West Nile Virus Prediction

Xuechao Wu, xuechaow@usc.edu

12/07/2015

1. Project Homepage

https://github.com/xuechaow/EE660_Final_Project

2. Abstract

West Nile Virus has been outbreak since 2006, according to Times.com. An prediction of next outbreak is demanded by the CRC. This paper uses data collected in Chicago Area to predict the next appearance of WNV in Chicago. This paper focus on 3 important features related to virus, which are Mosquito Species, Location and Time. Establish three individual models for their conditions and the output of WNV presence. Use another logistic regression to combine these three model together into a general prediction model. Use Random Tree Approach to compare with mixed Model. With an Etest = 0.69, the result is partly useful to predict the presence. A secondary target is to predict mosquito activities.

3. Problem Statement and Goals

Given weather, location, testing, and spraying data, this task requires to predict when and where different species of mosquitos will test positive for West Nile virus. Specifically, test input X contains Date and Location, try to solve to presence probability. The task has 2 stages, first use cross validation to build an accurate (low E_in) Model.

(1) Test Result is difficult to acquire.
Because this is a competition task, the final test result is unclear. There is only test result as probabilities. Pmtk only provides randomTree function with output without Probabilities.

(2) Significant amounts of preprocessing required.
The training data has only less than 13 features, but those features has high independency. Date and Weather has strong periodic patterns. Location Data has strong cluster features. And Mosquito need to be detected and translated into math features. The pre-processing of raw data is time consuming and easily corrupt into wrong data.

4. Literature Review

Several people on Leaderboard used Motion and Deep Learning Method. There are also Neural Network and Random Tree approaches with >0.73 accuracy.

5. Prior and Related Work

Prior and Related Work – None. Just a competition expired on Kaggle.

6. Project Formulation and Setup

The most tricky part and dangerous part is the features' patterns. Dimensions are complicated but there are several related features such as [Address trap # (Longitude, Latitude)], and [Date (or post-processed season #) Weather]. If treat those features independently, chances are model will over-fit with non-singular feature dimension.

To avoid such misleading approach, the basis of the entire model is constructed from three branch models, each weighing differently. Model feature selection is based on feature analysis. Each Model may fix one feature of another to simplify the compilation and reduce the feature reuse.

(1) Date and Trap Model.
Date indicates seasons. Seasons controls raindrops. Humidity influence Mosquito Activities. So date may influence significantly on Virus. Using timeline analysis on specific place to investigate relations between date and WNVpresence. Here use reprocess date data into season and weather periods. Use K-means Clustering method to semi-supervise this model.

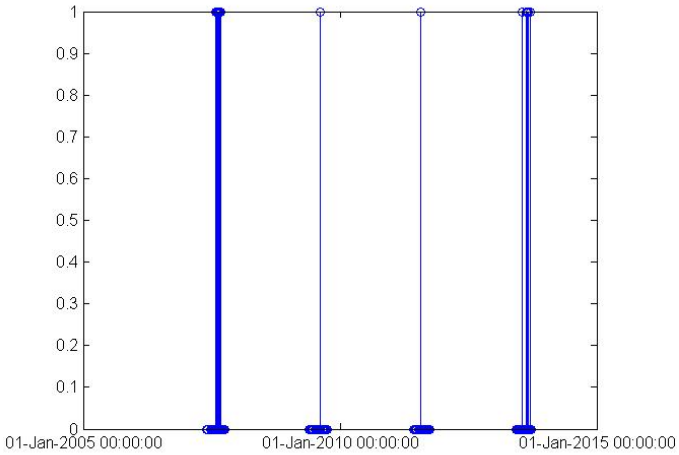
(2) Area Propaganda Model with Weather Indicator
Virus outbreak need media, or mosquitoes. So there is need to research on presence of virus in a fixed time window. K-means Clustering and N-Nearest Neighbor Recognition is applied.

(3) Species-Related Virus Break Model
Use a 2-D feature L2 regression to obtain within fixed area cluster.

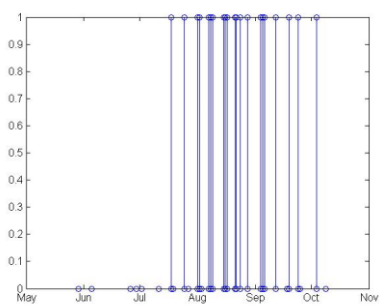
The final model will be added by those three model with different weights which need to be tuned. Use add rather than multiplication in order to reduce influence the high-order noise in duplicated features.

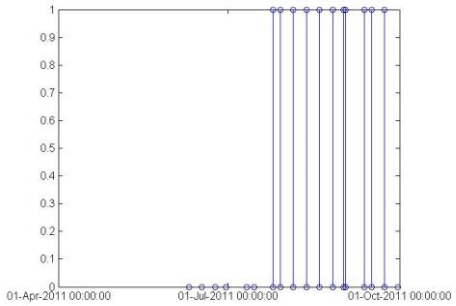
7. Methodology

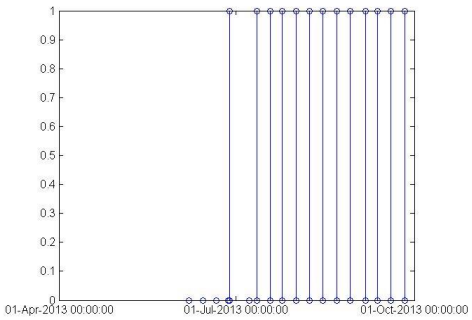
For Model 1, Logistic Regression would be best fit for a 2-class classifier. First of all analysis the appearance against time. The result are:

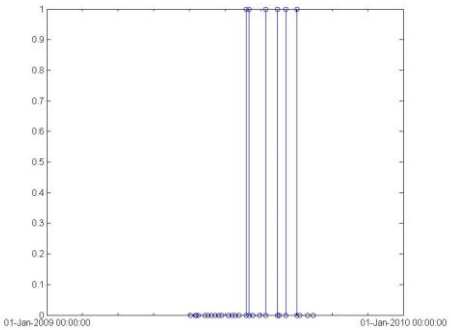


It tells the appearance follows a periodic regulation. So further mode analysis each year.









Train the model with Train Data Set.

For model 2, cluster near trap results together to predict the probability.

For model 3, use another logistic regress to get the classifier.

8. Implementation

8.1. Feature Space

Name	Value	Usable
ID	Number	Yes
Address	String	No
Date	Date/Number	Yes
Speicies	String	No
Trap ID	Number	Yes
NumMosquitos	Number	Yes
Latitude Longitutde	Number	Yes
Block Street Address address accuracy	Number	No
WNVpresented	Boolean	yes
Weather data and spray location	Boolean	no

8.2. Pre-processing of Training Data

Name	Description	Usable
ID	The id of the record	Yes
Address	The Full Address of Trap	No
Date	Date the test is performed	Yes
Speicies	The species of mosquitos	No
Trap ID	The id of the trap	Yes
NumMosquitos	The number of the mosquitos	Yes
Latitude Longitutde	Geography data	Yes
Block Street Address address accuracy	Social location description	No
WNVpresented	Whether the virus presents	yes
Weather data and spray location	Additional data	no

While in Test set, for consideration from the Website, the mosquito number is not available, and thus became a secondary prediction target.

Additional Data Set is for enhancing accuracy, including spray data and weather data.

Feature Extraction

Locations features in this dataset are extremely duplicated. We only use Latitude, Longitude, block number and Trap Number for location features. Replace the mosquito species with number from 1 to 8. Change Date to Uniform Time Value

Type	Feature Value
CULEX PIPIENS/RESTUANS	1

CULEX RESTUANS	2
CULEX PIPIENS	3
CULEX SALINARIUS	4
CULEX TERRITANS	5
CULEX TARSALIS	6
UNSPECIFIED CULEX	7
CULEX ERRATICUS	8

8.3. Training Process

After the pre-process. Most of the data is usable. But date number is way too larger than others, and we need to normalize date number to avoid overfitting.

9. Final Results

Best Prediction is 56%.

```
ans =

L2

errorRate =

   0.0521

errorRate_2 =

   0.0560
```