# Outlier Detection in Dengue Fever Rates

Research carried out by Liam Reid as part of Final Year Project and Dissertation: ‘Proposing new methods of detecting outliers in cloud resource data’.

## Abstract

The aim of this experiment is to test the effectiveness of a newly implemented technique of detecting outliers. This technique involves detecting outliers using an ensemble of ‘weak’ classifiers that work together and vote on whether a datapoint is an outlier or an inlier. This technique is tested on Dengue Fever rates in regions of Vietnam. Observations of generated graphs show that this technique is effective in detecting outliers. Obvious outliers and some subtle outliers can be detected using this method but on rare occasions an outlier is missed and there are many false alarms.

## Introduction

**Background Information**

This document contains details of newly implemented outlier detection techniques on Dengue fever rates in the Bac Lieu and An Giang regions of Vietnam. The data consists of details about the climate over a 20-year period. Data such as average temperature, rainfall, humidity and the fever rate itself will be considered in this experiment. The data is unlabelled. The techniques used will be unsupervised and will not require any training.

**How was the experiment carried out?**

This experiment was carried using a webapp developed using dash. The purpose of this application is to apply outlier detection algorithms to datasets defined by the user. The application will plot the data and generate scores for the chosen method. A score will not be generated for the detection methods on these datasets since they are unlabelled.

The outlier detection methods in the ensemble have been implemented using python. These methods work individually first to make a prediction with a confidence score. A voting system, also implemented using python, determines the final classification.

**Hypothesis**

The implemented outlier detection method an effective outlier detector for unlabelled datasets.

## Methods

The combined predictions of an ensemble of weak classifiers are used as the method to detect outliers in this experiment. These detectors generate a prediction for a piece of data (outlier or inlier), the combined predictions are combined to produce a final classification.

These techniques do not require training. This is beneficial for analysing this data since it is unlabelled. Evaluating the performance of these techniques will be difficult since there is nothing to compare the classifications against.

A quick observation shows that the time of year has an effect on the data. I.e., average temperature during summer months is higher than during winter months. To account for this, these datasets are split to months for the detection is performed. This means there will be no false positive predictions made when the timeseries moves from summer to winter, but the idea of concept drift is still considered year on year (due to factors like climate change).

There are 4 classification techniques in the Ensemble.

* **Moving Average**

This technique uses the average of the previous datapoints in the timeseries to classify the next. After the average is calculated, the standard deviation of the previous datapoints is calculated . The standard deviation is used as a threshold, if they next datapoint is less than or greater than the average calculated +/- the threshold then the datapoint is classified as an outlier.

* **Moving Median**

This technique follows the same steps as the previous except a median is calculated instead of the average.

* **Moving Boxplot**

This technique takes a number of the previous data points and generates a boxplot. The interquartile range is combined with the upper and lower quartiles to produce a threshold. If the next datapoint is outside the threshold then the datapoint is classified as an outlier.

* **Histogram**

This technique plots histograms of subsets of the data. If a range in the histogram has a height less than a defined threshold then the range is said to contain outliers. If a range has a height below the threshold, but the ranges beside it have a height higher than the threshold then it is considered a borderline inlier.

**Voting on a Final Classification**

These techniques run individually first, generating a prediction and a ‘confidence’ score. The confidence scores are combined to generate a final prediction. The formula behind this voting mechanism is described below.

Confidence is calculated by the distance between a datapoint and the threshold.

For predictions, -1(outlier) and 1(inlier), the above equation computes a minimum prediction of -n and a maximum of n. By visualising possible outputs on a spectrum, it can be said that an outlier score < 0 is likely to be an actual outlier.

N (max inlier)

-N (min outlier)

0

## Results

Here are the results of the outlier detection algorithm applied to the data. Red dots indicate outliers detected.

**An Giang**

|  |  |
| --- | --- |
| Graphical user interface, chart  Description automatically generated | Graphical user interface  Description automatically generated |
| I. Detection result for An Giang Dengue Fever Rate | II. Detection result for An Giang Average Humidity |
|  |  |
| Graphical user interface  Description automatically generated | Chart  Description automatically generated |
| III. Detection result for An Giang Total Rainfall | IV. Detection result for An Giang No. Raining Day |
|  |  |
| A picture containing different, various, colorful, several  Description automatically generated | Graphical user interface, application  Description automatically generated |
| V. Detection result for An Giang Average Temperature | VI. Detection result for An Giang Total Evaporation |

**Evaluation of An Giang Data**

The graphs above show that outliers have been detected in peaks and troughs throughout the timeseries, indicating that (possible) actual outliers have been detected. Most notably, a cluster of detections were made in graph I in the spike between 2000 and 2005. This correlates with the outlier detected in graph III indicating rainfall caused a spike in the fever rate. More notably, outliers were indicated in graph IV between 2010 and 2015 where there seems to be no raining days. In graph III, a spike in rainfall is detected, which correlates with the spike in evaporation detected in graph VI around the same time.

A great number of detections have been made in graph V compared with the rest, the detection technique may be ineffective against such unstable data. But a cluster of detections are shown in the 2005 to 2010 period. There is a trough in the timeseries which correlates with the detection made in graph II where an outlier is detected in average humidity. The outlier detected here does not look irregular when plotted since it is not major peak or a trough but could be expected to be an actual outlier because of the irregular temperature.

**Bac Lieu**

|  |  |
| --- | --- |
|  |  |
| I. Detection result for Bac Lieu Dengue Fever Rate | II. Detection result for Bac Lieu Average Humidity |
|  |  |
|  |  |
| III. Detection result for Bac Lieu Total Rainfall | IV. Detection result for Bac Lieu No. Raining Day |
|  |  |
|  |  |
| V. Detection result for Bac Lieu Average Temperature | VI. Detection result for Bac Lieu Total Evaporation |

**Evaluation of Bac Lieu Data**

The graphs above show that more outliers were detected here compared with the An Giang data, especially in graph V. Besides this, the outliers detected are in the peaks and troughs of the data