their new colonies within 15km.

Combining these two results, we can get an interesting network that changes over time, which we call an Evolving network.

The network is an undirected network (V,E), where V denotes the nodes and E denotes the edge. We denote the set of nodes generated at time t as $V^{(t)}$ and the set of edge generated at time t as $E^{(t)}$.

Algorithm 3 Evolving Network Generation Algorithm

Require: Initial $V^{(0)}$ and $E^{(0)}$
Ensure: The evolving network ($V^{(m)}$,$E^{(m)}$)
1: **repeat**
2: **Number:** At time $0 < t \leq m$, $X^{(t)} \sim$ Poisson($\sum_n \lambda_n$)
3: **Choose:** Random choose $X^{(t)}$ nodes in $V^{(t-1)}$
4: **Update:** For the chosen node in $V^{(t-1)}$, generate a new node from $N(\boldsymbol{\mu_v}, \hat{\boldsymbol{\Sigma}})$
5: **Edge:** Add a new edge between the chosen node and the new node in $E^{(t)}$
6: **until** t > m

We generate the evolving network and plot it (Figure 6).

The pink points are the predicted location of nests at 2020 Winter by our model. The evolving network could give us a perfect visualization.

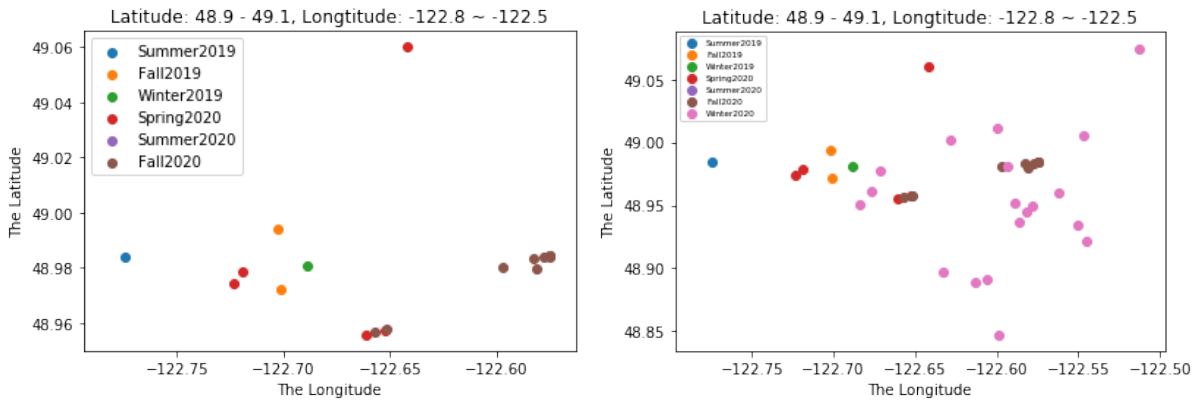

Figure 6: The Prediction of Location of the Nests from 2020 Fall to 2020 Winter

5 Classification Model

In the actual sighting report, the reporter will provide their text description, sighting location (latitude and longitude) and accompanying picture information. But most of the sightings will confuse it with other bee species, so we hope to establish an effective model to analyze and classify the sighting information to judge the possibility of classification errors.

The attachment provides two data sets, one contains textual information of sightings (notes), report date, sighting location, etc., and the other contains image data. We are going to divide the data set into two parts to build effective models respectively. For text and data information, we use SVM and RF (random forest) to predict by reconstructing features; for image information, we use CNN (convolutional neural network) to classify and predict.

Here list all the python libraries we are going to use in this section: **Pyecharts,imblearn,sklearn,jieba**

5.1 SVM & Random Forest

5.1.1 Location Filtering

At the beginning, it was a natural idea to use numerical features in the data as our features for training. After observing the location distribution of positive and negative samples in the data set, we found that positive samples only appear in a relatively small range of latitude and longitude, while report samples appearing in other ranges are basically negative samples or unrecognizable In the data preprocessing stage, we screened out all positive and negative samples in the range of 45-49 degrees north latitude and 121-124 degrees west longitude to facilitate future training. Now we have a total of 1615 samples.

It is not difficult to find that among all the current 1615 samples, there are only 14 positive samples, which gathered in a small region (see Figure 2). The ratio of negative and positive samples is close to 100:1. The data set is seriously unbalanced. In the subsequent steps, we will compare several different sampling methods for data expansion.

5.1.2 Text Vectorization

In the first part of this thesis, the modeling of the stochastic process of birth and death can roughly predict the trend of Asian wasp proliferation with the year. Therefore, the number of samples reported

within this area will gradually increase. Although the location data of latitude and longitude is the main source of our training model, we also noticed that the descriptive text information of the report is also an important data source.

Text vectorization is a common processing method in NLP (natural language processing). If we want to use notes as a feature, we need to measure the similarity of two texts, and that is what text vectorization do. By constructing a corpus (thesaurus), for any piece of text, it is split into word sets by word segmentation, and a vector is built to store the number of times each word in the dictionary appears in the piece of text. We can represent the text as a vector. So we can measure the similarity of two text vectors by calculating Euclidean distance, vector angle cosine or other metrics.

We have noticed that in this specific issue, the false reports of eyewitnesses often contain descriptions that are easily different from those of the real Asian hornets. For example, Asian hornets generally build nests in *forests*, but a considerable part of the sightings occur in *garages, homes, and buildings* (high probability of non-positive samples); Another example is that Asian hornets generally build nests on the ground, but individual witness reports are Eaves or places higher than the ground... There are many differences, so we decided to use two corpora to construct text vector features.

Our operation process is as follows:

1. Collect all negative samples from more than 4,000 samples, count all word frequencies, and screen out high-frequency and easily confused characteristic words as a negative corpus.
2. Select the unique characteristic words belonging to the Asian wasp from sources such as encyclopedias and papers as the positive corpus.
3. For each piece of text information, two vectors are respectively vectorized according to the two corpora.
4. Since the text information of the witness report is generally very short and the generated vector is very sparse, it is not convenient for us to train the model. Therefore, we sum the elements of each vector (that is, the sum of the number of times all words in the corpus appear in each text) to get the final The features we use for training.

The two word clouds in the figure below are a visualized positive and negative corpus.

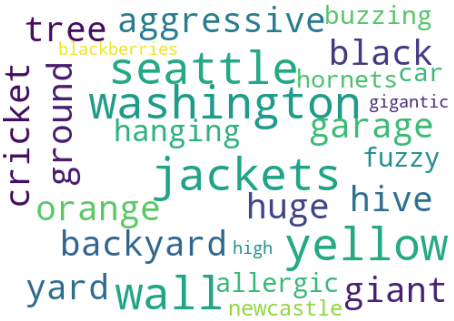

Figure 7: Negative Words Count

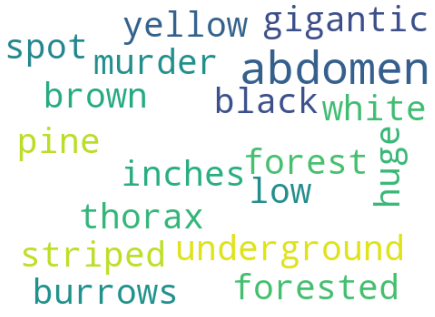

Figure 8: Positive Words Count

The histogram on the left shows us the negative corpus words extracted from more than four thousand sample texts, while the histogram on the right is the positive corpus. It can be found that because the corpus is small and the report content is short, most The text does not contain, or only contains one word.

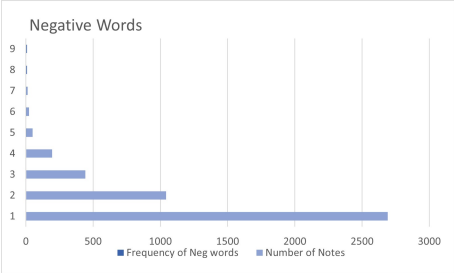

Figure 9: Negative Words Count

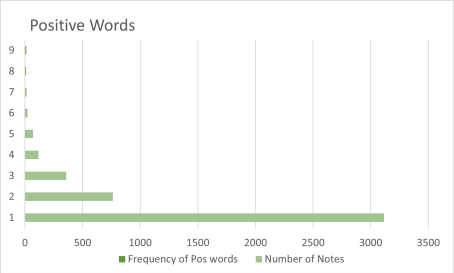

Figure 10: Positive Words Count

5.1.3 Sampling Method

Now it comes to the last step before we could finally train our classification model. After the text is vectorized, for each sample, in addition to the original latitude and longitude location information, we now have two more dimensional features (the sum of the frequency of positive and negative words). Note that the number of latitude and longitude is two dimensions greater than word frequency, so we first standardized the data set.

Next step is to augment the dataset in order to deal with the inbalance. We decided to use sampling methods to achieve that goal.Commonly used sampling methods include up-sampling, down-sampling and integrated sampling. Downsampling is an efficient method to reduce most samples, but this method requires that the original data set should be clustered into clusters. Observing Figure 5, we can find that most of the samples (negative samples) are evenly distributed, only in the lower right area. Obvious aggregation, so we do not use downsampling for processing.

We believe that the invasion process of the Asian wasp is a gradual spreading process from small to large. In any period of time, the positive samples confirmed by the report will gather within the spreading range, which inspired us to use the up-sampling method. The principle of the **SMOTE** up-sampling method is that for a sample in a minority class, the sample is expanded by randomly selecting any point on the line between the sample point and the nearest minority sample. The mathematical expression can be abbreviated as:

$$x_{new} = x + rand(0, 1) * |\hat{x} - x| \tag{14}$$

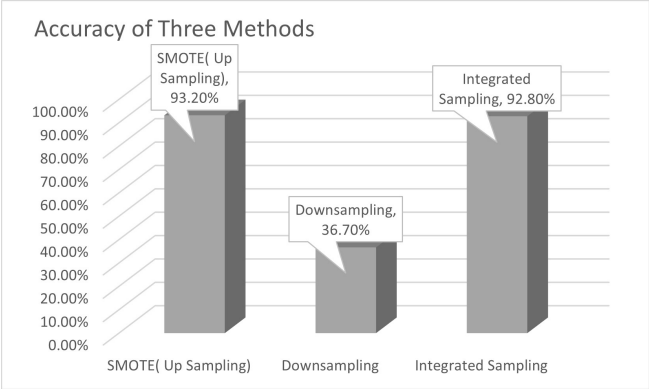

Figure 11: Accuracy for different sampling methods

We used the same linear kernel SVM and different sampling methods for training, and the results confirmed that upsampling has higher prediction accuracy than integrated sampling and downsampling (under the premise that the prediction model remains unchanged)

5.1.4 SVM For Classification

After using SMOTE sampling to augment dataset, we now first use SVM to do the classification.The reasons why we use SVM as one of our models are: First, SVM is suitable for training small-scale samples, and has better generalization and promotion capabilities than neural networks; secondly, its kernel function provides a series of s Choice. For some data sets, proper use of individual kernel functions may produce quite good classification results.

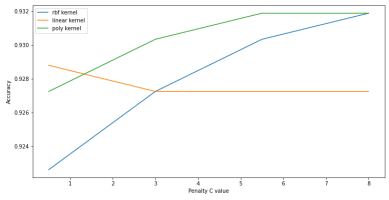

Figure 12: Accuracy varies with penalty constant C

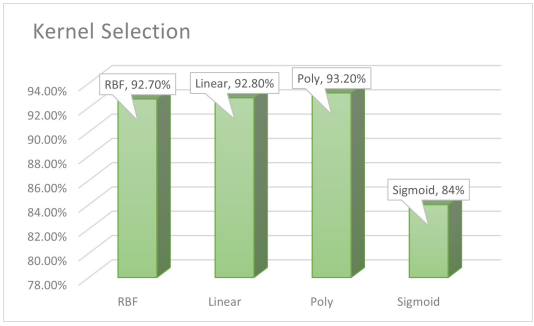

Figure 13: Accuracy of different kernels

The tuning process shows that the **linear kernel** function

$$K(x, y) = xy \tag{15}$$

and the **polynomial kernel**:

$$K(x, y) = (xy + c)^d \tag{16}$$

all perform better than the sigmoid kernel, and when the penalty constant C is low, which means that the model is underfitting, the poly kernel behaves better than the other kernels, so we suggest using the poly or linear kernel.

Different from the traditional two-class model logistic regression, SVM uses the Hinge Loss function with penalty items:

$$Loss = \sum_{i}^{N} [1 - y_i(wx_i + b)]_+ + \lambda |w|^2 \tag{17}$$

This loss function will give greater penalty weight to those significantly wrong classification samples, ensuring the accuracy of the classification results. The existence of the regularization term allows us to avoid overfitting by adjusting the penalty coefficient λ.And by using SVM for classification, our final predicting accuracy can reach to around 93%.

Although SVM has many advantages as our classification model, its biggest disadvantage is that it cannot output the probability of a classification and cannot judge the importance of each feature. These two types of information are very important to our predictive analysis, because When it comes to a new sighting report, we need to make a rough judgment on the possibility that the report is wrong (correct).

5.1.5 Random Forest for Classification

Although SVM provides kernel functions that can be selected and freely created, in current practical applications, it is not very clear how to choose kernel functions. At the same time, considering that if this species has a large-scale rapid spread in the future, the sharp increase in sample size will not be conducive to the update training of SVM, so we decided to find a better prediction model.

Random forest is a bagging algorithm used in ensemble learning. There are several decision trees in a random forest, and decision trees in the random forest are not related to each other. When a new input sample enters, each decision tree in the forest will make a judgment separately to see which category the sample belongs to (for the classification algorithm), and then see which If one category is selected the most, then predict that sample belongs to that category.

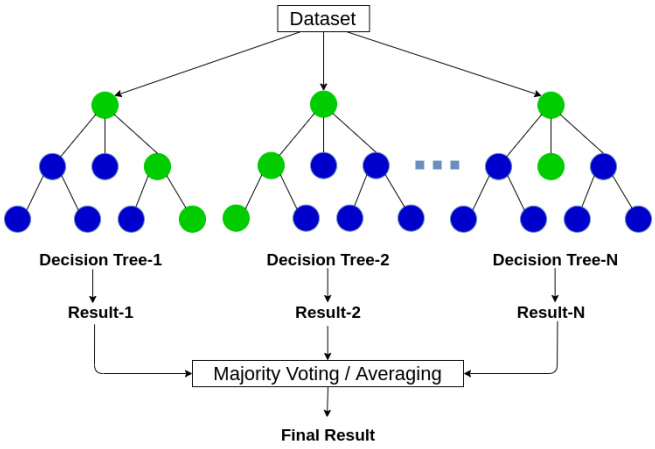

Figure 14: Bagging method (Google Image)

We use random forest mainly because of its following advantages:

1. When creating a random forest, unbiased estimation is used for generalization error, and the model has strong generalization ability.

2. The training speed is fast and can be made into a parallel method for training on large-scale data sets.

3. For unbalanced data sets, random forests can balance errors to a certain extent.

4. In the training process, random forest can detect the mutual influence between different features, and can output the importance of features. At the same time, we can output the classification probability we need.

We first adjust the parameters for the **number of base classifiers**. The increase of the base classifier will reduce the model overfitting, but will also increase the amount of calculation. The line graph on the left indicates that the model accuracy reaches its peak of 98.9% when the number of base classifiers reaches about 60, so we choose 60 as the final number of base classifiers.

The subsequent adjustment of the **maximum tree depth** showed that when the maximum tree depth is limited to 6, the prediction accuracy of the model can reach 99.2%. The **max_features** parameter in

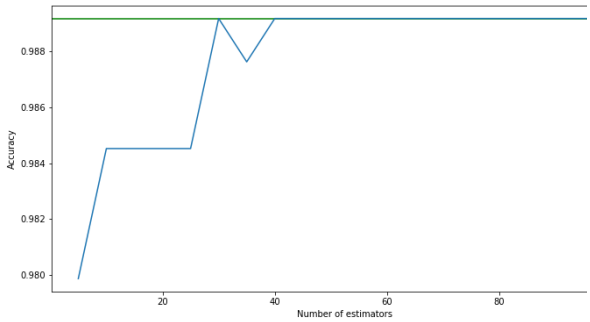

Figure 15: Changing number of estimators

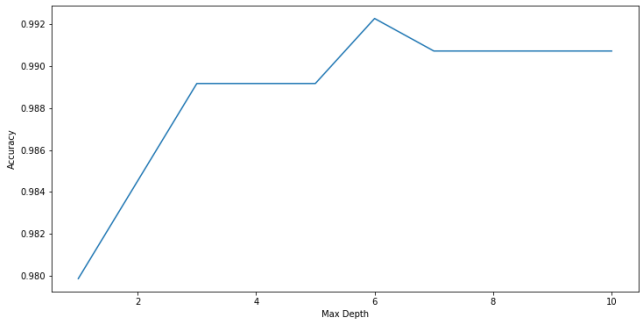

Figure 16: Changing max depth

random forest is used to control the maximum number of features used by a single decision tree. Since we only have four features in total, there is no need to adjust this parameter.

Now we have got the best random forest model, with 60 estimators and max tree depth of 6, we can out put the feature importance and probability of classification when we get a piece of new data.Simply call **clf.feature_importance_**, we get the importance ranking:[0.63,0.3,0.06,0.004], which means that the location information is the main criteria we use to judge whether the witness insect is Asian Hornets, and the latitude is most important. The notes provide very limited information.

Therefore, in the actual prediction, we also need to use the picture information taken by witnesses to further analyze it through the neural network.

5.2 CNN for Image Classification

To correctly identify whether an objective wasp belongs to the Asian giant hornet family, the provided visible data are of high importance. We are given a fairly large dataset including more than 3,000 images/videos of bees to help us build our model. It is intuitive that a Convolutional Nerual Network (CNN) (LeCun, Bengio, & Hinton, 2015) should be introduced.

5.2.1 Model Architecture

Since we only have 3,000 valid training samples and it is merely a binary classification problem, preventing series of problems like underfitting and computation overhead, it is reasonable to use a relatively lightweight model, thus, we decide to use *MobileNetV2* (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) with SOTA performance among tiny models. To adapt to the binary classification task, we further adapt *MobileNetV2* and visualization of our model is shown in Figure 17.

5.2.2 Class Imbalance

A very tricky problem that considerably affects our model performance is the data unbalance, with thousands of *Negative* samples via rarely a dozen of *Positive*. It brought some difficulties at very first step, making the training process very tough. We record the rudimentary results of the vanilla model without our optimization in .

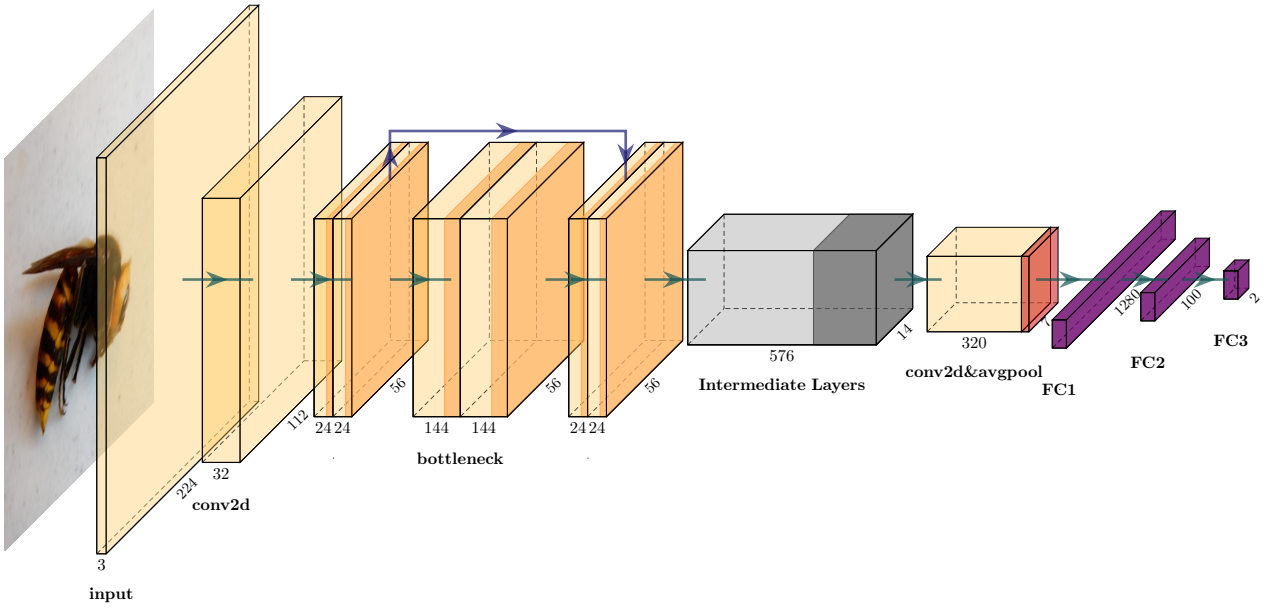

Figure 17: Structure of Our CNN Model (Intermediate Layers are Simplified by the Gray Box)

We come up with several techniques to tackle this problem:

- Data Level Augmentation
- Over Sampling
- Weighted CrossEntropy Loss

In Data Level Augmentation, we mainly use image stage operation to create my samples, which can, to some extent mitigate the lack of positive samples. We have included many common augmentation strategies: *Color*, *Brightness*, *Contrast*, *Rotation*, etc. Some examples are shown in Figure 18 and 19.

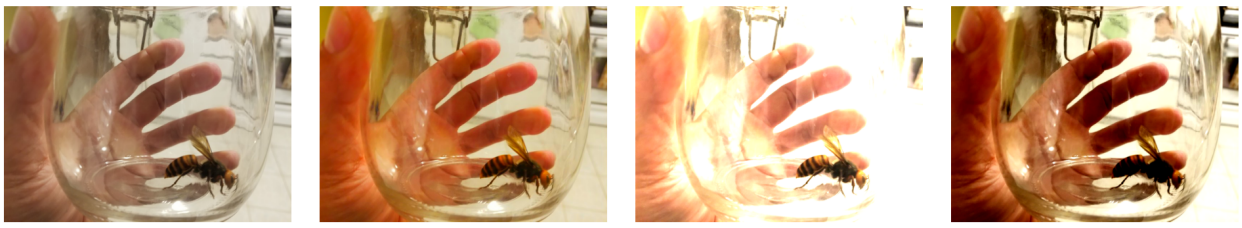

Figure 18: Data Augmentation by Image Enhance

As for Sampling, we also formulate some optimizations. We adopt the Over-Sampling (Ando & Huang, 2017) to make smooth the data unbalanced gap. Typically, as hyper-parameters, we assign a 5 / 1 ratio for positive / negative; better performance could be retrieved under better ratio.

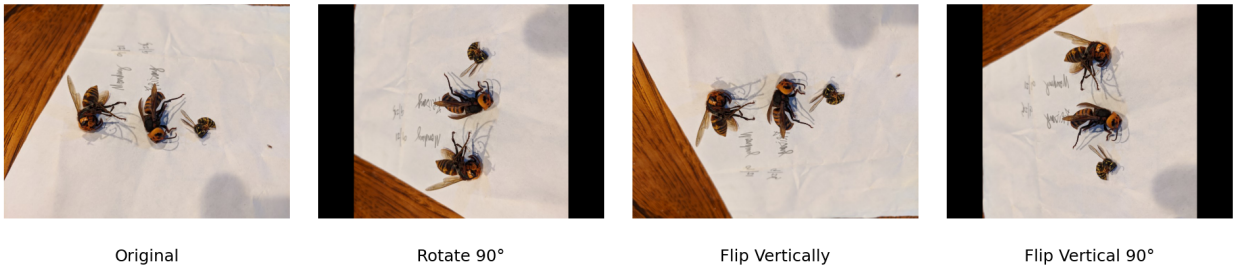

Figure 19: Data Augmentation by Physical Operation

In addition, though Data Augmentation and Over-Sampling could somehow enlarge the probability to include positive samples in a batch, it might not be capable in all scenarios. Instead of a simple CrossEntropy Loss Function:

$$\text{loss}(x, class) = -log(\frac{\exp(x[class])}{\sum_j \exp(x[j])}) = -x[class] + log(\sum_j \exp(x[j])) \tag{18}$$

We stick to the weighted CrossEntropy Loss Function, assigning particular weights to each class.

$$\text{loss}(x, class) = weight[class](-x[class] + log(\sum_j \exp(x[j])) \tag{19}$$

To save computational budgets and speeding training process, losses for each sample are averaged by weights for each mini-batches:

$$\text{loss} = \frac{\sum_{i=1}^{N} \text{loss}(i, class[i])}{\sum_{i=1}^{N} weight[class[i]]} \tag{20}$$

Trained and validated on Nvidia TITAN V VOLTA with *Pytorch*, the model performance is shown in Table: 3. Compared with baseline model (vanilla *MobileNetV2*) suffering from overfitting, our optimized model make fairly progress in picking out those suspect cases when evaluating under *Mean Recall* and *Precision*. Testing Time is the time needed for validated an image.

Optimization	Accuracy	Mean Recall	Precision	Testing Time
×	96.1%	20.7%	60.3%	0.028s
√	98.4%	83.2%	90.1%	0.029s

Table 3: CNN Model Performance

5.2.3 Iterative Update With Incremental Learning

Since new data might continue flowing in, it is essential to keep our model informed, serving to provide a more reliable prediction in the future. In Machine Learning (ML) terminology, this issue involves research topics of Life Lone Learning, Online Learning and Incremental Learning.

In Question 4's setting, the most straightforward approach is to retrain the model with the entire dataset (Joint Training), ensuring that all available samples from unknown Dataset $\mathcal{D}$ are seen by models. Regardless of its simplicity, this approach can actually preserve most of original knowledge learnt beforehand while successfully incorporating new data.

However, this method is very consuming, especially for a large and complex data or system. Many techniques arise in both the literature and industrial. Feature Extract (Parisi, Kemker, Part, Kanan, & Wermter, 2019) suggest that to promote efficiency, it is important to extract dominant samples from known dataset, which contains important information for a deep model. By only reusing these samples, can the model retain its previous knowledge.

6 Branching Process

In this section, we discuss what would constitute evidence that the pest has been eradicated in Washington State. We'll give a discrete version of the birth-death process we developed in section 4.1.

Define X_n as a random variable of the nests in n-th season and Y_i as a random variable of new colonies established by the i-th nest.

$$Pr(Y_i = k) = p_k, \quad k = 0, 1, 2, \ldots \tag{21}$$

And thus, the birth rate of a single nest $\lambda = EY_i = \sum_{k=0}^{\infty} k p_k$. If $X_{n-1} = z$, then $X_n = Y_1 + \cdots + Y_z$ and hence, $E(X_n|X_{n-1} = z) = z\lambda$. Thus,

$$E(X_n|X_{n-1}) = \lambda X_{n-1} \tag{22}$$

Therefore, by tower rule,

$$EX_n = E(E(X_n|X_{n-1})) = \lambda EX_{n-1} = \lambda^n EX_0 \tag{23}$$

If $\lambda < 1$, then $EX_n \to 0$ and hence $Pr(X_n > 0) \leq EX_n \to 0$. In this case, extinction occurs with probability 1.

By our previous model, the result indicates that if you can only find at most one new nest within 10km of the nest for two seasons, then you can be 95% confident that the pest has been eradicated in Washington State.

7 Strengths and Weaknesses

7.1 Strengths

- The birth and death process model can accurately give the specific number of pest colonies at a given time. This model gives the trend of the number of pests in the future, and accurately quantitatively analyzes the number of pests that were originally ambiguous.
- The bivariate normal distribution model can accurately give the approximate location of the new nest, and can give the confidence region. This model gives us 95% confidence in finding nests within 10km, which significantly reduces the difficulty of searching.

- The branching process model discusses the evidence of pest extinction and quantifies the rate of pest extinction. It provides a more precise method for controlling the growth of pests.
- Oversampling solves the problem of unbalanced data sets; the use of random forest algorithm can sort the importance of features, and can output the classification probability of samples. In addition, when the sample size is greatly expanded in the future, the parallel calculation of random forest can be greatly accelerated classification.

7.2 Weaknesses

- Due to lack of data, the branch process model is likely to have great contingency. There is an underfitting problem.
- Due to the lack of data, the fitting of the bivariate normal distribution has a large deviation, which reduces the accuracy of the model. It is found in the hypothesis test that the data does not have strong normality.
- Using machine learning models to train the classifier makes the overall interpretability poor. The notes in the eyewitness report contained too little available information; there were fewer words in the corpus, and the reconstruction features obtained by text vectorization were not effective enough.
- In Image Classification Task, some detailed aspects still need further improvement. Despite of the pertinent performance, our model or loss function is not specialized enough which indicates that there might exist some ways to further tune and boost it. Also, due to time limit and data shortage, we have not yet tested the aforementioned methods for continuous inflowing data, which deserves future experiments.

7.3 Improvement

- We can use more oversampling algorithms to get more data to train birth and death process models and bivariate normal distribution models.
- We can use the EM algorithm as an optimization algorithm to avoid the problem of missing data. At the same time, Bayesian inference can be used to alleviate the problem of lack of data.
- We can try to introduce Bayesian methods to classify data, such as using Bayesian neural networks, so that we will get the probability distribution of the classification results instead of just the classification labels.

8 The Link of Code

https://github.com/Aloofwisdom/MCM-2021-Problem-C.git

MEMO
TO: the Washington State Department of Agriculture
FROM: Team #2109268
DATW: Febuary 08, 2021
SUBJECT: Strategies to prioritize public reports of Asian giant hornet

---

Dear Sir/Madam,

According to your requirements, we have established models to predict the spread of Hornets and classify the reports of witnesses. We will summarize and report the results we got in this memo.

As an invasive species, the Asian Hornets may hunt and kill local bee colonies and pose a threat to people's daily lives. Given that there are very few officially confirmed reports, the first thing we need to do is to predict the number and spread of hornets. We assume that the reproduction of hornets obeys the birth and death process in the random process, and its diffusion on the map (two-dimensional plane) obeys the bivariate Gaussian distribution. The relevant parameters in the two models are estimated, and the corresponding Confidence interval. The results is as follows. 1. A single pest nest will generate approximately 0.64 new pest nests every quarter. 2. There is a 95% probability that a new pest nest will be found within 10km of the pest nest.

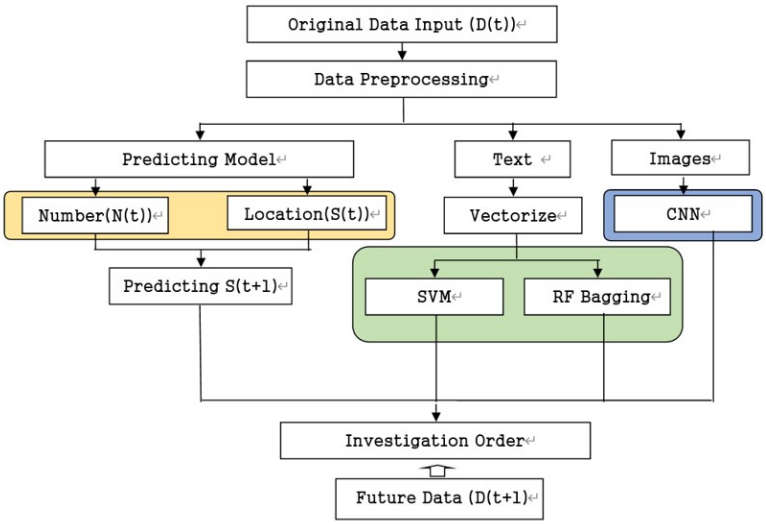

Figure 20: flow chart for our final result

For any time interval in the future, we can estimate the scope of activities of Asian Hornets through the above predicting model. Therefore, we hope to identify and classify the sighting report data generated within this scope to determine how many possibilities there are It is the real Asian Hornets, which leads to our second step.

The report provided by the witness can be split into two dimensions of information, text and pictures. For text information, we construct a positive and negative corpus, extract two new features from notes on the original latitude and longitude data, and then use SVM and random forest classifiers to train and adjust parameters respectively. The results show that the classification accuracy of SVM using poly kernel is 93%, and the random forest can also output the importance ranking and classification

probability of features when the accuracy reaches 99%, which undoubtedly provides for future report analysis Strong model support.

Based on the above model, we provide a strategy for reference to determine the order of investigation when obtaining batch reports in the future. This strategy is roughly divided into the following steps:

- For a specific time period in the future, first obtain the distribution range of hornets through the prediction model.You shall first investigate the reports within the prediction range.
- All samples that have been filtered by location are classified using three models we have trained. The classification results are sorted according to the number of positive classes from large to small, and the investigation is carried out in order according to the order. For two samples with the same number of positive classes, the investigation is conducted according to the priority of CNN>RF>SVM.
- Priority survey towards notes contain the samples with the largest number of words in the positive corpus.

If you can only find at most one new nest within 10km of the nest for two seasons, then you'll have 95% confidence to believe that the pest has been eradicated in Washington State.

We hope our suggestions and models can provide useful information for you!

Yours,sincerely
Team #2109268

References

Ando, S., & Huang, C. (2017). Deep over-sampling framework for classifying imbalanced data. *CoRR*, *abs/1704.07515*. Retrieved from http://arxiv.org/abs/1704.07515

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 05). Deep learning. *Nature*, *521*, 436-44. doi: 10.1038/nature14539

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54-71. Retrieved from https://www.sciencedirect.com/science/article/pii/S0893608019300231 doi: https://doi.org/10.1016/j.neunet.2019.01.012

Rafferty, J. P. (2019, February 7). Invasive species. *Encyclopedia Britannica*. Retrieved from https://www.britannica.com/science/invasive-species

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, *abs/1801.04381*. Retrieved from http://arxiv.org/abs/1801.04381

Skvarla, M. J. (2020, May 7). Asian giant hornets. *PennState Extension*. Retrieved from https://extension.psu.edu/asian-giant-hornets

Zhu, G., Gutierrez Illan, J., Looney, C., & Crowder, D. W. (2020). Assessing the ecological niche and invasion potential of the asian giant hornet. *Proceedings of the National Academy of Sciences*, *117*(40), 24646–24648. Retrieved from https://www.pnas.org/content/117/40/24646 doi: 10.1073/pnas.2011441117