Discussion

Machine learning versus statistical modeling

Anne-Laure Boulesteix*,[1] and **Matthias Schmid**[2]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Germany
[2] Department of Statistics, University of Munich, Germany

Received 9 October 2013; revised 4 November 2013; accepted 6 November 2013

This is a discussion of the following papers: "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory" by Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R. König, James D. Malley, and Andreas Ziegler; and "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications" by Jochen Kruppa, Yufeng Liu, Hans-Christian Diener, Theresa Holste, Christian Weimar, Inke R. König, and Andreas Ziegler.

Keywords: Logistic regression; Prediction; Reproducibility; Transportability; Tuning parameters.

The twin papers by Kruppa et al. (2014a, 2014b) give an illustrated overview of the use of machine learning methods for probability estimation. Their contribution is extremely important, since both aspects—prediction with machine learning and probability estimation—have been relatively neglected in the biometrical literature, although they play a major role in practice. A remarkable strength of their work is that theoretical properties of several important machine learning methods are summarized in a comprehensible form. Again, these theoretical issues have often been neglected in articles on machine learning in the biomedical literature. In this comment, we discuss a number of issues that however have to be carefully taken into account when machine learning methods are considered for predictive purposes in biometrical practice: the choice of the learning algorithm within the available candidates, the problem of parameter tuning, and computational transportability and reproducibility of the obtained prediction models.

1 The two cultures: stochastic versus algorithmic, explaining versus predicting

Breiman (2001) states in his seminal paper on the two cultures of statistical modeling: "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown." He also claims that the statistical community traditionally prefers the first view. In this perspective, biometricians certainly do not substantially differ from the general statistical community, a notable exception being the community of computational scientists with a background in statistics working at the interface between biometrics and bioinformatics, for example on omics data

*Corresponding author: e-mail: boulesteix@ibe.med.uni-muenchen.de, Phone: +49-89-7095-7598, Fax: +49-89-7095-7491

analysis. Medical statisticians often react with skepticism to the use of machine learning methods for different reasons, some of them discussed by Kruppa et al. (2014a, 2014b).

In a different but related approach, Shmueli (2010) distinguishes between the explanatory modeling perspective, the descriptive modeling perspective, and the predictive modeling perspective. In her framework, "predictive modeling [is] the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations", explanatory modeling is used for testing causal theory, and descriptive modeling aims at "representing the data structure in a compact manner".

The two frameworks by Shmueli (2010) and Breiman (2001) are connected. Roughly speaking, whereas the stochastic modeling approach considered by Breiman (2001) might fit well into the descriptive modeling perspective, the algorithmic approach might be preferred from a predictive point of view. These aspects are addressed by Kruppa et al. (2014b), for instance when they state that "the fundamental question is whether a clinician will trust the findings obtained with a fancy non-interpretable machine and use this in clinical routine". One may ask whether a clinician would accept to take a purely predictive perspective.

More generally, the question of the chosen perspective should probably be asked and answered much more clearly by statisticians and biomedical scientists in practice. Our experience is that practitioners sometimes come to us to develop a prediction rule for any medical condition or outcome while the project is in fact of descriptive aim (with no intention to develop a prediction tool to be applied to real patients in clinical settings for medical decision making). For example, the interest of clinicians is often in hazard ratios provided by Cox regression or odds ratios provided by logistic regression. They are analyzed with respect to effect size or statistical significance but not with respect to their predictive value. In this case, it is in our view questionable whether machine learning methods should be generally preferred over using more interpretable modeling approaches. If the purpose is predictive modeling, however, machine learning methods have a lot to offer. For example, if a practitioner wants to derive a prediction rule to be integrated in a medical electronic device, interpretability of the prediction rule is certainly not the priority (and anyway machine learning methods also provide measures that allow clinicians for assessing the importance of covariates for prediction purposes).

As clearly outlined in the two papers, no single prediction method performs universally best in all situations, which is in line with the well-known "no free lunch theorem" as described by Wolpert (2001). It is not easy to foresee which method will perform better on a particular dataset and even not easy to explain the respective performance of the methods once their results are known. In this perspective, there is no reason to restrict to a single prediction method if the goal is to achieve good prediction accuracy.

2 "Degrees of freedom" of the data analyst

One issue to consider when applying machine learning is that replacing a single standard method—logistic regression—by a wide range of other candidates (kNN, SVM, etc.) for handling the same issue increases the degrees of freedom of the data analysts.

By "degrees of freedom" of the analyst, we mean the amount of decisions and choices the data analyst has to make during the analysis (Simmons et al., 2011). Increasing the degrees of freedom of the analyst might also increase the risk of conscious or subconscious overoptimism and "fishing for significance", that is the risk that the analyst tries (many) different approaches and at the end chooses the approach that yields the best looking results.

At this stage it is important to distinguish between two distinct approaches for analysis. In the first approach, which we term "validation approach", the best method is typically selected from the set of candidate methods based on its prediction accuracy as estimated by cross-validation (CV) or a related resampling-based procedure. Afterwards, it is fitted again on the whole training set, and the resulting prediction rule is finally applied to an independent validation dataset for evaluation. This approach

adopted by Kruppa et al. (2014b) is correct and yields an unbiased estimate of the accuracy of the selected prediction rule.

In contrast, in the second approach, no independent dataset is available for validation, and the CV error estimates of the different candidate methods are the only results at hand. This implies a methodological problem. All candidate methods yield different CV accuracy estimates. Selecting the candidate yielding the best accuracy and reporting these results only could yield a biased accuracy estimate. It would be a form of fishing for significance resulting from the degrees of freedom of the analyst. That is because an optimization process takes place when selecting the "best" candidate. This issue is extensively illustrated through applications to high-dimensional small sample data in Boulesteix and Strobl (2009) and Jelizarow et al. (2010). A weighted-average approach based on decision theory is proposed by Bernau et al. (2013) to correct for the bias resulting from the optimization process.

An issue related to the increase of the "degrees of freedom of the data analyst" is the problem of tuning parameters. For example, in random forests various parameters can be controlled by the user, such as the number of candidate predictors considered at each split (often denoted as "mtry"), the minimal size of terminal leaves, the maximal depth of the trees, and so on (Boulesteix et al., 2012). Logistic regression does not involve any visible quantitative parameters like those of random forests. However, as noted by Kruppa et al. (2014a), the degrees of freedom of the data analyst is also considerable in logistic regression as far as variable selection or handling of interactions are concerned. Driving forward the comparison with random forests, one might say that variable selection and handling of interactions are intrinsic to random forests, as they are performed "automatically" and are (in)directly controlled by tuning parameters, while in logistic regression they would be done manually. Consequently, traditional methods bear the risk of *model misspecification*, which occurs when the degrees of freedom of a model are too small because they have been restricted manually by the data analyst. The proportional hazards assumption in Cox regression, for example, is often too restrictive and might lead to biased decision making (Schmid et al., 2013). Conversely, machine learning methods avoid model misspecification by allowing for a greater flexibility (i.e. by a larger number of tuning parameters). However, this flexibility comes at a big price in terms of computational costs.

All in all, we believe that the "degrees of freedom of the data analyst" is large in both model-based approaches such as logistic regression and machine learning. The difference is that in machine learning methods this happens through well-defined and quantitative parameters. This leaves space for potential dangers, typically: trying many values successively and presenting only the finest results. But this also makes the process of prediction rule construction more transparent as soon as methods such as CV are applied to choose the values of parameters—a procedure that is only rarely applied in the context of traditional methods like logistic regression.

Note, however, that even the choice of the CV procedure may be regarded as part of the degrees of freedom of the analyst. In fact, the results of CV may be variable and may highly depend on the particular random partition used for CV (Boulesteix et al., 2013). Also, there are numerous competing variants of CV (bootstrap resampling, k-fold CV, 0.632+ bootstrap, ...). These variants might strongly differ when applied to a single dataset. For example, the estimated prediction error obtained from k-fold CV is typically upwardly biased because learning is based on datasets that are smaller than the original sample. Conversely, increasing the value of k increases the similarity of the learning samples, so that the variability of the resulting prediction error estimate increases as well. A typical example is leave-one-out CV (with $k = n$), which results in error estimates with high variance. Bootstrap resampling, on the other hand, results in learning samples that are independent (conditionally on the data at hand) but also suffers from a relatively high bias since only 63.2% of the observations are used on average in each learning sample. To correct for this bias, Efron and Tibshirani (1997) proposed the 0.632+ bootstrap (which effectively is a weighted combination of the training and the test error). Still, this method retains a notable bias in situations with strong signal-to-noise ratio and small sample size (Molinaro et al., 2005). It is thus advisable to apply (different types of) CV several times and synthesize the obtained results to achieve better stability.

3 Computational transportability, reporting, and reproducibility

As noted by Kruppa et al. (2014b), computational transportability is an issue when applying machine learning in biometrical practice. By computational transportability, we mean the possibility from a technical point of view for other researchers to apply the proposed prediction model to their dataset or, in other words, the transport and exchange of computer programs allowing prediction between the developers of statistical methods and those who want to apply the methods and need to interpret and report their results.

Note that the term "transportability" is sometimes used to denote the generalization ability of a model, that is its ability to predict well in other settings such as in another hospital or another lab. This is a completely different aspect that we do not consider here, although it is also very important. To define computational transportability more precisely, let us consider the case of a researcher who wants to apply the prediction model proposed in a medical paper to make predictions for new patients. If the prediction model proposed in the medical paper was derived by logistic regression, all the researcher needs—besides a precise description of the covariates that is necessary to all methods—is the value of the coefficients of all covariates (including the intercept) of the model developed in the original paper and application, see for example To et al. (2006). Having that, the probabilities for the new patients are estimated by applying the function $f(x) = \exp(x)/(1 + \exp(x))$ to the linear predictor. The prediction model is thus (i) easily transportable in the sense that it does not take much time and effort to compute the estimated probabilities (any simple software such as Excel can be used), (ii) easy to report in the sense that the authors just need to present a few numbers that do not take much space and are easy to understand to anybody with basic statistical training.

In contrast to simple methods like logistic regression, prediction models derived by machine learning methods can often be applied by other researchers only if they have access to the corresponding software objects (or to the training data, as noted by Kruppa et al. (2014b) for the case of kNN). For example, suppose the prediction rule of interest was derived using the random forest algorithm. To apply it to own data, other researchers need the corresponding software object produced by the original researcher by applying the random forest algorithm to his training data. Strictly speaking, it would also be possible from a theoretical point of view to report a random forest in a paper—if the paper had a lot of pages and the researcher a lot of time to copy-paste these many pages into a software program! Indeed, a random forest can be seen as yielding a partition of the space of predictors. In principle, this partition could be described in a paper in terms of the involved predictors and thresholds. Of course, this is completely impossible in practice and nobody would do that.

For most machine learning methods, there is no other possibility to transport the prediction rule than to use the software object produced by the researcher who fitted it. It implies that this software object is made available in some form (e.g. from the journal's website, from the authors' homepage, from a data exchange platform, from the authors on request by email). While it would go beyond the scope of this letter to discuss the advantages and inconveniences of each approach, we point out that it is in general not trivial to make such an object available in the long term. Even if the software object remains publicly available, changes in the software program will be a major issue that may in practice impair the application of the prediction model after only a short period of time in the absence of adequate stable software infrastructure.

A solution to this problem could be the use of "hybrid" techniques such as gradient boosting with component-wise linear base-learners (Hothorn et al., 2010). Although gradient boosting—which is not discussed in the articles by Kruppa et al. (2014a, 2014b)—has its roots in the machine learning field, linear base-learners result in accessible prediction models with essentially the same interpretation as classical regression and classification models. Specifically, results obtained from linear gradient boosting are easily transportable.

An issue related to transportability is the *reproducibility* of published results. In case of a traditional model such as logistic regression, for example, researchers can usually choose among a variety of

equivalent implementations and thus reproduce (or validate) published results very easily. Machine learning methods, in contrast, are usually less transparent and unfortunately often poorly documented. Furthermore, if a given method is implemented in different softwares (which is not always the case), these variants often yield different results. Clearly, this makes reproducibility and external validation of a model difficult. Following the reproducible research policy of the Biometrical Journal (Hothorn et al., 2009), Kruppa et al. (2014a, 2014b) have documented their software code and also provide a publicly available implementation of their methods via the Random Jungle software (Schwarz et al., 2010).

4 Conclusion

We appreciate the successful effort of the authors to make machine learning methods more accessible to the biometrical community through two articles that contain both scientific contributions and tutorial-like presentations and illustrations of the considered methods. In our view, the choice of the learning method, parameter tuning, and computational transportability are some of the remaining challenges that machine learning methods will have to address to establish themselves as prediction tools in biometrical practice.

Conflict of interest
The authors have declared no conflict of interest.

References

Bernau, C., Augustin, T. and Boulesteix, A. L. (2013). Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics* **69**, 693–702.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 493–507.

Boulesteix, A.-L., Richter, A. and Bernau, C. (2013). Complexity selection with cross-validation for lasso and sparse partial least squares using high-dimensional data. In: *Algorithms from and for Nature and Life*. Springer, Berlin, DE, pp. 261–268.

Boulesteix, A. L. and Strobl, C. (2009). Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Medical Research Methodology* **9**, 85.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* **16**, 199–231.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research* **11**, 2109–2113.

Hothorn, T., Held, L. and Friede, T. (2009). Biometrical journal and reproducible research. *Biometrical Journal* **51**, 553–555.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K. and Boulesteix, A.-L. (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics* **26**, 1990–1998.

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.

Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307.

Schmid, M., Kestler, H. A. and Potapov, S. (2013). On the validity of time-dependent AUC estimators. *Briefings in Bioinformatics*, doi: 10.1093/bib/bbt059.

Schwarz, D. F., König, I. R. and Ziegler, A. (2010). On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* **26**, 1752–1758.

Shmueli, G. (2010). To explain or to predict? *Statistical Science* **25**, 289–310.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**, 1359–1366.

To, M., Skentou, C., Royston, P., Yu, C. and Nicolaides, K. (2006). Prediction of patient-specific risk of early preterm delivery using maternal history and sonographic measurement of cervical length: a population-based prospective study. *Ultrasound in Obstetrics & Gynecology* **27**, 362–367.

Wolpert, D. (2001). The supervised learning no-free-lunch theorems. In *Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications*. pp. 10–24.

