Outlier Detection: A Literature Review

Amay Avasthi

28 September, 2023

1 Abstract

Outlier detection is a critical task in various domains, including finance, cybersecurity, and healthcare. This literature review provides an overview of recent developments and techniques in outlier detection. We categorize existing methods, discuss their strengths and limitations, and highlight emerging trends. By examining this body of work, we aim to facilitate a deeper understanding of outlier detection techniques and their applicability in different domains.

2 Introduction

Outliers, data points that significantly differ from the majority of a dataset, can provide valuable insights or indicate errors in the data. Detecting outliers is a crucial step in data analysis, and it has applications in fraud detection, anomaly detection, and quality control, among others. In recent years, there has been a surge of interest in developing robust and efficient techniques for outlier detection due to the increasing availability of large-scale and high-dimensional datasets. This literature review aims to provide an overview of the various outlier detection methods and their performance across different domains.

3 Literature Review

3.1 Traditional Approaches

Early outlier detection methods, such as z-score, modified z-score, and the Tukey boxplot, focused on univariate or simple multivariate statistics. While these methods are straightforward, they may not perform well in high-dimensional datasets or those with complex relationships between variables

3.2 Statistical Methods

Statistical techniques like the Grubbs' test, the Dixon's test, and the Mahalanobis distance have been widely used for outlier detection. These methods are effective when underlying data distributions are well-understood. However, they may struggle with non-Gaussian data or when data assumptions are violated

3.3 Machine Learning-based Approaches

Machine learning methods, including isolation forests, one-class SVM, and k-nearest neighbors, have gained popularity in recent years for their ability to handle complex data structures and high dimensions. These methods are particularly useful when the data's underlying distribution is unknown or non-standard

3.4 Deep Learning Techniques

Deep learning-based outlier detection approaches, such as autoencoders and variational autoencoders, have demonstrated impressive results in capturing complex patterns in data. These techniques are advantageous when dealing with unstructured data like images and text

3.5 Ensemble Methods

Ensemble methods, which combine multiple outlier detection algorithms, have shown promise in improving overall detection accuracy and robustness. Techniques like feature bagging and model stacking can enhance outlier detection performance

4 Conclusion

Outlier detection is a crucial task in various domains, and its significance continues to grow with the proliferation of data. This literature review has provided an overview of traditional and contemporary outlier detection methods, including statistical, machine learning-based, deep learning, and ensemble techniques. Each approach has its strengths and weaknesses, making it essential to choose the most suitable method based on the specific characteristics of the data and the application domain. Future research should focus on developing hybrid techniques that leverage the advantages of multiple methods to enhance outlier detection accuracy and scalability.

References

1. Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.
2. Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21.
3. Dixon, W. J. (1950). Analysis of extreme values. The Annals of Mathematical Statistics, 21(4), 488-506.
4. Mahalanobis, P. C. (1936). On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India, 2(1), 49-55.
5. Liu, F. T., Ting, K. M., Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422).
6. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13(7), 1443-1471.