On the projection of machine learning scores to well-calibrated probability estimates

Johanna Schwarz and Dominik Heider

Department of Mathematics and Computer Science, University of Marburg

Abstract

Machine learning models, in particular computer-assisted clinical decision support systems (DSS), have been shown to be powerful tools to reduce medical costs and errors (Tsai et al., 2003) and have been applied in numerous fields, ranging from endoscopy (Dechêne et al., 2014) towards prediction of drug resistance in pathogens (Heider et al., 2014).

Nevertheless, these models typically have a caveat: many of them are perceived as black boxes by clinicians and, unfortunately, the resulting classifier scores cannot usually be directly interpreted as class probability estimates. Consequently, various calibration methods have been developed in the last two decades from the most basic approaches such as Platt scaling to more advanced Bayesian histogram binning strategies. However, finding the best-suited calibration method for a specific classification problem can be a tedious task as there is no easy-to-use tool available that allows a quick and comparative analysis of different calibration methods.

In this work, we present the R-package CalibratR, which can be used to automatically transform machine learning scores to well-calibrated probability estimates. Furthermore, we compare the calibration performance of four different state-of-the-art calibration methods, namely scaling, transforming, histogram binning and BBQ (Naeini et al., 2015) with our novel probability estimation method GUESS.

CalibratR was evaluated on simulated and real data sets where it successfully projected uncalibrated machine learning scores to reliable probability estimates and thus minimized the calibration error in the training and test sets. With the help of CalibratR, we were able to identify the optimal calibration method for each data set and application in a timely and efficient manner.

Using calibrated probability estimates instead of original classifier scores will contribute to the acceptance and dissemination of machine learning based classification models in cost-sensitive applications such as clinical research where easy-to-use yet reliable computer-assisted DSS are urgently needed to reduce preventable human errors.

References

A. Dechêne, C. Jochum, C. Fingas et al. (2014). Endoscopic management is the treatment of choice for bile leaks after liver resection. Gastrointest Endosc., 80(4):626-633.

D. Heider, J.N. Dybowski, C. Wilms, D. Hoffmann (2014): A simple structure-based model for the prediction of HIV-1 co-receptor tropism. BioData Min., 7:14.

M.P. Naeini, G.F. Cooper, M. Hauskrecht (2015): Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence, 2901–2907.

T.L. Tsai, D.B. Fridsma, G. Gatti (2003): Computer decision support as a source of interpretation error. The case of electrocardiograms. Journal of the American Medical Informatics Association, 10(5):478–483.