Boosting into a New Terminological Era

M. Schmid¹; O. Gefeller¹; T. Hothorn²
¹Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany;
²Institut für Statistik, Ludwig-Maximilians-Universität, München, Germany

A randomly selected issue of Methods of Information in Medicine will typically cover a widespread range of different topics representing the broad scientific spectrum covered by the journal. From time to time, however, a collection of papers under a more focused thematic umbrella has been published in the journal. Starting in 2003 with a series of six papers on practical experiences of health information systems [1–6] the term “special topic” has been used to highlight such a constellation. This term “special topic” is, however, easily and frequently confused with “special issue”, a term with a negative connotation, which is often used by journals for additional issues beyond their regular publication frequency that are published under relaxed conditions of peer review and contain second class manuscripts. To avoid this confusion in the future, the core editorial team of Methods of Information in Medicine has decided to change its terminology and employ from now on the term “focus theme” instead.

Papers submitted to a focus theme will be handled as regular original articles and are thus subject to the same rigorous reviewing process. After being eventually accepted for publication, they are collected in one issue of Methods of Information in Medicine and are being introduced by an editorial putting the papers in proper scientific perspective. Often guest editors – always supported by at least one member of the core editorial team – are involved in the editorial process of preparing such a collection of manuscripts for a focus theme and play a valuable role. For future guest editors a revised version of instructions describing their rights and duties has been prepared and is downloadable at the publisher’s website [7].

In this issue a series of four papers on boosting algorithms has been compiled as the first focus theme in the new terminological era. During the past years, boosting algorithms have developed into a powerful method for statistical model building and prediction in biomedical research [8]. Originally invented as a machine learning technique for predicting binary outcome variables with the help of ensembles of “weak learners” [9], boosting algorithms are nowadays used to build very general types of statistical regression models. The link between statistical modelling and the original notion of boosting as a machine learning technique was established by Friedman et al. [10] (“statistical view of boosting”), who showed that boosting can be interpreted as a method for fitting regression models in a stagewise fashion. Later, Bühlmann and Yu [11] used boosting algorithms to fit generalized additive regression models (as introduced by Hastie and Tibshirani [12]). Most notably, boosting algorithms can be modified such that they contain an intrinsic mechanism for variable selection and model choice (“component-wise learning” [11]). This is accomplished by stopping boosting algorithms before convergence (“early stopping”) such that prediction accuracy is optimized. With the help of this feature, boosting becomes a convenient method for analyzing high-dimensional data characterized by small sample sizes and large numbers of predictors. This issue is particularly important in the modern omics era of biomedical research, where such high-dimensional data structures are fre-
quently encountered and where the selection of a small set of relevant predictor variables is a key issue [13].

In low-dimensional settings, boosting algorithms are particularly suited for model choice, i.e., for selecting among different modelling alternatives for the predictor variables of interest [14]. Examples of modelling alternatives include linear effects, smooth nonlinear effects and random effects but also (possibly higher-order) interaction effects. Due to its component-wise learning mechanism, boosting is a convenient method not only for variable selection but also for model choice.

Nowadays, boosting is regularly used in biomedical applications, and it has proved to be competitive with other state-of-the-art techniques for statistical modelling. Most importantly, heuristic techniques for variable selection (such as stepwise variable selection, which is known to be unstable and biased) can be avoided if boosting algorithms are used to fit statistical models.

Methods of Information in Medicine has regularly covered new developments in boosting methodology and related fields in the past, see, for example [15–19] and will be open for future work in this area after this issue.

The four papers of this focus theme provide a summary of recent developments in boosting methodology and its applications to biomedical problems. They do not only cover extensions of classical boosting algorithms to new classes of statistical models but also methodological refinements of existing boosting algorithms. Ma et al. [20] present a new boosting approach for fitting accelerated failure time (AFT) models to high-dimensional survival data. Specifically, the proposed method results in sparse prediction rules for breast cancer prognosis that are derived using gene expression data obtained from multiple studies. Wang [21] develops a boosting algorithm for multi-class prediction problems. The new method, which is an extension of the HingeBoost algorithm for binary classification [22], has been specifically designed for the derivation of sparse prediction rules. This is accomplished by integrating the Twin Boosting methodology [23] into the new approach. Groll and Tutz [24] develop a new boosting algorithm to fit generalized additive mixed models for correlated and clustered responses. These correlation structures are frequently encountered in multicenter studies such as the Multicenter AIDS Cohort Study considered by the authors [25]. Using information criteria, Groll and Tutz are able to carry out variable selection via controlling the complexity of the boosting fit. Mayr et al. [26] develop a fully data-driven stopping rule for boosting algorithms. The new rule can be used to carry out early stopping (and thus optimization of prediction accuracy) in a more efficient way than conventional strategies for determining the optimal stopping iteration. The gain in efficiency is demonstrated using a set of gene expression data for the prediction of stage II colon cancer.

References