Priority prediction of Asian Hornet sighting report using machine learning methods

Yixin Liu

South China University of Technology

Contents

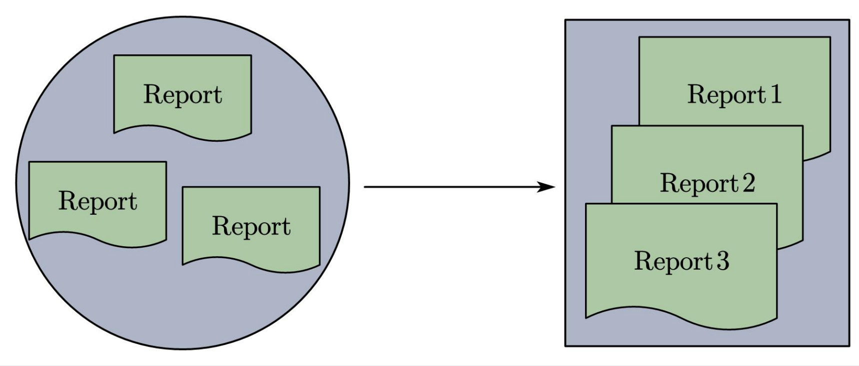
Contents

Background

- Asian hornet has caused **great damage** to the ecological environment.
- One of the key ways to control them is to locate the nest and destroy it by mobilizing the masses to submit **sighting reports**.
- Traditional manual screening methods are **inefficient and inaccurate**
- **How to automatically predict the priority of a sighting report** is an important issue.

Task Definition

- Priority Prediction of Sighting Report
  - Higher priority to the report with higher credibility.



Problem Definition

Problem Description

- Given a number of sighting report samples in a period of time, **predict the priority of the report** based on the information in the report samples.

Input & Output

- **Input**: Feature matrix of sighting reports $X \in R^{n \times k}$
- **Output**: Prioritization of sighting reports $Y \in R^{n \times 1}$

Performance Metric

- Classification accuracy (the higher the better)

Contents

Motivation

Motivation

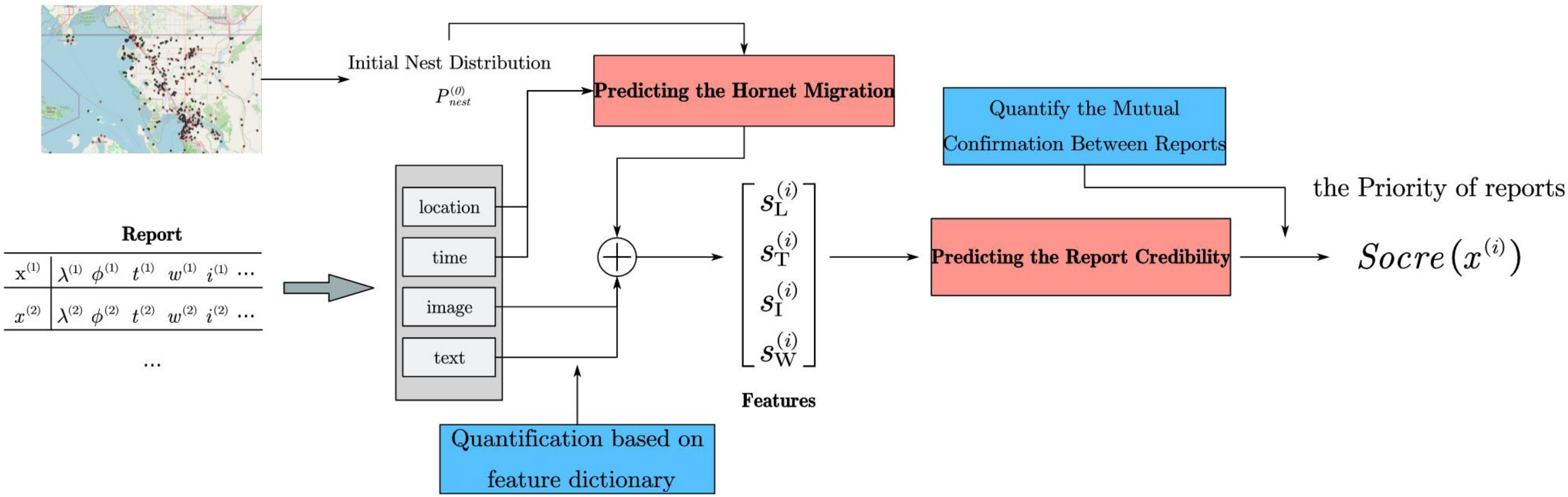
- Traditional manual methods are **not efficient and accurate**
- **A variety of rich information** in the sighting report can help predict the priority
- Combine **machine learning** to mine the features in the sighting report to realize automatic prediction

Challenges

- **Feature Engineering:** To extract the key features from numerous information in a report is challenging.
- **Priority mapping:** To map the credibility of a single report to its priority in multiple reports is challenging.

Contents

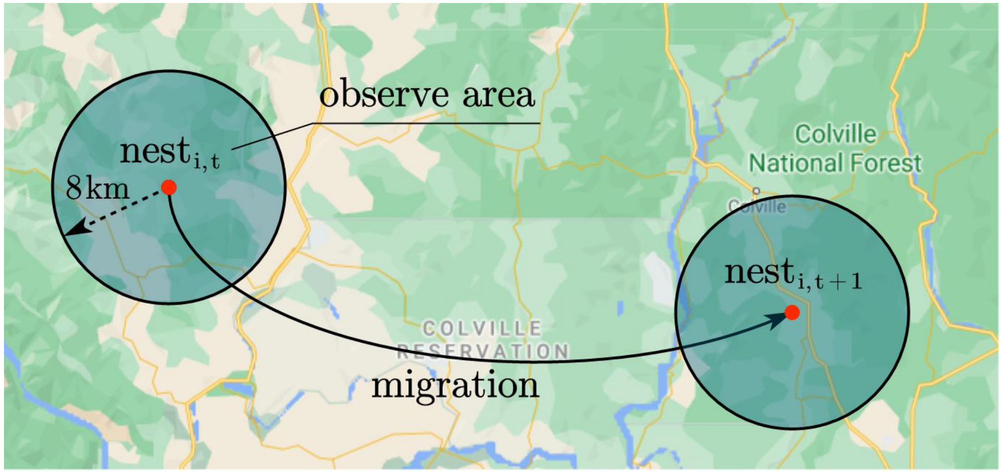
Our Approach



- Construct rich features from four dimensions: **Location, Time, Image and Text**
- Using **Logistic Regression** to predict the credibility of the report
- Determined the repots' priority by further considering the **Mutual Influence**

Features Extraction

1. Location Feature s_{L}

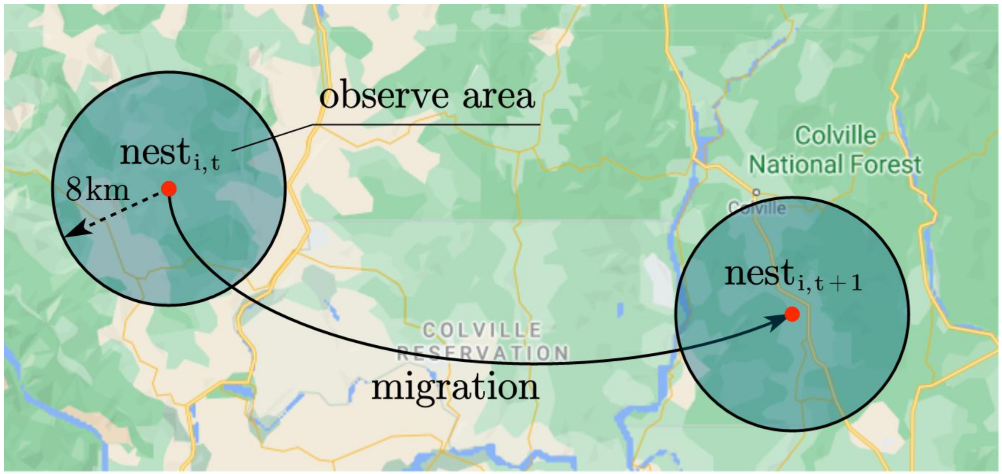


- **Assumption:** Initial distribution $P_{x,y}^{(0)}$, migration month $\mathbb{T}$
- **Nest Migration**

$$P_{\text{nest}}^{(t+1)}(\lambda,\phi) = \begin{cases} P_{\text{nest}}^{(t)}(\lambda,\phi) & ,t \notin \mathbb{T} \\ \sum_{(\lambda',\phi')} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d-30)^2}{2\sigma^2}} P_{\text{nest}}^{(t)}(\lambda',\phi') & ,t \in \mathbb{T} \end{cases}$$

Features Extraction

1. Location Feature s_{L}



- **Observe hornets**

$$P_{\text{observe}}^{(t)}(\lambda', \phi') = \sum_{(\lambda, \phi)} e^{-\beta_1 d}\, P_{\text{nest}}^{(t)}(\lambda, \phi), \forall t$$

- **Location Feature**

$$s_{\mathrm{L}}(\lambda', \phi', t) = P_{\text{observe}}^{(t)}(\lambda', \phi')$$

Features Extraction

2. Time Feature s_{T}

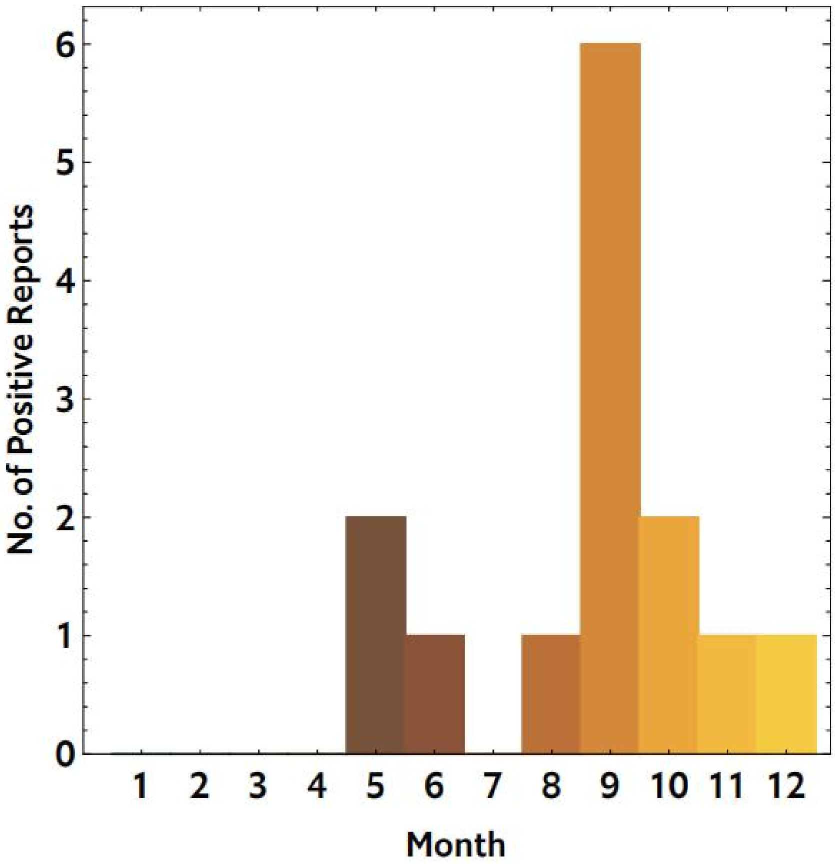


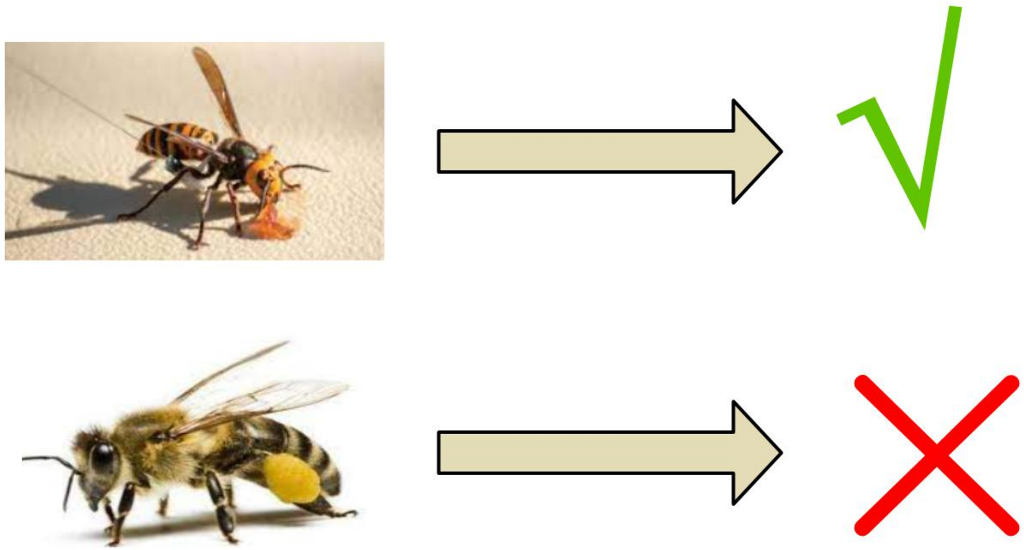
Figure 7: Positive Reports per Month

$$s_{\mathrm{T}} = \begin{cases} c_T, & T \in \mathbb{T} \\ 0, & \text{otherwise} \end{cases}$$

where c_T is the number of positive reports in month T over the historical data.

Features Extraction

3. Image Feature s_{I}



$$s_{\mathrm{I}} = \begin{cases} n_{\mathrm{I}}, & \text{image(s) provided} \\ 0, & \text{otherwise} \end{cases}$$

where n_{I} denotes the number of image(s) attached to the report.

4. Word Feature s_{W}

TABLE I
THE DICTIONARY OF KEY CHARACTERISTICS OF *V. mandarinia*

	Asian giant hornet	Other confusing hornets
Nest Location	"Underground","forests", "burrows","roots","trunks", ···	"Time limbs", "house eaves", "exposed", "lawns", ···
Body Appearance	"Yellow heads", "black thorax", "striped abdomens", "giant", ···	"Small", "black and white colored", ···

$$s_{\text{W}} = \frac{1}{n_{\text{W}}} \sum_{i=1}^{n_{\text{W}}} (q_i - k_{\text{i}}) + \beta_2 \log(n_{\text{W}} + 1)$$

where q_i is the word frequencies of word w_i in the dictionary of Asian giant hornet feature, k_i is the word frequencies in the confusing dictionary, n_{W} is the text length of report.

Classification Problem

- **Optimization Problem with Cross Entropy**
  - $x^{(i)}$: the sighting report $x^{(i)} \sim u(\cdot)$, where $u(\cdot)$ is a certain distribution that the report data is sampled from.
  - y: the report category label, $y \in \{0,1\}$
    - $y = 1$ denotes that the report confirms the existence of a nest
    - $y = 0$ means that the report is a false positive
  - $p^{(i)}$: the probability that report $x^{(i)}$ belongs to the true positive

$$\max \mathbb{E}_{x^{(i)} \sim u(\cdot)}\left[y^{(i)} p^{(i)} + \left(1 - y^{(i)}\right)\left(1 - p^{(i)}\right)\right]$$

  - $\phi\left(x^{(i)}\right)$: feature representation

$$\phi\left(x^{(i)}\right) = \left[s_L^{(i)}, s_T^{(i)}, s_I^{(i)}, s_W^{(i)}\right]^T$$

  - f: fit a classifier f solve the above Equation by $p^{(i)} = f\left(\cdot | \phi\left(x^{(i)}\right); \theta\right)$

Classification Prediction

- **Feature Representation**

$$\phi\left(x^{(i)}\right) = \left[s_L^{(i)}, s_T^{(i)}, s_I^{(i)}, s_W^{(i)}\right]^T$$

- **Logistic Regression**

$$p\left(x^{(i)}\right) = \frac{1}{1 + e^{-\theta^T \phi\left(x^{(i)}\right)}}$$

- **Loss Function with *Weighted Cross-Entropy***

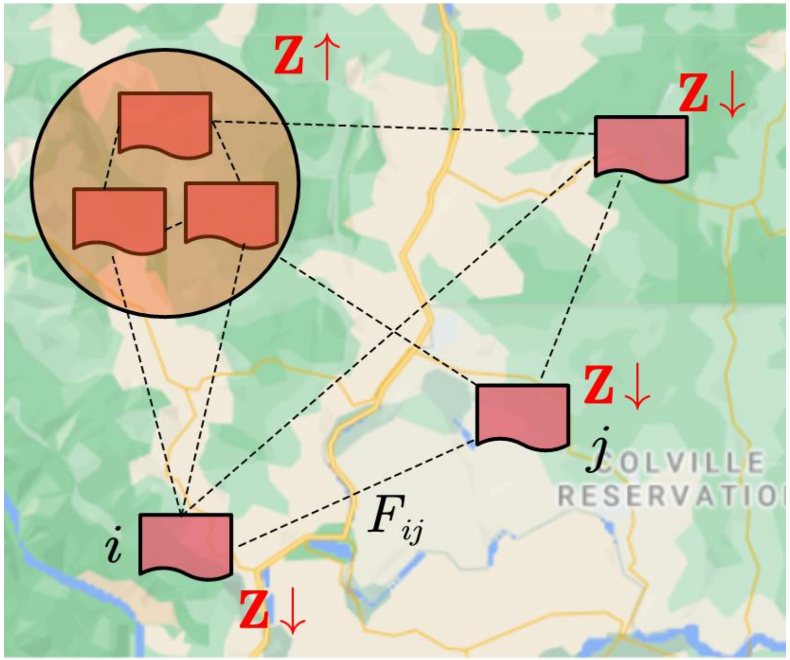
$$H(p,y) = \tau y \log(p) + (1-y)\log(1-p)$$

$$\min J(\theta) = \frac{1}{n}\sum_{i=1}^{n} H\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \frac{\beta_3}{2}||\theta||_2^2$$

- **Stochastic Gradient Descent**

$$\theta' \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$

Priority Prediction



- **Mutual Influence Factor**

$$F_{i,j} = \begin{cases} e^{-\lambda d_{ij}}, & \text{if } t_i = t_j \\ \dfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}\left(d_{ij} - 30\Delta t_{i,j}\right)^2}, & \text{otherwise} \end{cases}$$

- **Priority Evaluation**

$$Z_i = \sum_{j=1}^{n} F_{i,j}\, p_j$$

The greater the Z-value of a report, **the more likely** it is to be positive, and thus the **sooner it should be investigated**.

Contents

Dataset & Cleaning

- **Dataset**
  - **From:** WSDA (Washington State Department of Agriculture)
  - **Ranging:** from 2019 to 2021
  - **Data Cleaning:** 4,400 -> 4,355
  - **Highly Unbalanced:** only 0.3%(14/4355) is positive report

Detection Date	Notes	Lab Status	Lab Comments	Submission Date	Latitude	Longitude
2020-2-29	I'm not sure what this is, but it was the biggest looking wasp/hornet I've ever seen, at least an inch long. Sorry for the poor quality picture, I went inside to get a glass to catch it and it flew away.	**Negative ID**	This is a large fly that mimics bees! Thanks for submitting it.	2020-2-29	48.729596	-122.480035
2019-10-30	Hornet specimen sent to WSU	**Positive ID**		2020-1-15	48.971949	-122.700941

Parameter Setting

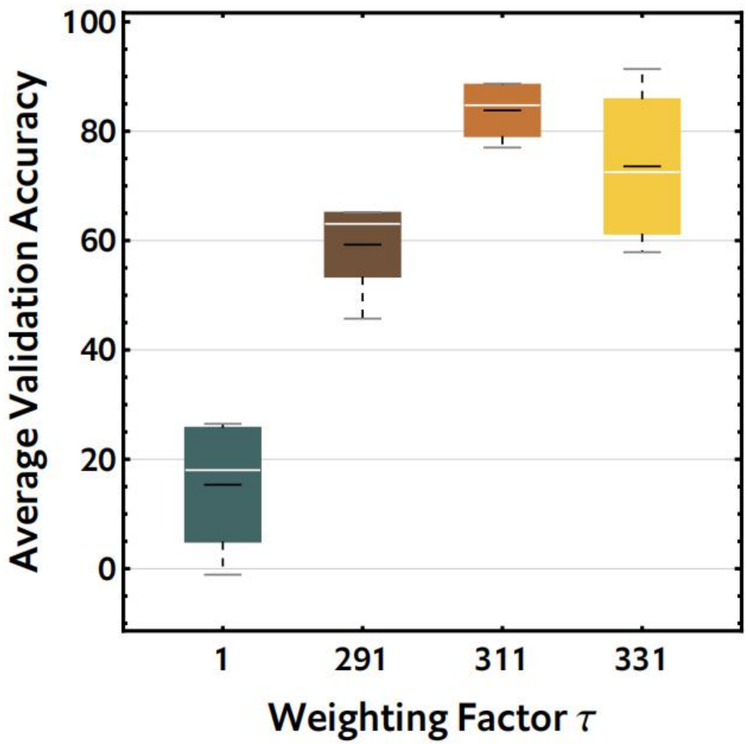
- **Parameter Setting**
  - **The initial distribution**: the confirmed sighting report locations of the data set from **March 2019 to April 2020**.

$$P_{\text{nest}}^{(0)}(\lambda,\phi) = \begin{cases} 1, & \text{if}(\lambda,\phi) \in \mathbb{O}_{positive} \\ 0, & \text{otherwise} \end{cases}$$

  - …….

Results of Classification Prediction

- **The influence of balance factor τ**



- On the **highly unbalanced** dataset, we achieve weighted accuracy of **83.5%.**

Results of Priority Prediction

- **Priority Prediction with the Mutual Influences**
- **Comparison**
  - **Baseline:** directly take $p^{(i)}$ as Priority
  - **Ours:** consider mutual influence $F_{i,j}$

- Our method outperforms the baseline method.

Contents

Conclusion

Contributions

- We formalized the problem of priority prediction of sighting reports into a two-category problem.
- To characterize a sighting report, we construct rich features from four dimensions: location, time, image and text.
- We purpose a machine learning model based on logistic regression to predict the credibility of the report, and determined their priority based on the relationship among the reports.

Thank You

Thank You

Q&A