Machine learning with limited data challenges and opportunities in rare diseases

This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@808343d on May 25, 2020.

Authors

- **Jineta Banerjee**
  0000-0002-1775-3645 · jaybee84
  Sage Bionetworks · Funded by ['Neurofibromatosis Therapeutic Acceleration Program', "Children's Tumor Foundation"]

- **Robert J Allaway**
  0000-0003-3573-3565 · allaway · allawayr
  Sage Bionetworks · Funded by ['Neurofibromatosis Therapeutic Acceleration Program', "Children's Tumor Foundation"]

- **Jaclyn N Taroni**
  0000-0003-4734-4508 · jaclyn-taroni
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Casey Greene**
  0000-0001-8713-9213 · cgreene
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Justin Guinney**
  0000-0003-1477-1888 · jguinney
  Sage Bionetworks · Funded by ['Neurofibromatosis Therapeutic Acceleration Program', "Children's Tumor Foundation"]

Abstract

Substantial technological advances have dramatically changed biomedicine by making deep characterization of patient samples routine. These technologies provide a rich portrait of genes, cellular pathways, and cell types involved in complex phenotypes. Machine learning is often a perfect fit for the types of data now being generated, and Nature Methods routinely has reports of machine learning methods that extract disease-relevant patterns from these high dimensional datasets. Often, these methods require a large number of samples to identify reproducible and biologically meaningful patterns. With rare diseases, biological specimens and consequently data, are limited due to the rarity of the condition. In this perspective, we outline the challenges and emerging solutions for using machine learning in these settings. We aim to spur the development of powerful machine learning techniques for rare diseases. We also note that precision medicine presents a similar challenge, in which a common disease is partitioned into small subsets of patients with shared etiologies and treatment strategies. Advances from rare disease research are likely to be highly informative for other applications as well.

Introduction

This is an introductory section...

Techniques that build on prior knowledge and indirectly related data are necessary for many rare disease applications

> This section will highlight promising approaches for analyzing rare disease data to extract biological insights. We will discuss techniques like transfer learning, representation learning, cascade learning, integrative analysis, and knowledge-graph creation and use that leverage other knowledge and data sources to construct testable hypotheses from rare diseases datasets with limited sample sizes 1–8.

Techniques and procedures must be implemented to manage model complexity without sacrificing the value of machine learning

> Inherent challenges posed by low sample numbers in rare diseases are further aggravated by disease heterogeneity, poorly defined disease phenotypes, and often a lack of control (i.e. normal) data. Machine learning approaches must be carefully designed to address these challenges. We discuss how to implement methodological solutions like bootstrapping sample data, regularization methods for deep learning, and hyper-ensemble techniques to minimize misinterpretation of the data9,10.

Techniques to manage disparities in data generation are required to power robust analyses in rare diseases

> Rarity of patients leads to heterogeneity in sample collection, causing disparities in the data. We will discuss how rigorous normalization and methodologies capturing sample-wise gene-set level information can help appropriate integration of disparate data points to power machine learning approaches11–13.

Conclusions

> We will conclude by discussing the potential of the above-mentioned approaches in rare diseases and other biomedical areas where data is scarce.

References