Using Clustering and Robust Estimators to Detect Outliers in Multivariate Data

A.M. Pires[1] and C.M. Santos-Pereira[2]

[1] Department of Mathematics and Applied Mathematics Centre (CEMAT), IST, Technical University of Lisbon, Avenida Rovisco Pais - 1049-001, Lisboa, Portugal.
[2] CEMAT/IST and Department of Mathematics, Universidade Portucalense Infante D. Henrique, Rua Dr. António Bernardino de Almeida, 541-619, 4200-072, Porto, Portugal

Keywords: Outliers, Clustering, Discriminant analysis.

1 Introduction

Outlier identification is important in many applications of multivariate analysis. Either because there is some specific interest in finding anomalous observations or as a pre-processing task before the application of some multivariate method, in order to preserve the results from possible harmful effects of those observations. It is also of great interest in discriminant analysis if, when predicting group membership, one wants to have the possibility of labelling an observation as "does not belong to any of the available groups". The identification of outliers in multivariate data is usually based on Mahalanobis distance. The use of robust estimates of the mean and the covariance matrix is advised in order to avoid the masking effect (Rousseeuw and von Zomeren, 1990; Rocke and Woodruff, 1996; Becker and Gather, 1999). However, the performance of these rules is still highly dependent of multivariate normality of the bulk of the data. The aim of the method here described is to remove this dependency. The first version of this method appeared in Santos-Pereira and Pires (2002). In this talk we discuss some refinements and also the relation with a recently proposed similar method (Hardin and Rocke, 2004).

2 Methodology

Consider a multivariate data set with n observations in p variables. The basic ideas of the method can be described in four steps:

1. Segment the n points cloud (of perhaps complicated shape) in k smaller subclouds using a partitioning clustering method with the hope that each subcloud (cluster) looks "more normal" than the original cloud.
2. Then apply a simultaneous multivariate outlier detection rule (Davies and Gather, 1993)to each cluster by computing Mahalanobis-type distances from all the observations to all the clusters. An observation is considered an outlier if it is an outlier for every cluster. All the observations in a cluster may also be considered outliers if the relative size of that cluster is small (our proposal is less than $2p+2$, since for smaller number of observations the covariance matrix estimates are very unreliable).
3. Remove the observations detected in 2 and repeat 1 and 2 until no more observations are detected.
4. The final decision on whether all the observations belonging to a given cluster (not previously removed, that is with size greater than $2p + 1$) are outliers is based on a table of between clusters Mahalanobis-type distances.

There is no need to fix k in advance, we suggest to use an AIC based criterion to select k. This criterion can also be used to select the clustering method (step 1) as well as the location-scatter estimators (step 2).

3 Simulation study

In order to evaluate the performance of the above method and to compare it with the usual method of a single Mahalanobis distance we conducted a simulation study with several distributional situations and

- Three clustering methods: k-means, *pam* (partioning around medoids, from Kaufman and Rousseeuw, 1990) and *mclust* (model based clustering for gaussian distributions, from Banfield and Raftery, 1992).
- Three pairs of location-scatter estimators: classical $(\bar{\mathbf{x}}, \mathbf{S})$; Reweighted Minimum Covariance Determinant (Rousseeuw, 1985) with an approximate 25% breakdown point, and $\mathrm{OGK}_{(2)}(0.9)$ (Maronna and Zamar, 2002).

4 Application to discriminant analysis

The application of the proposed methodology to the "pen-based automatic character recognition data" (available from http://www.ics.uci.edu/∼mlearn /MLSummary.html) is used to illustrate this aspect.

References

J.D. Banfield and A.E. Raftery (1992). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–822.

C. Becker and U. Gather (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, 94, 947–955.

P.L. Davies and U. Gather (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88, 782–792.

J. Hardin and D. Rocke (2004). Outlier detection in multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics and Data Analysis*, 44, 625–638.

L. Kaufman and P.J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

R. Maronna and R. Zamar (2002). Robust estimates of location and dispersion for high dimensional data sets. *Technometrics*, 44, 307–317.

D.M. Rocke and D.L. Woodruff (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91, 1047–1061.

P.J. Rousseeuw (1985). Multivariate estimation with high breakdown point. In: W. Grossman, G. Pflug, I. Vincze and W. Werz, editors, *Mathematical Statistics and Applications, Volume B*, pp. 283-297. Reidel, Dordrecht.

P.J. Rousseeuw and B.C. von Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.

C.M. Santos-Pereira and A.M. Pires (2002). Detection of outliers in multivariate data: a method based on clustering and robust estimators. In: W. Härdle and B. Rönz, editors, *Proceedings in Computational Statistics: 15th Symposium Held in Berlin, Germany, 2002*, pp. 291–296, Physica-Verlag, Heidelberg, .