Challenge of Dengue Fever Prediction competition

Akiyuki Ishikawa
May 4, 2017

Proposal Review: https://review.udacity.com/#!/reviews/392445

About this proposal

I will challenge the "DengAI: Predicting Disease Spread" competition as my capstone project.

The competition site is here:

https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/

Domain Background

Preventing infectious diseases is a major challenge in order to keep our life healthy.
Dengue fever is one of infectious diseases, which are most commonly transmitted to people through the bites of infected mosquitos[1][2][3].
To predict spreading of the disease by using data is one of the challenging themes for us[4][5][6].

Dengue disease occurs tropical and sub-tropical parts of the world.
However, global warming effect may expand the area where the mosquitos can live and the influence spreads to the area where the people have not previously been harmed by this disease like Japan which is my country.

Knowing how predict the number of dengue cases, we can prepare and prevent proliferation of mosquitoes which serve as a medium for the diseases by spraying insecticide or removing water from locations where mosquito larvae live.

Motivation

Since this dataset contains multiple interesting type of data such as satellite data and time series data,
I think this is a good themes for not only checking my understanding of the techniques I learned in this course but also studying more techniques such as time series data analysis.

Problem Statement

In this project, I will construct my model to predict time series of the number of dengue cases in weeks. The model consumes past or current environmental features and gives the prediction of the number of Dengue disease cases.

In addition to creating a model, I want to analyze the dataset to answer the following questions.

- How the environmental information affects the number of the dengue infection cases?
- Are there any characteristic variations depending on the cities?
- What is the most important feature to predict dengue cases?

Datasets and inputs

The dataset is provided by drivendata's competition site:
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/data/
(or http://dengueforecasting.noaa.gov)

- Training Data Features:
  The features for the training dataset.
- Training Data Labels:
  The number of dengue cases for each row in the training dataset.
- Test Data Features:
  The features for the testing dataset

Details of data are as follows[7]:

- City and date indicators
  - city – City abbreviations: sj for San Juan and iq for Iquitos
  - week_start_date – Date given in yyyy-mm-dd format
- NOAA's GHCN daily climate data weather station measurements
  - station_max_temp_c – Maximum temperature
  - station_min_temp_c – Minimum temperature
  - station_avg_temp_c – Average temperature
  - station_precip_mm – Total precipitation

    - station_diur_temp_rng_c – Diurnal temperature range
- PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)
    - precipitation_amt_mm – Total precipitation
- NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)
    - reanalysis_sat_precip_amt_mm – Total precipitation
    - reanalysis_dew_point_temp_k – Mean dew point temperature
    - reanalysis_air_temp_k – Mean air temperature
    - reanalysis_relative_humidity_percent – Mean relative humidity
    - reanalysis_specific_humidity_g_per_kg – Mean specific humidity
    - reanalysis_precip_amt_kg_per_m2 – Total precipitation
    - reanalysis_max_air_temp_k – Maximum air temperature
    - reanalysis_min_air_temp_k – Minimum air temperature
    - reanalysis_avg_temp_k – Average air temperature
    - reanalysis_tdtr_k – Diurnal temperature range
- Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements
    - ndvi_se – Pixel southeast of city centroid
    - ndvi_sw – Pixel southwest of city centroid
    - ndvi_ne – Pixel northeast of city centroid
    - ndvi_nw – Pixel northwest of city centroid

The example of the features is here:

feature	value
city	sj
year	1990
weekofyear	18
week_start_date	1990-04-30
ndvi_ne	0.1226
ndvi_nw	0.103725
ndvi_se	0.1984833
ndvi_sw	0.1776167
precipitation_amt_mm	12.42
reanalysis_air_temp_k	297.572857143

reanalysis_air_temp_k	297.572857143
reanalysis_avg_temp_k	297.742857143
reanalysis_dew_point_temp_k	292.414285714
reanalysis_max_air_temp_k	299.8
reanalysis_min_air_temp_k	295.9
reanalysis_precip_amt_kg_per_m2	32
reanalysis_relative_humidity_percent	73.3657142857
reanalysis_sat_precip_amt_mm	12.42
reanalysis_specific_humidity_g_per_kg	14.0128571429
reanalysis_tdtr_k	2.6285714286
station_avg_temp_c	25.4428571429
station_diur_temp_rng_c	6.9
station_max_temp_c	29.4
station_min_temp_c	20
station_precip_mm	16

I will use these features as inputs. I will do some preprocessing since these features have some missing values. I think I can fill the missing value by using mean value or valid values of the nearest time from the missing data.
Furthermore, I remove some unimportant features in order to simplify the model and to understand what feature is important.

Solution Statement

The solution of this project is creating a model which predicts the number of Dengue disease cases. This model consumes past or current environmental features and gives the prediction of the number of Dengue disease cases. The model is constructed by using a supervised learning method.

In order to get a good model, I should do some preprocessing such as filling missing data and process outliers properly and specify the subset of important features.
After some preprocessing, I want to try the following method

- gradient boost tree (ex. xgboost[8])
- RNN (ex. keras[9])
- Some time series analysis (ex. prophet[10])

Benchmark Model

Benchmark analysis is given in the following page:
https://shaulab.github.io/DrivenData/DengAI/Benchmark.html.
In this page, a benchmark model is given by NegativeBinomial model which is one of generalized linear model.
I reproduce the result of the above page, and use the result as my benchmark.

Evaluation Metrics

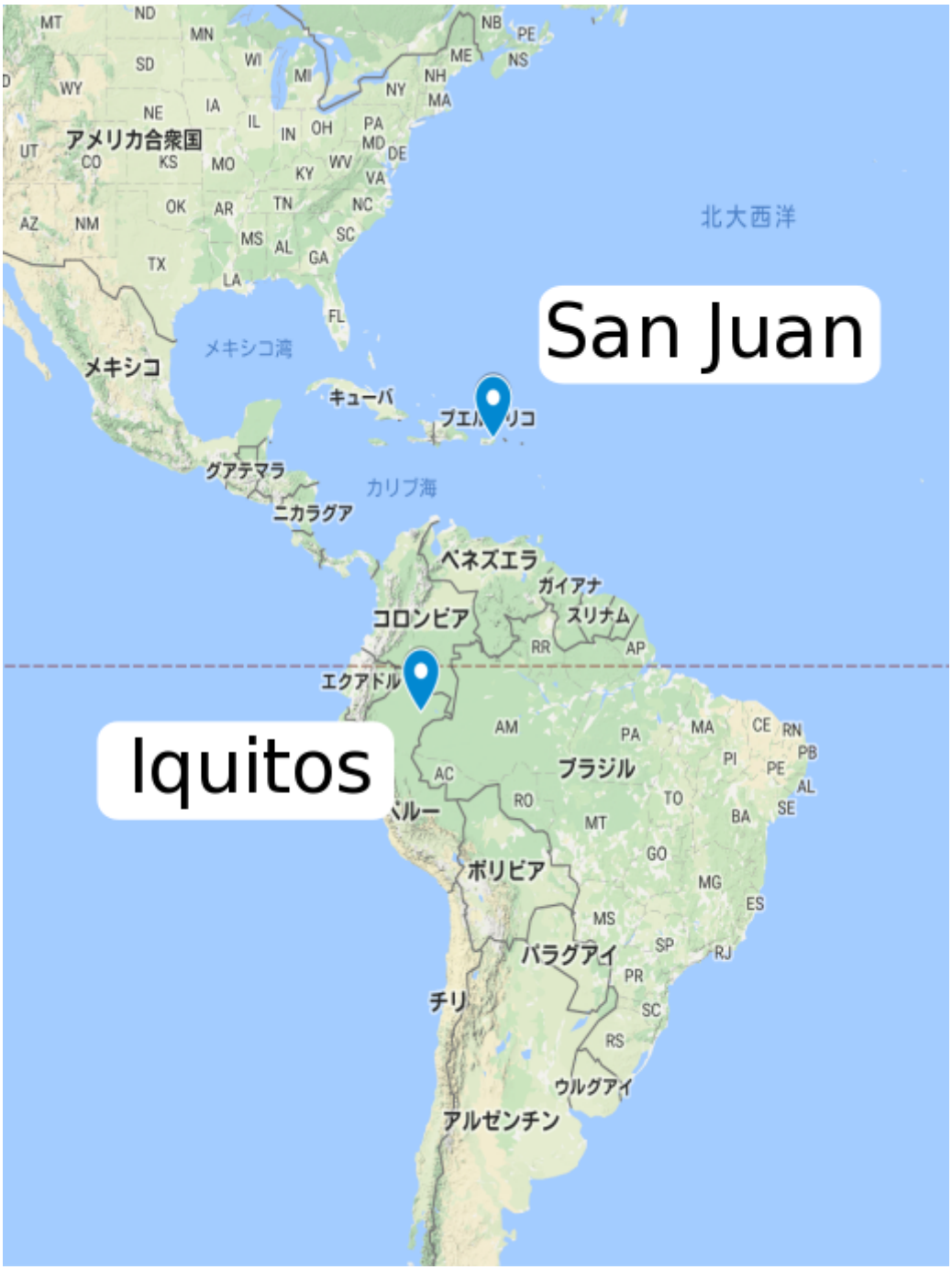
In this competition "Mean absolute error" is used[11].

$$\frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|,$$

where f_i is i-th prediction and y_i is true value.

Project Design

1. Split data by cities
   This dataset contains data from two cities **Iquitos** (Peru) and **San Juan** (Puerto Rico).
   These locations are very different.



It is reasonable to expect that the characteristic of relation between features and target value is also different. Therefore I split data by cities and create two different models.

2. Feature selection
   Correlation between each features and target (count of dengue cases) gives some information which feature is important for the prediction.
   As other approaches, feature selection technique can be used to specify which features are important[12].
3. Select validation set
   Since this is a time series problem, I will choose last one year training data as a holdout validation set.

4. Create models and estimate
   Using preprocessed data, I will train the model.
   Here, I will try some different approaches such as Gradient boost tree and RNN and etc.
   The performances of the models are estimated by using holdout validation set.
   According to the result of holdout validation, the model hyper-parameters are tuned.
5. Apply model to test data and submit the result
   The models are applied to test data and submit the result, because the target values of the test data are not published.

---

1. https://www.wikiwand.com/en/Dengue_fever ↩
2. https://www.cdc.gov/Dengue/ ↩
3. www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/ ↩
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4816319/ ↩
5. https://www.omicsonline.org/open-access/dengue-fever-prediction-a-data-mining-problem-2153-0602-1000181.php?aid=62375 ↩
6. https://www.kaggle.com/c/predict-west-nile-virus ↩
7. https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#features_list ↩
8. https://github.com/dmlc/xgboost ↩
9. https://github.com/fchollet/keras ↩
10. https://github.com/facebookincubator/prophet ↩
11. https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#mae ↩
12. http://scikit-learn.org/stable/modules/feature_selection.html ↩