Marcus Lehr
Veer Kumar

Project Checkpoint #2

Methods Justification and Data Exploration

Project Recap

We will be exploring the utility and limitations of a novel univariate outlier detection algorithm- tentatively called the QQ outlier method. The method identifies points which are both distant from the mean and have an outsized influence on the distribution of the data, as determined by QQ plots. This method was developed in the context of quantitative data being captured in clinical studies to ascertain data points that warrant manual review. It was found to work well empirically; however, the characteristics and limitations of the algorithm have not been formally explored.

Methods Justification

We propose to evaluate the novel outlier algorithm via a number of methods. Firstly, we will seek to evaluate the algorithm's accuracy. We will do this using publicly available data sets curated for anomaly detection development (described below). The reasons for selecting these data sets are two-fold- first the clinical data set used to develop the algorithm has usage restrictions and second we would like to assess the utility of the algorithm when exposed to different types of data so we can better understand the generalizability of the algorithm. Accuracy will be assessed using methods commonly applied to binary classification models, which largely consists of confusion matrix derived metrics (Agrawal, 2023; Vujović, 2021). A confusion matrix describes the relationship between actual classes and predicted classes (Table 1).

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 1. Confusion matrix. TP = True positives, TN = True negatives, FP = False positives, FN = False negatives

To further interpret the confusion matrix counts, we will look at several common metrics derived from these numbers (Table 2). Of particular importance to us will be overall accuracy and sensitivity, because the algorithm should be as accurate as possible while prioritizing the minimization of false negatives. Some number of false positives will lead to additional burden, but missing values that could be erroneous are of higher consequence in this context. The confusion matrix and relevant derivative metrics can be calculated in R using the associated functions in the caret package.

Metric	**Formula**
Accuracy	(TP + TN) / (TP + FP + TN + FN)
Sensitivity (aka True Positive Rate (TPR) or Recall)	TP / (TP + FN)
Specificity	TN / (TN + FP)
Positive Predictive Value (PPV)	TP / (TP + FP)
Negative Predictive Value (NPV)	TN / (TN + FN)

Table 2. Metrics derived from confusion matrix.

Another common classifier metric is the ROC AUC curve. Specifically, in our case it would be useful to look at precision-recall curves, as these are known to be more informative than standard ROC AUC curves for binary classification problems with class imbalances which are inherent to anomaly detection (Vujović, 2021). However, these curves are used to evaluate models which output a class probability. Our QQ outlier method has a binary output and thus such metrics cannot be made without making modifications to the algorithm. While modifying the algorithm is certainly within the scope of work proposed, would require adding a parameter. As is, our algorithm is parameters-less and this is desirable for the sake of generalizability and consistency. As a substitute for ROC curves will use the Kappa statistic, which measures the improvement of a model relative to random classifier (Vujović, 2021).

After performing the above analysis on the entire data sets, we will investigate the robustness of the algorithm to small sample sizes by repeating the analysis on subsamples of the data. This will be done via bootstrapping at fixed values of N using the boot package in R. The expected result will be a distribution for the confusion matrix metrics described above at each value of N. This can be used to generate maximum likelihood estimates (MLEs) and confidence intervals (CIs). The goal will be to use these results to assess the minimum number of samples needed for reasonable performance. A note about this method is the number of bootstrap iterations will need to be fairly high as outliers are rare and we will be taking relatively small samples. If the number of iterations becomes prohibitive, we can oversample the outliers as an alternative. This should drastically reduce the number of iterations required however it should be noted that in his context we often expect the number of outliers to be zero so the algorithm should be quite robust to that environment. However, while class imbalance is a common challenge in machine learning (Abraham, 2013) there are no learned parameters in the QQ outlier method. Therefore, the main concern with this issue is simply the computational time of our bootstrapped metrics.

Finally, we will compare these performance metrics to other comparable outlier detection methods. Additionally, we will compute the theoretical complexity and observed run times of the algorithms. Theoretical complexity can be determined via code analysis (Mejia, 2020) and confirmed by timing at different sample sizes. Observational run times can be readily computed using the built-in system.time() function. In terms of other methods, we will compare to a couple of basic statistical methods (Seo, 2006; Maas, 2019), a distance-based method (S. Papadimitriou, 2003), and a clustering-based method (Haizhou Du, 2016).

Data Description

The datasets employed in this project were sourced from the Compendium of Microarray Anomaly Detection Data Sets, as made available through the Characterizing Systematic Anomalies in expression Data (CSAX) (Noto, 2014). Each dataset within the compendium is characterized by three primary components: an expression matrix denominated as "matrix," a labels file designating the labelled outliers termed "metadata," and a summary documentation that entails all the references and notes labeled as "README." Since we are dealing with multivariate datasets, we would look to project these down to reduce their dimensionality using methods like Principal Component Analysis (PCA) (Ali, 2017) or Uniform Manifold Approximation and Projection (UMAP) (McInnes, 2018). Further, to test whether our data follows a specific probability distribution we will perform goodness of fit tests to see how well the statistical model fits the observed data using the vcd package in R.

1) Smokers:

This dataset focuses on understanding the molecular changes that occur in the bronchial epithelium of smokers, both those diagnosed with lung cancer and those without the diagnosis. With a total of 192 samples, the primary task is to distinguish samples with a positive lung cancer diagnosis (Anomaly) from those with a negative diagnosis (Normal). In this context of distinguishing between normal and positive lung cancer diagnoses. Since the number of cases with negative diagnosis greatly outnumbers the positive diagnosis. Traditional methods may exhibit bias toward the majority class, leading to skewed results. This is why it is essential to address class imbalance while building our novel outlier detection algorithm. As mentioned previously in methods, oversampling the outliers could be seen as a potential alternative to address this imbalance.

2) ATRT:

The ATRT dataset is centered around the classification of brain tumors. The tumor samples were from brain tumor resections at Children's Hospital Boston. Specifically distinguishing between two main types: rhabdoid tumors and non-rhabdoid tumors. Among the brain tumor samples, there is a specific focus on identifying a rare and highly aggressive type known as Atypical Teratoid Rhabdoid Tumor (ATRT). The dataset contains a total of 48 samples derived from brain tumor tissues. It categorizes samples into "Normal," representing non-rhabdoid tumors, and "Anomaly," representing atypical teratoid rhabdoid tumors (ATRT). AT/RT tumors are exceptionally rare and predominantly affect the central nervous system, primarily in the brain and spinal cord.

3) Breast.er

The dataset under examination comprises 129 samples sourced from breast cancer patients, with the primary objective of distinguishing between estrogen receptors ER- and ER+ breast cancer. The current study (Wang, 2005) (Ali, 2017)presents a 76-gene signature for lymph-node-negative breast cancer, offering a notable sensitivity of 93% for detecting patients at risk of distant metastases within 5 years. This high sensitivity is a strong point, minimizing false negatives, which is essential for early intervention. However, the signature demonstrates a specificity of 48%, indicating a need for improvement in distinguishing patients unlikely to develop distant metastases. The findings highlight the gene signature's clinical potential and the room for specificity enhancement, underscoring its value in risk assessment and the scope for refining diagnostic accuracy.

4) Synthetic data

In addition to the above real data sets, we will generate some synthetic data sets from different distributions and inject outliers to determine the algorithms robustness to different data distribution types.

References

Abraham, S. M. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*, 332-340.

Agrawal, S. K. (2023, 09 29). *Metrics to Evaluate your Classification Model to take the right decisions*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Ali, M. U. (2017). Using PCA and Factor Analysis for Dimensionality Reduction of Bioinformatics Data. *International Journal of Advanced Computer Science and Applications(IJACSA)* .

Haizhou Du, S. Z. (2016). Novel clustering-based approach for Local Outlier Detection. *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 802-811). San Francisco, CA: IEEE.

Maas, Y. (2019). Outlier detection in non-Gaussian distributions. *Bachelor's Thesis*. TU Delft.

McInnes, L. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimensionality Reduction. *arXiv preprint arXiv:1802.03426*.

Mejia, A. (2020, 10 03). *How to find time complexity of an algorithm?* Retrieved from adrianmejia: https://adrianmejia.com/how-to-find-time-complexity-of-an-algorithm-code-big-o-notation/

Noto, K. (2014). CSAX: Characterizing Systematic Anomalies in eXpression Data. *RECOMB*.

S. Papadimitriou, H. K. (2003). LOCI: fast outlier detection using the local correlation integral. *Proceedings 19th International Conference on Data Engineering* (pp. 315-326). Bangalore, India: IEEE.

Seo, S. (2006, 04 26). A Review and Comparison of Methods for Detecting Outliers. *Masters Thesis*. University of Pittsburgh.

Vujović, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, Volume 12, Issue. 6.

Wang, Y. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*.