

Assessing the stability of unsupervised learning results in small sample size problems

Möller U

Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, D-07745 Jena
Ulrich.Moeller@hki-jena.de

Introduction and research problem Statistical learning from data [1] is very common in biomedical research. Learning from unlabelled data (called unsupervised analysis) is particularly challenging, because the type of structure to be identified is (considered to be) unknown. The results of unsupervised learning algorithms depend on selected criteria and explicit or implicit model assumptions. Therefore, it is difficult to assess how strongly such results represent true data structure. Artefacts or biases of single algorithms can be suppressed by searching for consensus results of conceptually different methods (ensemble approach) [2]. Another critical issue is the sample size problem. Post-genomic data generated by high-throughput technologies (e.g., DNA microarrays or protein mass spectrometry) are increasingly used to classify diseases or disease subtypes, most of all, cancer. Here sample sizes are small (usually < 100), but the dimensionality is high (often > 1000). Dimension reduction techniques, such as principal component analysis, are applicable to alleviate mathematical problems (e.g., estimating covariance matrices for fitting mixture models). However, the problem of generalizing from a single and sparse sample to the underlying population still remains. To obtain statistical confidence but to avoid the assumption of a (possibly inappropriate) hypothesis, data resampling has been proposed as an attractive alternative in practical statistics [3]. Although several resampling techniques were introduced into gene expression data analysis, the strengths and weaknesses of these methods versus each other are largely unknown. A recent study presented a comparison of resampling methods for supervised learning [4]. Some authors speculated about possible artefacts of bootstrapping in applications to class discovery [5]. However, clear evidence of the relative contribution of different resampling schemes to unsupervised learning still has to be obtained. This would be important for method selection, because the analysis of multiple resamples is time-consuming and the particular way of resampling potentially affects the performance of all subsequent data analysis and modelling steps.

Materials and Methods The performance of resampling schemes was investigated based on stochastic models previously used for algorithm benchmarking [6]. To estimate the practical relevance of the model results, we considered microarray data sets, each representing different tumor types (e.g., leukemia subtypes or cancer tissues from breast, prostate, lung, and colon) [5]. The genes in these data sets were selected to represent the phenotype so that the clinically established phenotype classes ('model') could be used as a gold standard against which to test unsupervised learning results. For a further generalization, biological data sets of the widely used UCI repository were included in the study [7].

A method is presented to assess the performance of resampling schemes in a cluster validation context [8]. For this purpose, a crucial model parameter was chosen: the number of clusters C . To consider the fact that clustering performance is data-specific, different methods were used for clustering and for estimating C , providing 458 estimates of C for each original sample. This strategy aimed at a robust model identification based on any given sample. Resampling performance was defined by the strength of consensus among the parameter estimates for a set of resample partitions representing the same population. General appropriateness of the relevant analysis steps (original sampling, resampling, clustering, and model parameter estimation) was ensured by considering those results where the model was identified based on the original data set and the consensus result of the resamples (i.e., strongly biased performance results due to poorly configured steps were excluded).

Results The performance of unsupervised mixture model identification via data resampling depends on the resampling scheme used. A resampling scheme, called perturbation, characterized by adding noise onto the data, exhibited the best performance throughout the benchmark data as well as the methods used for clustering and cluster validation. The second best performance was measured for subsampling (randomly selecting 80% of the data). Both methods clearly outperformed the common bootstrap method (drawing N items with replacement, where N is the original sample size).

Discussion The performance ranking of resampling schemes for cluster validation, obtained in our study, may serve as a useful guidance for method selection in future studies. Our investigation also demonstrated that the amount of change made to the original sample is a relevant parameter. Note that a perturbation technique can utilize the information of the entire sample, whereas the expected loss of original items in a bootstrap sample is ≈ 0.368 . Therefore, a suitable control of the induced amount of change can improve the performance. As a conclusion from this study we present a novel method where an automatic control is implemented in a perturbation scheme.

Literatur

- [1] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data mining, inference, and prediction. New York: Springer; 2001
- [2] Strehl A, Gosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *J Machine Learning Research* 2002; 3: 583-617.
- [3] Lunneborg CE. Data analysis by resampling - concepts and applications. Pacific Grove: Duxbury Press; 2000.
- [4] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 2005; 21: 3301-7.
- [5] Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003; 52: 91-118
- [6] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* 2002; 3: RESEARCH0036
- [7] Blake C, Merz C. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Science, 1998.
- [8] Möller U, Radke D. Performance of data resampling methods for robust class discovery based on clustering. *Intelligent Data Analysis* 2006; 10(2), in press