Design & Innovation Project (DIP)

Project Report

CNN for Graph Structured Data

Project Group: E207

School of Electrical and Electronic Engineering
Academic Year 2018/19
Semester 2

Table of Contents

1. Purpose / Project Objectives

Dengue is a mosquito-borne disease and the symptoms are severe and is sometimes life-threatening. There is no effective treatment of dengue yet. “Currently, the mainstay of dengue prevention strategy relies on surveillance systems to allow early case detection, larval and breeding site source reduction” (Yung, Chan, Thein, Chai, & Leo, 2016). The objective of our Design and Innovation Project is to implement machine learning to predict dengue outbreak clusters based on environmental factors in previously unlabelled location clusters in mainland Singapore. The National Environment Agency (NEA) monitors and records the number of dengue cases in various localities on weekly basis. This information is meant to inform residents and alert them to take precautions against mosquito breeding. The aim of our project is to construct a neural network and to train it to output predictions in the form of probabilities. To do so, this project makes use of Graph Convolutional Networks (GCN) which is a type of neural networks. The data which we have deemed vital for predicting probability of the number of dengue cases are average monthly rainfall, humidity levels and temperature. This project could potentially be applied to predicting and uncovering previously unidentified or incompletely identified location clusters at the risk of dengue, with the help of these learned features, starting off from a relatively smaller number of labelled nodes at the start of training. Predicting the dengue outbreak clusters in unlabelled locations will potentially be helpful to health and environment agencies in monitoring occurrences of dengue across Singapore.

2. Project Summary

The problem we have identified was that the dengue virus is still a threat to residents of Singapore despite the efforts to ensure a clean hygienic public environment, and there are still outbreaks of the dengue virus. In 2018 alone, there were 3285 cases of dengue reported, a marked increase of 20% from the previous year. Dengue is transmitted by the *Aedes*, primarily the *Ae. aegypti* and *Ae. albopictus*. After a period of low dengue, in 1986, there was a resurgence of dengue. This has been attributed to lower herd immunity, virus transmission outside homes and an increase in the age of infection. Resources are continuously allocated to fight its spread and treating dengue as vector control costs approximately $50 million per year in Singapore. Reducing and preventing the occurrence of dengue is a priority of the Ministry of Health (MOH) and the NEA. As a result, dengue is a notifiable disease in Singapore, wherein hospitals and clinics must report cases of dengue at each occurrence. There are public campaigns i.e. *Mozzie Wipe-out* aimed at raising awareness for residents on preventing mosquitoes from breeding in order to minimise occurrences of dengue. Having a forecast that is classified in location clusters will be helpful in guiding government agencies in allocating resources and interests in preventive and treatment measures. While these campaigns have contributed in heightening public awareness, dengue outbreaks have not ceased, and the number of dengue cases is still significant. Our project utilises past data that primarily factor in machine learning techniques that will output predictions regarding the number dengue cases and locations of dengue clusters. This data includes past occurrences of dengue and will be the basis upon which equations modelled by the GCN to make predictions.

As of present, studies have already been conducted on methods that provide forecasts of potential dengue clusters across Singapore, however these studies are based on statistical analysis rather whereas our project utilised machine learning techniques. The goal of the project was to perform an overall prediction using a large sample of data while not necessarily determining uncertainty of the forecast, therefore machine learning is more suited.

When it comes to machine learning to train a network, Convolutional Neural Networks (CNN) is more known. However, for this project, we focused on GCN to train a model that can make the desired predictions because data in non-Euclidean structures cannot be processed by a CNN. Furthermore, GCNs are very powerful. Graphs are used to solve many real-life problems by representing networks. The location clusters across Singapore can be represented as a network therefore graphs are suitable data structures. We selected non-linear temporal and spatial factors as features to contribute to the forecast of dengue cases in a location cluster. Following the paradigm of machine learning, we aimed to train a suitable model, test the model, and evaluate the results through machine learning. We also determined the specific features and training methods that improved the performance of the model.

We started off by classifying the dengue cases into location clusters as well as the temporal and spatial factors attributed to those clusters that we have identified as significant in breeding *Aedes* mosquitoes. The temporal and spatial factors that we have focused on are rainfall, humidity levels and temperature. Since our project involves predicting forecasts for dengue cases, we utilised recorded dengue cases from 2017 to 2018. The data of the abovementioned factors was featurized with the help of node embeddings, which is the method used to encode predictive information. For the purpose of this project, we classified the location clusters for each of the data factors based on the location of each station recording rainfall data. By determining the nearest stations for each location of the other factors, the data for each of the remaining factors (i.e. number of cases, humidity level and temperature) was implemented as features. For the purpose of testing, we focused on factors that have significant correlations with dengue transmission.

For this project, we used a Graphics processing unit (GPU) and libraries such as TensorFlow, NumPy and Pandas. In the following sections, we have described our methods of data collections, the workings of GCN, the software libraries and model implemented and the results after testing. This report also discusses the selection of each feature and finally our analysis and evaluation of the results and findings.