Multivariate Outlier Detection.

Difficult to handle: univariate vs. multivariate

Previous outlier detection approaches focus at one variable. A more advanced way is to consider the data set as a whole. We consider two algorithm of multivariate outlier detection: Mahalanobis distance and feature-bagging based outlier detection.

The feature-bagging based outlier detection can be considered as an ensemble method. The ensemble method compares results of several outlier detection algorithms. Every outlier detection algorithms uses a small subset of random variables to detect outliers. Every detected outlier ascribes a probability of being an outlier. The probabilities of being an outlier are compared to find outliers with the highest probability to be an outlier.

We considered applying the HighDimOut package. The HighDimOut package provides a function called Func:FBOD and Func.ABOD.

The Func.FBOD uses bagging based outlier detection method with the help of local outlier factor (LOF). The local outlier factor describes how remote a sample is. The degree of isolation depends of the local distance to neighbors. The result of Func.FBOD is a vector which contains the score of feature-bagging based outlier detection based on local outlier factor.

The Func.ABOD uses angle based outlier detection algorithm. The used angle based outlier detection algorithm calculates the angle variance of an object to the neighbors. The angle variance is used to decide if an object is an outlier or not. An outlier has a low angle variance.

We applied Func:FBOD and Func.ABOD on a subset with 100 samples. The Func.ABOD calculated still after two hours. The results of the Func:FBOD seemed to be an improvement. Nevertheless, compared to our univariate outlier detection the computational time consumption increased for both algorithms. We have to consider the computational time consumption as a factor of using this multivariate outlier detection due to our limited computational resources as. Our limited computational resources lead to the decision of using univariate instead of multivariate outlier detection for the second.