Fight for Washington State: Can Artificial Intelligence Beat the Asian Giant Hornet?
为华盛顿州而战: 人工智能能打败亚洲大黄蜂吗?

2023 年 9 月 21 日

摘要

Recently, the Asian giant hornet has been observed in Washington State, which may cause damage to the ecosystem in the future. Therefore, the Washington State Department of Agriculture (WSDA) has provided large amounts of observations on the species, hoping to get our assistance

最近，在华盛顿州发现了亚洲大黄蜂，这可能会对未来的生态系统造成破坏。因此，华盛顿州农业部 (WSDA) 提供了大量的物种观测资料，希望得到我们的帮助。

For problem 1, we propose two metrics: the resource competition coefficient and the environmental friendliness to construct a time-step difference equation, and simulate the population dispersal of the Asian giant hornet. We predict the distribution of nests in Washington State within 10 years, and gain the range of activities of an Asian giant hornet by adding noise. The results show that if no measures are taken against the spread of Asian giant hornets, the number of the species will show an approximate exponential growth at the initial stage in Washington State. To evaluate the accuracy of the model, we use **Logistic Growth Model （Logistic 增长模型）** to test the accuracy of the model. The loss is 0.076, and the fastest growing year is the seventh year.These results indicate that we need to control the number of nests early.

针对问题一，我们提出了资源竞争系数和环境友好度两个指标，构建了一个时间步长差分方程，并模拟了亚洲大黄蜂的种群扩散。我们预测了 10 年内华盛顿州的巢分布，并通过添加噪音获得了亚洲大黄蜂的活动范围。结果表明，如果不采取措施防止亚洲大黄蜂的扩散，该物种的数量将在华盛顿州的初始阶段呈现近似指数增长。为了评估模型的准确性，我们使用 Logistic 增长模型来检验模型的准确性。亏损 0.076，增长最快的年份是第七年。这些结果表明，我们需要及早控制巢的数量。

In problem 2, we divide it into feature extraction and image classification. For the former, we establish a model based on auto-encoder to condense images information, of which the minimum testing loss dips to 0.0274. For the latter, we build three models, **the Binary Logistic Regression（二元逻辑回归）**, **the Support Vector Machine（支持向量机）**, and **the Convolutional Neural Network（卷积神经网络）** (CNN).The highest accuracy of the three models on the testing set is 0.5303, 0.5758, and 0.8030 respectively, so we choose CNN as our classification model. Finally, we summarize the features of negative images from three aspects: species characteristics, subject definition and background softness.

在问题二中，我们将其分为特征提取和图像分类。对于前者，我们建立了一个基于自编码器的模型来压缩图像信息，该模型的最小测试损失降至 0.0274。对于后者，我们建立了三个模型，二元逻辑回归，支持向量机和卷积神经网络 (CNN)。三种模型在测试集上的最高准确率分别为 0.5303、0.5758 和 0.8030，因此我们选择 CNN 作为我们的分类模型。最后，从物种特征、主体清晰度和背景柔和度三个方面总结了负面图像的特征。

In terms of problem 3, we conduct **Agglomerative Clustering （聚类）** according to the latitude and longitude of each sighting with unverified or unprocessed status, and divide them into 5 classes. Then we define the priority of a region based on the positive probabilities of input images in this area. Through analysis, we find that Seattle is located in the center of the highest probability area.

对于问题三，我们根据每个未验证或未处理状态的目击事件的纬度和经度进行聚类，并将其分为 5 类。然后根据输入图像在该区域的正概率定义该区域的优先级。通过分析，我们发现西雅图位于最高概率区域的中心。

With regard to problem 4, we update the model with different intervals to find optimal update time interval. Selected indicators include the testing loss of auto-encoder and the accuracy of CNN in testing dataset. The optimal time interval for updating models is defined as the **abscissa （横坐标）** corresponding to the **extremum （极值）** of the **numerical derivative（数值导数）** of the time-varying indicators. Finally, we get the following conclusions: the auto-encoder needs to be updated every four months, and the CNN needs to be updated every three months.

关于问题四，我们用不同的时间间隔更新模型，以找到最佳的更新时间间隔。选择的指标包括自动编码器的测试损失和 CNN 在测试数据集中的准确度。更新模型的最佳时间间隔定义为为与时变指标数值导数极值相对应的横座标。时变指标数值导数极值所对应的尾数。最后，我们得出以下结论：自动编码器需要每四个月更新一次。需要每四个月更新一次，CNN 需要每三个月更新一次。

As for problem 5, we design a number of variables based on observational data to characterize changes in the number of nests. We eliminate the influence of the image

recognition model based on **Bayesian inference （贝叶斯推理）**. When K **converges to （收敛于，趋近于）** 0, we think that the pest has been eradicate.

至于问题五，我们根据观测数据设计了一些变量来描述巢数量的变化。我们根据贝叶斯推理排除了图像识别模型的影响。的影响。当 K 收敛到 0 时，我们认为害虫已被消灭。

Last but not least, we summarize the suggestions and write a memorandum for the WSDA, to assist relevant departments in biological control.

最后，我们总结了这些建议，并为 WSDA 撰写了一份备忘录、协助相关部门开展生物防治工作。

1 Introduction

1.1 Problem Background

The Asian giant hornet is the largest hornet in the world, which is native to East Asia, South Asia, mainland Southeast Asia and some far east parts of the Russian. Recently, it was discovered in the American Northwest at the end of 2019, and there had been lots of sightings in 2020. In fact, the invasion of the Asian giant hornet is not an isolated incident. In 2004, it first appeared in Europe, and then began to spread to Spain, Belgium, Portugal and Italy, rapidly. According to European studies, the propagation speed of Asian giant hornet can reach 49.5 kilometers per year.

亚洲大黄蜂是世界上最大的大黄蜂，原产于东亚、南亚、东南亚大陆和俄罗斯远东地区。最近，它于于 2019 年底在美国西北部被发现，并在 2020 年被大量发现。在 2020 年。事实上，亚洲大胡蜂的入侵并非个案。2004 年它首次出现在欧洲，随后开始蔓延到西班牙、比利时、葡萄牙和意大利、迅速蔓延。根据欧洲的研究，亚洲大黄蜂的传播速度可达 49.5 公里。

1.2 Clarifications and Restatements

In this problem, we are given the data of observations on the Asian giant hornet, which is perceived by local citizens, including the time, location, latitude and longitude, and the corresponding observation photos. Washington State Department of Agriculture divided observations into four status based on photos, **namely （即是）**: positive, negative, unverified, and unprocessed. We will solve the following problems based on the historial observations:

在这个问题中，我们得到了关于亚洲大黄蜂的观测数据。包括时间、地点、经纬度和相应的观测照片。华盛顿州农业部根据照片将观察结果分为四种状态，即：正面、负面、未核实和未处理。我们将根据历史观测资料解决以下问题：

1. Based on positive observations and corresponding latitudes and longitudes, combined with

the biological characteristics of the Asian giant hornet, forecast the short-term spread of the species and analyze the accuracy of the prediction.

根据正面观测结果和相应的经纬度，结合亚洲大胡蜂的生物学特征结合亚洲大胡蜂的生物学特征，预测该物种的短期传播，并分析预测的准确性; 并分析预测的准确性

2. Establish a model for judging whether a photo contains at least an Asian giant hornet, and analyze the image features that make the model output “Negative” results;

   建立一个模型来判断一张照片是否至少包含一只亚洲大黄蜂、并分析使模型输出” 否定” 结果的图像特征；

3. Based on the above classification model, define the government’s priority in handling citizen observations, for the reason that some unprocessed or unverified observations are most likely to be positive and require the government to explore;

   根据上述分类模式，确定政府处理公民观测数据的优先次序。公民意见，因为一些未经处理或未经核实的意见很可能是积极的，需要政府去探索因为一些未经处理或验证的观测结果很可能是积极的，需要政府进行探索；

4. Consider that more sightings have been observed, determine the update method and update frequency of the model;

   考虑到已观察到更多的目击事件，确定模型的更新方法和更新频率；

5. According to the model above, analyze what indicators of Washington State has achieved, it can be said that the pest has been “eliminated”;

   根据上述模型，分析华盛顿州实现了哪些指标、可以说已经” 消灭” 了害虫

6. Submit a memorandum to to the Washington State Department of Agriculture to comprehensively supplement our research results and take corresponding protective measures at the same time.

   向华盛顿州农业部提交备忘录，全面补充我们的研究成果，同时采取相应的保护措施。

1.3 Our Work

In problem 1, we construct and simulate a time-step difference equation with two metrics: the resource competition coefficient and the environmental friendliness. We use this model to predict the distribution of nests in Washington State. Then, we gain the range of activities of a single Asian giant hornet by adding noise. After that, we use the Logistic Growth Model to test the accuracy of the model.

在问题一中，我们用两个指标构建并模拟了一个时步差分方程：资源竞争系数和环境友好程度。我们用这个模型来预测华盛顿州的鸟巢分布。然后，我们通过添加噪声来获得单只亚洲大黄蜂的活动范围。之后，使用逻辑增长模型来检验模型的准确性。

We divide problem 2 into two subproblems: feature extraction, and image classification. In order to reduce the impact of sample imbalance, we apply image flipping and Borderline-SMOTE methods for data augmentation first, and then divide the data into the **training set（训练集）** and the **validation set（测试集）** (testing set). Next, we utilize auto-encoder to collect key features of images. To prevent **over-fitting（过度拟合）**, we store the best performance model in the testing set **for subsequent use（供后续使用）**. Next, we establish three models, the Binary Logistic Regression, the Support Vector Machine, and the Convolutional Neural Network. After comparing their accuracy on the testing set, we choose CNN as our image classification model. Finally, we summarize the main features of having negative labels from three aspects: species characteristics, subject definition and background softness.

我们将问题 2 分成两个子问题：特征提取和图像分类。为了减少样本不平衡的影响，我们首先使用图像翻转和边界线-SMOTE 方法进行数据扩充，然后将数据分为训练集和验证集（测试集）。训练集和验证集（测试集）。接下来，我们利用自动编码器收集图像的关键特征。图像的关键特征。为防止过度拟合，我们将性能最佳的模型存储在测试集中，以供后续使用。存储在测试集中，以备后续使用。接下来，我们建立了三个模型，分别是二元逻辑回归、支持向量机和卷积神经网络。在比较了它们在测试集上的准确性后，我们选择了卷积神经网络作为图像分类模型。最后，我们从物种特征、主体清晰度和背景柔和度三个方面总结了有负面标签的主要特征。

As for problem 3, we take the scope of activities of one-time inspection by the government into account, and conduct Agglomerative Clustering according to the latitude and longitude of each sighting. Then we define the priority of a region based on the **positive probabilities（正向概率）** of input images.

With regard to problem 4, we discuss the effect of updating the model with different time intervals on the testing results, including the loss of auto-encoder and the accuracy of CNN. The optimal time interval of updating models is defined as the abscissa corresponding to the extremum of the numerical derivative of the function, which is the maximum value for auto-encoder, and the opposite for CNN.

In terms of problem 5, the number of nests, the range of spatial distribution and the positive ratio in observations can all reflect the growth of pests. We design a number of variables to characterize changes in the number of nests. Since the image recognition model has a certain error probability, we have made a correction based on Bayesian inference.

关于问题四，我们讨论了以不同时间间隔更新模型对测试结果的影响，包括自动编码器的损失和 CNN 的准确性。对测试结果的影响，包括自动编码器的损失和 CNN 的准确性。更新模型的更新模型的最佳时间间隔定义为函数数值导数的极值（即自动编码器的最大值）所对应的横座标。值，而 CNN 则相反。

就问题五而言，巢穴数量、空间分布范围和观测数据中的正比例都能反映观测数据的增长情况。比值都能反映害虫的增长情况。我们设计了一些变量来来描述虫巢数量的变化。由于图像识别模型有一定的误差概率，我们根据贝叶斯推理进行了修正。

2 Reasonable Assumptions

Assumptions about the data provided.

Assumptions about the behavior of Asian giant hornet

The environment is similar in different regions, so **fitted hyperparameters（拟合超参数）** in a certain small area can be regarded as constants. Therefore, hyperparameters can be promoted throughout Washington State.

不同区域的环境是相似的，因此在某个小区域内的拟合超参数可以被视为常数。因此，超参数可以在整个在整个华盛顿州推广。

3 Problem 1: Propagation Simulation Based on Yearly Time-Step Difference Equation

4 Problem 2: Image Recognition Model Based on Auto-Encoder and Convolutional Neural Network