Research on the spatio-temporal characteristics of hornets' nests based on K-means clustering

Li Mijia¹*, Ma Lei¹, Cai Zhejun²
¹ College of Science, Donghua University, Shanghai 200000, China
² School of Computer Science and Technology, Donghua University, Shanghai 200000, China
*duoerxs@163.com

Abstract-Asia giant hornet is a predator of European honeybees, invading and destroying their nests. Nowadays, the State of Washington is suffering the attack of this hornet. Due to the potential severe impact on local honeybee populations, it will probably cause a great deal of anxiety.

The most informative factors we consider are Detection Date, latitude and longitude. We propose a "time-space-state model" to evaluate the information evaluated by these two indicators. The time-space-state model can be divided into two kinds: the general model of time discretization and space continuum; The general model of space discretization and time semi-continuity. In the previous model, we mainly use K-means algorithm to determine the nest position in discrete time. In the latter model, we use the improved K-means algorithm to obtain the distribution model of hornets with seasonal variation.

Through research, we have discovered the temporal and spatial laws of hornets’ nests, which are of great significance for maintaining the stability of agricultural production.

1. Introduction

Biological invasion is a global problem in agriculture, food production and biodiversity. In invasive species, some hornets have been known to have a serious impact on honeybees, such as Asian giant hornets, which are considered as agricultural pests can destroy the entire European bee colony in a short time[1]. Therefore, discovering the temporal and spatial laws of hornets’ nests is of great significance for maintaining the stability of agricultural production.

2. Data preprocessing

To better understand the data, we carried out the target data visualization and analyzed the relationship among Detection Date, Latitude, Longitude and Lab Status. The specific results are shown in the figure 1:
Figure 1: Data visualization results

We found time outliers in the Negative ID, Positive ID, and Unprocessed categories. And the amount of Positive ID and Unprocessed data is very small. The longitude and latitude regions showed a significant centralization trend, indicating that sighting reports were relatively concentrated in a certain region.

Raw data is prone to loss, noise, outliers and other problems due to resources and collection methods. Data cleaning is used to find and deal with these problems, turning "dirty data" into "clean data" and improving the quality of data. Data cleaning is an indispensable part in data preprocessing, which mainly includes missing value processing, outlier detection and processing. Through the screening of eigenvalues, we found that there are 13 abnormal data of Detection Date, which accounted for 0.29% of the 4406 data reviewed. To facilitate subsequent data application, we deleted these abnormal data.

If a feature data has a large proportion of missing values or large data differences, it means that the feature contains fewer valid values, and it has less impact on Lab Status prediction. Therefore, Lab Comments are deleted and not used as input features.

According to our observations on Notes, most witnesses posted more casual statements and less relevant to Lab Status, so we do not consider Notes as a characteristic.

Due to the different image formats given, we use python to clean all files in 2021MCM_ProblemC_Files into image formats, and then divide them into four folders according to the four attributes of Lab Status, named positive, negative, unverified, and unprocessed. From 2021MCM_ProblemC_I images by GlobalID.xlsx we can get the sighting event dataset, and from 2021MCMProblemC_DataSet.xlsx we can get the picture record dataset. In order to obtain all the characteristic items of each sighting completely, we merged these two files. In addition, since multiple pictures may correspond to one sighting event, we match such pictures and obtain a cross data set (the number of sightings without pictures).

In order to process the image data more conveniently later, we first regularize the images. Due to the serious lack of the positive data set, we crawled the picture of the Asian Hornet from the given background information and the authoritative website[3]. Then, we use ImageDataGenerator for data enhancement. ImageDataGenerator is an image generator in keras, which generates batches of tensor image data through real-time data enhancement and can iterate in a loop. We perform rotation transformation, flip transformation, zoom transformation, translation transformation, noise disturbance, and cross-cut transformation on all collected positive images, with the purpose of expanding the dataset size and enhancing the generalization ability of the model. As a result, the positive data set is expanded to approximate the number of the negative data set, which is 2070 images.

With the comprehensive consideration of the area and shape change after projection, we use the Miller coordinate system as the projected coordinate system. Miller's cylindrical projection was developed by O.M. Miller in 1942. [9] It uses the tangent of the developable surface of the cylindrical projection at the equator. Compared with the Mercator projection, it greatly improved the distortion of the area and shape.
We import the latitude and longitude matrix of 4440 records into anaconda, and realize the mutual conversion between latitude and longitude and Miller coordinate system by programming. The corresponding x, y coordinates have replaced the original longitude and latitude columns in the data set (Longitude-x, Latitude-y) (unit: meter). **Subsequent discussions on the spatial position are based on the projected coordinate system and will not be repeated.**

3. **Time-space-state model**

   According to the habits of the wasp colony, we can get the following information:
   
   - We approximately think that the cluster point is the nest location.
   - Each hornet is not more than 8 kilometers away from its nest, that is, the location of each hornet and its clustering center is not more than 8 kilometers.
   - The first generation of nests (here refers to the location of the nest in 2019) and the second generation of nests (here refers to the location of the nest in 2020) do not exceed 30 kilometres, allowing multiple same-generation nests to exist within a certain range.

3.1. **General model of time discretization and spatial continuity**

   Based on the above mentioned characteristics of this bee colony, it is not difficult to find that the location of its nests is replaced by years, and each nest exists for up to one year (here we assume one year). Taking into account its spatial characteristics, we can use coordinate-based clustering algorithm (an algorithm that automatically divides a pile of unlabeled data into several categories) to determine the nest position in this discrete time.

   Negative ID records are wrong sightings (from error identification or other), which have no significance in analyzing the propagation model of hornets; Unverified records may contain some factual records, but this factor is discussed at the end of this article. A preliminary discussion is made on the verified records.

   Based on data preprocessing, we know that Lab Status is positive for a total of 14 data, including 5 in 2019 and 9 in 2020.

   With sample distance as heuristic rules, K-means clustering model is undoubtedly a preferred spatial clustering. By repeating the process of moving the class center point, to move the class center point to the average position of its contained members, then reclassify its internal members, and finally reach the global stability. The model flow chart is shown in Figure 3.

   Under the above assumptions, we cluster the five data in 2019 and the nine data in 2020 respectively.
to calculate the cost function. The results are shown in Figure 4:

![Figure 4: k-means model flow chart](image)

It can be intuitively seen that the corresponding curve turning points in 2019 and 2020 both occur when \( k=2 \), so the optimal clustering number is 2. Then, \( k=2 \) is taken as the number of clustering to cluster the data of two years respectively, and the results are put into a picture for visual analysis. As shown in Figure 5:

![Figure 5: Cluster Center Exploration](image)

3.2. The general model of spatial discretization and time semi-continuity

Given the approximate nest location distribution, we wanted to further study the seasonal distribution characteristics of the number of wasps (Positive ID) within the nest area (8 km).

According to the law of large numbers, when the sample size is sufficient, the overall distribution characteristics can be simulated through the distribution of samples. The verified Positive ID has only 14 records, far from meeting the requirement of large samples. Therefore, it is necessary to first classify unverifiable records (over 2000). Taking 2020 as an example, the discussion of time factor is added on the basis of the original model, and clustering is carried out according to the three indicators of time, \( X \) and \( Y \) (the number of positive IDs is used as the number of clusters). This clustering process of data set containing only time-space features is transformed into spatio-temporal clustering. The standard of spatio-temporal clustering is proximity principle. The proximity principle is that the sample belongs to the category represented by the nearest cluster point.

Since Positive's correct sightings were both in 19 and 20 years old, only the first two were analyzed. We divide all data (not including Negative) into 2019, 2020 and others by year. Among them, there are 107 data in 2019, 2206 data in 2020 and 58 other data. We delete other data because it has no practical significance in this model. In this model, the data of 2019 and 2020 are mainly discussed. Since in this model, we want to analyze the distribution characteristics of wasp population with seasonal changes, we left five useful feature items: GlobalID, Detection Date, Lab Status, Latitude, and Longitude.

To eliminate the influence of dimension between time and space coordinates, all features are unified
into a roughly identical numerical range; So that different levels of indicators can be compared and weighted.

Matrix $A$ is denoted as the normalized and centralized matrix, and the dimension of this matrix is $2313 \times 5$. Each row represents a record. Columns 1 – 3 are the normalized and standardized results of $T, X, Y$, respectively. where $T$ represents the interval between Detection Date and January 1 of the same year, for example, $T$ corresponding to the original 2019/1/2 is 1; $X$ and $Y$ represent the transformed coordinates of Latitude and Longitude respectively. Column 4 is an integer variable with values 1,2 and 3 (corresponding to Positive ID, Unverified, Unprocessed, respectively).

K-Means algorithm is a commonly used clustering algorithm, but its algorithm itself has certain problems, and it is generally difficult to comprehensively consider the training time and model accuracy. For the clustering process involving a large amount of data (2313 records, 4 features), K-means is obviously not perfect. As an improvement of K-means clustering method, Mini Batch Kmeans uses Mini Batch (batch processing method to calculate the distance between data points. The advantage is that it is not necessary to use all the data samples in the calculation process, but to extract some samples from different types of samples to represent their respective types of calculation, so as to deal with large data sets. [4] [5]

It can be seen from the figure above that, compared with K-means algorithm, Mini Batch Kmeans can greatly reduce the calculation time while keeping the clustering accuracy as far as possible.

We also need to confirm the number of clustering, and we directly call this method through the Python machine learning sklearn library. At the same time, we cluster the six possible clustering numbers from 7 to 10 respectively, and gather the clustering results and contour coefficient scores in a single figure.

![Figure 6: Clustering results and contour coefficients](image)

According to the above figure, we can see that the optimal contour coefficient corresponds to the number of clustering centers, $K=7$, and the contour coefficient is 0.42. In this case, the simulation fitting is good.

Thus, we get seven categories in the case of optimal clustering. Through observation, we found that 9 groups of Positive records in 2020 were respectively in the following categories:

| The Ith class | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|--------------|---|---|---|---|---|---|---|
| Number of positive | 4 | 1 | 0 | 1 | 1 | 2 | 2 |

We want to find records that are closer to reality, so we need to measure the relative proximity of each event. For each record, the relative proximity can be calculated as follows:

- Locate the record’s class.
- Lock the number of Positive ID in the current class and the spatio-temporal characteristics of each.

If the number is 0, then the proximity is infinity.

- Since de-dimensionalization has been carried out, the distance index between this record and each Positive ID is directly measured by Euclidean distance.
- Take the smallest distance index as the relative proximity of the record to the real event.

All records are traversed as described above, sorted in ascending order according to their relative proximity. Thus, we obtained 100 record features that were most similar to the Positive ID situation, including 14 true and 86 unverifiable data. According to the time series characteristics of the 100
samples, a rough positive distribution model with seasonal variations can be established.

Rough observations of correct sightings and analysis of colony behavior suggest a rough distribution of the population, which should be almost invisible in winter, increasing in spring, summer, and early September after spawning, then decreasing by the end of November. The activity around the different nests is uniformly distributed independently, so only the relative position of each record point to the nearest nest needs to be considered. Since the breeding pattern of the swarm is the same from year to year, you only need to consider the month and day at the time. A histogram is drawn using the distribution of the 100 correct sightings (86 with data enhancement) in different months.

Figure 7: histogram of the relationship between the correct number of sightings and month

4. Conclusion
We first constructed a general model of discretization in time and continuity in space. Based on the characteristics of the bee colony, we found that the location of its nests changes in units of years. At the same time, considering its spatial characteristics, we used a coordinate-based clustering algorithm to determine the nest position in this discrete time.

Next, we constructed a general model of spatial discretization and time semi-continuity. With the help of the approximate nest distribution obtained by the previous model, we use the Mini Batch Kmeans method to obtain the seasonal variation of the number of wasps within the nest activity range.

With the help of the above TSS model, we can find the spatiotemporal information that fits the actual hornets' nests, which is of great significance for maintaining the stability of agricultural production.

References
[1] ping Zhu,Javier Gutierrez Illan, Chris Looney, David W. Crowder. Assessing the ecological niche and invasion potential of the Asian giant hornet[J]. Proceedings of the National Academy of Sciences, 2020, 117(40).
[2] 2021MCM_ProblemC_Vespamandarina.pdf
[3] https://www.bigbadtoystore.com/Product/VariationDetails/128949
[4] https://blog.csdn.net/cht5600/article/details/76014573
[5] Wang Xiaoning, Wang Xiao. Questionnaire segmentation design based on K-means clustering algorithm [J]. Research World, 2021(02): 74-81.