CS4622 - Machine Learning

DengAI: Predicting Disease Spread

Final Project Report

Team Members:

140004E	Abeygunawardhana H. A. W.
140263U	Jayatilake S. A. P.
140462E	Perera P. D. I. T. S. K.
140623B	Thewa Hettige S. P.

Git Link: https://github.com/umstek/DengAI

Introduction

Dengue fever is a deadly viral disease transmitted by mosquitoes. Since spreading of this disease is positively correlated with climate factors such as temperature, humidity and precipitation amount, If we have enough historical data related to the dengue disease and the climate changes, it is possible to use the correlation of the data to predict an epidemic before it happens.

"DengAI: Predicting Disease Spread" is a competition hosted by DrivenData Foundation. The competition is about "predicting the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, and more" according to DrivenData.

This document presents the approach we took to produce competitive results when compared to results obtained by other data scientists participating in the DengAI competition.

Methodology

The downloaded data first was interpolated (effective than others) to impute missing values. The data and various statistics such as mean, standard deviation were visualized to help understand the dataset. Since there was a significant difference of the data from two cities, they were separated, and the separated datasets were used for further analysis/actions.

A cross-correlation analysis was done for all features of the dataset and redundant features that correlated for more than 95% were removed. All temperature values were converted into Celsius and station temperatures were merged using average into one feature and same was done for all NDVIs. Diurnal temperature feature looked totally random, so it has been removed. In San Juan, similar-looking reanalysis temperatures were merged to create one feature. (Data was smoothed for some models using a rolling mean/median.)

These data were then run through a 5-fold cross validation grid search to find the best models and hyperparameters. The prominent models AdaBoost, Boosted Trees, and Random Forest were used to predict results (Weka, Orange3, statsmodels and sklearn were used). They got results of MAE higher than 24. For some of the predictions, all the features were used while for others, only the features that maximally correlated with the target variable have been used.

Since the results were not sufficiently good, a time series analysis and prediction using R's STL decomposition has been done but the results were worse. Finally, seasonal component of the target variable has been modeled with linear regression and the prediction from that has been added to the prediction calculated from the trend taken from the 3 features were common for two cities that mostly correlated with the target variable. This yielded the best result so far.

Results and Analysis

Model	**Description**	**Typical MAE when submitted**
DNN	Deep (with multiple hidden layers) neural networks with all/some features	24 to 29
Random Forest	Random Forest Regressors with few relevant features	24+
AdaBoost/Boosted Trees	Different software/algorithms of boosting with all features, one week lagged, or most relevant features	24+

Boosted Trees with STL	Automatically selected most relevant STL-decomposed seasonal and trend components lagged to maximize correlation with total cases.	25 to 35+
Linear regression (dual)	1. A linear model trained with week of year feature and total cases target so it effectively predicts the seasonal pattern of total cases. 2. Another linear model trained with year-smoothed (effectively the trend component) 3 most relevant (correlated with target) features that are common for both cities as features and the remaining difference from the above predictor and actual total cases as target. Results from the above two models added together.	Below 22, 19.3798 (best result we have so far)

All models yielded improved results when smoothed with a rolling mean of approximately 5 weeks. It looks like although a week is a significant amount of time, it still can have some amount of noise. Although random forest and boosted trees models seem to give good results, they were not the best match for the problem. This can be observed from the fact that removing 95% correlated features increase MAE. Neural networks, specially deep networks need more data than a mere 2000. Neural networks have a lot of hyperparameters to tune so not suitable for a problem like this. The predictions using STL decomposed features did not work as expected. But the predictions with learner created to predict seasonal and trend components worked very well. Plain statistical methods seem not to work in the problem. Although it is obvious mosquito life-cycle and periods of incubation of virus must lag the effects of rain and temperature, these did not matter for the best model to work. The seasonal component was more dominant and the smoothing might have brought the lagged effects forward. The popular belief is that rain causes the spread of dengue but here we do not notice a large effect of any precipitation parameters. Rather, station temperature parameter looks like the most dominant one.

Conclusion

- Data preprocessing has a huge impact on the results of machine learning algorithms.
- Although there are a huge number of attributes, only few of them will be useful for predicting the results.
- Data conversion and normalization is required for more accurate predictions.
- New attributes can be created using the attributes that are already available,which will increase the accuracy.
- Although there are complex and sophisticated models to do regression, simple models are less likely to overfit for the unseen data.