Machine learning methods for rare diseases

This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@15c7d28 on June 11, 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

As with common diseases, genomic and transcriptomic data from rare diseases can suffer from artifacts introduced by batch, processing methodology, sequencing platform, or other non-biological phenomena. The consequences of these non-biological artifacts are amplified in rare diseases which often have few samples and heterogenous phenotypes. Furthermore, because datasets are many times pieced together from multiple small studies, disease phenotype or other important biological characteristics are often confounded by the previously mentioned "batch" factors. A key consideration here is, if possible, active dialogue with the data generators or experts in the field who may have unexpected insights into potential sources of variation. One example of the value of this, experienced by the authors, occurred when studying tumors associated with the disease neurofibromatosis type 1. These datasets were, unbeknownst to the computational biologists, generated from samples obtained with vastly different surgical techniques (laser ablation and excision vs standard excision), resulting in substantial biological differences that are a consequence of process, not reality. One might expect, in this example, that this technical decision would result in profound changes in the underlying biology, such as the activation of heat shock protein related pathways, unfolded protein responses, and so on. Consequently, careful assessment of confounding factors and implementation of normalization methods is important to identifying biologically meaningful features within a dataset. Assessment of confounding factors and heterogeneity in rare disease datasets is perhaps most easily performed using unsupervised learning approaches such as clustering and dimensionality reduction. Clustering methods like k-means clustering or hierarchical clustering can be used to characterize the structure present in many different types of data such as genomic or imaging data. [1,2]. Similarly, a variety of dimensionality reduction methods are can be used to visualize sample heterogeneity and potential confounding variables, including multidimensional scaling (MDS), principal components analysis (PCA), t-distributed stochastic neighbor embedding (tSNE), and uniform manifold approximation and projection (UMAP), among many others. [3,4,5,6] All of these methods can be effectively used to identify batch effects and other structure in the data (cite), though some, like tSNE and UMAP, have parameters, such as perplexity (number of nearest neighbors), that can substantially affect the output, and thus the interpretation, of the analysis [6,7]. Therefore, successful application of these methods requires a sufficient understanding of the underlying method and parameter sweeping to get a clear picture of the structure of the underlying data. Dimensionality reduction techniques are not restricted to 'omic' data - they can also be used in rare disease applications to characterize the structure and heterogenity of imaging data [8], mass cytometry data [9], and others. Once the nature of the non-biological heterogeneity has been established, different techniques can be used to correct the differences. Common approaches to ameliorate non-biological effects include reprocessing the raw data using a single analysis pipeline if the data are obtained from different sources, application of batch correction methods [10,11], normalization of raw values (e.g. z-scores, trimmed mean of M-values [12]). It can also be helpful to be fatalistic, in some sense, when working with rare disease data. For various reasons including ethical considerations, limited funding, and limited biospecimen availability, experimental design and the resulting data will be less-than-ideal - for example - when batch variables and biological variables are confounded. In these cases, it may be prudent to take a step back, re-evaluate the data, and identify methods that can operate within these constraints.

Conclusions

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

draft

> this is a test file to differentiate draft-branch from master

References

1. **Clustering cancer gene expression data: a comparative study**
Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir, Alexander Schliep
BMC Bioinformatics (2008-11-27) https://doi.org/dqqbn6
DOI: 10.1186/1471-2105-9-497 · PMID: 19038021 · PMCID: PMC2632677

2. **Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis**
Sonal Kothari, John H. Phan, Todd H. Stokes, Adeboye O. Osunkoya, Andrew N. Young, May D. Wang
IEEE Journal of Biomedical and Health Informatics (2014-05) https://doi.org/gdm9jd
DOI: 10.1109/jbhi.2013.2276766 · PMID: 24808220 · PMCID: PMC5003052

3. **Multidimensional Scaling**
Michael A. A. Cox, Trevor F. Cox
Springer Berlin Heidelberg (2008) https://doi.org/dg9m4f
DOI: 10.1007/978-3-540-33037-0_14

4. **Principal component analysis: a review and recent developments**
Ian T. Jolliffe, Jorge Cadima
Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2016-04-13) https://doi.org/gcsfk7
DOI: 10.1098/rsta.2015.0202 · PMID: 26953178 · PMCID: PMC4792409

5. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

6. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv (2018-12-07) https://arxiv.org/abs/1802.03426

7. **How to Use t-SNE Effectively**
Martin Wattenberg, Fernanda Viégas, Ian Johnson
Distill (2016-10-13) https://doi.org/gffk7g
DOI: 10.23915/distill.00002

8. **dSNE: a visualization approach for use with decentralized data**
D. K. Saha, V. D. Calhoun, Y. Du, Z. Fu, S. R. Panta, S. M. Plis
bioRxiv (2020-04-06) https://doi.org/ggzp52
DOI: 10.1101/826974

9. **Sensitive detection of rare disease-associated cell subsets via representation learning**
Eirini Arvaniti, Manfred Claassen
Nature Communications (2017-04-06) https://doi.org/gf9t7w
DOI: 10.1038/ncomms14825 · PMID: 28382969 · PMCID: PMC5384229

10. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2007-01) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

11. **svaseq: removing batch effects and other unwanted noise from sequencing data**
Jeffrey T. Leek
Nucleic Acids Research (2014-12-01) https://doi.org/f8k8kf
DOI: 10.1093/nar/gku864 · PMID: 25294822 · PMCID: PMC4245966

12. **A scaling normalization method for differential expression analysis of RNA-seq data**
Mark D Robinson, Alicia Oshlack
Genome Biology (2010) https://doi.org/cq6f8b
DOI: 10.1186/gb-2010-11-3-r25 · PMID: 20196867 · PMCID: PMC2864565