Outlier Detection

Parisa Suchdev

May 5, 2023

The Cardiotocography dataset contains measurements of fetal heart rate (FHR) and uterine contractions (UC) recorded during labor. Outlier detection is a common task in data analysis that involves identifying observations that deviate significantly from the expected behavior of the majority of the data points. In this context, outliers in the Cardiotocography dataset may represent abnormal or risky conditions during labor that require medical attention.

The approach used in this analysis involves applying Principal Component Analysis (PCA) to reduce the dimensionality of the data and then using various clustering-based outlier detection algorithms, such as one-class SVM, isolation forest, LOF, and DBSCAN, to identify potential outliers.

The PCA analysis was performed first, which explains the variability of the data in terms of two principal components. These two components capture about 32% of the total variance in the dataset. When the PCA components are plotted, it appears that outliers are centered and surrounded by data points.

One-Class Support Vector Machine (SVM): One-Class SVM is a popular method for outlier detection. It learns a decision boundary that separates the data points from the origin in a high-dimensional space, using a kernel trick. The performance of the One-Class SVM depends heavily on the choice of the kernel function and the value of its hyperparameters. In your experiment, One-Class SVM without PCA detected 113 outliers, while with PCA, it detected 110 outliers. Although the difference in the number of outliers detected is small, the F1 score of One-Class SVM with PCA is slightly better than that without PCA.

Isolation Forest: Isolation Forest is another popular method for outlier detection that works by isolating outliers in a tree structure. It randomly selects a feature and a split value and uses it to split the data into two parts. It repeats this process until the data points are isolated, and outliers are those points that require fewer splits to be isolated. In your experiment, Isolation Forest without PCA detected 160 outliers, while with PCA, it detected 436 outliers. The F1 score of Isolation Forest without PCA is slightly better than that with PCA. However, Isolation Forest with PCA can clearly distinguish outliers from the rest of the data points, making it a suitable method for detecting outliers in high-dimensional data.

Local Outlier Factor (LOF): LOF is a density-based method for outlier detection that measures the local deviation of a data point from its neighbors. It compares the density of the surrounding data points with the density of the data point, and outliers are those points with a much lower density. In your experiment, LOF without PCA detected 213 outliers, while with PCA, it also detected 213 outliers. The F1 score of LOF with PCA is slightly better than that without PCA. However, LOF is not suitable for detecting outliers in datasets with a high degree of dimensionality because the notion of "locality"

becomes ambiguous, and it becomes harder to define an appropriate radius for the nearest neighbors.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN): DBSCAN is another density-based method for outlier detection that groups together data points that are close to each other in terms of distance, while marking those that lie alone as outliers. It has two important parameters: epsilon, which defines the radius of the neighborhood, and min_samples, which defines the minimum number of points required to form a cluster. In your experiment, DBSCAN without PCA detected 431 outliers, while with PCA, it did not find any outliers. This could be because the outliers were projected onto the lower-dimensional space defined by the PCA components, and therefore, the clustering algorithm did not identify them as outliers. DBSCAN is suitable for detecting outliers in datasets with a high degree of dimensionality, but it requires careful selection of its hyperparameters, and it is sensitive to the density of the data points.

Each of the outlier detection methods you used has its strengths and weaknesses. One-Class SVM and Isolation Forest are suitable for detecting outliers in high-dimensional data, while LOF and DBSCAN may struggle with such datasets. Isolation Forest with PCA can clearly distinguish outliers from the rest of the data points, while One-Class SVM with PCA has a slightly better F1 score than that without PCA. DBSCAN requires careful selection of its hyperparameters and may struggle to detect outliers in lower-dimensional spaces defined by PCA components.