NONPARAMETRIC METHOD FOR OUTLIERS DETECTION

Jean-Michel NGUYEN

PIMESP, CHU Nantes - France

Typical microarray datasets are large. The larger the sample size and the standard deviation are, the higher is the probability to have an extreme value that must not be considered as outlier.
All methods currently used for detecting outliers are based on a ratio of the N/D form where N is a measurement of the distance between the outlier and the other values and D is a measurement of the spread of the sample. This distance is calculated using the mean or the nearest value and standardised by the SD. These methods re-quire a distribution hypothesis.
Authors consider as erroneous the application of nonparametric methods within the context of outlier detection because the definition of outliers requires having some hypothesis of the data generation process. Consequently, it is somewhat surprising to investigate outliers, without any assumption, except if such approaches are based on robust and nonparametric statistical information. A nonparametric approach makes no assumptions about the basic data-generating mechanism.
In this paper we propose a new descriptive statistical parameter called “Upper Sample Limits” (USL) which generalizes the concept of maximum natural sample numbers. USL are a new type of information relating to position parameters for ordinal values. USL is defined by the minimization of two specified functions and defined as a « virtual range » of a sample. We describe the properties of USL and develop a test of discordancy based on the USL for an univariate outlier detection. The results show that the threshold of detection of outliers increases with the sample size and the standard deviation of the sample. We compare the results of the novel method with MAD and Grubbs methods, using truncated normal distributions. No distribution hypothesis is required, thus making this method nonparametric.
Comparing our results with the MAD test, we noticed that the threshold of rejection decreased with the SD of the sample and increased with the sample size. These results are not surprising. The larger the SD , the higher the probability of obtaining a large value that must not be considered as an outlier. The larger the sample size , the more significant and robust the information contained in the USL . These properties enables our method to be applied to a large collection of data sets such as those obtained from microarrays.