Computer Science and Application 计算机科学与应用, 2021, 11(6), 1802-1822
Published Online June 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.116186

Hans汉斯

胡蜂预测与防护系统的设计

张瀚文[1]，张文浩[1]，郭金翟[1]，张恺雯[2]，徐　阳[1]

[1]曲阜师范大学，计算机学院，山东　日照
[2]曲阜师范大学，地理与旅游学院，山东　日照
Email: rzzhw8811@163.com, zhangkaiwen228@163.com

收稿日期：2021年5月28日；录用日期：2021年6月21日；发布日期：2021年6月28日

摘　要

为解决当今世界很多地区出现虫灾的问题，本研究以2021年1月大规模出现在美国华盛顿州亚洲大黄蜂(胡蜂)为例，通过胡蜂出现地理位置，使用灰色系统预测GM(1,1)结合ARIMA预测下一次胡蜂出现的位置来构建虫灾预测与防控系统。该系统具有在灾害区域中预测蜂群的动态、对目击者上传的图像等信息进行核实、实时进行系统更新、预测灾害结束时间的功能。研究对当地用户的评论与图像信息使用CNN结合SVM进行核实，对数据进行准确的筛选来确定虫灾严重的区域。同时，分析两次目击胡蜂的时间进行，对模型更新时间进行预测得到变动的区间。最终通过分析整个蜂群的动态，预测灾害的结束时间，从而提高病虫害防治工作情况的动态、信息化管理。

关键词

时间序列预测，灰度预测，图像分类，卷积神经网络，半监督支持向量机

Design of Vespa Forecast and Protection System

Hanwen Zhang[1], Wenhao Zhang[1], Jindi Guo[1], Kaiwen Zhang[2], Yang Xu[1]

[1]IT Academy, Qufu Normal University, Rizhao Shandong
[2]School of Geography and Tourism, Qufu Normal University, Rizhao Shandong
Email: rzzhw8811@163.com, zhangkaiwen228@163.com

Received: May 28th, 2021; accepted: Jun. 21st, 2021; published: Jun. 28th, 2021

Abstract

In order to solve the problem of insect plagues in many parts of the world today, this study uses the large-scale appearance of Asian hornet (Vespa) in Washington State in the United States in January 2021 as an example. The geographical location of the wasp is used to predict GM (1,1) us-

文章引用: 张瀚文, 张文浩, 郭金翟, 张恺雯, 徐阳. 胡蜂预测与防护系统的设计[J]. 计算机科学与应用, 2021, 11(6): 1802-1822. DOI: 10.12677/csa.2021.116186

ing the gray system, combining with ARIMA to predict the location of the next wasp, build a pest prediction and prevention system. The system has the functions of predicting the dynamics of the bee colony in the disaster area, verifying information such as images uploaded by witnesses, updating the system in real time, and predicting when the disaster will end. The research uses CNN combined with SVM to verify the comments and image information of local users, and accurately screens the data to identify areas with severe pests. At the same time, the time of two sightings of the wasp is analyzed, and the model update time is predicted to obtain the range of change. By analyzing the dynamics of the entire bee colony and predicting the end time of the disaster, we can improve the dynamic and information management of the pest control work.

Keywords

Time Series Prediction, Gray Prediction, Image Classification, Convolutional Neural Network, Semi-Supervised Support Vector Machine

1. 引言

1.1. 研究背景

胡蜂是世界上最大的黄蜂种类，其蜂巢出现的速度很快。胡峰是欧洲蜜蜂的掠食者，入侵并破坏它们的巢穴。少量的胡峰能够在短时间内摧毁整个欧洲蜜蜂群落，从而被认为是农业中其他害虫的掠食者[1]。胡峰的生命周期与其他许多黄蜂相似，受精后的蜂后在春天出现，并产生一个新的群体，秋天，新蜂后离开巢，在土里过冬，等待春天的到来[2]。一只新蜂后的筑巢距离估计为 30 公里。由于胡蜂对当地蜂群存在严重的潜在影响，胡蜂的存在会引起很大的危害[3]。华盛顿州已经建立了求助热线和一个网站，供人们报告这些胡峰的目击情况，然后根据这些来自公众的目击情况，决定如何优先分配有限的资源进行防治，以进行后续调查[4] [5]。虽然一些报告已被确定为胡蜂，但许多其他目击者却被证明是其他类型的昆虫。

1.2. 研究目的

本文旨在通过绘制蜜蜂分布图，得到“簇”，计算每个“簇”的平均向量，进而计算“簇”中每个样本到中心的距离。根据距离进行时间序列分析，得到预测结果的置信区间和距离簇中心的距离。

通过情感分析和图像识别对数据进行量化，采用半监督支持向量机(TSVM)对数据进行分类[6]，改进卷积神经网络(CNN)结合 SVM 对图像进行分析，通过设置阈值确定可信度高的目击信息，然后利用这些数据在地图上进行可视化。

利用通过 CNN + SVM 分类获得的正确样本用于计算时间序列。由于时间序列是季节性的，因此使用季节性 ARIMA 预测并得到结果。对于报告的可信度，可以使用我们的预测模型来获得可信度得分。因此，将该问题转化为时间序列的近似预测问题。

2. 模型

2.1. 模型假设

假设胡蜂在一个地区的运动是不规则的布朗运动。假设实验室状把阳性 ID 作为正例，阴性 ID 作为

反例，且数据准确。人类不会通过预防和治疗措施来控制和消灭胡蜂，但是只有通过胡蜂迁徙才能实现消灭一个地区的胡蜂。文本分数的计算仅假设用户的情感因素。

2.2. 符号说明

文中所涉及到的符号在此做统一交代。主要是将有规律性的变量用同一字母不同下标标记，阐述文中经常出现的特殊符号或变量，如表 1 所示。

Table 1. System symbol description
表 1. 系统符号说明

符号	含义
$Dist(i,j)$	i 、 j 两点之间的距离
X_i	第 i 次累加生成序列
Z_i	X_i 与 X_i 结果序列的均值生成序列
D_l	标记的数据集
D_u	无标记的数据集
ξ_i	第 i 个样本点对应的松弛因子
C_l	有标记数据集的平衡因子
C_u	无标记数据集的平衡因子
u_+、u_-	正负例样本数

3. 系统功能

3.1. 蜂群的局部变动预测功能

通过查阅资料，胡蜂的运动主要分为两种：一是蜂巢确定后胡蜂在局部区域的运动。二是胡蜂在来年春天的大规模迁徙。这里对胡蜂局部变动预测功能进行阐述。

首先，绘制一个胡蜂分布图如图 1 所示，根据地图上的每个“簇”，计算出一个簇的簇中心向量，通过计算得到一个簇的簇中心向量，并计算出所有的样本到簇中心的距离。

其次，使用 ARIMA 模型进行初步预测，得到预测结果的置信区间。根据 AIC，BIC 检验图像，确定使用 ARIMA(2.1,1)模型得到置信区间和预测结果。考虑到真正胡蜂观测数据数量较少，因此采用灰度预测算法预测下一个运动的距离。最终的结果是，以聚类中心向量为圆心，预测距离为半径的圆。

最后，检测预测半径是否在 ARIMA(2,1,1)的置信区间内，对模型进行评价。

根据纬度和经度，两点之间的距离可按下式计算：

设第一点 A 的经度和纬度为(LonA, LatA)。第二点 B 的经度和纬度是(LonB, LatB)，根据 0 度经度线，经度在东方是正的，经度在西方是负的。那么这两个点以上处理后算作(MLonA，MLatA)和(MLonB, MLatB)。然后根据三角推导，可以得到以下公式来计算两点之间的距离：

$$C = \sin(MlatA)*\sin(MlatB)*\cos(MlonA - MlonB) + \cos(MlatA)*\cos(MlatB)$$

$$Dist(A,B) = R*\arccos(C)*\frac{\pi}{180}$$

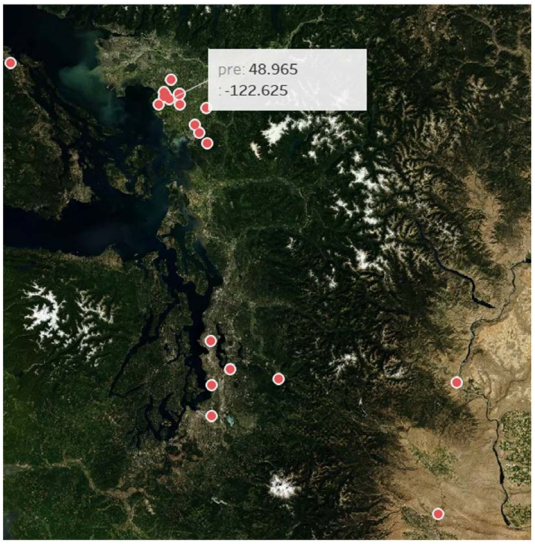


Figure 1. Hu Feng's distribution map
图 1. 胡峰的分布图

可以发现当前时间段，美国西海岸的蜜蜂可以分为两个“簇”，首先我们假设蜜蜂的移动具有布朗运动的性质，再对蜜蜂进行分簇后，每个簇首先计算中心向量，以第一个簇为例，计算结果如下：

纬度	经度
48.92826662	−122.57398731

对于每个数据点计算与簇均值向量的距离，相关距离计算如下：

坐标点	X1	X2	X3	X4
距离	10.212	10.466	27.845	7.046

使用时间序列[7]对簇中样本的距离进行预测，首先对距离序列按照时间排序后，进行时间序列的平稳性检测，并使用 ADF 和单位根检验发现时间序列不具有平稳性，因此通过差分进行平稳序列的转化操作，对一阶差分得到的平稳时间序列使用自相关和偏自相关检验，确定 AR 和 MA 的参数，ACF 与 PACF 的结果图如图 2：

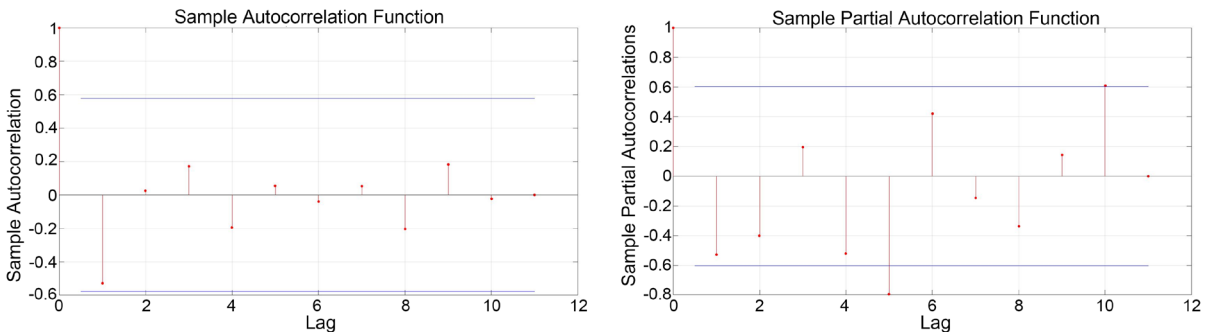


Figure 2. ACF and PACF inspection results
图 2. ACF 和 PACF 检验结果

可以从图像中看到 ACF 从 1 开始是一个明显的截尾，因此选定 MA 为 1 阶。PACF 为拖尾，因此通过 AIC 和 BIC 进行暴力确定阶数，最后得到的模型为 ARIMA(2,1,1)。

为了确定时间序列是否具有随机性，因此进行白噪声检验，判断得到残差是否服从标准正态分布，得到的检验图像如图 3 所示：

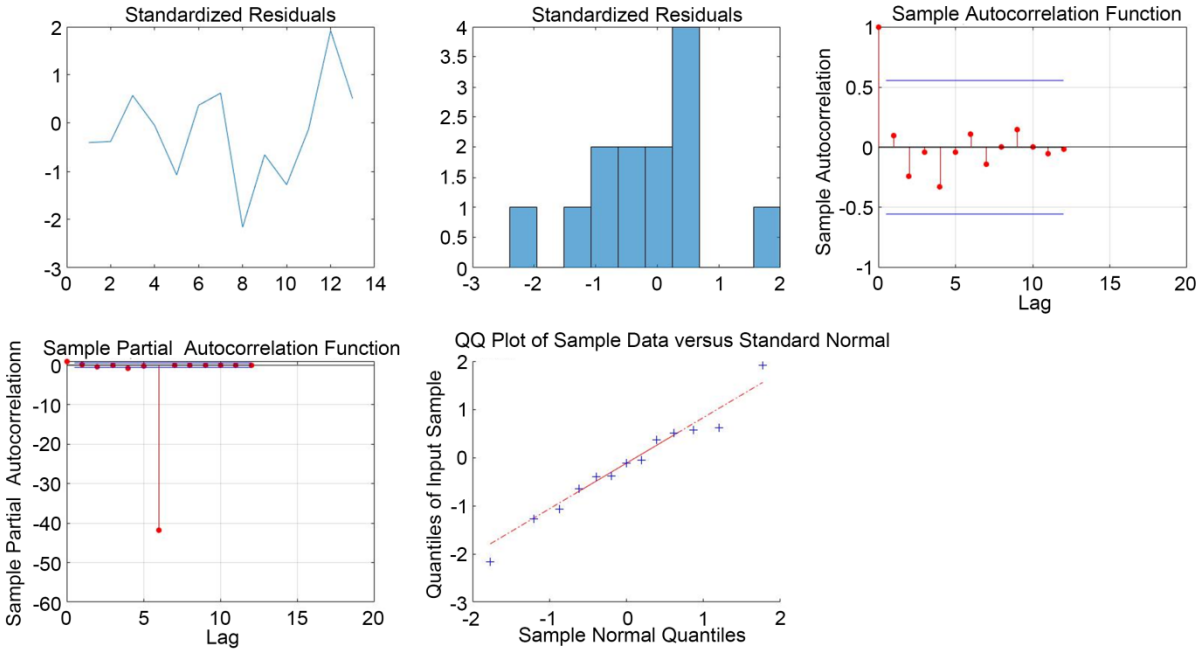


Figure 3. Residual error test chart
图 3. 残差检验图

从图像中可以看到经过处理的距离时间序列的残差服从正态分布，因此通过白噪声检验。相关误差和检验结果如表 2 所示：

Table 2. Relevant errors and test results
表 2. 相关误差和检验结果

Parameter	Value	Standard Error	T Statistic
Constant	0.992096	3.80169	0.260962
AR{1}	−1.39158	0.144046	−9.66071
AR{2}	−0.914579	0.190583	−4.79885
MA{1}	1	0.351666	2.8436

最后使用 D-W 检验，检验时间序列的自相关性，D-W 检验结果：1.7228。因此不存在一阶的自相关性。可以使用该时间序列得到预测的结果，具体预测的距离图像和置信区间如图 4 所示。

然后我们使用灰色预测算法预测距离蜜蜂在下一个春天到来之前的运动，灰色预测揭示了事物动态关联的特征与程度。由于以发展态势为立足点，因此对样本量的多少没有过分的要求，也不需要典型的分布规律，计算量少，且不致出现关联度的量化结果与定性分析不一致的情况。而且能利用微分方程来充分挖掘系统的本质，精度高。使用的 GM(1,1)原理如下：

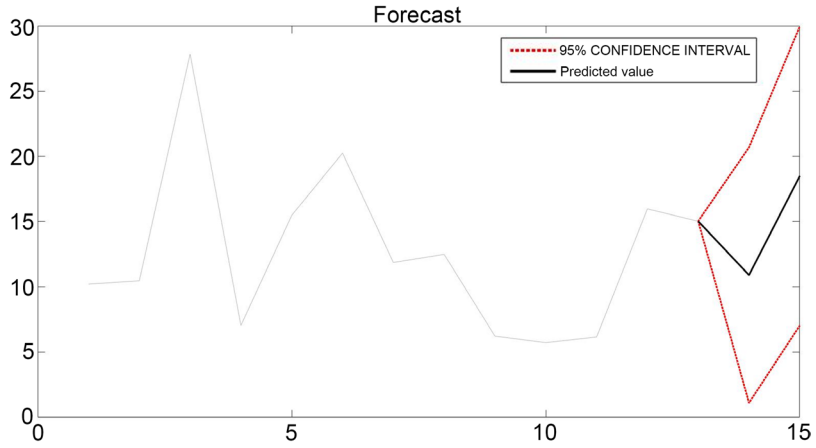


Figure 4. ARIMA(2,1,1) forecast result graph
图 4. ARIMA(2,1,1)的预测结果图

设有非负序列：

$$X_0=\left(x_0(1),x_0(2),\cdots,x_0(n)\right)$$

$$X_1=\left(x_1(1),x_1(2),\cdots,x_1(n)\right)$$

对 x_0 做一次累加得到：

其中 $x_1(k)=\sum_{i=1}^{k}x_0(i);\ k=1,2,\cdots,n$ 。

x_0 的紧邻均值生成序列为：

$$Z_1=\left(z_1(2),z_1(3),\cdots,z_1(n)\right)$$

其中 $z_1(k)=\frac{1}{2}\left(x_1(k)+x_1(k-1)\right);\ k=1,2,\cdots,n$ 。

则定义 GM(1,1)的灰微分方程为：

$$d(k)+az^{(1)}(k)=b$$

进而有：

$$\begin{bmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \cdots & \cdots\\ -z^{(1)}(n) & 1\end{bmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ \cdots\\ x^{(0)}(n)\end{pmatrix}$$

$$a=[a,b]^{\mathrm{T}}=\left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}}Y$$

最终的响应序列为：

$$\hat{x}_1(k+1)=\left(x_0(1)-\frac{b}{a}\right)\mathrm{e}^{-ak}+\frac{b}{a};\ k=1,2,\cdots,n$$

使用 matlab 进行 GM(1,1)对簇 1 的距离半径进行预测，得到结果为 9.5003。误差为 4.1355%因此可以 95%的接受该假设，并且经过检测该预测半径在簇 1 的置信区间内部，因此下一时刻蜜蜂运动到的地点距离簇均值向量的距离为 9.5，绘制分布图如图 5 所示：

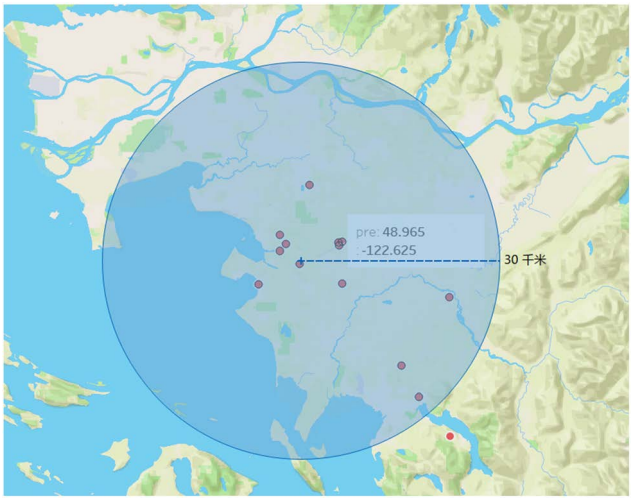
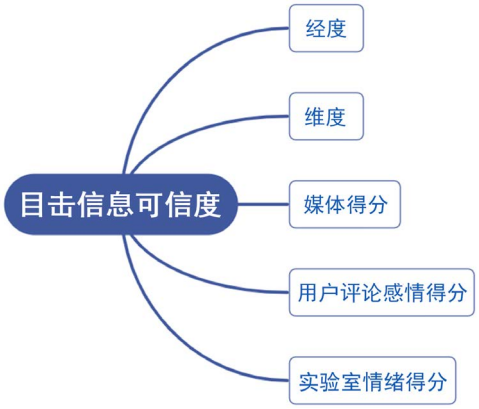


Figure 5. Visualization of GM(1,1) prediction results
图 5. GM(1,1)预测结果的可视化

3.2. 信息可信度计算功能

在图像处理方面，首先选择了 CNN 卷积网络进行图像识别[8]。针对数据的特点，对 CNN 模型进行改进，调整其激活函数，用支持向量机代替卷积网络中的全连接层。对于图像数据，采用预训练网络提高模型精度，增强图像质量。采用适用于非平衡数据的 ROC 曲线和 P-R 曲线，更合理地评价了模型的精度。



在情感分析方面，TextBlob 是一个用 Python 编写的开源文本处理库，可用于执行许多自然语言处理任务，如词性标记、名词成分提取、情感分析、文本翻译等。

最后，分析数据，Lab Status 由肯定、否定、UNV 和 UNP 四个类别组成，我们使用 UNP 作为测试集，其余数据作为训练集，这样的训练可以出现非标签样本。因此，将 TSVM 用于半监督学习。数据集被划分并随机抽样。建立多个支持向量机对数据进行平均，考虑到数据的不平衡性，采用 CNN 和支持向量机相结合的方法得到预测概率。

3.2.1. 图像识别

TSVM 是一种经典的半监督[9]支持向量机，是在具有未标记样本的训练集上训练得到 SVM，然后预测测试集的分类模型，半监督 SVM [10]的基本假设为“低密度分割”，TSVM 尝试对于训练集中的每一个未标记样本进行穷举，然后得到间隔最大化的超平面，TSVM 基本原理如下：

假设训练集包括，有标记样本数据集：$D_l=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$、无标记样本数据集：$D_u=\{x_{u_1},x_{l+2},\cdots,x_{l+u}\}$，其中 $y_i\in\{-1,+1\}$，$l\ll u$，$l+u=m$。TSVM 的学习目标为 D_u 中的样本给出预测标记 $\hat{y}=(\hat{y}_{l+1},\hat{y}_{l+2},\cdots,\hat{y}_{l+u})$，$\hat{y}_i\in\{-1,+1\}$，使得

$$\min_{w,b,\hat{y},\xi}\frac{1}{2}\|w\|_2^2+C_l\sum_{i=1}^{l}\xi_i+C_u\sum_{i=l+1}^{m}\xi_i$$
$$\text{s.t. } y_i\left(w^{\mathrm{T}}x_i+b\right)\geq 1-\xi_i,\ i=1,2,\cdots,l$$
$$\hat{y}_i\left(w^{\mathrm{T}}x_i+b\right)\geq 1-\xi_i,\ i=l+1,l+2,\cdots,m$$
$$\xi_i\geq 0,\ i=1,2,\cdots,m$$

其中，(w,b) 确定了一个划分超平面；ξ 为松弛变量，$\xi_i(i=1,2,\cdots,l)$ 对应于有标记样本，$\xi_i(i=l+1,l+2,\cdots,m)$ 对应于未标记样本；C_l 和 C_u 用于平衡模型的复杂度、有标记样本和未标记样本重要程度的折中参数。

对于得到的数据，指标有：纬度、经度、实验室评论情感得分、评论情感得分、图片得分，以及标记信息，在对未标记样本进行标记指派以及调整的过程中，发现数据集存在类别不平衡的问题，反例数据样本要明显多于正例样本个数，因此这里使用 boostrap 方式对反例进行随机采样，多次运行代码对结果取平均值，运行的结果如图 6 所示：

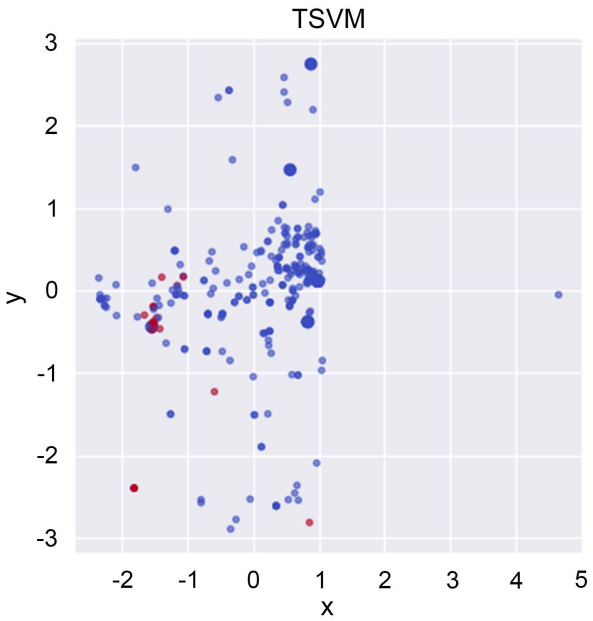


Figure 6. TSVM prediction result graph
图 6. TSVM 预测结果图

其中，散点图的横纵坐标依次经过正则化和无量纲化处理的纬度和经度坐标。散点图中的样本点为训练集中未标记样本和原数据集中没有预测的样本即 Lab Status 为 UnProcessed 的样本，红色样本为预测为正例的样本，蓝色为预测为反例的样本。

使用 TSVM 得到的精度接近为 0.9987，接近与 1.0，极有可能会产生过拟合的问题，并且由于数据集构建的随机性，即使是使用 bootstrap 实现，也具有不完整性，没有充分利用数据，因此我们使用卷积神经网络进行预测，将 SVM 替换 CNN 的全连接层。

使用 TSVM 作为最终的分类问题，考虑到样本不平衡性很大，如果使用全部样本进行预测会出现以下两个问题：

1) 训练时间太长，对于新报告的数据不能及时做出响应，导致模型的实时性较差。

2) 目前数据样本的不平衡太过于严重，直接使用 TSVM 会使得模型的精度降低。

将 TSVM 优化模型中的 C_u 项拆分为 C_u^+ 和 C_u^- 两项，分别对应基于伪标记而当作正、反例使用的未标记样本，并在初始化时使得：

$$C_u^+ = \frac{u_-}{u_+} C_u^-$$

其中，u_+ 与 u_- 为基于伪标记而当做正、反例使用的未标记样本数。

对数据集进行切分，使得新产生的数据的样本不平衡性尽可能的少。这种方法需要保证样本的随机性。

我们选择随机采样法降低训练集数据的不平衡性，这个我们取从反例样本中随机进行 10 次选取，每次选取 100 个反例样本，对得到的结果计算模型的泛化误差，最后将 10 次实验得到的泛化误差取平均值。用泛化误差的均值近似本模型的好坏。

使用 TSVM 对 UnProcessed 以及未标记 Unv 的样本部分的分类结果，如表 3 所示：

Table 3. Classification results
表 3. 分类结果

样本全局 ID	样本标签
{1E2B3656-E2CD-4DA9-8CEF-FDE70664643B}	0
{26DDF8E2-DA0C-4F87-A65A-233115BAFCCD}	0
{23756338-4E29-4F92-ADE0-F0375321FB8B}	0
{5BBFCFBA-27A6-46AB-9440-06FB025C2EEE}	1
{3E50801D-9DBB-43DE-8D32-31CFA88C74D9}	1

构建不同数据集，得到的分类结果散点图绘制如图 7 所示：

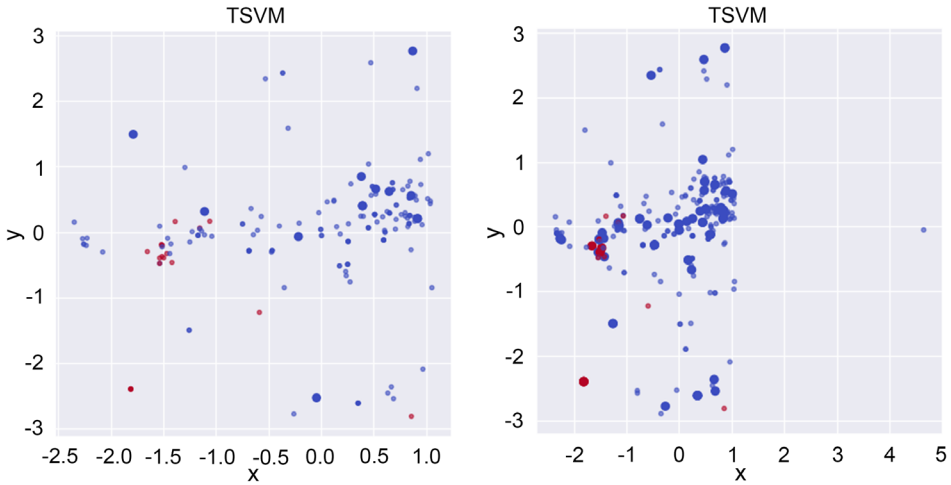


Figure 7. Different classification scatter plots of different training sets
图 7. 不同训练集的不同分类散点图

对应三个随机采样得到的不同数据集，模型的精度依次为：1.0、0.987、0.924。

但考虑到 TSVM 虽然通过随机采样得到多个数据集进行预测，但是依然无法保证数据的随机性，并且得到的结果是一个 0~1 值，而不是概率，因此模型的准备性没有 CNN、神经网络的准确性高。因此我们使用卷积神经网络进行预测。

卷积神经网络有着强大的学习能力和泛化能力，由于其对高维数据处理无压力、能够自动提取图像特征、较高的分类精度等优点，有助于图像的分类检测[11]。典型的卷积神经网络模型有 GoogLeNet、LeNet、AlexNet 以及 VGGNet 等。CNN 的结构包含许多其他类型的层，比如卷积层、池化层、全连接等，简要结构如图 8 所示[12]：

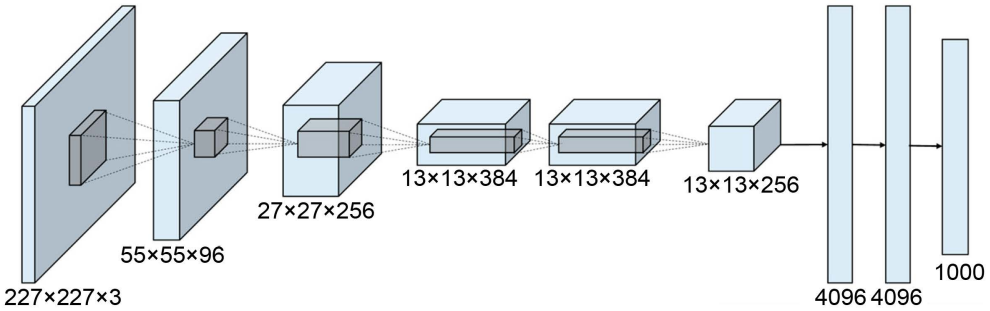


Figure 8. Brief structure of CNN
图 8. CNN 简要结构

我们将数据划分为训练集与测试集，通过模型的分类检验，发现分类的结果效果并不理想，这有可能和所给的数据偏差大与模型本身不契合数据所导致，如图 9 所示：

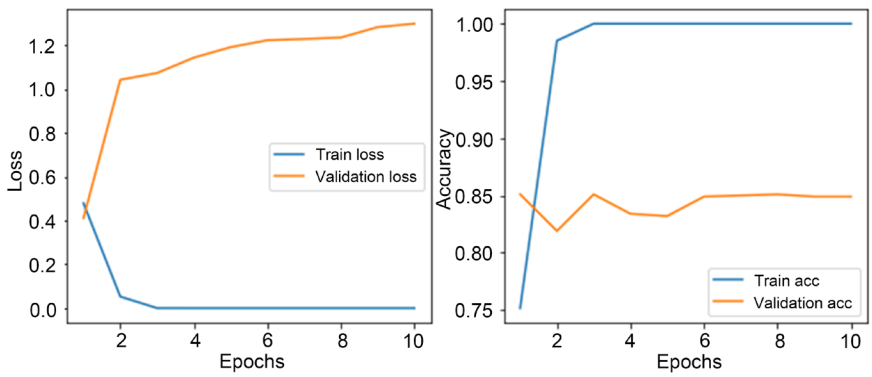


Figure 9. Direct use of CNN prediction results
图 9. 直接使用 CNN 预测结果图

首先对模型进行改进，针对模型中的激活层，使用 sigmoid 函数的效果并不理想，经过测试，发现使用 RELU 作为激活函数效果最好，RELU 可以有效解决梯度消失问题，其计算速度非常快，只需要判断输入是否大于 0。SGD 的求解速度远快于 sigmoid 和 tanh [12]。

由于所给的图片数据差距较大，本文使用了图像增强方法在 Keras 中，可以利用图像生成器很方便地定义一些常见的图像变换。将变换后的图像送入训练之前，可以按变换方法逐个看看变换的效果。

想要将深度学习应用于小型图像数据集，通常不会贸然采用复杂网络并且从头开始训练，因为训练代价高，且很难避免过拟合问题。相对的，通常会采用一种更高效的方法——使用预训练网络，如图 10 所示。

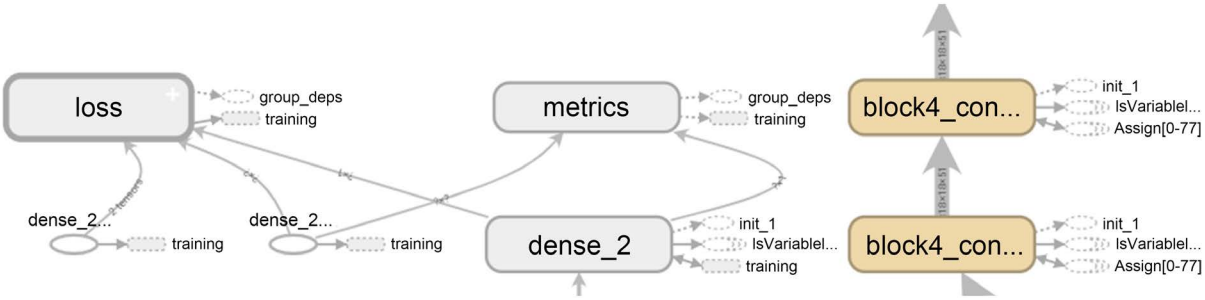


Figure 10. Schematic diagram of pre-training
图 10. 预训练示意图

预训练网络的使用通常有两种方式，一种是利用预训练网络简单提取图像特征，之后可能会利用这些特征进行其他操作；另一种是对预训练的网络进行裁剪和微调，以适应自己的任务[6]数据直方图如图 11 所示。

第一种方式训练代价极低，因为它就是简单提取个特征，不涉及训练；缺点是保存提取出来的特征需要占用一定空间，且无法使用图像增强。

第二种方式可以使用图像增强，但训练代价也会大幅增加。本文使用 keras 中自带的 VGG16 模型提取图像特征，然后以这些图像特征为输入，训练一个小型分类器。

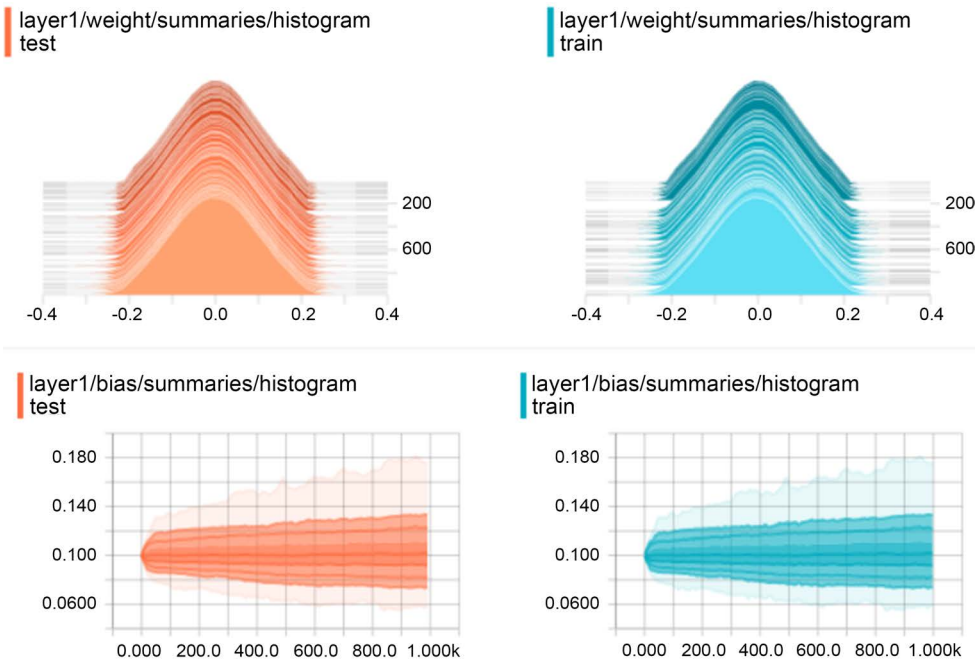


Figure 11. Data histogram
图 11. 数据直方图

通过预训练的 VGG 网络卷积基，来简单的提取了图像的特征，并用这些特征作为输入，训练了一个小分类器。这种方法好处在于简单粗暴，特征提取部分的卷积基不需要训练。但缺点在于，一是别人的模型是针对具体的任务训练的，里面提取到的特征不一定适合自己的任务；二是无法使用图像增强的方法进行端到端的训练。因此，更为常用的一种方法是预训练模型修剪 + 微调，好处是可以根据自己任务需要，将预训练的网络和自定义网络进行一定的融合；此外还可以使用图像增强的方式进行端到端的训练。过程为：

首先，在已经训练好的基网络上添加自定义网络；

其次，冻结基网络，训练自定义网络；

最后，解冻部分基网络，联合训练解冻层和自定义网络。

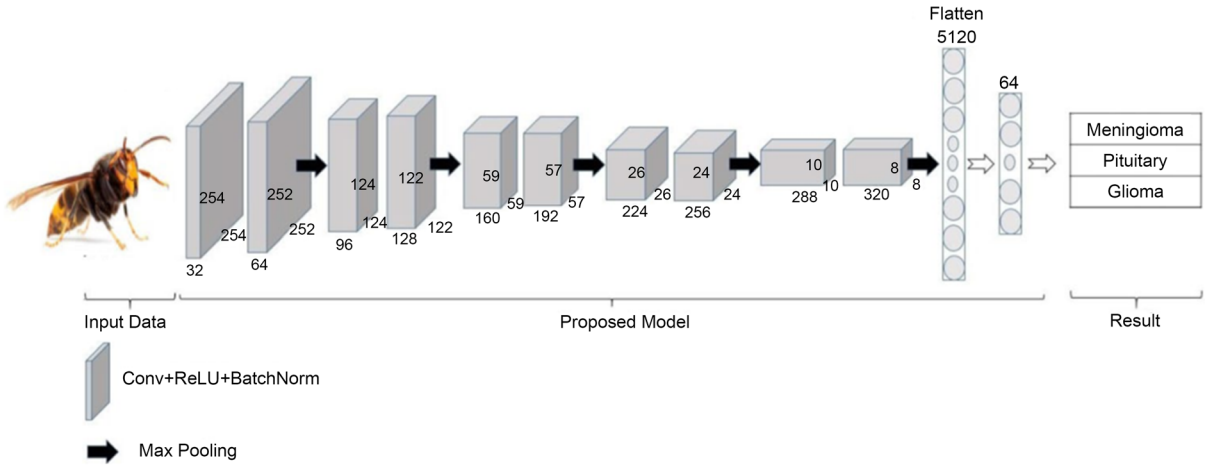


Figure 12. Model establishment
图 12. 模型的建立

在使用了上述的模型如图 12，改进方法之后再对模型进行训练，发现得到的结果相比之前优化了很多，如图 13、图 14 所示。

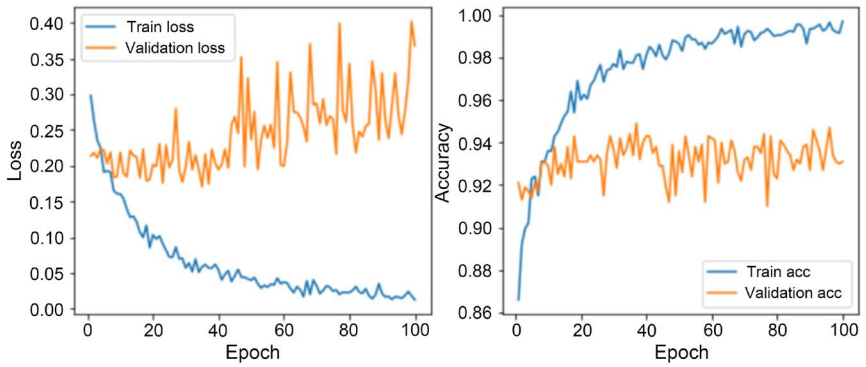


Figure 13. Model effect after optimization
图 13. 优化后的模型效果

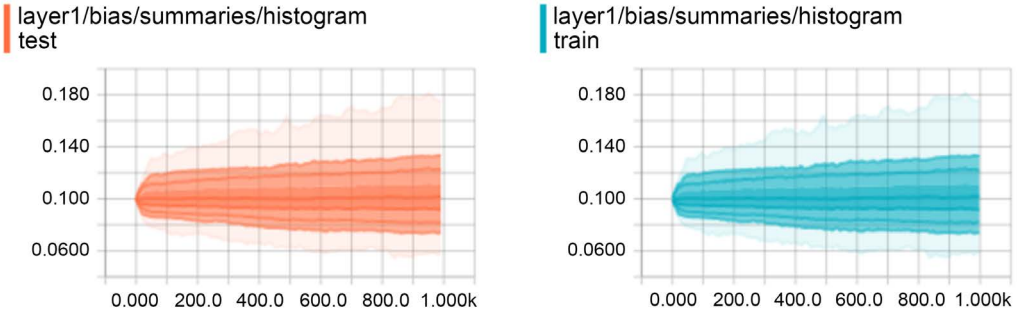


Figure 14. Neuron output distribution
图 14. 神经元输出分布

将卷积神经网络的优势与支持向量机的稳定性相结合，利用训练好的卷积层与池化层提取图片的特征，放入支持向量机中进行训练，进行分类操作。其意义在于利用 SVM 来替换卷积网络中的全连接层，如图 15 所示。

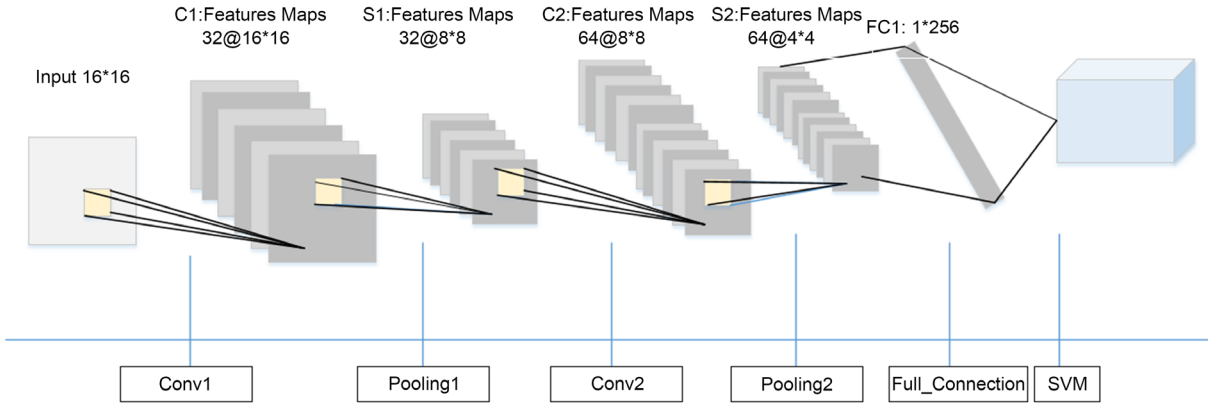


Figure 15. Combination of SVM and CNN
图 15. SVM 与 CNN 结合

使用卷积作为特征提取以及 SVM 作为分类器的具体理由如下，如图 16 所示。

1) 由于卷积和池化计算的性质，使得图像中的平移部分对于最后的特征向量是没有影响的。从这一角度说，提取到的特征更不容易过拟合。而且由于平移不变性，所以平移字符进行变造是无意义的，省去了再对样本进行变造的过程。

2) CNN 抽取出的特征要比简单的投影、方向，重心都要更科学。不会让特征提取成为最后提高准确率的瓶颈、天花板。

3) 可以利用不同的卷积、池化和最后输出的特征向量的大小控制整体模型的拟合能力。在过拟合时可以降低特征向量的维数，在欠拟合时可以提高卷积层的输出维数。相比于其他特征提取方法更加灵活。

4) 非线性映射是 SVM 方法的理论基础，SVM 利用内积核函数代替向高维空间的非线性映射；

5) 对特征空间划分的最优超平面是 SVM 的目标，最大化分类边际的思想是 SVM 方法的核心；

6) 支持向量是 SVM 的训练结果，在 SVM 分类决策中起决定作用的是支持向量。

	Conv1	Pooling1	Conv2	Pooling2	Full_Connectin1	SVM
input size	16*16	16*16	8*8	8*8	4*4*64	1*256
filter	5*5	2*2	5*5	2*2	-	-
output size	16*16	8*8	8*8	4*4	256	-

Figure 16. Feature map size change
图 16. Feature map 大小变化

3.2.2. 情感分析

文本情感分析是指用自然语言处理(NLP)、文本挖掘以及计算机语言学等方法对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。情感分析是文本分类的一个分支，通过对带有情感色彩(褒义贬义/正向负向)的主观性文本进行分析，以确定该文本的情感倾向。

情感分析应用的领域非常广泛：比如说商品的评论挖掘、电影推荐、股市预测等。其中自然语言工具包(NLTK)是最受欢迎的自然语言处理库(NLP)，背后有非常强大的社区支持。

TextBlob 是一个开源的文本处理库，它可以用来执行很多自然语言处理的任务，比如，词性标注、名词性成分提取、情感分析、文本翻译等。

我们使用情感分析得到 lab comments 中每个句子的得分，考虑到存在口语化的回答，因此取所有句子的最小值来确定此 comment 的情感得分。

最后得到的部分结果如表 4 所示：

Table 4. Emotion score
表 4. 情感得分

That looks like a bumble bee.	−0.1777777777778000
That looks like a yellow jacket.	0.1363636363636363500
That's a hummingbird.	0.0000000000000
of our yellowjackets, but I cannot quite tell which one from these images.	0.500000000000000
Positive lD	0.600000000000000000
Unverified	0.000000000000000000
Negative lD	0.200000000000000000

3.2.3. 模型评估

由于处理的是样本不平衡问题，使用传统的损失去评估可能会导致不准确，因此本文使用了 P-R 曲线以及 ROC 曲线来评判模型的准确性。

在衡量学习器的泛化性能时，根据学习器的预测结果对样本排序，按此顺序逐个把样本作为正例进行输出，每次计算测试样本的真正率 TPR，和假正率 FPR 并把这两项作为 ROC 的纵轴和横轴。其中真正率衡量实际值为正例的样本中被正确预测为正例的样本的比例，假正率表示实际值为负例的样本中被错误的预测为正例的样本的比例 ROC 曲线有一个巨大的优势就是，当正负样本的分布发生变化时，其形状能够基本保持不变，而 P-R 曲线的形状一般会发生剧烈的变化，因此该评估指标能降低不同测试集带来的干扰，更加客观的衡量模型本身的性能。

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

ROC 曲线接近于(1,0)点，表明模型泛化性能越好，越接近对角线的时候，表明此时模型的预测结果为随机预测结果。

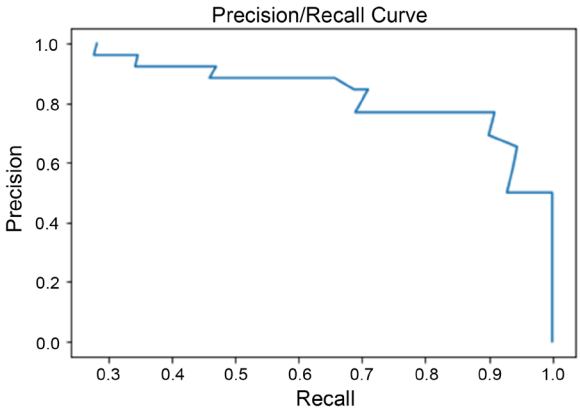


Figure 17. P-R curve
图 17. P-R 曲线

PR 曲线常被用在信息提取领域，同时当我们的数据集中类别分布不均衡时我们可以用 PR 曲线代替 ROC。PR 曲线的横轴代表查全率，实际上就是真正率，纵轴代表查准率，表示预测为正例的样本里实际也为正例的样本所占的比例，如图 17 所示。当 PR 曲线越靠近右上方时，表明模型性能越好，与 ROC 曲线类似，在对不同模型进行比较时，若一个模型的 PR 曲线被另一个模型的 PR 曲线完全包住则说明后者的性能优于前者。如图 18 中橘色线代表的模型要优于蓝色线代表的模型，若模型的 PR 曲线发生了交叉，则无法直接判断哪个模型更好。因此可以使用 F1 指标进行评估，也就是查准率和查全率的加权平均：

$$F1 = 2*\frac{P*R}{P+R}$$

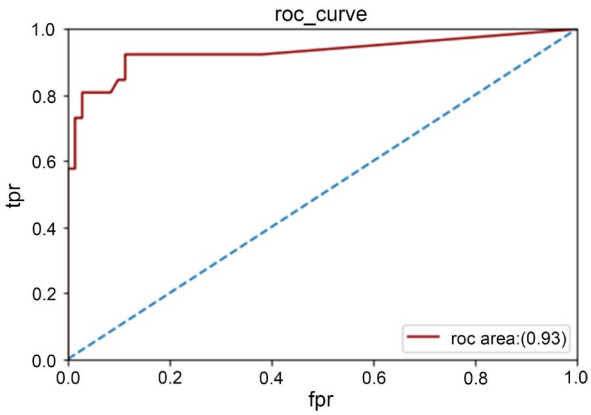
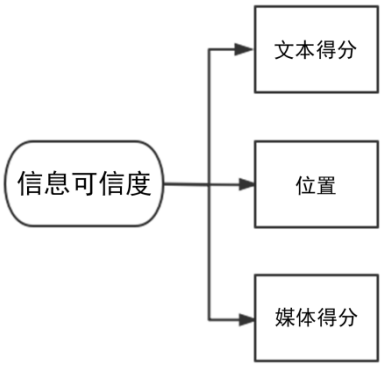


Figure 18. ROC curve of the algorithm
图 18. 算法的 ROC 曲线

3.3. 目击者信息可信度模型

根据第一问的模型，可以得出胡峰的传播模型以及其随时间转移的规律。再结合第二问对目击者文本信息和图片信息处理的模型，可以对目击信息的可信度进行加权评分。如图下所示：



首先我们分析的数据集中正例得到这些目击信息的加权得分，根据这十四个信息的得分确定一个阈值，通过这个阈值来决定新得到的的目击信息是否具有考察的价值。

以数据集中未分类的数据为例，通过对文本信息的评分以及图片信息进行分类得到概率进行加权组合，得到目击信息可信度指数，将数据排序，再从里面选取符合阈值的数据，如图 19 所示。

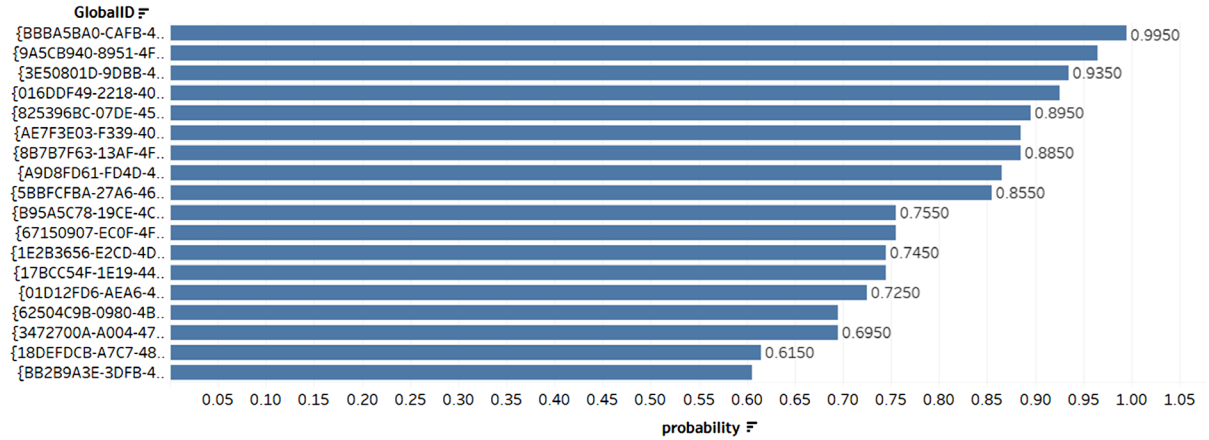


Figure 19. Sighting information credibility ranking
图 19. 目击信息可信度排名

将这些点从地图上展示，如图 20、图 21 所示，红色是符合阈值的目击信息，蓝色是正例。根据第一问的位置预测，本文将这些点分为两类，一类是基于之前的黄蜂可能出现点的需要去格外关注的目击信息，一类是可能再新地点开始繁殖并活动的区域，这些虽然与之前的地理位置联系并不大，但是可能成为一个新的胡峰巢穴，也需要重点去考察。

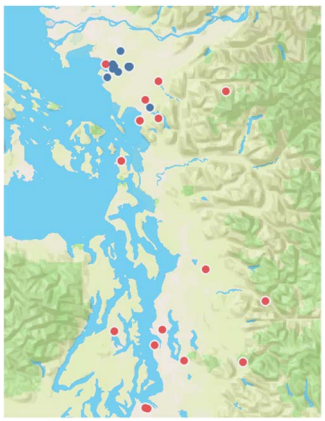


Figure 20. Comparison of positive and unverified
图 20. Positive 与 unverified 比较

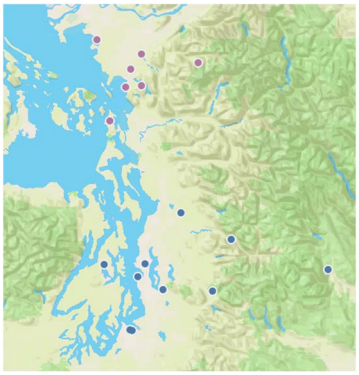


Figure 21. Two types of more important points
图 21. 两类比较重点的点

3.4. 模型的更新

3.4.1. 如何更新模型

华盛顿州用户每提交一次报告，模型就更新一次，因此模型更新的时间随着用户提交新记录的时间而变动。我们通过使用 CNN 结合 SVM 的方式对所给的图像进行分类，对于 Lab Status 值为 Positive ID 的记录进行分析。通过计算前后两次记录的时间差得到一个时间间隔序列，因此通过研究该序列的稳定性来预测下一次模型更新所需要的时间。

考虑到蜜蜂在每年的春季会进行一次迁徙，因此这里对第一问的时间序列进行改进，将季节性因素引入，并观察时间序列的初始情况如图 22 所示：

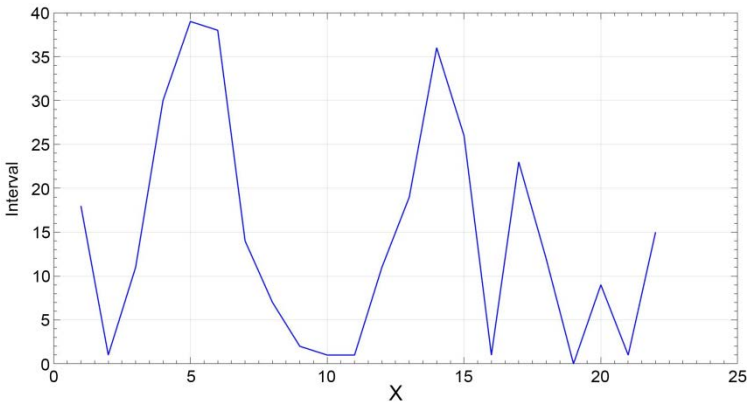


Figure 22. Initial image of time interval sequence
图 22. 时间间隔序列初始图像

从图中也可以明显看出时间间隔具有季节性，即在春季时间间隔会增大，在一年中的其他时间段，时间间隔变化较小。但由于初始的时间序列不具有平稳特性，对时间序列做一阶差分，后绘制 ACF 和 PACF 曲线如图 23 所示：

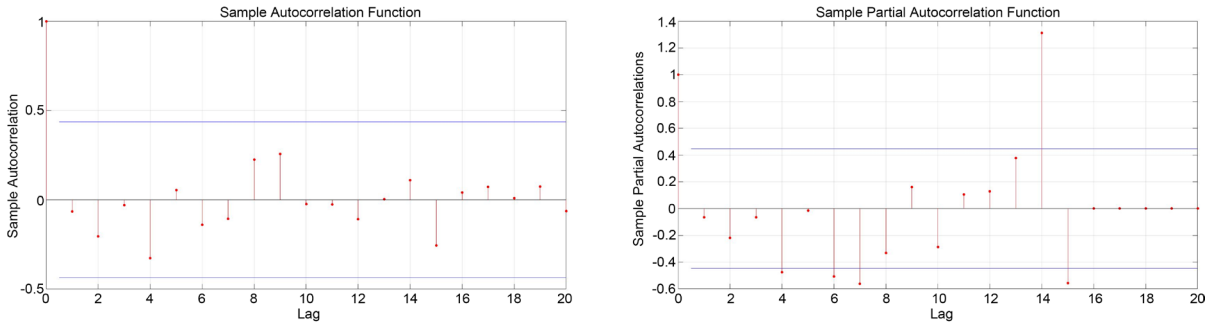


Figure 23. ACF and PACF curves
图 23. ACF 和 PACF 曲线

从曲线的截尾和拖尾现象，以及使用 AIC 和 BIC 对 ARIMA 进行调参，最终确定使用 ARIMA(2,1,2)。

对得到的残差进行检验，结果如图 24 所示。

从残差检验的图像中可以发现残差的数学期望为 0，方差近似为 1，近似服从标准正态分布，因此通过白噪声检验。

最终得到预测的时间序列图像如图 25 所示。

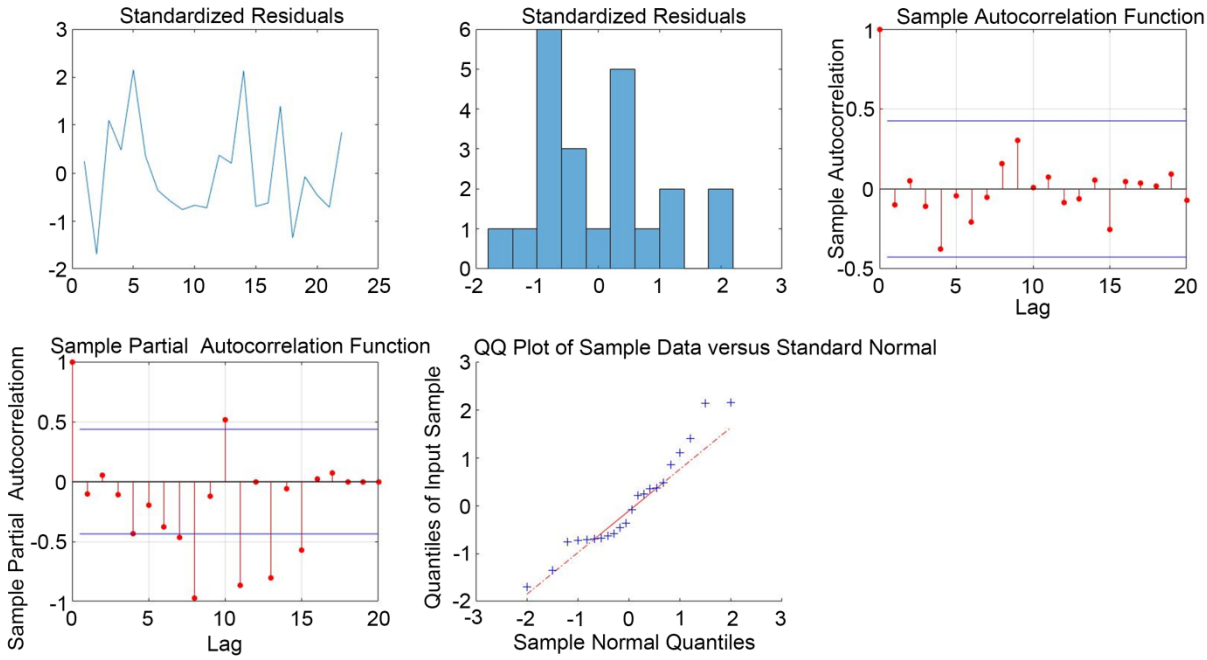


Figure 24. Residual testing of seasonal time series
图 24. 季节性时间序列的残差检验

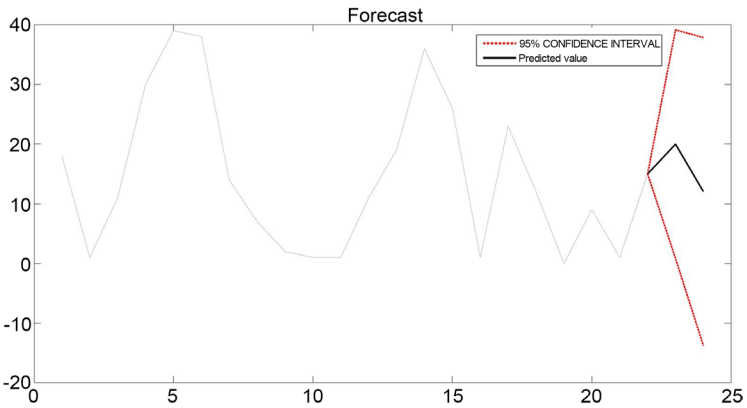


Figure 25. Seasonal time series forecast results
图 25. 季节性时间序列预测结果图

通过季节性时间序列得到预测结果稳定在 10.12 到 19.779 之间，因此将模型的更新时间确定在 10.12 到 19.779 之间。

3.4.2. 需要多久胡蜂才会灭绝

假设人类不会采取一些防控手段对蜜蜂进行控制和消除，只能通过蜜蜂自己的迁徙实现一个区域内的根除。

通过对问题进行分析，根据我们的模型害虫根除的现象有如下两点：华盛顿州内部新报告数减少、新报告数的可信度很低。如果满足这两个现象则大概率蜜蜂发生迁徙，华盛顿州内蜜蜂的数量大幅度降低甚至被根除。

其中，华盛顿州内部新报告数减少，可以通过我们模型的时间间隔进行衡量，一旦时间间隔出现季节性的涨幅，认为蜜蜂发生了向内陆的迁徙。对于新报告可信度的预测，可以通过我们的 CNN +

TSVM 判断模型得到一个预测概率，当这个预测概率 < 0.05 时，我们可以认为这个报告是不真实的，因此不予统计。

因此，当新报告出现的频率降低，报告的可信度下降时，我们可以认为华盛顿州的蜜蜂发生迁徙，即被根除。

对于报告的可信度可以使用我们的预测模型得到可信度评分，因此主要问题转换为在华盛顿州刻画报告时间间隔长短的问题。这里我们通过统计每天华盛顿州提交的目击报告数目，根据从 2019 年 9 月开始的报告数目挖掘有效信息。

首先我们对数据进行预处理，去除 2019 年 9 月之前的数据，这些数据是在蜜蜂还没有被发现时具有的，因此属于脏数据。

根据每个月份中的箱线图找到对应的异常值，进行异常值的剔除操作，并且通过箱线图也可以反应出蜜蜂数量在华盛顿州的 2020 年 7 月到 2020 年 9 月较多。

从图像中可以看出在华盛顿 2019 年 9 月开始呈现的趋势与蜜蜂在华盛顿州的变化情况十分类似，在 2020 年 7 月到 2020 年 9 月蜜蜂在华盛顿出现的次数增多，在 10 月以后气温降低不适宜蜜蜂居住，持续到 2021 年春天蜜蜂在华盛顿州都会下降，因此考虑对每天的报告数进行拟合，应用时间序列模型进行预测后一个月蜜蜂的下降以及蜜蜂在华盛顿州是否会被根除。

每天的报告数经过 Python 绘制折线图效果如图 26、图 27 所示。

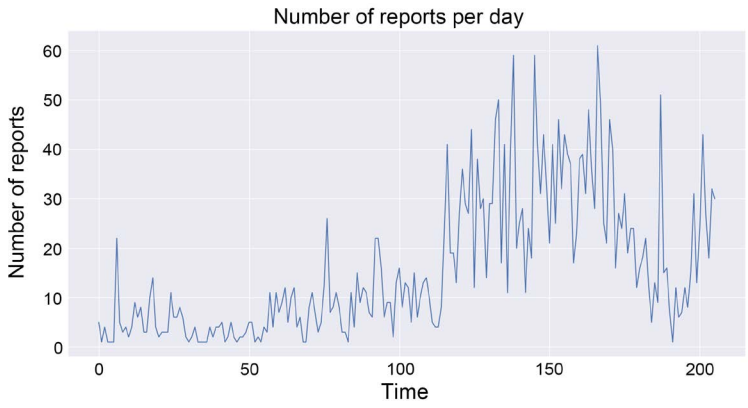


Figure 26. Line chart of daily report
图 26. 每天报告的折线图

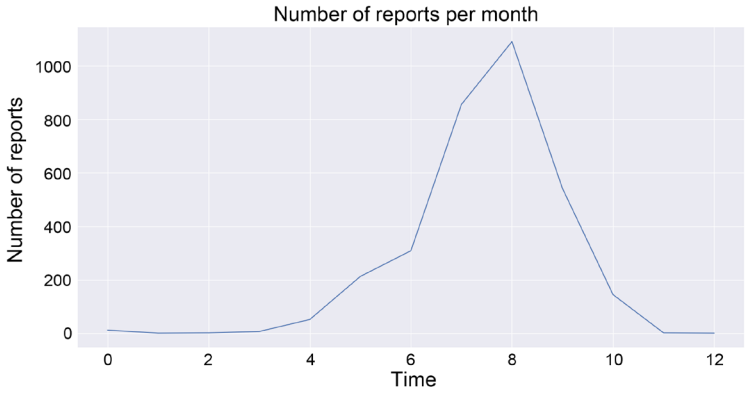


Figure 27. Line chart of monthly report
图 27. 每月报告的折线图

从图中可以看出，当亚洲大黄蜂进入华盛顿州时，报告数处于一种低峰状态，随着时间的推移报告数上升，但对于每个不同的阶段，报告数都较为平稳，因此使用我们第一问的时间序列模型对华盛顿州的报告数进行预测，用预测每天的报告数来近似时间间隔的大小，如果时间间隔增大，则每天的报告数应该会减少；如果时间间隔减小，则报告会愈发频繁，每天的报告数就会增多。

因此对于蜜蜂在华盛顿州存在的多少可以通过每天的报告数进行反应，报告数越多蜜蜂就越多，报告数越少蜜蜂就越少。

通过使用第一问的时间序列模型，经过 AIC、BIC 调参确定使用 ARIMA(0,1,2)模型。

我们使用得到的时间序列模型预测华盛顿州后一个月的日报告数，得到最终的预测结果如图 28 所示：

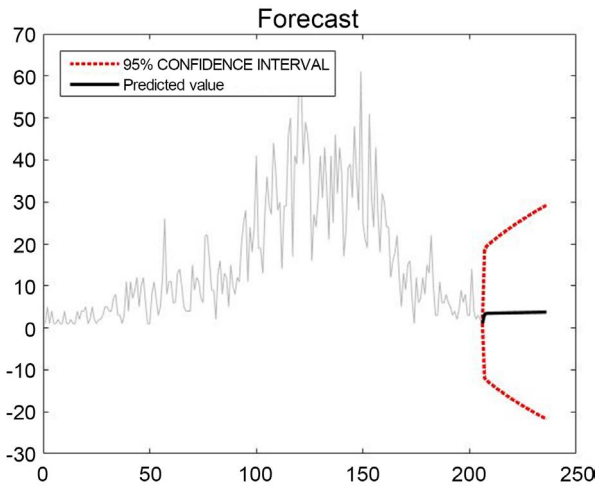


Figure 28. ARIMA(0,1,2) forecast results
图 28. ARIMA(0,1,2)预测结果

后一个月开始会出现一个小幅度的上涨，然后趋近于 0，表示在后一个月内每日的新报告几乎为 0，因此我们有 95%的把握接受这个结论。即在后 30 天中蜜蜂会从华盛顿州继续向内陆地区迁徙，华盛顿州的蜜蜂将在 1 个月内大幅度降低在月中或者月末达到一个“被根除”的状态。

4. 结论与讨论

用灰色预测去探究亚洲黄蜂的时空分布，对样本量要求不高，不需要典型分布型，计算量小，关联度定量结果与定性分析结果无差异。卷积神经网络模型可用于自动特征提取。优化后的算法能很好地适应样本分类不平衡的情况。亚洲黄蜂的生物学行为是季节性的。针对这一特点，我们在模型的时间序列和其他部分加入了季节的影响，较好地解决了生物季节性运动的问题。通过使用半监督支持向量机，在不知道没有标记的情况下分类是正确的，但可以惩罚它们，S3VM 和熵正则化假设可以很好地分离样本类。

但是，在图像识别中，没有很好的算法从视频中提取出最合适的帧进行检测。一些图像可能使用亚洲黄蜂网地图，没有数据库来确定该图像是否是网络地图并且已经发布。同时，灰色预测精度较低，不可能得到非常具体的预测结果，基于指数率的预测没有考虑系统的随机性，中长期预测精度较差，并且 CNN 分类需要很大的样本量和良好的速度硬件支持。因此，对于胡蜂空间上变动的预测和模型的更新时间还有待深入讨论。

基金项目

国家级大学生创新创业训练计划项目(项目编号: X202010446034X)。

参考文献

[1] 宋星富, 张丽亨. 蜜蜂几种敌害昆虫的习性及防治措施[J]. 中国蜂业, 2012, 63(10): 30.

[2] 杜辉林. 再谈胡蜂的习性、行为及防治[J]. 蜜蜂杂志, 2016, 36(7): 26-27.

[3] Alaniz, A.J., Carvajal, M.A. and Vergara, P.M. (2021) Giants Are Coming? Predicting the Potential Spread and Impacts of the Giant Asian Hornet (*Vespa mandarinia*, Hymenoptera: Vespidae) in the USA. *Pest Management Science*, **77**, 104-112. https://doi.org/10.1002/ps.6063

[4] Washington State University (2020) Scientists Predict Potential Spread, Habitat of Invasive Asian Giant Hornet.

[5] 罗文华, 曹兰, 杨金龙, 等. 胡蜂的防治方法[J]. 蜜蜂杂志, 2020, 40(8): 16-17.

[6] 余同瑞, 金冉, 韩晓臻, 李家辉, 郁婷. 自然语言处理预训练模型的研究综述[J]. 计算机工程与应用, 2020, 56(23): 12-22.

[7] 朱家明, 段寒冰, 王子健, 张浚铃. 基于 ARIMA 模型对北京垃圾分类对策的计量分析[J]. 中央民族大学学报(自然科学版), 2020, 29(2): 49-56.

[8] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9): 2508-2515+2565.

[9] 王晓昆, 温显斌. 基于样本选择策略的 SAR 图像半监督分类方法[J]. 天津理工大学学报, 2020, 36(3): 26-32.

[10] 宋静. 基于稀疏孪生支持向量机的人脸识别[J]. 信息技术, 2020, 44(7): 121-124.

[11] Ferraz, C.T., Barcellos, W., Pereira Jr., O., *et al*. (2021) A Comparison among Key Frame Extraction Techniques for CNN Classification Based on Video Periocular Images.

[12] 朱斌, 刘子龙. 基于新型初始模块的卷积神经网络图像分类方法[J]. 电子科技, 2021, 34(2): 52-56.