SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
Department of Information and Library Science
Bloomington

Predicting Dengue Fever Epidemics: A Neural Network Approach

Carlos Sathler
cssathler@gmail.com

Abstract

Dengue is a disease caused by four types of related viruses transmitted by a mosquito, most commonly the Aedis aegypti. The disease is considered an alarming threat to the health of populations spanning millions of people living in tropical and subtropical areas of the globe where the mosquito thrives. A large number of quantitative studies have confirmed that the incidence of dengue is positively correlated with climatic conditions, specifically, temperature, humidity and precipitation levels. These quantitative studies invite the question: how well can we predict future cases of the disease based on climate variables that are included in weather forecasts? To answer this question several departments of the U.S. Federal Government have joined efforts to create the Dengue Forecasting project, which makes climate and dengue data available to data scientists at large and challenges them to submit predictive models to help forecast future dengue epidemics. In our project we create predictive models using neural networks and evaluate their performance against more commonly used machine learning models

Background

The Centers for Disease Control (CDS) estimates one third of the world population is exposed to the dengue virus and is at risk of contracting the disease [1]. The World Health Organization (WHO) in their publication “Global strategy for dengue prevention and control 2012-2020”10 asserts that “Dengue morbidity can be reduced by implementing improved outbreak prediction and detection through coordinated epidemiological and entomological surveillance” [2].

Models that quantitatively predict incidence of the disease based on climate data can potentially serve as one of the many tools to survey the risk of impending Dengue outbreaks. In 2015 the US Department of Commerce released the Dengue Forecasting project inviting data scientists to develop predictive models to forecast dengue using climate related data [3].

In our project we explore this data and create predictive models using a machine leaning approach and deep learning in particular. We submit our results to the “DengAI: Predicting Disease Spread” competition from DrivenData [4] to compare our model’s performance, particularly the performance of our deep learning models, against the results of other aspiring data scientists.

Methodology

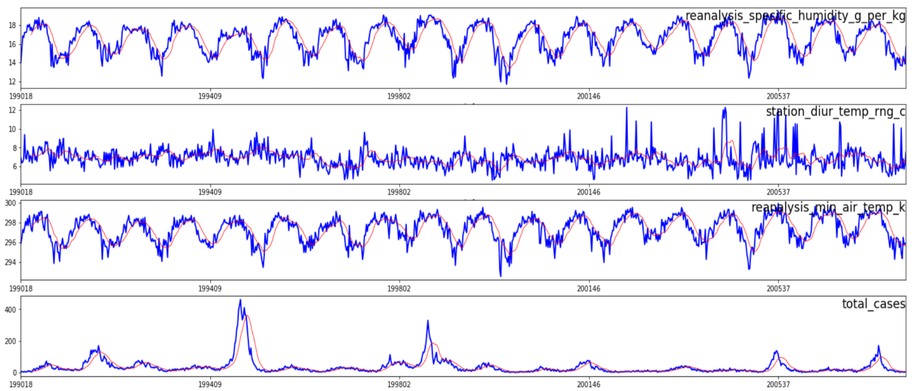
We followed a typical machine learning process:

1. Obtain data
2. Perform exploratory data analysis
3. Choose a learning model
4. Preprocess the data for machine learning model
5. Tune the model
6. Run the model on the test dataset
7. Capture predictions
8. Iterate for better results

Dataset

The dataset contained data for the cities of Iquitos and San Juan. The data consisted of temperature (10 features), precipitation (3 features), humidity (2 features) and vegatation (4 features). Below we show time series plots for 3 randomlly selected features of the dataset, along with total dengue cases. The red lines represent moving avarages and the feature name appears on the top right corner of each plot. Several data cleaning and pre-procesing steps were performed before running our models, such as imputing of null values. A couple features were found to be strongly correlated and two of them could be eleminated, but in general, no temparature, precipitation, or vegetation feature showed strong correlation with total number of dengue cases.

However, we did find strong correlation between total cases of dengue across sussessive weeks, particularly for the city of San Juan. Therefore, many of our models explored autoregression.

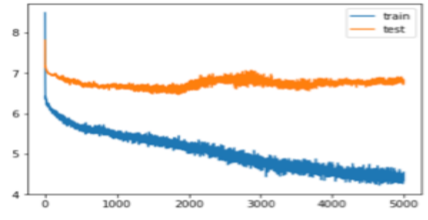


Results

We submitted a total of 45 predictions to the DengAI competition. Most of these were predictions obtained with neural network models. Our models included multilayer perceptron implementations (MPL), Long-Short Term Memory (LSTM) recurrent networks and Gated Recurrent Unit (GRU) networks. We experimented with architectures with hidden layers ranging from 1 to 5 and each time with a wide range of node numbers, epochs, learning rates, batch sizes, optimizer options, activation functions, regularization settings, and dropout factors. We experimented different time windows ranging from 1 to 16 weeks, explored models with auto regression, as we tried a few different pre-processing options, including the removal of outliers. We also ran classification models by discretizing dengue cases into 10 and 50 categories, with little impact to our results, which always scored above the 30 Mean Absolute Error (MAE) mark.

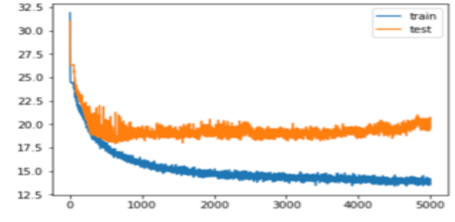
The models we implemented invariably plateaued too early, as evidenced in the sample training vs. validation loss plots below. In the end more traditional models performed better. A summary of our results is shown in the bar plot "Mean Absolute Error Score". Our best score, 22.8077 (not on the bar chart) was obtained with Bayesian Ridge regression for San Juan predictions, and weekly average for Iquitos. That placed us in 160th place in the competition, out of 2,362 participants.

```
Final loss train: 4.548939554615224
Final loss valid: 6.773314711025783
```



Iquitos

```
Final loss train: 13.777417920757067
Final loss valid: 20.05314829450477
```



San Juan

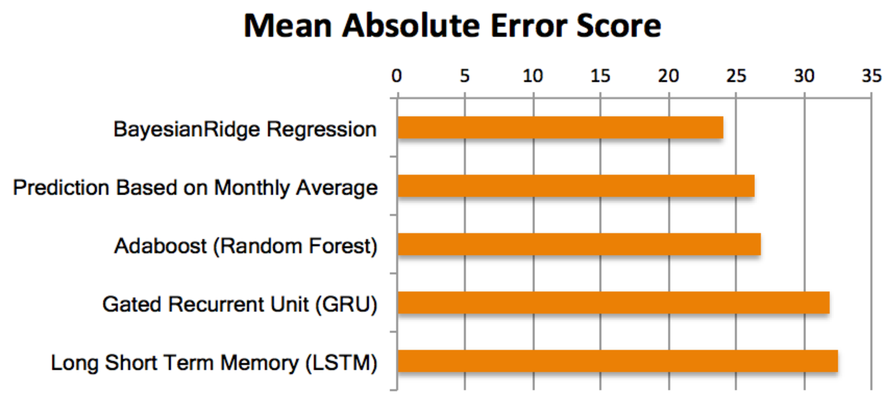
Discussion

In his article "Neural networks for time series processing” [5] Georg Dorffner shows that neural networks are theoretically capable of approximating any “reasonable” time series function but that limited data, and a variety of other problems, such as local minima and overfitting will impair neural networks’ ability to produce results comparable to those obtained by linear models [5 p.11-12]. It is possible tht our models would have performed better with more data.

Should we have additional time for this project, it would make sense to switch to an analytical approach, and experiment with more complex regression models, such as poisson. Still, it would be well worth it to evaluate more advanced neural network architectures on the project data. For example, it would be interesting to frame our problem as a sequence-to-sequence (seq2seq) problem, which would allow us to explore the effect of different time windows not only for past periods, but also for predictions beyond a single week.

Finally, a full exploration of neural networks wouldn’t be complete without experiments with deep learning and convolutional networks. Particularly Undecimated Fully Convolutional Neural Networks (UFCNN) have been reported to outperform other neural networks [6], such as the ones we explored in our project. It would be useful to create models for our problem using these architectures.

While our results were not competitive, we believe our project offers a positive contribution to a well worthy cause, mainly through judicious testing and documentation of several neural network approaches to dengue prediction. Further work would be needed to determine if more sophisticated neural network architectures (such as seq2seq and UFCNN) are capable of producing better, more competitive results for dengue prediction.



References

[1] “Dengue.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Jan. 2016, www.cdc.gov/dengue/index.html.

[2] Global strategy for dengue prevention and control World Health Organization - Geneva: World Health Organization, 2012

[3] US Department of Commerce, NOAA, National Weather Service. “Dengue Forecasting.” Dengue Forecasting, NOAA's National Weather Service, dengueforecasting.noaa.gov/.

[4] DrivenData. DengAI: Predicting Disease Spread, DrivenData, www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/.

[5] Dorffner, Georg. "Neural networks for time series processing." Neural network world. 1996.

[6] Gamboa, John Cristian Borges. "Deep Learning for Time-Series Analysis. "arXiv preprint arXiv:1701.01887 (2017).