Marcus Lehr
Veer Kumar

Project Introduction

Overview

For our project, we have chosen to investigate a homegrown outlier detection technique. It was created in the context of quantitative data being captured in clinical studies to ascertain data points that warrant manual review. However, the applicability to other types of data, accuracy, efficiency, and statistical assumptions have not been investigated.

Detecting outliers holds substantial significance in the field of medicine, as these data points carry significant and often crucial information (Chandola et al., 2009). In healthcare, techniques for identifying outliers are employed to spot unusual patterns in patient records, which can potentially contain valuable information such as emerging disease symptoms. To aid in these efforts, we hope to contribute a new tool to the arsenal of outlier techniques by more thoroughly investigating the utility of our method and comparing it with other established methods. Additionally, we may propose adjustments to the algorithm to achieve a more robust mechanism to detect anomalies in patient data.

Problem description

When conducting human studies the data collected are inherently precious. This is because human studies are more expensive compared to studies with model organisms and there are regulatory pressures to keep sample sizes as small as possible for humane reasons. Because of these reasons, having to remove subjects from the analysis due to poor data quality is a major concern. In order to address this, data quality checks have been implemented with the goal of flagging outliers for manual review. These checks are done in real time during the study to improve the odds that invalid data points can be recovered or re-meausred. Data recovery becomes more difficult or impossible as time passes due to the time sensitivity of certain types of data, or the risk of losing contact with a participant after the study is over. Additionally, while the quality control checks should be biased against false negatives, too many false positives will burden study staff and participants. Thus, an outlier algorithm is required that is resilient to small sample sizes and different distribution characteristics. Robust performance at small sample sizes will allow monitoring to begin as soon as possible after the study starts. Furthermore, the algorithm should be highly generalizable. Each study collects different types of data and samples different populations of interest and thus the underlying distribution characteristics of any given data measure cannot be predicted in advance.

To meet these requirements, a 1-dimensional algorithm seems most suitable. Many simple outlier methods exist for this type of data, typically relying on the data distribution characteristics. Some common methods include using a Z-score or a

multiple of the IQR range as a thresholding mechanism. Initially, such methods were explored but found to be inadequate for multiple reasons. First was that choosing an appropriate threshold to maximize accuracy proved to be challenging. Furthermore, the ideal threshold is likely to be different for different types of data and determining that threshold requires some knowledge of what the data distribution will look like. Second, distribution characteristics are sensitive to outliers. Consequently, very large outliers can mask more subtle ones. This can often be resolved with multiple iterations as removing the more extreme outlier will the expose other ones, however this sort of behavior is not ideal due to the time sensitive nature of the data. Finally, outliers based simply on distribution characteristics are prone to false positives. For example, if using a Z-score threshold, the Z-scores will change along with the standard deviation of the data. The standard deviation will contract as extreme values are removed and raise the Z-scores of the remaining data. The trouble is that when using the observed data to define the distribution characteristics, there will always be data at the fringes of that distribution.

Algorithm description

To address this problem, an algorithm was developed which looks at changes in the distribution rather than using a distribution metric to create a threshold. The algorithm first generates normal quantiles for the data, which would typically be used to create a quantile-quantile (QQ) plot, and a linear regression is fit to the data and its slope is captured. Next, the most extreme data point- as determined by the data point with the largest absolute Z-score- is removed and the process is repeated. This process is repeated until all data points have been removed (without replacement). The absolute difference in slope after each removal is then calculated and a linear model is created against the number of points that have been removed. Cook's distance is then calculated to determine the influence of each data point on the model. Any data point which has a Cook's distance higher than the mean Cook's distance is flagged as an outlier. Finally, because data points become highly influential as the remaining sample size decreases, only the initial outlier flags are kept. Once a data point fails to meet the outlier criteria all remaining values, i.e. those with a lower Z-score, are ignored.

Proposed work

While this algorithm was found to produce satisfactory results in practice the statistical foundations remain ambiguous. For instance, because normally distributed quantiles are being used it's unclear if a normal distribution is required or how robust it is to non-normal data. Additionally, while the method appears to resolve the issues associated with other methods and produce reasonable results with small sample sizes, the accuracy has not been formally investigated. Furthermore, the applicability to other data types has not been investigated. To address this, we will test the algorithm using publicly available datasets with manually curated outliers (e.g.

https://github.com/GuansongPang/ADRepository-Anomaly-detection-datasets). We will use bootstrapping to determine the accuracy of outlier detection at different sample sizes. This will be performed on different types of data with different distributions. During this testing we may propose adjustments to the algorithm to improve accuracy or efficiency.

We will follow up on this by comparing the algorithm's performance to other published outlier detection algorithms. For example, an algorithm introduced by Papadimitriou et al. called the Local Correlation Integral (LOCI) (Papaimitriou et al, 2003). Similar to the proposed method LOCI is parameter-less, which is highly desirable from a generalizability standpoint. It would also be useful to test against more efficient clustering algorithms such as density peak clustering (Haizhou et al, 2016).

References:

Chandola, Varun & Banerjee, Arindam & Kumar, Vipin. (2009). Anomaly Detection: A Survey. ACM Comput. Surv.. 41. 10.1145/1541880.1541882.

Papadimitriou, Spiros & Kitagawa, Hiroyuki & Gibbons, Phillip & Faloutsos, Christos. (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral.. Proceedings - International Conference on Data Engineering. 315-326. 10.1109/ICDE.2003.1260802.

Haizhou Du, Shengjie Zhao, Daqiang Zhang and Jinsong Wu, "Novel clustering-based approach for Local Outlier Detection," 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), San Francisco, CA, 2016, pp. 802-811, doi: 10.1109/INFCOMW.2016.7562187.