Bootstrap evaluation of unsupervised statistical learning and applications

Berthold Lausen
University of Essex

ABSTRACT

The talk provides an overview on estimating confidence limits of estimated clusters by unsupervised learning methods. The problem is illustrated by the task to estimate hierarchical clustering by distance data of high dimensional observations without class labels (unsupervised). Ultrametric and additive tree metric are mathematical models of phylogenetic inference or hierarchical clustering of distance data. Paul O. Degens introduced an additive measurement error model for distance data which was used by Lausen & Degens (1986, 1988a) to develop a three objects variance estimator which provides a point estimate of the variance parameter without estimating the overall phylogenetic tree as an ultrametric or additive tree. Estimating the unknown location parameter, ultrametric or dendrogram, the three-objects variance estimator is used to compute parametric bootstrap estimates of the probability to observe the estimated clusters (Lausen & Degens, 1988b). The approach is applied in the context of user segmentation based on online behavioural data (Hadjiantoni et al, 2022). We discuss and compare the parametric bootstrap approach with other recent suggestions of confidence limits for point estimates by unsupervised statistical learning.

REFERENCES

Hadjiantoni S, Yang H, Long Y, Petraitytė R, Lausen B (2022), User segmentation based on online behavioural data via ensemble predictions and clustering at IFCS2022 Classification and Data Science in the Digital Age, 17th conference of the International Federation of Classification Societies (IFCS),19-23 July 2022, Porto.

Lausen B, Degens PO (1986), Variance estimation and the reconstruction of phylogenies. In Degens PO, Hermes HJ, Opitz O (eds): Die Klassifikation und ihr Umfeld, Indeks-Verlag, Frankfurt.

Lausen B, Degens PO (1988a), Evaluation of the reconstruction of phylogenies with

DNA-DNA-hybridization data. In Bock H-H (ed.): Classification and Related Methods of Data Analysis: Proceedings of the First Conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, F.R.G., 29 June-1 July, 1987, North-Holland, Amsterdam.

Lausen B, Degens PO (1988b), Bootstrap evaluation in hierarchical cluster analysis. In Diday E (ed.): Data analysis and informatics V, Versailles.