Probability estimation and machine learning—Editorial

Martin Schumacher*

Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Freiburg, Germany

Received 11 April 2014; revised 25 April 2014; accepted 28 April 2014

Most of this issue of the *Biometrical Journal* is dedicated to the special topic of probability estimation. The topic is introduced and detailed by the two comprehensive twin papers by Kruppa et al. (2014a, 2014b) on theory and applications compiled by a group of authors who had recently worked together, focusing in particular on machine learning methods. This work is then discussed in five accompanying commentaries (Binder, 2014; Boulesteix and Schmid, 2014; Shin and Wu, 2014; Simon, 2014; Steyerberg et al., 2014) which have been invited by the editors of the journal, and a reply is finally given by Andreas Ziegler (2014), the communicating author of the twin papers.

Solicited by the editors, this Editorial would like to encourage the readers of the *Biometrical Journal* to look into probability estimation and machine learning in more detail. When going through the following papers, readers of the *Biometrical Journal* may get the impression that, finally, machine learning techniques have arrived in the journal. However, it is surely not the first time that there were contributions concerning this topic; a simple search using the term "machine learning" identifies over 50 contributions over the past ten years where this method played a role. What is new, however, is that the articles by Kruppa et al. (2014a, 2014b) give a comprehensive overview of theoretical as well as practical aspects of probability estimation using machine learning methods. In the spectrum of methods, they follow more or less what is considered valuable in recent textbooks, such as in the popular Hastie-Tibshirani-Friedman book on "Elements of Statistical Learning" (2009) and others (Berk, 2010; Malley et al., 2011). So, for example, nonparametric regression, random forests, k-nearest neighbors and support vector machines are covered. What makes the overview special is that it concentrates on probability estimation and not merely on classification as mostly done in the literature, including the textbooks mentioned above (Hastie et al., 2009; Berk, 2010; Malley et al., 2011). The importance of probability estimation in a biomedical context has been recognized very early—see for instance the five papers on "The measurement of performance in probabilistic diagnosis" published by J. Hilden, J. D. Habbema and B. Bjerregaard between 1978 and 1981 (Habbema et al., 1978; Hilden et al., 1978a, 1978b; Habbema et al., 1981a, 1981b)—but it has not attracted sufficient attention in the past. In terms of diagnostic procedures, for example, it may be much more adequate—also in terms of the attached uncertainty—to present a diagnosis of a specific disease in terms of an estimated probability rather than as a simplified yes/no answer.

From that reasoning it is evident that the assessment of predictions in terms of estimated probabilities is of outstanding importance. With machine learning methods in mind, measures of prediction error must be able to adequately handle all kinds of predictions. That means they should not rely on specific model assumptions but should treat predictions, whether they are derived from a statistical model, a machine learning algorithm, or even from an expert guess, in an equal and unbiased manner. The Brier (1950) score is such a measure that has this and additional desirable properties—see for instance

*Corresponding author: e-mail: ms@imbi.uni-freiburg.de, Phone: +49-761-203-6662

the review article by Gerds et al. (2008)—the Brier score is also the preferred choice in the two articles by Kruppa et al. (2014a, 2014b).

Besides the need for optimally tuned machine learning approaches that are exemplarily presented in the two articles, validation of the predictions in independent test data is the method of choice. With this approach one can investigate whether the estimated probabilities can be used beyond the particular study (the training data) where they have been derived from. Often, however, adequate test data can only be gathered with enormous effort. In such a situation, bootstrap or other cross-validation techniques may guide further development in that they can be used to reliably estimate the prediction error. An example is the famous 0.632+ estimator developed by Efron and Tibshirani (1997) that can also be applied to the Brier score (Gerds and Schumacher, 2007). It involves a term called the "noninformation error" that reflects some kind of worst-case scenario. Especially when probability estimation is derived via machine learning methods the noninformation error provides valuable information on the potential amount of overfitting and resulting overoptimism that can be inherited when these techniques are not properly tuned. So we find ourselves in a similar situation as with, for example, regularized regression models such as the Lasso (Tibshirani, 1996) or boosting (Binder et al., 2011). This makes a unifying view of all these approaches as flexible statistical models useful.

I would like to invite the readers of the *Biometrical Journal* to use the opportunity to get familiar with theoretical as well as practical aspects of machine learning techniques and with probability estimation per se. I hope that this Special Topic will stimulate further scientific discussion and encourage future submissions dealing with comparative investigations on the use of machine learning as well as traditional statistical methods in biomedical applications.

Martin Schumacher

References

Berk, R. A. (2010). *Statistical Learning from a Regression Perspective*. Springer, New York.

Binder, H. (2014). What subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction? *Biometrical Journal* **56**, 584–587.

Binder, H., Porzelius, C. and Schumacher, M. (2011). An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biometrical Journal* **53**, 170–189.

Boulesteix, A.-L. and Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal* **56**, 588–593.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .623+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Gerds, T. A., Cai, T. and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal* **50**, 457–479.

Gerds, T. A. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63**, 1283–1287.

Habbema, J. D. and Hilden, J. (1981a). The measurement of performance in probabilistic diagnosis. IV. Utility considerations in therapeutics and prognostics. *Methods of Information in Medicine* **20**, 80–96.

Habbema, J. D., Hilden, J. and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine* **17**, 217–226.

Habbema, J. D., Hilden, J. and Bjerregaard, B. (1981b). The measurement of performance in probabilistic diagnosis. V. General recommendations. *Methods of Information in Medicine* **20**, 97–100.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning; Data mining, Inference and Prediction* (2nd edition). Springer, New York.

Hilden, J., Habbema, J. D. and Bjerregaard, B. (1978a). The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 227–237.

Hilden, J., Habbema, J. D. and Bjerregaard, B. (1978b). The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 238–246.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal* **56**, 534–563.

Kruppa, J., Liu, Y., Diener, H.-C., Holste, T., Weimar, C., König, I. R. and Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal* **56**, 564–583.

Malley, D. J., Malley, K. G. and Pajevic, S. (2011). *Statistical Learning for Biomedical Data*. Cambridge University Press.

Shin, S. J., and Wu, Y. (2014). Variable selection in large margin classifier-based probability estimation with high-dimensional predictors. *Biometrical Journal* **56**, 594–596.

Simon, R. (2014). Class probability estimation for medical studies. *Biometrical Journal* **56**, 597–600.

Steyerberg, E. W., van der Ploeg, T. and Van Calster, B. (2014). Risk prediction with machine learning and regression methods. *Biometrical Journal* **56**, 601–606.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Ziegler, A. (2014). Rejoinder. *Biometrical Journal* **56**, 607–613.

