Discussion

What subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction?

Harald Binder*

Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center, Johannes Gutenberg University Mainz, 55101 Mainz, Germany

Received 2 October 2013; revised 6 December 2013; accepted 6 December 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: Added value; External knowledge; Logistic regression; Risk prediction; Uncertainty.

When talking about prediction in a machine learning context, often the focus is on binary class predictions, that is, without providing predicted probabilities. While binary predictions might be sufficient in some areas, Kruppa et al. (2014a) point out many areas of medicine where probability predictions are used. In the following, I am focussing on why these applications motivate probability predictions, what other kinds of medical questions arise in that context, and where the techniques illustrated by Kruppa et al. (2014b) might have difficulties when addressing these questions.

1 Why probability predictions?

In a medical context there will typically be a health-care provider, institution, or individual, who will use probability predictions as one element of a medical decision making process. In such a setting it is essential to be aware of and to communicate uncertainty (Gigerenzer et al., 2007), a task for which probabilities are an important building block. Therefore, it is laudable that the two papers by Kruppa et al. (2014a) and Kruppa et al. (2014b) specifically focus on illustrating how machine learning techniques can provide probability predictions and discuss the properties of different approaches.

Another obvious use for predicted probabilities is to help health-care providers in assigning a weight to the predictions in the overall decision process. For example, a set of markers that provides an inconclusive prediction, for example, most predicted probabilities around 0.5, might be given less weight than a set of markers that nicely assigns most individuals to high- and low-risk groups, say with

*Corresponding author: e-mail: binderh@uni-mainz.de, Phone: +49-6131-17-3938, Fax: +49-6131-17-2968

probabilities smaller than 0.2 or larger than 0.8. There are several other subject matter questions that may arise from the need to determine the contribution of a prediction to the decision process, some of them discussed in the following.

2 Some subject matter questions

Can we (reasonably) predict clinical events based on specific patient characteristics?

A fundamental question in the use of statistical approaches for risk prediction, diagnosis, or in a prognostic/predictive setting is whether specific patient characteristics are useful at all for prediction. The main danger then is to miss or misjudge an important predictor of clinical events that could be useful for arriving at medical decisions. Therefore, the choice between statistical models and machine learning approaches on a global level, and the decision for a specific technique within these broad classes of approaches, is guided by the need to extract a maximum of prediction performance. An approach with competitive prediction performance will make it easier to distinguish between settings without good candidate markers and settings with good candidate markers available. While it is unreasonable to expect an approach to always obtain better prediction performance than other approaches, the search for a general purpose tool that performs well in many situations seems reasonable.

Kruppa et al. (2014a) highlight the consistency properties of many machine learning approaches, which might qualify them as contenders for such a general purpose tool that could augment or replace established approaches such as logistic regression. The question remains in what kind of situation such potential theoretical advantages provide tangible benefit. In the application paper of Kruppa et al. (2014b), it takes seven variants of machine learning approaches in three applications to even identify one setting where logistic regression only offers mediocre performance and is not among the top performers.

It should also be noted that a straightforward combination of univariate testing together with a regularized regression approach was among the top performers in the recent IMPROVER Diagnostic Signature Challenge (Tarca et al., 2013). In particular, regularized regression is an example of how traditional approaches, such as logistic regression, can be adapted to deal with problems, such as a large number of candidate markers, that might at first sight appear to require machine learning techniques. For regularized regression techniques, such as the LASSO, very efficient implementations are available (Friedman et al., 2010), which makes their use nearly as fast and convenient as traditional regression approaches.

While there certainly are settings where the use of machine learning approaches might be advantageous, it has already been observed that “the apparent superiority of more sophisticated methods may be something of an illusion” (Hand, 2006), as the studies evaluating the performance of these methods tend to ignore sources of uncertainty in real applications, such as fine nuances of structure in the training data often being subject to change in real validation data.

Should we measure a new marker in addition to what we already have?

When looking at predicted probabilities for an individual, a natural question is whether or to what extent the uncertainty can be reduced by measuring additional markers, for example, whether individuals from a group with intermediate predicted risk can be divided into a high- and low-risk group based on additional markers. It is surprisingly difficult to answer such added-value questions with respect to prediction performance of more sophisticated approaches. For example, added value can easily be underestimated when comparing the performance of an approach trained only with established markers to training with new and established markers together. As illustrated for a specific risk prediction approach in Binder and Schumacher (2008), the established markers might need to be given a special role in the training process. Unfortunately, uniform treatment of all potential predictors provided seems to be a feature of many machine learning approaches, making it difficult to assign such a special role to established predictors. In contrast, this is implemented in a rather straightforward way for regression modeling approaches by considering the effect of the new markers adjusted for established predictors.

Can we obtain predictions that take into account subject matter knowledge?

Distinguishing between established predictors and new candidate markers might be seen as a specific kind of subject matter knowledge that is to be incorporated into probability predictions. A further step might be to more generally incorporate subject matter knowledge, thus increasing the weight the predictions can and should be given in the medical decision process. Naturally, constraints imposed when using generally uncertain medical knowledge could be at odds with consistency. While the latter is highlighted as important by Kruppa et al. (2014a), health-care providers might favor an approach that incorporates medical knowledge over a consistent approach, if the latter property provides no tangible benefit for an application at hand. Ideally, the uncertainty concerning the usefulness of specific medical knowledge for probability predictions should be incorporated. Such uncertainty could, for example, be taken into account by a tuning parameter that governs the degree of incorporating external knowledge. This tuning parameter should then automatically be chosen for optimizing probability predictions. While corresponding extensions have been suggested for some machine learning approaches (Johannes et al., 2010, e.g.), a regression modeling framework allows to express external knowledge as constraints on model parameters in a straightforward manner (Binder and Schumacher, 2009; Kim et al., 2013). Therefore, users of machine learning approaches might have to work harder for incorporating external knowledge. Such potential extra difficulties in data analysis would need to be justified by some benefit in another area.

3 Concluding remarks

As nicely illustrated by the two papers of Kruppa et al. (2014a) and Kruppa et al. (2014b), machine learning approaches have reached an important milestone by providing probability predictions. However, the medical context that motivates such a kind of predictions naturally leads to the subject matter questions indicated above. With respect to these additional questions, it seems that the machine learning approaches discussed by Kruppa et al. (2014a) and Kruppa et al. (2014b) still have some way to go until they can potentially be as useful for providing answers as regression modeling approaches.

Conflict of interest
The author has declared no conflict of interest.

References

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* **9**, 14.

Binder, H. and Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics* **10**, 18.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M. and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest* **8**, 53–96.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science* **21**, 1–14.

Johannes, M., Brase, J. C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., Sültmann, H. and Beißbarth, T. (2010). Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics* **26**, 2136–2144.

Kim, S., Pan, W. and Shen, X. (2013). Network-based penalized regression with application to genomic data. *Biometrics* **69**, 1–12.

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.

Tarca, A. L., Lauria, M., Unger, M., Bilal, E., Boue, S., Kumar Dey, K., Hoeng, J., Koeppl, H., Martin, F., Meyer, P., Nandy, P., Norel, R., Peitsch, M., Rice, J. J., Romero, R., Stolovitzky, G., Talikka, M., Xiang, Y., Zechner, C. and IMPROVER DSC Collaborators (2013). Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics* **29**, 2892–2899.

