Machine Learning in identification and control of biological disasters

Journal of Computer Science IJCSIS

Vol. 19 No. 4 APRIL 2021 International Journal of Computer Science and Information Security (IJCSIS)

Machine Learning in identification and control of biological disasters

Chen Ling, Xingjian Lyu, Tian Zhenyu, Gabriela Mogos[4]
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
[4] Gabriela.Mogos@xjtlu.edu.cn

***Abstract*— Biological invasion is a very common phenomenon. All alien species must have sites to grow and reproduce, which allows them to be called "colonizer" in a general sense. They would have adverse effects on the agricultural ecosystem by reducing the growth and output of the desired species [1].**

In this case, the invasion of wasps caused serious potential adverse effects on bee populations in Europe, as well as on the other local biological populations. To reduce this adverse effect, avoiding invalid identify and reducing the possibility of the biological invasion, we construct model to identify wasps.

In this paper, we use Grey Forecast Model and Convolution Neural Network Model to improve the accuracy of identification and better control the disaster.

I. Introduction

Researchers and practitioners apply such algorithms to data for two main reasons: to predict new data and to better understand existing data. To predict something about new data, researchers collect data, apply an algorithm, and analyzes the resulting model to gain insights into the data.

In our study, firstly, we construct Grey Forecast Model to predict the tendency in November and December and evaluate it by posteriori test. Then we use the natural language processing in python to clean the data and leave the needed terms. After that, we use convolution neural network for image processing.

Finally, we find that variance has a positive correlation to the severity of the disaster and discuss the variance to evaluate the eradication of vespa mandarinia.

II. Detailed Assumption and Notations

A. Assumption

1. The number of reports in one year is not periodic.
2. The disaster about vespa mandarinia is short-term.
3. The number of reports has some relations with respect to time [i.e., it is not random].
4. The population distribution in the state is fixed for the recent several years.
5. The result of separating negative to Negative1 and Negative2 by NLP is accurate.

B. Notation

1. *Negative1*: It is bees but not vespa mandarinia while be negative.
2. *Negative2*: It is not a kind of bee while be negative.

III. Grey Forecast Model

Since the harmful effect of vespa mandarinia is short-term, the relatively early data cannot well represent the current situation and predict the future situation. In this case, we only select the data of the last few months for prediction.

Simultaneously, due to the long interval between detection date and submission date, the data in November and December are not of reference significance. Therefore, we use the Grey Forecast Model to predict the development of the next two months, through the data of four months from July to October.

A. Principle of Grey Forecast Model

Let $X^{(0)}$ be a set:

$$\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(N)\}$$

Let $X^{(1)}$ be a set:

$$\{x^{(0)}(1), x^{(0)}(1)+x^{(0)}(2), \ \cdots, \ x^{(0)}(1)+x^{(0)}(2)+\cdots+x^{(0)}(N)\}$$

If we write it as $x^{(1)} = \{\sum_{i=1}^{n} X^{(0)}(i): n = 1,2,\cdots N\}$,

Then $x^{(k)} = \{\sum_{i=1}^{n} X^{(k-1)}(i): n = 1,2,\cdots N\}$.

By building a differential equation to discrete sequences, it can be in the form:

$$\frac{dx}{dt}+ax=u''$$

if we build a first-order differential equation.

By the definition of derivative:

$$\frac{dx}{dt} = \lim_{\Delta t\to 0}\frac{x(t+\Delta t)-x(t)}{\Delta t},$$

when $\Delta t \ll t$ let Δt be 1, approximately.

We have written it into the discrete form:

$$\frac{\Delta x}{\Delta t} = x(k+1) - x(k) = \Delta^{(1)}(x(k+1)).$$

Since $\frac{\Delta x^{(1)}}{\Delta t}$ is with respect to the value of two times, let $x^{(i)}(i)$ be $\frac{1}{2}\left[x^{(i)}(i)+x^{(i)}(i-1)\right]$, where i=2,3, …, N.

$$x^{(i)}=\frac{1}{2}\left[x^{(i)}(i)+x^{(i)}(i-1)\right],\ (i=2,3,\cdots N)$$

and

$$x^{(1)}(k+1)=\frac{1}{2}\left[x^{(1)}(k+1)+x^{(1)}(k)\right]$$

Solving

$$\begin{cases} \frac{\Delta x}{\Delta t}=x(k+1)-x(k)=\Delta^{(1)}(x(k+1)) \\ \Delta^{(1)}\left(x^{(1)}(k+1)\right)=x^{(1)}(k+1)-x^{(1)}(k)=x^{(0)}(k+1) \\ x^{(1)}(k+1)=\frac{1}{2}\left[x^{(1)}(k+1)+x^{(1)}(k)\right] \\ \Delta^{(1)}\left(x^{(1)}(k+1)\right)+a(x(k+1))=u \end{cases}$$

it follows that:

$$x^{(0)}(k+1)=a\left[-\frac{1}{2}\left(x^{(1)}(k)+x^{(1)}(k+1)\right)\right]+u.$$

Transform it into matrix:

$$\begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(N) \end{bmatrix}=\begin{bmatrix} -\frac{1}{2}\left[x^{(1)}(2)+x^{(1)}(1)\right] & 1 \\ -\frac{1}{2}\left[x^{(1)}(3)+x^{(1)}(2)\right] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}\left[x^{(1)}(N)+x^{(1)}(N-1)\right] & 1 \end{bmatrix}\begin{bmatrix} a \\ u \end{bmatrix}$$

Let y=$\left[x^{(0)}(2),x^{(0)}(3),\cdots,x^{(0)}(N)\right]^T$

$$B=\begin{bmatrix} -\frac{1}{2}\left[x^{(1)}(2)+x^{(1)}(1)\right] & 1 \\ -\frac{1}{2}\left[x^{(1)}(3)+x^{(1)}(2)\right] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}\left[x^{(1)}(N)+x^{(1)}(N-1)\right] & 1 \end{bmatrix},U=\begin{bmatrix} a \\ u \end{bmatrix}$$

We have y=BU and

$$\widehat{U}=\begin{bmatrix} \hat{a} \\ \hat{u} \end{bmatrix}=(B^TB)^{-1}B^Ty$$

B. Precision Test and Evaluation

We have the residual error:

$$e(k)=x^{(0)}(k)-\hat{x}^{(0)}(k),k=1,2,\cdots,n.$$

Let $x^{(0)}$ and $E=\{e^{(k)}\}_{k=1}^{n}$ be two sequences,

we have the residual error respectively:

$$\begin{cases} S_1^2=\frac{1}{n}\sum_{k-1}^{n}\left[x^{(0)}(k)-\bar{x}\right]^2 \\ S_2^2=\frac{1}{n}\sum_{k-1}^{n}\left[e(k)-\bar{e}\right]^2 \end{cases},$$

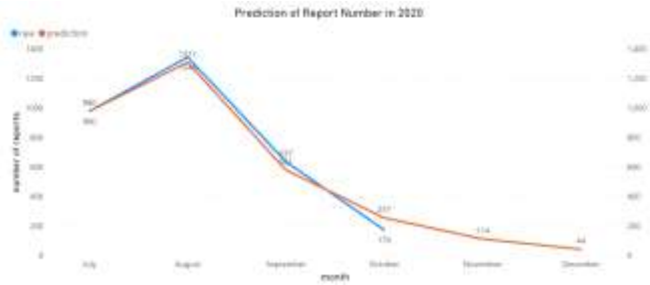
where $\bar{x}=\frac{1}{n}\sum_{k-1}^{n}x^{(0)}(k)$ and $\bar{e}=\frac{1}{n}\sum_{k-1}^{n}e(k)$

.

C. Evaluation

By calculating the posterior error, the posterior error ratio of pass the precision test. [Note: here we use code to examine the data and thereby no more explanation]

D. Results of the Model



In this graph, the blue line represents the origin data [i.e., total count by month from July 2020 to October 2020] and the orange one represents the data after fitting, which simultaneously predicts the number of reports in November and December.

E. Model Results

In this model, we utilize the data for four months from July to October in 2020 and forecast the trend for the latter two months.

The total count strictly decreases after August and the rate of decreasing tends to be smaller month by month.

Hence, we predict that in November the total count is 114 and in December is 44. Note that since Gray Forecast Model based on small number of data is not suitable to predict for a long-term tendency, it cannot be precise if forecasting the trend in 2021.

IV. Convolutional Neural Network Model

To calculate the likelihood of mistaken classification (mistake other hornets for Vespa mandarinia, defined as Negative1) and prioritize the investigation of the reports most likely to be positive sightings, we design a CNN deep learning model to do classification.

A. Data Cleaning

Drop All Data with Wrong Format Data *in the Submission Date Column*

```
raw["Detection Date"] = raw["Detection Date"].astype(str)
sorted_raw = raw.sort_values("Detection Date").reset_index(drop=True)
```

```
sorted_raw.head(5)
```

```
sorted_raw.drop(index=[i for i in range(10)],inplace=True)
```

Drop Rows with Null Value.

```
dataset = dataset.dropna().reset_index(drop=True)
```

Drop all Records without Submitted Image or Movie.

```
selected_file_format
['image/jpg', 'image/png', 'video/quicktime', 'video/mp4']
dataset = dataset[dataset["FileType"].isin(selected_file_format)].reset_index(drop=True)
```

***Drop all Records Labelled** with "Unverified ID".*

Those Data, which cannot be Judged by Experts, can be Noise of our Model.

```
selected_lab = ['Negative ID','Positive ID']
raw_new_lab = raw_new_lab[raw_new_lab["Lab Status"].isin(selected_lab)].reset_index(drop=True)
```

B. *Data Labelling*

Since negative label Negative1 mentioned above is not given, natural language processing is introduced to separate the negative label into Negetive1 and Negative2 (mistake other species instead of hornets for Vespa mandarinia).

Convert to Lower Case

Remove Punctuation

Remove Stopwords

Lemmatization

Lemmatization is the process of converting a word to its base form. This has great advantages than stemming, for lemmatization considers the context.

In natural language processing, lexical features and structural features are very important [2], word will be converted to its base form, but stemming just removes the last few characters, which may give rise to incorrect meanings and spelling errors.

Tokenize and Assign Part of Speech

Since the lab comments have an almost fixed format, which means noun words can indicate whether this report showing the specific situation (Negative1).

Generate Top Noun Words and Filter Them

```
noun=[]
for record in df["Lab Comments_tag"]:
    for word in record:
        if(word[1]=="NN"):
            noun.append(word[0])
```

```
pd_noun = pd.Series(noun).value_counts()
```

```
pd_noun.head(10)
```

```
wasp        695
sawfly      488
digger      468
horntail    359
hornet      259
wood        256
killer      202
look        158
cicada      151
please      121
dtype: int64
```

Manually Labelling

C. Prepare for Model Input

Enhance Picture Labelled Positive

As can be seen below, the distributions of different labels are quite different, especially the positive label.

```
dataset["Lab Status"].value_counts()

Negative2     1800
Negative1     1394
Positive        14
Name: Lab Status, dtype: int64
```

Therefore, we use image enhancement technology to expand those positive images from 14 to 200.

Convert movies

Since people submitted files with different formats, we convert the movie formats to image formats by extracting key frame. We also ignore files with other formats because they are trivial to the consequence.

```
dataset["FileType"].value_counts()

image/jpg                                                                  3032
image/png                                                                    75
video/quicktime                                                              76
video/mp4                                                                     9
application/pdf                                                               5
application/vnd.openxmlformats-officedocument.wordprocessingml.document       3
application/x-zip-compressed                                                  2
application/octet-stream                                                      1
Name: FileType, dtype: int64
```

D. Apply CNN model

The recognition ability of convolutional neural network is highly reliable and effective [3], so we choose this model to train our dataset.

Convert Image to Array Using Tensorflow

```
import tensorflow.compat.v1 as tf
for i in range(len(cnn_data)):
    img = tf.gfile.FastGFile(cnn_data['FileName'][i],'rb').read()
    with tf.Session() as sess:
        X.append(tf.image.decode_jpeg(img).eval())
```

Build the Model

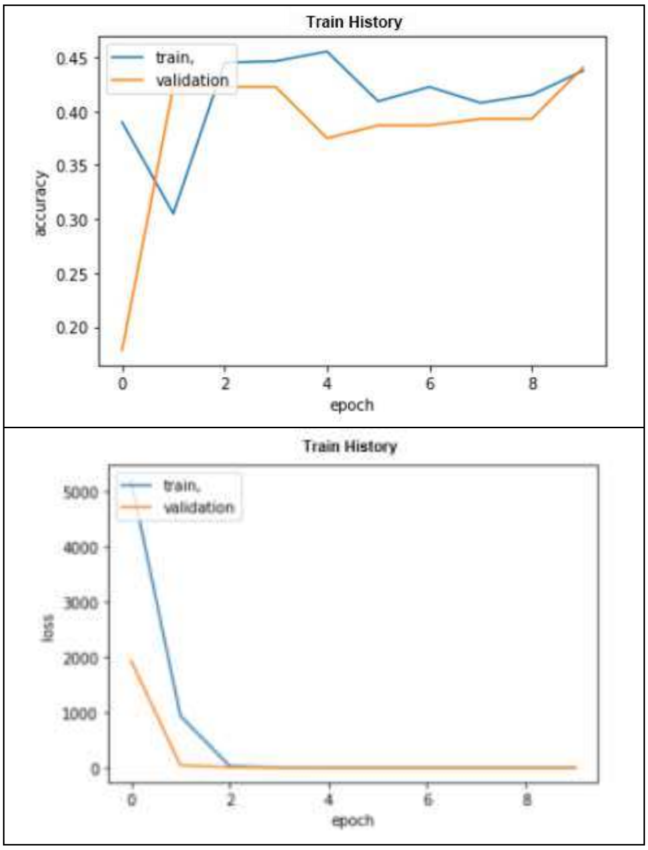
```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_12 (Conv2D)           (None, 256, 256, 16)      416
_________________________________________________________________
max_pooling2d_12 (MaxPooling (None, 128, 128, 16)      0
_________________________________________________________________
conv2d_13 (Conv2D)           (None, 128, 128, 36)      14436
_________________________________________________________________
max_pooling2d_13 (MaxPooling (None, 64, 64, 36)        0
_________________________________________________________________
dropout_12 (Dropout)         (None, 64, 64, 36)        0
_________________________________________________________________
flatten_6 (Flatten)          (None, 147456)            0
_________________________________________________________________
dense_12 (Dense)             (None, 128)               18874496
_________________________________________________________________
dense_13 (Dense)             (None, 3)                 387
=================================================================
Total params: 18,889,735
Trainable params: 18,889,735
Non-trainable params: 0
_________________________________________________________________
None
```

conv2d: in convolutional networks, neural networks that use convolution in place of general matrix multiplication in at least one of their layers.
For a specific kind of weight matrix W.

$$h = \sigma(W^T x + b)$$

maxpooling2d : Pooling layer is used to progressively reduce the spatial size of the representation to reduce the amount of features and the computational complexity of the network.
dropout: used to avoid overfitting.
flatten: bridge convolutional layer to other layers.

E. Model Evaluation



It shows that the model performance is not satisfied our expectations, and the accuracy is around 45%.

It may be caused by the limited number of positive images since using image enhancement may not be enough.

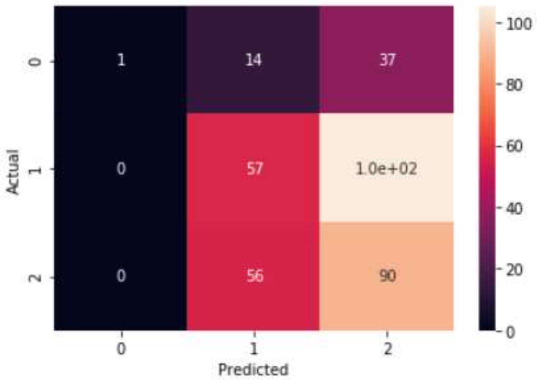
Moreover, by focusing on the matrix, the chance for recognizing a positive image is comparably lower.

F. Model Results

The model output is a 3*1 matrix which represents the possibility of being classified to different classes.

The second parameter (Negative1) can display the possibility of a mistaken classification of the report.

The first parameter (Positive) can be used to prioritize reports when multiple reports come together to the institution.



0 – Positive
1 – Negative1
2 – Negative2

V. Conclusion

Although we devote a lot to the *CNN Model*, its accuracy is unsatisfied.

Therefore, other classification models are introduced to distinguish records between positive and negative. Those models are based on latitude, longitude, and detection month. We hope the *CNN model* could evolve and become more accurate as the positive reports increases, however, those other models can be referred currently to prioritizing the investigation in addition to the *CNN*.

We can conclude that all classification models perform well on training set, but when it comes to the f1 score, they all get low marks. *RandomForest* has highest f1 score. It is because that *RandomForest* is good at dealing with unbalanced sample, which is consistent with the situation.

References

[1] FA Bazzaz, Life history of colonizing plants: some demographic, genetic and physiological features. In: Mooney HA, Drake JS (eds) Ecology of biological invasions of North America and Hawaii, Springer-Verlag, New York, pp 96–110, 1986.
[2] M.E. Santholma, Marianne, Grammar sharing techniques for rule-based multilingual NLP systems. In: Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA). Tartu (Estonia), 2007. https://archive-ouverte.unige.ch/unige:3455.
[3] F. Takeda and S. Omatu, "High speed paper currency recognition by neural networks," in IEEE Transactions on Neural Networks, vol. 6, no. 1, pp. 73-77, Jan. 1995, doi: 10.1109/72.363448.
[4] M. Matsuura, Ecological study on vespine wasps (Hymenoptera: Vespidae) attacking honeybee colonies. I. Seasonal changes in the frequency of visits to apiaries by vespine wasps and damage inflicted, especially in the absence of artificial protection. Applied Entomology and Zoology, 23(4): 428–440, 1988.

IJCSIS Call For Papers 2021-2022

https://sites.google.com/site/ijcsis/

The topics suggested by the journal can be discussed in term of concepts, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete ***unpublished papers***, which are ***not under review in any other conference or journal*** in the following, but not limited to, topic areas. All tracks are open to both research and industry contributions.

Ad Hoc & Sensor Network **Ad hoc networks for pervasive communications** **Adaptive, autonomic and context-aware computing** **Advanced Computing Architectures and New Programming Models** **Agent-based middleware** **Autonomic and self-managing middleware** **B2B and B2C management** **BioInformatics** **Bio-Medicine** **Biotechnology** **Broadband and intelligent networks** **Broadband wireless technologies** **Cloud Computing and Applications** **Collaborative applications** **Communication architectures for pervasive computing** **Communication systems** **Computational intelligence** **Computer and microprocessor-based control** **Computer Architecture and Embedded Systems** **Computer Business** **Computer Vision** **Computer-based information systems in health care** **Computing Ethics** **Context-awareness and middleware** **Cross-layer design and Physical layer based issue** **Cryptography** **Data Base Management** **Data Mining** **Data Retrieval** **Decision making** **Digital Economy and Digital Divide** **Digital signal processing theory** **Distributed Sensor Networks** **E-Business** **E-Commerce** **E-Government** **Emerging signal processing areas** **Enabling technologies for pervasive systems (e.g., wireless BAN, PAN)** **Encryption** **Energy-efficient and green pervasive computing** **Event-based, publish/subscribe, and message-oriented middleware** **Evolutionary computing and intelligent systems**	**Knowledge based systems** **Knowledge management** **Location Based Services** **Management information systems** **Medical imaging** **Micro/nano technology** **Middleware Issues** **Middleware services and agent technologies** **Mobile and Wireless Networks** **Mobile Computing and Applications** **Mobile networks and services** **Multimedia Communications** **Multimodal sensing and context for pervasive applications** **Multisensor fusion** **Natural Language Processing** **Network management and services** **Network Modeling and Simulation** **Network Performance; Protocols; Sensors** **Networking theory and technologies** **Neural Networks** **Neuro-Fuzzy and applications** **Open Models and Architectures** **Open Source Tools** **Operations research** **Optical Networks** **Pattern Recognition** **Peer to Peer and Overlay Networks** **Perception and semantic interpretation** **Pervasive Computing** **Performance optimization** **Positioning and tracking technologies** **Programming paradigms for pervasive systems** **Quality of Service and Quality of Experience** **Real-time computer control** **Real-time information systems** **Real-time multimedia signal processing** **Reconfigurable, adaptable, and reflective middleware approaches** **Remote Sensing** **RFID and sensor network applications** **Scalability of middleware** **Security and risk management** **Security middleware**

Expert approaches **Fuzzy algorithms** **Fuzzy logics** **GPS and location-based applications** **Green Computing** **Grid Networking** **Healthcare Management Information Technology** **Human Computer Interaction (HCI)** **Image analysis and processing** **Image and multidimensional signal processing** **Image and Multimedia applications** **Industrial applications of neural networks** **Information and data security** **Information indexing and retrieval** **Information Management** **Information processing** **Information systems and applications** **Information Technology and their application** **Instrumentation electronics** **Intelligent Control System** **Intelligent sensors and actuators** **Internet applications and performances** **Internet Services and Applications** **Internet Technologies, Infrastructure, Services & Applications** **Interworking architecture and interoperability**	**Security, Privacy and Trust** **Security Systems and Technolgies** **Sensor array and multi-channel processing** **Sensor fusion** **Sensors and RFID in pervasive systems** **Service oriented middleware** **Signal Control System** **Signal processing** **Smart devices and intelligent environments** **Smart home applications** **Social Networks and Online Communities** **Software Engineering** **Software engineering techniques for middleware** **Speech interface; Speech processing** **Supply Chain Management** **System security and security technologies** **Technology in Education** **Theoretical Computer Science** **Transportation information** **Trust, security and privacy issues in pervasive systems** **Ubiquitous and pervasive applications** **Ubiquitous Networks** **User interfaces and interaction models** **Virtual reality** **Vision-based applications** **Web Technologies** **Wired/Wireless Sensor** **Wireless technology**

ISSN 1947-5500