What did we learn from the IMPROVER Diagnostic Signature Challenge?

Adi L. Tarca\textsuperscript{1,2,*}, Roberto Romero\textsuperscript{2}, Julia Hoeng\textsuperscript{3}, Manuel Peitsch\textsuperscript{3}, Gustavo Stolovitzky\textsuperscript{4}

1. Department of Computer Science, Wayne State University, Detroit, MI, USA
2. Perinatology Research Branch, NICHD/NIH, Bethesda, MD, and Detroit, MI, USA
3. Philip Morris International, Research & Development, Neuchâtel, Switzerland
4. IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
*Contact author: atarca@med.wayne.edu

Keywords: outcome prediction, microarray data, crowdsourcing

The IMPROVER Diagnostic Signature Challenge (DSC) [1,2], was designed by scientists from Philip Morris International's (PMI) Research and Development department and the IBM's Thomas J. Watson Research Center, with the goal of assessing the robustness of methods currently in use for outcome prediction using high-dimensional biological data. In this double blind crowdsourcing competition funded by PMI, public microarray datasets were suggested for developing prediction models in four disease areas (5 endpoints total). The scientific community participated in the challenge, with 54 teams submitting predictions on new sets of samples generated by organizers. The predictions were ranked using three different performance metrics. The team AT & RR received the best performing entrant award, being ranked 2\textsuperscript{nd} in three of the four scored sub-challenges and 12\textsuperscript{th} on the fourth one.

In this work, we will present the main results from the IMPROVER DSC including ranking stability analyses of the participating teams and identification of modeling factors that explained the models success and performance variability. The approach of the best overall team is also presented including an \(R\) package called \texttt{maPredictDSC}, that implements their classification pipeline. The main function of the package starts with raw microarray data files and a class label for each training sample, and returns fitted models and their predictions on the test samples. In addition, the package allows to explore 26 other combinations of preprocessing, features selection and classification methods. Using performance data from the 27 different models produced by \texttt{maPredictDSC} as well from the models submitted in the challenge we have concluded among others that: i) no fixed classification pipeline works best for all datasets ii) the endpoint explains most of the variability in the performance data, of iii) the importance of various steps involved in the classification is dataset and metric dependent iv) classical discriminant analysis methods seemed to perform at least as well as emerging prediction algorithms specifically designed for high-dimensional data provided that proper tuning was made to the specificities of each dataset.

The use of crowdsourcing to validate research building blocks of interest for both academia and the industry proved to be promising by allowing a large body of computational work to be conducted in a few months rather than years.

References