Berthold Lausen (Department of Mathematical Sciences, University of Essex)
Kaloyan Stoyanov (Profusion Ltd.)
Henrik Nordmark (Profusion Ltd.)
Aris Perperoglou (Department of Mathematical Sciences, University of Essex)

Ensembles of selected classifiers and clusters

Topic 2 – Learning more from what we already know

Keywords: Ensembles, classification, clustering

Introduction

We review methods to use ensembles of selected classifiers to achieve classification rules with increased accuracy (Gul et al. 2016, Khan et al. 2016). Feature selection methods are often used as preprocessing method. For example after preprocessing microarray data with 500 000 probes and 22 125 features (probesets) which represent genes, we use a proposal to improve feature selection of microarray data based on a proportional overlapping score (Mahmoud et al. 2014).

We investigate ensemble methods for cluster analysis. Using ensemble concepts Stoyanov (2015) developed an R package to use hierarchical clustering as preprocessing for k-means clustering. In addition we discuss proposals to use nonparametric and parametric bootstrap resampling and distance based variance estimation (Felsenstein 1985; Lausen & Degens, 1987; Degens, Lausen & Vach 1990) to derive a statistical evaluation of clusters.

Conclusions

Degens, P.O., Lausen, B., Vach, W. (1990), Reconstruction of phylogenies by distance data: Mathematical framework and statistical analysis, Lecture notes in biomathematics 84, 9-42.

Felsenstein, J. (1985), Confidence limits on phylogenies: an approach using the bootstrap, Evolution, 783-791

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., Lausen, B. (2016), Ensemble of a subset of kNN classifiers, Advances in Data Analysis and Classification, online first

Khan, Z., Gul, A., Mahmoud, O., Miftahuddin, M., Perperoglou, A., Adler, W., Lausen, B. (2016), An ensemble of optimal trees for class membership probability estimation. In: Wilhelm, A., Kestler, H. A. (eds.), Analysis of Large and Complex Data, Springer-Verlag Berlin.

Lausen, B., Degens, P.O. (1988), Evaluation of the reconstruction of phylogenies with DNA-DNA hybridization data, in: Bock, H.H. (ed.), Proceedings First conference of the international federation of classification societies (IFCS), North Holland, Amsterdam, 367-374.

Mahmoud, O., Harrison, A.P., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M., Lausen, B. (2014), A feature selection method for classification within functional genomics experiments based on the proportional overlapping score, BMC Bioinformatics 15 (1)

Stoyanov, K. (2015), Hierarchical k-means clustering and its application in customer segmentation. Master dissertation, Department of Mathematical Sciences, University of Essex, UK.