CrossMark

EDITOR'S NOTE

Editorial: special issue on machine learning for health and medicine

Jenna Wiens¹ · Byron C. Wallace¹

Received: 8 September 2015 / Accepted: 23 September 2015 / Published online: 22 December 2015 © The Author(s) 2015

The exponential proliferation of heterogeneous health-related information presents unprecedented opportunities for improving patient care. Health-related data arise from diverse sources including, but not limited to, individual patient health records, genomic data, data from wearable health monitors, online reviews of physicians, clinical literature, and medical imagery. Currently, clinicians and other domain experts—including physicians and biostatisticians—are overwhelmed by the volume and variety of the available data. Transforming these data into actionable knowledge presents a barrage of pragmatic and technical challenges.

Members of the machine learning community, broadly construed and including researchers in core machine learning, statistical natural language processing, computer vision and other related sub-fields, have been leading the way in developing methods to turn massive amounts of data into actionable knowledge with the goal of ultimately improving patient care. While there is a long history of work at the intersection of machine learning and health, recently there has been a resurgence in the area due to the increased availability of data and computational power, and the potential of the latter to capitalize on the former. Reflecting this excitement, health-related workshops have taken place in recent years at several machine learning related conferences, including NIPS, AAAI, KDD, and ICML (amongst others). These workshops have attracted a diverse set of participants from related sub-communities. This special issue is meant to provide a unified forum for a sample of high-quality work in this interdisciplinary area, thus showcasing the promise of machine learning for data-driven healthcare.

Machine learning in the context of healthcare applications is especially challenging due to a combination of issues. First, there is the sheer volume and variety of the data. As mentioned above, machine learning in healthcare involves many data types, from waveforms to unstructured text. Second, research in this area spans the entire learning pipeline from problem formulation, to feature engineering/selection, to model learning, and output. At each step along the pipeline one may encounter challenges related to a variety of issues including: confounding, missing data, class imbalance, temporal consistency, and task heterogeneity. Finally, it is not enough to develop accurate models; to have impact, the technology must



Austin, TX, USA

be adopted by clinicians or biomedical researchers. Consequently, model interpretability (which may be overlooked in other contexts) is often imperative. We note that many of the challenges mentioned above arise in other contexts and are fundamental issues in machine learning. While these problems are not unique to our context, they do arise more frequently (and often jointly) in healthcare, compared to other domains.

This special issue is organized to reflect the diversity of work in this field. The work presented here spans the spectrum of healthcare from individual genomics (performing variable selection over genomic data) to public health (supporting the production of clinical evidence syntheses). Despite the variability in modalities (and hence methods), reoccurring methodological challenges have emerged. The papers in this special issue address several of these characteristic challenges including: handling data sparsity, ensuring model interpretability, performing variable selection over high dimensional data, mitigating class imbalance, untangling causality and modeling temporal dynamics.

"Multi-Task Seizure Detection: Addressing Intra-Patient Variation in Seizure Morphologies" by Alex Van Esbroeck, Landon Smith, Zeeshan Syed, Satinder Singh and Zahi Karam looks to increase the accuracy of early detection of epileptic seizures from continuous electroencephalographic (EEG) data. To achieve robust seizure detection, they attempt to account for inter-patient and intra-patient variation in seizure morphologies using a multi-task learning approach. This multi-task model includes patient-specific components and a shared component; the latter enables the model to generalize well. They demonstrate that their approach is able to substantially reduce the number of false positives produced by the model.

"Learning (Predictive) Risk Scores in the Presence of Censoring due to Interventions" by Kirill Dyagilev and Suchi Saria aims to overcome confounding variables to predict the severity of sepsis from electronic health records. The authors use a ranking algorithm to achieve this, using pairwise clinical comparisons regarding disease severity at different times (and across patients) to avoid interventional confounding. This approach outperforms existing state-of-the-art methods for severity score prediction and the predictions were found to be consistent with clinical expectations.

"Supersparse Linear Integer Models for Optimized Medical Scoring Systems", authors Berk Ustun and Cynthia Rudin aim to create a highly tailored, accurate and sparse scoring system for sleep apnea using small coprime integer coefficients. A challenge here is that the model must be both accurate and sparse; the latter is necessary for achieving interpretability. This motivates their development of a new method, which they term a Supersparse Linear Integer Model (SLIM). This model is now in use at the Massachusetts General Hospital Sleep Laboratory, demonstrating its value in practice.

"Learning Classification Models of Cognitive Conditions from Subtle Behaviors in the Digital Clock Drawing Test" by William A. Souillard-Mandar, Randall Davis, Cynthia Rudin, Rhoda Au, David J. Libon, Rodney Swenson, Catherine C. Price, Melissa Lamar and Dana L. Penney aims to build a system for automatically screening patients for cognitive impairment using data recorded by digital pen during a "clock drawing test". This poses several challenges, including the high amount of subjectivity inherent to assessments, and the need for the screening model to be interpretable. They propose a system that combines both state-of-the-art sensor technology (digital pen) and state-of-the-art machine learning methods to learn a model focused on sparsity and understandability. They achieve better discrimination in terms of screening and diagnosis than existing clinical scoring systems.

"A Dynamic Ensemble Approach to Robust Classification in the Presence of Missing Data" by Bryan Conroy, Larry Eshelman, Cristhian Potes and Minnan Xu-Wilson presents an approach for early detection of hemodynamic instability of patients in the ICU. Missing data is inherent to this problem, and the authors therefore address this via a two-stage approach



that uses an ensemble of low-dimensional classifiers induced on the available data and then combining the outputs of these using a dynamic weighting scheme that depends on the pattern of missing data. They compare their approach to alternative strategies to handling missing data, and show the advantages of their approach in light of the practical constraints of their application.

"Learning to Identify Relevant Studies for Systematic Reviews using Random Forest and External Information" by Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani considers the challenging task of semi-automating the identification of published biomedical evidence for inclusion in systematic reviews. This poses the challenge of class imbalance, as there are far fewer relevant than irrelevant articles for any given review. The primary contribution of this work is the exploitation of "external" features, including citation information. They demonstrate that this approach achieves high recall (i.e., is able to identify the studies of interest) while reducing the number of irrelevant articles; in turn, this may reduce the workload involved in producing systematic reviews, and ultimately enable evidence-based care.

Finally, "Accelerating a Gibbs Sampler for Variable Selection on Genomics Data with Summarization and Variable Pre-selection combining an Array DBMS and R" by David Sergio Matusevich, Wellington Cabrera and Carlos Ordonez considers the general task of variable selection in (very) high dimensional data. More specifically, they target the task of identifying medically significant variables. Their primary contribution is a set of algorithmic optimizations that accelerate sampling methods (specifically, Gibbs sampling), thus enabling the use of Markov Chain Monte Carlo (MCMC) methods for variable selection in very big datasets. They demonstrate the practical utility of their method in terms of reduced running time.

We thank all authors for their contributions to this special issue. The high quality nature of the submissions made it possible to showcase a variety of work taking place at the intersection of machine learning and health, without any one modality, application, or method dominating. We would also like to thank the reviewers, for their thoughtful feedback and quick turnaround during the multiple rounds of reviewing. Finally, we thank Dragos Margineantu and the rest of the team at the Machine Learning Journal for the opportunity to oversee this special issue, and for the help we received throughout the process.

Our goal for this issue is to increase the visibility of this interdisciplinary field, while showcasing the breadth of machine learning research challenges that present in healthcare. We hope that the work presented here will inspire more machine learning researchers to join us in this uniquely challenging field. We believe the potential societal impact of this work will only continue to grow.

