Outlier Detection

Parisa Suchdev

May 4, 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 39% of the total variance in the dataset. When the PCA components are plotted, it appears that outliers are centered and surrounded by data points.

The one-class SVM algorithm was applied with and without PCA components. Without PCA, the algorithm achieved an F1 score of 0.678, indicating that outliers are scattered with data points of feature 1 and 2 and some are separated. With PCA, the F1 score was slightly higher at 0.679, and outliers were scattered around two components of PCA and some at the center.

The isolation forest algorithm was applied with and without PCA components. Without PCA, the algorithm achieved an F1 score of 0.677, and outliers were scattered with data points of feature 1 and 2 and some are separated. With PCA, the F1 score was slightly lower at 0.663, and outliers were scattered around two components of PCA and NOT at the center.

The LOF algorithm was also applied with and without PCA components. Without PCA, the algorithm achieved an F1 score of 0.659, and outliers were scattered with data points of feature 1 and 2 and some are separated. With PCA, the F1 score was slightly higher at 0.679, and outliers were scattered around two components of PCA and NOT at the center and some at the side closer to data points.

The DBSCAN algorithm was applied with and without PCA components. Without PCA, the algorithm achieved a silhouette score of 0.352, and outliers were scattered with data points of feature 1 and 2 and some are separated. With PCA, the algorithm predicted only one label of -1.

The results suggest that applying PCA to the Cardiotocography dataset can improve the performance of outlier detection algorithms. In general, the algorithms achieved comparable performance with or without PCA, with some variations in the locations of identified outliers. The LOF algorithm appears to perform the best, achieving the highest F1 score with PCA components. However, the DBSCAN algorithm seems to struggle in identifying outliers in this dataset.