Abstract

Considerable efforts are spent in the diagnostic research on finding biomarker panels that have a high potential to accurately identify a complex disease at an early stage.

This thesis addresses the realisability of specific requirements which a diagnostic rule should comply with in order to be accepted and useful within diagnostic workflows. Major aims in the process of rule building for diagnostic purposes are beside the high accuracy also the simplicity and interpretability of diagnostic rules. They have to provide accurate and reproducible results in order to be reliable. They have to be simple for an easy assessment in the diagnostic practice and good interpretable for a high acceptance by medical practitioners.

A simultaneous accomplishment of these quality standards is difficult due to the trade-off between accuracy and model complexity.

For instance *Logic Regression* might be a suitable method for diagnostic classification problems as it provides very simple and interpretable discriminant rules. These are defined as *and-or* combinations of binary predictors. However a performance loss is expected due to the necessity to dichotomize continuous predictors.

Advantages and disadvantages of simple and easy interpretable classification models (e.g. Logic Regression) when compared to established but more complex and powerful ones (e.g. Regularized Discriminant Analysis, Random Forests) are highlighted and discussed.

Another major challenge is to ensure the fair comparison of classification algorithms and diagnostic rules in order to select the most promising candidates. Regarding a general diagnostic task the algorithm should be selected that leads to the most stable and unbiased results. Regarding some special diagnostic question the most accurate discriminant rule should be selected. Adequate designs to evaluate and optimize classification algorithms and rules are presented.

This thesis deals also with the problem of an accurate estimation of rules and of their performance in the context of a heterogeneous target population but non-representative training data. Learning the diagnostic rule on some excerpt of the target population with different observed subclass prevalences than the true ones might be a source of severe bias regarding both the selected rule and its estimated accuracy.

Four weighting classification algorithms that account for the subclass prevalence structure of the target population during the processes of rule building and rule validation are presented. Their feasibility over various practical settings is assessed both empirically and theoretically.

All investigated methods are applied on some real data sets of rheumatoid arthritis cases and controls provided by Roche Diagnostics GmbH, Penzberg. Supplementary information is gained with simulated data.

Key-words: Diagnostic rule, biomarker data, logic regression, regularized discriminant analysis, subclass structure, weighted classification algorithms, optimization designs