Individuals and Organizations Completing Research in the IMEDS Lab

IBM Thomas J. Watson Research Center and University of Illinois at Urbana-Champaign

Scope of Research:

Proposal to request access to IMEDS Laboratory data from Kush R. Varshney, IBM Thomas J. Watson Research.

Research Objectives and Aims:

Similar to the research by Cynthia Rudin et al. utilizing IMEDS data, we are developing new statistical algorithms to learn simple scoring systems from data.

There is a growing belief that in the face of high complexity, checklists and other simple scorecards or algorithms can significantly improve people's performance on decision-making tasks. An example of such a tool in medicine, the clinical prediction rule, is a simple decision-making rubric that helps physicians estimate the likelihood of a patient having or developing a particular condition in the future. An example of a clinical prediction rule for estimating the risk of stroke is the $CHADS_2$ score using which a health worker determines which of five diagnostic indicators a patient exhibits and adds the corresponding points together. The higher the total point value is, the greater the likelihood the patient will develop a stroke. This rule was manually crafted by health workers and notably contains few conditions with small integer point values and is extremely interpretable by people. We would like to learn clinical prediction rules that generalize accurately from large-scale electronic health record data rather than relying on manual development. The key aspect of the problem is maintaining the simplicity and interpretability of the learned rule: similar to the hand-crafted version rather than a complicated, uninterpretable 'black-box' model. Such transparency is critical for trust and adoption by users.

Scope and Proposed Approach:

We will build upon our recent research on the supervised learning of interpretable classification rules using Boolean compressed sensing ideas. With the same goal as Rudin et al., we will develop a method for learning interpretable clinical prediction rules using sparse signal representation techniques. In that previous work of ours, the form of the classifier was a sparse AND-rule or OR-rule whereas here, we would like to find a sparse set of medical conditions or features with small integer coefficients that are added together to produce a score. Such a model is between the "1-of-N" and "N-of-N" forms implied by OR-rules and AND-rules. The problem has a close connection to the semiquantitative group testing problem.

It is common practice in statistical machine learning research to compare the results of new algorithms with the results of old algorithms on exactly the same data set. Therefore, it is critical for us to access IMEDS data in order to perform sound empirical studies in comparison to Rudin et al.

Impact:

There are many available algorithms for the supervised learning problem, but models from ones that typically perform the best, such as kernel support vector machines, random forests, and neural networks are black boxes in the sense that it is difficult for humans to interpret them. In contrast, early heuristic approaches such as decision lists and decision trees have a high level of interpretability and are still widely used by analytics practitioners for this reason despite being less accurate. Our impact will be to develop a method that produces highly accurate outputs that are well-received, trusted, and adopted by human decision makers. In addition, the clinical prediction rules themselves that the algorithm learns from the electronic medical records will provide valuable decision support for health care workers.

Experience:

Amin Emad, Dmitry Malioutov and Kush Varshney are the individuals involved in this study, and are requesting access to the IMEDS lab. Amin is a fifth-year PhD student in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He has published research papers in several top journals, including IEEE Transactions on Information Theory, PLOS ONE, and IEEE Transactions on Communication. Dmitry and Kush are both research staff members at IBM's T. J. Watson Research Center with more than 5 years of industrial and applied machine learning research experience each (after completing their PhDs). They have also published in several top journals, including Journal of Machine Learning Research, IEEE Transactions on Signal Processing, IEEE Transactions on Information Theory, Quantitative Finance, and Big Data.

Timeline:

We expect the entire research project (methodological developments, empirical studies, IMED lab experiments, publication) to take less than 12 months, and are thus requesting access until April 1, 2016, with access beginning as soon as possible.