Report on West Nile virus forecasting challenge

Summary

This research work aims at predicting the total number of neuroinvasive diseases (e.g. meningitis, encephalitis, or myelitis) due to prevalent West Nile Virus (WNV) for each county in contiguous United States for upcoming years. This prediction will allow the Department of Public Health of respective areas to proactively react when and where the WNV transmission occurs.

The experiment involves 19 years of historical data ranging from 2000 to 2018 which includes the total number of deaths due to neuroinvasive disease, weather data (temperature, precipitation, humidity, etc.) for each of 3109 counties in the United States. To better characterize the accuracy, the model is also included with Spatial relation with respect to the total number of cases by considering Neighbouring County Average deaths due to neuroinvasive disease.

Prediction is accomplished by developing a Long Short Term Memory (LSTM) recurrent neural network. LSTM is robust and well-suited to classifying and processing based on time series data. To include the temporal effect, the input parameter involved the past 5 years data to predict the current year; for each prior year used, the features included were weather (temperature, precipitation, humidity, and gini index of precipitation) and spatial variable (neighboring county average wnv cases). The target variable was a 15-element vector, and each element describes the probability of WNV cases and points to a particular bin. Each bin was defined by an inclusive minimum and a non-inclusive maximum number of cases of the disease.

A huge amount of collective and collaborative effort was made in developing this prediction model and as a result the model has an accuracy of 73% on unseen data that means the model can predict 73 times correctly out of 100 observations. However, while the current precision is relatively low, it outperforms the trivial prediction which has higher accuracy but misses all high number cases. Such a problem is caused by highly unbalanced data that most samples have 0 counts, which is typical in modeling epidemic diseases. Therefore, we would rather sacrifice accuracy so that valid warnings of potential outbreaks could be predicted.

Studies/Researches

There have been several studies and researches on WNV and its transmission, since the activity of human illness due to WNV is majorly in summers, which implies that climatic factors play a major role in it. Climatic change influences the emergence of vector-borne diseases such

as malaria, dengue and West Nile virus (WNV) by altering their rates, ranges, distribution and seasonality (Paz, Shlomit, 2015). Weather conditions (in particular temperature, precipitation and humidity) affect the survival and reproduction rates of the vectors, their habitat suitability, distribution and abundance (Paz, Shlomit, 2015).

Additionally, efforts have been made in discovering that suitable environments also permit species of mosquitoes to thrive. In the Los Angeles area, an increase in avian seroprevalence influenced numbers of reported human cases of West Nile neuroinvasive disease (Kwan et al. 2012). Housing unit density, neglected swimming pools, mean per capita income, increased mosquito breeding sites and ditches, and housing average age were additional risk factors for Orange County, California (Liao et al. 2014).

Methodology

Computational Resources

For the scope of this particular research, there were several resources put into use -

- In general, **Python 3.6** as a programming tool for data preparation, cleaning, manipulating and building model was employed along with its packages - Numpy, Pandas, Scikit-Learn, Keras, Scipy - and **Jupyter Notebook** was used as an IDE.
- **CyberGIS-Jupyter** was used for the preparation of weather data.
- **HAL (Hardware-Accelerated Learning)** Cluster was used for running the model. (This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.)

Data Sources/Dataset description

The weather data was obtained from the North American Regional Reanalysis group. The grain of the table dictates the average value of several variables like precipitation, temperature, etc. for each county for each year.

Data Cleaning

The data obtained was not in the digestible format for the predictive model and therefore needed some manipulation wrangling procedures before feeding as an input parameters.

Weather Data

Weather data for the years 2000 to 2018 was obtained from the North American Regional Reanalysis (NARR). It consists of 19 NetCDF files at a daily temporal resolution, and a 32 square kilometer spatial resolution. Three sets of files were downloaded: accumulated

precipitation (apcp), air temperature at the Earth's surface (air.sfc), and relative humidity at two meters above the Earth's surface (rhum.2m).

First, these files were converted to GeoTIFF format using the Geospatial Data Abstraction Library's (GDAL) gdal_translate function. Then, these files were re-projected from EPSG:5070 to EPSG:4326, to match the projection of our county shape file, using the same function.

A land mask, also provided by NARR, was applied using GDAL's raster calculator script. After masking, we have 6940 GeoTIFF files: one for each day of the time period, including leap days.

Using the Python module rasterstats' zonal_stats function, we input each of these daily files, along with the county shapefile. Make sure to include the parameter all_touched=True, which is recommended because the county sizes are small compared to the grid resolution of our raster files. We save the output DataFrame to a .csv file. All 6940 .csv files are then combined into one file.

Gini Coefficient of Inequality

We use the Gini coefficient of inequality to measure disparity in precipitation across the counties, for each year. This coefficient was originally developed to measure income and wealth distribution, but it was found to be useful in precipitation inequality as well.

Each year's data was summed weekly. Then, we used Python Spatial Analysis Library's (PySAL) Gini coefficient function, resulting in yearly inequality coefficients for each county.

Feature Engineering

In this model, there were several features constructed or introduced as an input parameter to LSTM network, to include the spatial relationship in the model.
This features includes:
1. Neighbouring County Average - The value of this feature is the average number of cases for a particular year in the neighbouring county with respect to the current county in consideration.

2. Outbreak Inclusion - The value is either 0 or 1, where 0 means if no outbreak happened ever and 1 means if there was any outbreak or even one case of neuroinvasive disease was reported.

3. Urban/Rural - The value is either 0 or 1, where 0 means if the county is Urban and 1 means if the county is Rural.

Handling Unbalanced Data

The input dataset constructed using WNV cases data was highly unbalanced with regard to the bin number (target class) assigned to every county, with most of the counties lying in bin 1. In

order to address the problem of unbalanced target classes, oversampling of the instances belonging to minority classes was performed such that the total number of instances lying in each class becomes the same, and equals to the number of instances belonging to bin 1. It should be noticed that the validation set was randomly drawn before oversampling the training set so that the training and validation set are unrelated.

Neural Network Modeling (LSTM)

To take into account the temporal effect of the prior years on the prediction for current year, training an LSTM model seemed to be the best choice. Since the original input dataset obtained from CDC consisted of data ranging from years 2000 to 2018 and the final predictions were to be made for year 2020, it clearly indicated that the target year in the training set should differ by a value of at least 2 from the prior years to be used for prediction. For example, if we use 2 prior years for prediction of bins' probabilities for year 2018, then 2015 and 2016 data should be used; 3 prior years would mean using data from 2014, 2015 and 2016. To design the training dataset, all possible years are used as the target class and the prior years features/data are used for the prediction of bin probabilities for those target classes. For instance, if we are using 2 prior years then we use year 2018 bins as target classes and the predictors as data from 2015 and 2016, year 2017 is also used as target year with predictors being year 2014 and 2015 data features, and so on and so forth until 2003 is used as the last possible target year. We made different training datasets by changing the number of prior years used for prediction, varying the number from 2 to 15, and consequently used those multiple training datasets for training various LSTM models.

Input Features

Two types of training datasets, with different feature vectors, are used for prediction.

The first type of training dataset corresponds to the one used for counties for which the weather data (temperature, precipitation, humidity and gini index) was available. For every prior year used in training dataset, the feature vector included: temperature, humidity precipitation, gini index, Neighbouring County Average, outbreak inclusion, number of WNV cases and urban/rural nominal variables.

For the rest of the 7 counties, namely 'Massachusetts-Nantucket', 'Virginia-Lancaster', 'Maryland-Calvert', 'Washington-San Juan', 'Virginia-Northumberland', 'South Dakota-Oglala Lakota', and 'Wisconsin-Ozaukee' for which the weather data was missing, another training set was created that had the feature vector comprising of: Neighbouring County Average, outbreak inclusion and number of WNV cases.

As a result of two datasets for two different classes of counties (with and without weather data), two different LSTM models are developed and the results of the two are then combined to give the final predictions.

Cross-Validation

5 fold cross validation is used for evaluating the model's performance. The best performing model used a value of number of prior years as 5, and gave an accuracy of 0.73 on the validation dataset.

Hyperparameter Tuning

The hyperparameters used in the LSTM model are chosen intuitively and are not tuned. The chosen network structure and the hyperparameters used are as follows:

Layer 1: Batch Normalization layer

Layer 2: Dense layer, number of output nodes=4, activation= "tanh"

Layer 3: LSTM, number of output nodes=15, activation= "tanh", dropout = 0.2, recurrent_dropout = 0.2.

Layer 4: Dense layer, number of output nodes=15, activation= "softmax"

Future Directions

The model will be further improved by adding the socio-economic features and also tuning the hyperparameters using the "Hyperas" python package, the results obtained after adding additional features and tuning the model will be updated in the next submission.