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) Specy-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

Describe the entire procedure you followed to get your model, train it, and evaluate it. This will include (to be stated in the sequence used): the hypothesis set (or you may even have multiple hypothesis sets at different levels), at what point the dataset was brought into the picture, when its various subsets were used (validation or test sets, etc.), training procedure, and model selection (if any), and validation and test procedures.

You might consider using a flow chart to illustrate your framework. Details in each step could be relatively brief, as they can be covered in the next section.

8. Implementation

Report your implementation details and results here. You can follow the steps described in Section 6 “Methodology”.

Typically the following steps are covered:

8.1. Feature Space

Describe the dataset you used, explain the meaning and data type (integer, real, string or binary, categorical, etc.) of each feature.

8.2. Pre-processing of Training Data

Main dataset consists of different data sets, which are: Training Set, Testing Set, and Additional Set. Training set includes

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

Describe how you train your model, the classifiers or regression processes you use, and the parameters you chose.

If a parameter is chosen by heuristics, state so. If a parameter is chosen by some model selection or validation process, you can cover it in the next subsection.

Analyze the complexity of your hypothesis set. Give the number of samples you have and the dimension of the pre-processed feature space. Explain what you did to avoid overfitting and underfitting.

8.4. Testing, Validation and Model Selection

You must run enough tests to show that your method works. Perform cross validation if needed.

If you have multiple potential models in the beginning, explain how you perform model selection in detail.

Report testing results here. In the case of classification, you can choose two salient features and plot your decision boundaries w.r.t. them. In the case of regression, you can plot the resulting regression function in 3D (as a function of two salient features) or in a few 2D plots (each as a function of one salient feature).

If you have a large dataset, it might be reasonable to use validation process to tune your model. If you did that, state so.

9. Final Results

Describe and document your final results. Different from the results in Section 7.4., the results here should be the final performance of your system(s) and an estimate of its out of sample performance. Figures, plots, and/or tables can be useful.

If you are working on an online competition, report the performance of your best submission and compare it to others on the leader board. If you want to compare your results with other work, do so here or in the Interpretation section below.

10. Interpretation

Why do you think the results came out the way they did? What has been learned from them? Anything particularly noteworthy or unexpected?

11. Summary and conclusions

Briefly summarize key findings, and optionally state what would be interesting or useful to do next.

12. Reference

Never forget to cite your references.