Real-Time Disease Prediction based on Weather Data: A Case on Life Sciences Analytics

Garima Makkar
Senior Business Analyst,
Tata Consultancy Services,
Bangalore
+91-9811144199
garima.makkar@tcs.com

Abstract:

Weather is often considered as an unmanageable factor, which has direct consequences on human mortality rates, mental injury, physical health and other health outcomes. Unstable and warm climate plays a major role in driving the global emergence, regeneration and redistribution of communicable diseases like dengue. There is a worldwide pandemic about dengue, which is known to be the most dominant arthropod-borne infection in humans. According to 1995 World Health Organization (WHO) Report, around 50 million of dengue infection cases occur globally every year. One of the important factors which contribute to the spread of dengue is climate change. Events like rainfall, humidity, temperature etc. have well-defined role in the transference cycle. Any changes in these events can lead to increase in incidence of this disease.

The impact of global epidemic on mankind facilitates the need for developing early warning systems (EWS) on infectious disorders with respect to the climate change. Past studies in this context takes only the historical weather statistics into account. However because of increasing incidence and climate variability, these traditional systems are likely to get outstripped. In this paper, a new methodology is proposed to predict the number of dengue cases that are likely to occur on the basis of real time weather data. Our analysis is universally applicable, and enables comprehensive scenarios of daily dengue cases to be explored using real-time weather API, enabling dengue control measures to be effectively targeted, timed and implemented.

Keywords*: Weather, Real-time, Disease, Dengue, Early Warning System, Climate, Forecast*

1. Introduction

Climate variation and weather patterns play an important role in global emergence, regeneration and redistribution of various infectious disorders. There has been an acute outburst of arthropod-borne infectious diseases and their likely geographic extension which poses an alarming danger to human health these days. One such vector-borne disease is dengue, which is considered as the most widespread abroviral disease in humans around the world. According to WHO Report, every year there are around 50-100 million dengue incidences, and more than 500,000 cases are hospitalized. One of the important reasons behind the dengue transmission is climate change. Such vector-borne diseases (like malaria, dengue, lyme disease etc.) are quite sensitive towards meteorological variables such as precipitation, temperature and humidity. For instance, warm temperature is important to gonotrophic cycle and feeding behaviour of adult dengue vectors, as well as the rate of viral simulation and speed of larval development. Similarly, rainfall-induced standing water is considered critical for dengue vectors to breed. Thus all in all it can be said that events like rainfall, temperature, wind speed etc. have well defined role in the transference cycle and any change in these events can lead to increase in the incidence of this disease.

The impact of global epidemic on mankind facilitates the need for developing early warning system (EWS) on infectious disorders with respect to the climate change. Past studies incorporates only the historical weather statistics into account. However because of increasing uncertainty and climate variability, the traditional systems in this context are getting outstripped. Also, weather prediction which is an attempt by meteorologists to forecast the state of atmospheric parameters such as humidity, temperature, rainfall etc., is considered as one of the most complex task in today's world. Understanding how factors like weather and climate affects the disease occurrence in a specific geographic region is an eternal part of disease forecasting. Hence, models based on weather data helps to predict where and when the human cases are most likely to occur. Such information assists to target the control resources and restricted prevention and may finally decrease the burden of diseases.

Keeping this in mind, we propose a Real-time Disease Prediction framework that analyses the five-day weather forecast data being extracted using an API key and provides a number of dengue cases that are likely to occur in these upcoming five days. Given the necessary weather data, we have used a supervised machine learning algorithm called Random Forest to carry out this real-time disease prediction analysis. Our analysis as explained in this paper is divided into following sections. The Section 2 discusses some of the past works done in context of disease and weather patterns. Section 3 gives a brief about our problem statement while Section 4 explains our proposed methodology. The result and conclusion are explained in Section 5,6 and 7.

2. Literature Review

This section gives a brief about some of the research works that have been done in context of weather and disease so far. Basically, these studies can be divided into two groups: 1) Theoretical work that describes the nature of the impact and 2) Empirical work which explains the impact of weather variation on human health. The following is the description of few of these works explaining the impact and magnitude of climate variability on human population.

Many researchers have explored the responsiveness of mosquito cases and its transmission patterns with respect to changes in meteorological variables. For instance, authors like Jetten et. al., 1997 and Reiter et. al., 2001, suggested that the viral development and its transmission occur more frequently and more rapidly at warmer temperatures. Another consensus by Yang et. al., 2009, suggests that when the temperature is between 27°C & 30°C then the spread of virus is at its peak. Lambrechts et. al., 2011, presented a thermodynamic model showing how daily temperature fluctuations affect the vector pathogen interactions and why short-term variations in temperature are important when studying the disease transmission dynamics.

Also, various empirical models have been developed for estimating the weather effects on different infectious diseases. Disease prediction using time series analysis has been explored by many. For example, Enscore et. al., 2002, studied the relationship between climatic variables and human plaque incidences using Poisson regression and concluded that the variations in plaque risk can be estimated by temperature and time-lagged amount of late winter precipitation. Similar to this work, Hii et. al., 2012, applied time series Poisson multivariate regression model to predict dengue cases in Singapore over the period 2000-2010. Seasonality, autoregression, trend and various lag times were considered in their analysis to find the optimal dengue forecasting period using cumulative rainfall and weekly mean temperature as the only independent variables. Authors like Tong et. al., 2009; Abeku et.al. 2004; Guo-Jing et. al., 2006 etc., all have proposed similar time-series framework analysis for various other vector-borne diseases and in different geographic regions.

Disease prediction using Support Vector Machine (SVM) has also been performed by few of the researchers. Wu et. al., 2009, employed Support Vector Machine regression to forecast the number of disease cases based on the climatic factors. Their methodology resulted in a strong correlation between the monsoon seasonality and dengue virus transmission. Hales et. al., 1999; Gagnon et. al., 2001; Cazelles et. al.,2005 etc. are some of the other disease forecasting works reporting varying associations and lagged effects between weather and disease cases.

Thus, there has been limited set of recent studies which developed site-specific multivariate regression models using different combinations of weather variables to predict different vector-borne diseases (e.g., malaria, dengue etc.) in different regions of the world. But almost all of them are based on historical weather data and none of them forecast these incidences in real time. Our analysis presents a research on real time disease prediction based on weather statistics. Since weather data is complex, non-linear in nature so the traditional methods aren't effective and efficient to solve such problems. The proposed methodology evaluates the developed models by tuning different combinations of parameters to predict the disease (in our case it is Dengue) in real time given the five day weather forecast .The criteria used for model selection is Root mean square error. Contrary to similar research work, the data model and methodology suggested in this paper resulted in higher accuracy and better performance (i.e. reduced computational complexity).

3. Problem Statement

The objective here is to create an Early Warning system that will predict the dengue incidences in real time given the five-day weather forecast. Our analysis focuses on weather as the fundamental factor behind dengue epidemics which may allow us to reduce the timeframe of high risk dengue infection.

4. Methodology

The following section explains the methodology that is being followed to carry out the necessary experimentation for our analysis. The first part describes the dataset which is being utilised, followed by the steps performed to predict the dengue cases on the basis of weather conditions in real-time.

4.1 Data

Our main purpose is to predict the number of dengue cases that are likely to occur using current and future weather conditions of Alabat city, Philippines. For this purpose, we have used the data which is a merger of two different datasets: - 1) Dengue cases with respect to historical weather conditions and 2) five days forecasted weather data based on API key. Former dataset has been taken from kaggle which is the predictive analytics competition platform while the latter is from Openweathermap, a service providing historical, current and future weather conditions for each and every city based on the API key. Using this amalgamation of datasets, we performed the following steps to build our predictive model.

4.2 EDA

Exploratory data analysis is considered as the first and foremost step in any data analysis procedure. Basically, this approach employs a set of techniques to tackle various tasks such as detection of outliers, spotting missing data, maximising insights in a dataset, uncovering underlying relationships etc. Most of these techniques are graphical in nature with a few quantitative techniques, helping to display the data to speak for itself. And one of the best ways for data analysts to present their analysis outside the industry is through visualisations. Thus, keeping this is mind, we also performed EDA so as to see the relationship or correlation (if any) present between the variables/features in our raw dataset. For example, we plotted a bar plot to see the monthly dengue cases in Alabat for consecutive two years. And it was seen that in the months of January and February, the number of dengue cases reported were the highest. Similarly, scatter plots are being used to see the relationship/correlation between: 1) number of dengue cases versus temperature mean and 2) number of dengue cases versus humidity. From these we got to know how dengue number varies with different weather elements.

After knowing all vital information about our dataset, we are prepared to carry out the next step of data preparation stage , called data cleaning which is explained in the below section.

4.3 Data Cleaning

Most often after data has been collected, data screening, should take place to make sure that data to be analysed is as ‘precise’ as it can be. Basically, data cleansing is the process of finding and correcting inaccurate data from a database arising due to the corruption or faulty record of the data. The purpose of this step is to identify incomplete, inaccurate or unrelated data points which are then replaced, altered or deleted. Inaccurate or inconsistent data can cause a number of issues which can lead to drawing of incorrect conclusions. Therefore, data cleaning becomes a chief constituent in data analysis situations.

Keeping this in mind, we performed the following assessments for our analysis:-

i. **Missing Value Treatment**: Missing data which refers to an empty data value being stored for a variable in an observation is a common phenomenon in real world problems. The presence of missing values in a dataset is likely to affect the data insights as well as the performance of our predictive model, making it a crucial step of data exploration and data preparation stage. Being such an excruciating pain, it

becomes important to handle them effectively so as to reduce bias and to produce powerful models. The question now is how to handle any missing data point in our dataset? In general, there are various methods to deal with them such as deletion, imputation, prediction model methods etc. In our dataset also we had missing values for which we used all these methods, for example, missing values in columns like minimum temperature, maximum temperature etc. were replaced by min/max of previous three days value of these columns. Similarly, missing point in snowfall data column was replaced by sum of snowfall in last three days. Likewise, all the missing data points were being taken care of.

ii. Second assessment which we performed is **feature engineering**, which is an exercise of extracting extra information from the existing dataset. This step itself is divided into two parts: - Variable transformation & Variable creation. Former is usually done to change the scale, distribution or relationship of a variable while the latter process helps in generating new variables out of the existing variables. So for our case, we generated new variables such as last three days temperature, last three days rainfall etc. The purpose of these newly generated variables is that the dengue infected region is likely to get affected by these variables as well, along with the present day weather elements. Hence this step helped to highlight the hidden relationship of a variable with respect to our target variable.

So these above mentioned data preparation steps helped us to fetch useful information out of our raw dataset. Now using this information, we'll discuss the application of machine learning algorithms to our dataset in the next subsection.

4.4 Model Development

With the help of weather API key and pre-processed data obtained from above step, we'll solve our problem statement using a supervised machine learning algorithm called Random Forest. The following is the description of this algorithm.

Random Forest: - a supervised machine learning algorithm

Used for both regression and classification tasks, random forest is an ensemble learning technique that function by creating a multiple of decision trees during training period and outputting the class that is mean (in case of regression) or mode of the classes(for classification) of individual trees. In our experiment, we'll be using this supervised learning method for regression purpose since we are dealing with a labelled data here. To solve our problem as mentioned in section 3, we first divided our pre-processed dataset into train and test followed by tuning of parameters and finally applying random forest algorithm in R which is a programming software tool for statistical computing and graphics.

5. Results

The application of Random forest technique gave the expected number of dengue occurrences against different weather conditions using which we calculated the prediction score of our benchmark model .This score is then used to calculate the following metrics for our model:-

i. Variable Importance Plot: This plot identifies the important features that are closely associated with the target variable and contribute more for variation of the outcome variable. Pressure, rainfall etc. are the topmost factors impacting our analysis of dengue prediction.

ii. Root mean Square error: This performance metric tells how well our random forest model is able to predict the test set outcomes. Model with smallest value of RMSE is chosen as an optimal model in terms of performance.

With these two metrics, we were able to find the expected number of dengue cases given the weather conditions of Alabat region for the test dataset.

6. Live Demo model

The above section explains our methodology for predicting the dengue cases given the two years weather data for Alabat region. We'll apply the same methodology to determine the number of cases for the upcoming five days given five day weather forecast. Thus, this live model shows two things: 1) Current weather conditions for the Alabat city on a map and 2) Expected dengue cases in Alabat for next five days.

7. Conclusion

A precise Early Warning System to forecast imminent epidemics strengthens the power of preventive measures against any infectious disease like dengue. Here, we developed weather based prediction models which provide timely early warning in Alabat, preparing the city to adopt effective control measures and medical operations during outbreaks. We found out pressure, rainfall and temperature as the key meteorological factors responsible for the dengue incidences in Alabat. The same study can be leveraged to other vector-borne diseases and other geographic regions. Thus, this experimentation is universally applicable, and enables comprehensive scenarios of daily dengue cases to be explored using real-time weather API, enabling dengue control measures to be effectively targeted, timed and implemented.

References:-

1) Jetten, Theo H., and Dana A. Focks. "Potential changes in the distribution of dengue transmission under climate warming." *The American journal of tropical medicine and hygiene* 57.3 (1997): 285-297.
2) Reiter, Paul. "Climate change and mosquito-borne disease." *Environmental health perspectives* 109.Suppl 1 (2001): 141.
3) Yang, H. M., et al. "Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue." *Epidemiology & Infection* 137.8 (2009): 1188-1202.
4) Lambrechts, Louis, et al. "Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti." *Proceedings of the National Academy of Sciences* 108.18 (2011): 7460-7465.
5) Enscore, Russell E., et al. "Modeling relationships between climate and the frequency of human plague cases in the southwestern United States, 1960-1997." *The American Journal of Tropical Medicine and Hygiene* 66.2 (2002): 186-196.
6) Hii, Yien Ling, et al. "Forecast of dengue incidence using temperature and rainfall." *PLoS neglected tropical diseases*6.11 (2012): e1908.
7) Tong, Shilu, et al. "Climate variability and Ross River virus transmission." *Journal of Epidemiology & Community Health*56.8 (2002): 617-621.
8) Abeku, T. A., et al. "Effects of meteorological factors on epidemic malaria in Ethiopia: a statistical modelling approach based on theoretical reasoning." *Parasitology* 128.6 (2004): 585-593.
9) Yang, Guo-Jing, et al. "A growing degree-days based time-series analysis for prediction of Schistosoma japonicum transmission in Jiangsu province, China." *The American journal of tropical medicine and hygiene* 75.3 (2006): 549-555.
10) Wu, Yan, et al. "Mining weather information in dengue outbreak: predicting future cases based on wavelet, SVM and GA." *Advances in Electrical Engineering and Computational Science*. Springer, Dordrecht, 2009. 483-494.
11) Hales, Simon, et al. "El Niño and the dynamics of vectorborne disease transmission." *Environmental Health Perspectives*107.2 (1999): 99.
12) Gagnon, Alexandre S., Andrew BG Bush, and Karen E. Smoyer-Tomic. "Dengue epidemics and the El Niño southern oscillation." *Climate Research* 19.1 (2001): 35-43.
13) Cazelles, Bernard, et al. "Nonstationary influence of El Nino on the synchronous dengue epidemics in Thailand." *PLoS medicine* 2.4 (2005): e106.