Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 4041736, 3 pages
https://doi.org/10.1155/2017/4041736

Editorial

Predictive Modelling Based on Statistical Learning in Biomedicine

Olaf Gefeller,[1] Benjamin Hofner,[2] Andreas Mayr,[1,3] and Elisabeth Waldmann[1]

[1]*Department of Medical Informatics, Biometry & Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Waldstr. 6, 91054 Erlangen, Germany*
[2]*Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany*
[3]*Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany*

Correspondence should be addressed to Olaf Gefeller; olaf.gefeller@fau.de

Received 4 July 2017; Accepted 4 July 2017; Published 28 September 2017

Twenty years ago, the journal Computational and Mathematical Methods in Medicine was launched under its previous title Journal of Theoretical Medicine. During those years at the end of the last century, the understanding of machine learning technology and its potential combination with statistical modelling approaches was in its infancy. The modern term "statistical learning" for this fusion of methodology from different scientific areas could already be found in the scientific literature (see Vapnik [1, 2]), but its meaning was slightly different from today. The famous textbook by Hastie et al. [3] popularised the term in its current meaning when being published in its first edition in 2001. During recent years, considerable research has been devoted to exploring this combination of state-of-the-art statistical methodology with machine learning techniques. Such an approach provides many practical advantages, particularly regarding data situations frequently encountered in modern biomedical research characterized by large numbers of potential features or variables. In such situations, the primary aim is often to obtain sparse and explanatory models, which can be generalized effectively. Via statistical learning approaches, interpretable prediction rules leading to accurate forecasts for future or unseen observations can be deduced from potentially high-dimensional data.

This special issue is devoted to this evolving research area at the intersection between different scientific branches. It attracted a broad spectrum of methodological contributions regarding different types of algorithms and fields of biomedical application. Out of sixteen submissions that were rigorously evaluated by international experts, nine made it into this issue. The compilation of papers in this special issue consists of one review paper and eight original research articles.

In their review titled "An Update on Statistical Boosting in Biomedicine" A. Mayr et al. give an overview of recent developments in the evolving area of statistical boosting algorithms. In doing so, they update and expand earlier reviews of this specific area of statistical learning research [4–6]. For the first time, recent methodologic research on boosting functional data and on the application of boosting techniques in advanced survival modelling is reviewed. Modern biomedical applications of this type of statistical learning are also sketched to provide an overview not only of recent methodologic improvements but also of practical implementation of boosting in answering biomedical research questions.

In the paper titled "A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data" A. Bommert et al. propose a way to select models based on multiple important criteria: prediction accuracy as well as sparsity and stability of the model. For model stability, the authors investigate, analytically and in a simulation study, various stability measures and conclude that the Pearson correlation has the best properties. In another simulation study with various learning approaches such as random forests, support vector machines, lasso regression, and boosting in combination with a variety of filter methods preselecting features, they investigate Pareto

fronts and conclude that it is possible to find models with a stable selection of only a few features without losing much predictive accuracy.

The paper titled "Correcting Classifiers for Sample Selection Bias in Two-Phase Case-Control Studies" by N. Krautenbacher et al. aims at improving results of models based on stratified data. It gives a very detailed explanation of the general problem of the resulting selecting bias when aiming at capturing more information by using a higher proportion of individuals with rare outcomes and a thorough summary of existing methods. Furthermore, two novel approaches are presented which outperform the state-of-the-art methods when being used in the random forests context and perform equally well when being used for logistic regression.

The paper titled "Integration of Multiple Genomic Data Sources in a Bayesian Cox Model for Variable Selection and Prediction" by T. Treppmann et al. is the only Bayesian contribution to the special issue. The authors integrate information from different sources to improve variable selection performance and prediction ability in the context of high-dimensional survival analysis. In order to achieve their goal, they combine Lee et al.'s approach [7] with George and McCulloch's Gibbs sampler [8]. Basically, the latter approach allows variable-specific penalties for the lasso-type approach of the former. In their biomedical application, the authors use information from copy number variation data to improve a model based on gene expressions.

In their paper titled "Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies" S. Friedrichs et al. describe a framework to incorporate genetic pathways, that is, gene-interaction networks, in prediction models for the analysis of genome-wide association studies (GWAS). The approach adapts a boosting method with specific kernel-based learners. The authors show that their approach identifies important genetic factors while evading the issue of multiple testing. As the genetic interaction networks can be biologically interpreted, the approach facilitates understanding the biological processes involved in disease susceptibility. Furthermore, it enables the prediction of clinical outcomes for new patients and thus constitutes a powerful tool in the analysis of GWAS data.

The article titled "Probing for Sparse and Fast Variable Selection with Model-Based Boosting" by J. Thomas et al. proposes a fundamentally new concept to select the optimal number of iterations for statistical boosting algorithms. This so-called stopping iteration is the main tuning parameter for these kinds of algorithms and represents the classical trade-off between variance and bias. Typically it is selected based on resampling procedures focusing on the predictive risk and therefore on prediction accuracy. The authors propose focusing on the variable selection properties of the algorithm: they incorporate additional noninformative probes (shadow variables) for each candidate variable and stop the algorithm once the first of these probes was selected. This new approach is considerably faster than resampling, because the model is fitted only once without additional tuning. In large-scale simulations, the authors show that their approach leads to sparser models with less false positives than traditional methods to determine the stopping iteration.

The focus of the paper titled "Nonparametric Subgroup Identification by PRIM and CART: A Simulation and Application Study" by A. Ott and A. Hapfelmeier lies on traditional machine learning technology. Classification and Regression Trees (CART) have been introduced more than 30 years ago by Breiman et al. [9] and have made their way into the standard repertoire of methods to identify homogenous subgroups in high-dimensional data situations. The Patient Rule Induction Method (PRIM), which has been developed for the same purpose in biomedical applications based on a computational idea of Friedman and Fisher [10], has attracted some interest but is less often used in practice. Ott and Hapfelmeier compare the two strategies by means of an exhaustive simulation study. In particular, they show in which scenarios PRIM outperforms CART. The manuscript covers also an application using a clinical data set in which the two approaches produce similar results. However, the authors also demonstrate in their application that CART, although simpler to implement, is a rather static technique, whereas PRIM can be flexibly tuned by the user.

In their paper titled "IPF-LASSO: Integrative L_1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data" A.-L. Boulesteix et al. focus on the problem of integrating high-dimensional molecular, genetic, or other "omics" data from different sources or modalities together with clinical variables into one prediction model. They adapt the classical lasso (which leads often to very similar solutions to statistical boosting approaches; see Hepp et al. [11]) by introducing different L_1 penalties for the modalities in order to account for their different importance. The application of the approach is illustrated via the development of prediction models for the survival of cancer patients based on clinical variables, microarray gene expressions, and somatic copy number alterations.

In their interesting application-oriented paper titled "Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods" Y. Wu et al. compare different machine learning approaches in their discriminatory performance on voice pattern data from patients with Parkinson's disease and healthy controls. Their novel contribution consists of suggesting a new method of feature selection from voice patterns subsequently processed by machine learning algorithms. Their results show superiority of classification performance of their approach termed "interclass probability risk method" over traditional competitors.

Acknowledgments

We express our appreciation to all authors for their informative contributions and to all reviewers for their support and constructive criticism making this special issue possible. The first and the third authors' work on this editorial was supported by Deutsche Forschungsgemeinschaft (DFG) (http://www.dfg.de, Grant no. SCHM 2966/1-2). Support of the Interdisciplinary Center for Clinical Research (IZKF) of the Friedrich-Alexander University Erlangen-Nürnberg

via Project no. J49 (grant to Andreas Mayr) and Project no. J61 (grant to Elisabeth Waldmann) is also gratefully acknowledged.

Olaf Gefeller
Benjamin Hofner
Andreas Mayr
Elisabeth Waldmann

References

[1] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[2] V. N. Vapnik, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control, Wiley- Interscience, New York, NY, USA, 1998.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[4] P. Bühlmann and T. Hothorn, "Boosting algorithms: regularization, prediction and model fitting," *Statistical Science*, vol. 22, no. 4, pp. 477–505, 2007.

[5] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms: from machine learning to statistical modelling," *Methods of Information in Medicine*, vol. 53, no. 6, pp. 419–427, 2014.

[6] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "Extending statistical boosting: an overview of recent methodological developments," *Methods of Information in Medicine*, vol. 53, no. 6, pp. 428–435, 2014.

[7] K. H. Lee, S. Chakraborty, and J. Sun, "Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data," *International Journal of Biostatistics*, vol. 7, no. 1, p. 21, 2011.

[8] E. I. George and R. E. McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, Mass, USA, 1984.

[10] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 123–143, 1999.

[11] T. Hepp, M. Schmid, O. Gefeller, E. Waldmann, and A. Mayr, "Approaches to regularized regression - A comparison between gradient boosting and the lasso," *Methods of Information in Medicine*, vol. 55, no. 5, pp. 422–430, 2016.