Identifying extreme observations: a robust distance-based function

C. Arenas[1], I. Irigoien[2], F. Mestres[3]

[1]carenas@ub.edu, Department of Statistics, University of Barcelona
[2]itziar.irigoien@ehu.es, Department of Computer Sciences and Artificial Intelligence, University of the Basque Country
[3]fmestres@ub.edu, Department of Genetics, University of Barcelona

Nowadays, a major challenge to diagnose cancer, psychiatric disorders or other diseases is the treatment and interpretation of actual data. For example, it is usual to observe the expression of thousands genes in a small number of samples and the detection of outliers genes is a very important issue to understand the data. Furthermore, it has recently highlighted the need to integrate both, gene expression (continuous data) with clinical/pathological data (usually categorical and ordinal data) in order to capture the information, which is lost in independent genomic or clinical studies. Thus, models useful for any type of features and for any size/dimensional data sets are desirable.

In Irigoien et al., (2013) a distance-based depth function was defined. We modified this function in order to obtain a more robust one, which is useful to find extreme observations, outliers or noise in multidimensional data sets among not necessarily continuous features. We demonstrated that for this new function, the sensitivity curve is bounded and the breakdown point is just above 25%, showing the robustness of the method. Furthermore, the proposed method presents several advantages over existing methods. It only uses the distance values between observations and prior knowledge of any distribution model on data or estimation of parameters is not required; it deals with data sets involving a mixture of numerical, ordinal, binary and categorical features; the method offers a ranking by assigning each observation an outlier classification factor reflecting its degree of outlyingness and it is robust in front of the masking effect. Application on real data obtains good results, showing the utility of the method

Keywords: Biomedical data; data depth; high dimensional data; mixed features.