If outliers exist contaminated else not contaminated

Jesse Wood[1] Bach Hoai Nguyen[1] Bing Xue[1] Mengjie Zhang[1]
Daniel Killeen[2]

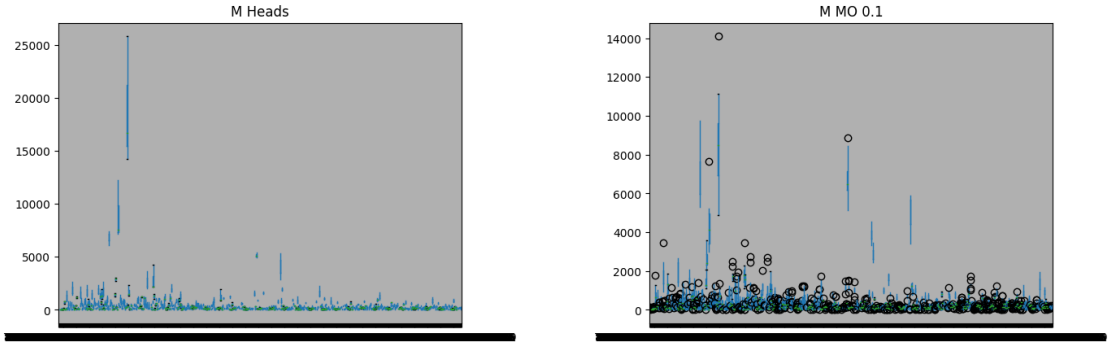
Victoria University of Wellington, Te Herenga Waka, Wellington, New Zealand
{jesse.wood, hoai.bach.nguyen, bing.xue, mengjie.zhang}@ecs.vuw.ac.nz
New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand
daniel.killeen@plantandfood.co.nz

Abstract

Occam's razor prescribes the simplest solution that works is the best. This relates to the machine learning phenomena of overfitting. Where complex models overfit and learn noise in the training set, then fail to generalize to the test set. However, in this peculiar case, Occam's razor eliminates machine learning altogether. Leading to a zero-shot algorithm for contamination detection in marine biomass, that doesn't require any training whatsoever. Outlier thresholding via Grubbs' test, combined with a conditional if statement, gives a robust solution for detecting mineral oil contamination in marine biomass, in less than 25 lines of Python code. This work presents a contamination detection algorithm for detecting mineral oil contaminants in marine biomass with rapid evaporative ionisation mass spectrometry.

1 Introduction

This method came from exploratory data analysis when looking at box-and-whisker plots for fish contaminated with mineral oil. Notice that only MO-contaminated samples contained fliers, outside LQR - 1.5(IQR) or UQR + 1.5(IQR). Visually once could easily classify the entire dataset with 100% accuracy, just looking for circles, which represent fliers, on the box-and-whisker plots.



(a) not contaminated (b) contaminated

Figure 1: Not contaminated (left) and contaminated (right) with mineral oil

2 Theory

Proposed in 1950, the Grubbs' test [1] detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or fewer since it frequently tags most of the points as outliers.

Grubbs' test is defined for the following hypotheses:

1. H_0: There are no outliers in the data set
2. H_a: There is exactly one outlier in the data set

Grubbs' test can be written as

$$G = \frac{\max_{i=1,\ldots,N} |Y_i - \bar{Y}|}{\sigma}$$

where $\bar{Y}$ and σ denote the sample mean and standard deviation, respectively.
The Grubbs' test makes the following two assumptions:

1. Data is **univariate**
2. Data is drawn from a **normal distribution**

Univariate - the intensity is measured at discrete mass-to-charge ratios. This can be viewed as a series of observations of only one single attribute, the intensity on the y-axis, and the mass spectrograph measures this attribute, at different mass-to-charge ratios. Therefore, rapid evaporative mass spectrometry (REIMS) can be considered a univariate dataset. Thus, the first assumption holds true.

Normal distribution - central limit theory implies data drawn from any distribution with 30 or more samples, can be approximated as a probability distribution that is normally distributed [2]. With 1024 features on the x-axis, all independently drawn samples from the population, it is safe to say the 30-sample limit, has been exceeded. It is safe to treat the rapid evaporative ionisation mass spectrometry (REIMS) dataset as a normal distribution. Thus, the second assumption is approximately true.

3 Dataset

Real-world application of performing mineral oil contamination detection on marine biomass with rapid evaporative ionisation mass spectrometry (REIMS) dataset. The dataset contains 306 instances, each with 1024 features, which are mass spectrographs, whose x-axis measures mass-to-charge ratio (m/z), and whose y-axis measures intensity.

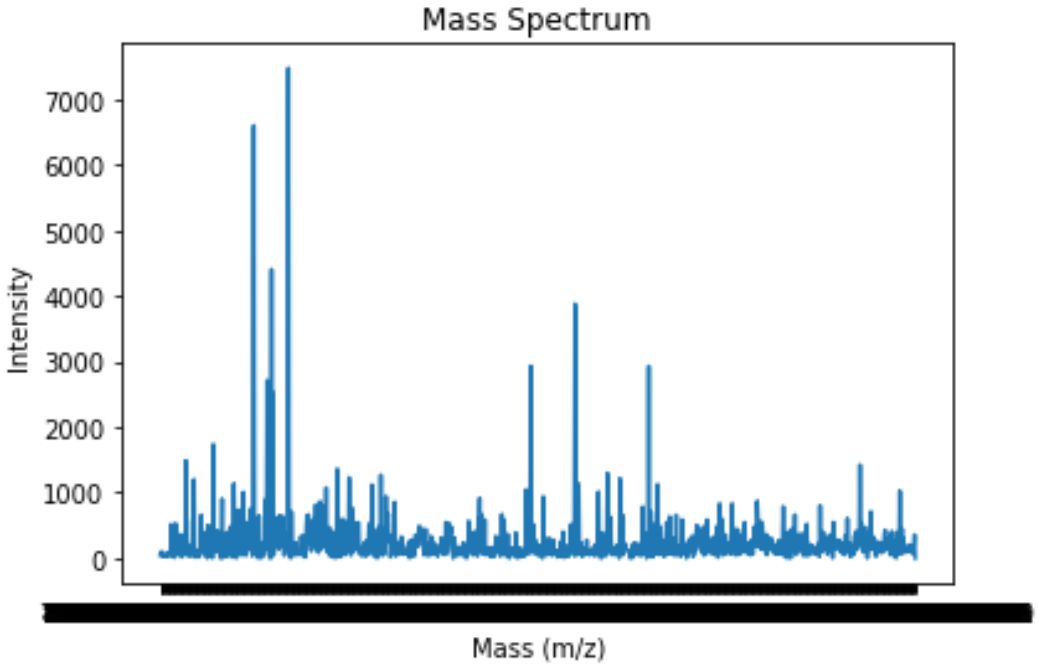


Figure 2: Rapid Evaporative Ionisation Mass Spectrometry (REIMS)

4 Method

With less than 25 lines of code (excluding comments) here is an implementation for zero-shot inference training-free Grubbs-test-based algorithm for mineral oil (MO) contamination detection algorithm (CDA) in marine biomass (MB) with rapid evaporation ionisation mass spectrometry (REIMS), called *Slick Oil* - a contamination detection algorithm to combats *Oil Slicks*.

Python program: *Slick Oil* - a contamination detection algorithm to combat *oil slicks*

```
    def fit(data, threshold = 1E-66):
        outlier_idxs = []
        for i in range(fish_data.shape[1]):
            # On your left.
            if i == 0:
                continue
            feature = fish_data.iloc[:,[i]]
            feature = np.array(feature).flatten()
            outlier_idxs.append(
                grubbs.max_test_indices(
                    feature, alpha=threshold))
        # I am speed!
        outlier_idxs = np.array(outlier_idxs)
        return outlier_idxs

    def predict(index, outlier_idxs):
        outliers = 0
        r = outlier_idxs.shape[0]
        for outlier_idx in range(0, r):
            if index in outlier_idxs[outlier_idx]:
                outliers += 1
        outlier_exists = outliers > 0
        return 1 if outlier_exists else 0
```

5 Results

100% accuracy on test dataset with zero-shot training-free inference, Grubb's test with a threshold smaller than or including $\alpha = 1 \times 10^{-66}$, which is a very very small number, with a simple conditional if statement to check if outliers exist.

6 Conclusion

> Make everything as simple as possible, but not simpler.
>
> – Albert Einstein

References

[1] F. E. Grubbs, *Sample criteria for testing outlying observations.* University of Michigan, 1949.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.