Machine learning methods for rare diseases

Authors

- **Jineta Banerjee**
  ⓘD [0000-0002-1775-3645](#) · ◯ [jaybee84](#)
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Robert J Allaway**
  ⓘD [0000-0003-3573-3565](#) · ◯ [allaway](#) · 🐦 [allawayr](#)
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Jaclyn N Taroni**
  ⓘD [0000-0003-4734-4508](#) · ◯ [jaclyn-taroni](#)
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Casey Greene**
  ⓘD [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**
  ⓘD [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.

Regularization

Machine learning algorithms are optimized to find patterns among data points and prioritizes the strongest patterns that exist in a dataset. Given a limited dataset with strong pre-existing technical differences between groups of samples, this optimization may lead to the model learning technical differences thus lowering its predictive accuracy [1]. For example, in a set of 1000 samples where 700 samples are from one healthcare site and 300 from another, it is likely that there will remain site-specific differences between them even after normalization of the samples. If the site-specific differences are more pronounced than the underlying patterns differentiating the samples, any machine learning model trained with these data will preferentially learn the site-specific differences to classify the samples, and rank them higher than the underlying patterns leading to a model showing high prediction accuracy of training data (termed low bias in model). When new test data points are introduced to the model, possibly coming from a third site, the model is unable to locate the earlier differences in the new data points and fails to classify them accurately causing a significant drop in accuracy of the model (termed high variance in model prediction). Such a model is termed "overfitted" to its training data. Overfitting can lead to misinterpretation of the site-specific differences as true patterns in the limited data points and thus needs to be minimized. Minimization of overfitting can be accomplished by cross-validation and regularization methodologies. While cross-validation aims to reduce the variance in prediction, regularization adds a small amount of bias to the initial model to minimize its dependance and sensitivity to training data. Regularization makes models less reliant on training data by adding a penalty (determined by crossvalidation), and then minimizes the error between the model's prediction and ground truth of the test data. Regularization can not only minimize overfitting but can additionally help in predicting outcomes using a limited number of samples.

Regularization can be of three main types, each with their particular strengths and weaknesses. (1) Ridge regression aims to minimize the magnitude of the features, but in models that try to select the most important features for accurate prediction of sample labels, ridge regression shrinks all features equally, but cannot completely remove unimportant features. Thus in presence of many correlated parameters (eg. gene expression networks), ridge regression may not be ideal in reducing the feature space. (2) LASSO or least absolute shrinkage and selection operator regression on the other hand works well for selecting few important features since its effect can minimize the magnitude of some features more than the others. Thus it helps in selecting most important features while the magnitude of irrelevant features are shrunk to 0 and eventually removed. This selection attribute of LASSO (in a sample set of size "n", LASSO can select "n" features for the model) may be an advantage in reducing model complexity, but a disadvantage in cases where identification of all possible collinear features is important (eg. all biomarkers correlating to a particular disease phenotype) [2]. (3) Elastic-Net regression is a combination of LASSO and ridge regression[3]. Both of the methodologies when applied together helps to select most useful features, specially where there are a lot of correlated features. In this setup, LASSO leads to selection of one of the correlated features and reduces the others to 0 (grouping of features), and the magnitude of the selected features are then minimized through ridge regression.

Any supervised learning implementation in rare disease would require robustness towards feature selection from a small number of samples, i.e. the features selected by a model as important should be stable in view of new data points added to a dataset, even though their relative importance may change due to additional evidence. This robustness is mostly acquired through the combination of various regression strategies. Since machine learning applications in rare disease are infrequent, combination strategies used for rare variant discovery and immune cell signature discovery can serve as good case studies to examine. Many deleterious genomic variants can be extremely rare due to the constant selection pressure working against them. Since the frequency of a rare variant is so low (less than 1%) applying routine statistical procedures that were extensively developed for common variant

association, to analyze a low minor allele frequency (MAF) seem inappropriate [4]. For its feature selection attribute, LASSO has been widely applied in microarray and GWAS data for common variants. But since LASSO by itself is too stringent for rare variants, it has been employed along with group penalties to help identify rare variants/ low frequency predictors [5]. Variations of LASSO have also been implemented to aggregate or group the occurrance of rare variants together by gene or chromosome location [6,7,8]. In this strategy, a 0–1 dummy variable was created for each SNP based on the presence or absence of the rare variant. Then linear combinations of the selected dummy variables were considered by using the LASSO procedure. Even though most of the dummy variables were 0, their linear combination was more likely to be nonzero thus leading to increased signal to noise ratio for the rare variants. Only those linear combinations that were non-zero in at least 5% of the subjects were then included to ensure that the new markers were not rare [7,9]. While ridge regression is not usually utilized for feature selection, adaptive ridge regression has been utilized to help combine rare variants into a single score analogous to feature engineering for increasing the signal of rare variants[10]. Another variation of LASSO included its integration with the probabilistic logistic bayesian approach to identify a protective rare variant in lung cancer[11]. Xu et al. on the other hand combined the feature selection methods with a generalized pooling strategy, and evaluated the performance of these hybrid approaches for detection of rare genetic variants[12]. Another interesting approach is the sparse-group LASSO approach which incorporates prior knowledge into the regularization[13]. This approach works well for a scenario where only few genes in a pathway are true predictors of a phenotype, where it helps select the driving genes in a pathway of interest.

Alternatively, Elastic-net regression (a combination of LASSO and ridge regression) has also been used to reduce the feature space in various types of cancer datasets [14,15]. In cases where the number of features were far greater than the number of samples, elastic-net has usually been found to outperform the other regression approaches [3]. A variation of the elastic-net regression was used for identifying immune cell signatures in an RNA-seq dataset where the number of cells sampled were far fewer than number of genes profiled [**???** 10.1186/s12859-019-2994-z]. This two-step regularized logistic regression technique included a pre-filtering phase to select the optimal number of genes and then implemented elastic-net regularization for gene selection. The second step generated gene signatures for individual cell types using selected genes from first step and then implemented a binary regularized logistic regression for each cell type against all other samples.

Still to add: techniques in deep learning eg. Deep and shallow architecture: https://ieeexplore.ieee.org/document/7863293

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. [16,17]. 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. [18,19,20,21] 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 [21,22]. 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 [23], mass cytometry data [24], 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 [25,26], normalization of raw values (e.g. z-scores, trimmed mean of M-values [27]). 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. **Definitions, methods, and applications in interpretable machine learning**
   W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yu
   Proceedings of the National Academy of Sciences (2019-10-29) https://doi.org/ggbhmq
   DOI: 10.1073/pnas.1900654116 · PMID: 31619572 · PMCID: PMC6825274

2. **Regularization**
   Jake Lever, Martin Krzywinski, Naomi Altman
   Nature Methods (2016-09-29) https://doi.org/gf3zrr
   DOI: 10.1038/nmeth.4014

3. **Regularization and variable selection via the elastic net**
   Hui Zou, Trevor Hastie
   Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2005-04)
   https://doi.org/b8cwwr
   DOI: 10.1111/j.1467-9868.2005.00503.x

4. **Statistical analysis strategies for association studies involving rare variants**
   Vikas Bansal, Ondrej Libiger, Ali Torkamani, Nicholas J. Schork
   Nature Reviews Genetics (2010-10-13) https://doi.org/dn4jtz
   DOI: 10.1038/nrg2867 · PMID: 20940738 · PMCID: PMC3743540

5. **Association screening of common and rare genetic variants by penalized regression**
   H. Zhou, M. E. Sehl, J. S. Sinsheimer, K. Lange
   Bioinformatics (2010-08-06) https://doi.org/c7ndkx
   DOI: 10.1093/bioinformatics/btq448 · PMID: 20693321 · PMCID: PMC3025646

6. **Identification of Grouped Rare and Common Variants via Penalized Logistic Regression**
   Kristin L. Ayers, Heather J. Cordell
   Genetic Epidemiology (2013-09) https://doi.org/f5cw72
   DOI: 10.1002/gepi.21746 · PMID: 23836590 · PMCID: PMC3842118

7. **A LASSO-based approach to analyzing rare variants in genetic association studies**
   Jennifer S Brennan, Yunxiao He, Rose Calixte, Epiphanie Nyirabahizi, Yuan Jiang, Heping Zhang
   BMC Proceedings (2011-11-29) https://doi.org/bjcndj
   DOI: 10.1186/1753-6561-5-s9-s100 · PMID: 22373373 · PMCID: PMC3287823

8. **Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data**
   Bingshan Li, Suzanne M. Leal
   The American Journal of Human Genetics (2008-09) https://doi.org/d4jpcb
   DOI: 10.1016/j.ajhg.2008.06.024 · PMID: 18691683 · PMCID: PMC2842185

9. **Comparison of statistical approaches to rare variant analysis for quantitative traits**
   Han Chen, Audrey E Hendricks, Yansong Cheng, Adrienne L Cupples, Josée Dupuis, Ching-Ti Liu
   BMC Proceedings (2011-11-29) https://doi.org/b9mf4x
   DOI: 10.1186/1753-6561-5-s9-s113 · PMID: 22373209 · PMCID: PMC3287837

10. **Adaptive Ridge Regression for Rare Variant Detection**
    Haimao Zhan, Shizhong Xu

PLoS ONE (2012-08-28) https://doi.org/f36tm5
DOI: 10.1371/journal.pone.0044173 · PMID: 22952918 · PMCID: PMC3429469

11. **An Improved Version of Logistic Bayesian LASSO for Detecting Rare Haplotype-Environment Interactions with Application to Lung Cancer**
Yuan Zhang, Swati Biswas
Cancer Informatics (2015-02-09) https://doi.org/ggxxfp
DOI: 10.4137/cin.s17290 · PMID: 25733797 · PMCID: PMC4332044

12. **Multiple Regression Methods Show Great Potential for Rare Variant Association Tests**
ChangJiang Xu, Martin Ladouceur, Zari Dastani, J. Brent Richards, Antonio Ciampi, Celia M. T. Greenwood
PLoS ONE (2012-08-08) https://doi.org/f35726
DOI: 10.1371/journal.pone.0041694 · PMID: 22916111 · PMCID: PMC3420665

13. **A Sparse-Group Lasso**
Noah Simon, Jerome Friedman, Trevor Hastie, Robert Tibshirani
Journal of Computational and Graphical Statistics (2013-04) https://doi.org/gcvjw8
DOI: 10.1080/10618600.2012.681250

14. **Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification**
Zakariya Yahya Algamal, Muhammad Hisyam Lee
Computers in Biology and Medicine (2015-12) https://doi.org/f73xvj
DOI: 10.1016/j.compbiomed.2015.10.008 · PMID: 26520484

15. **Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification**
Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, Hai Zhang
BMC Bioinformatics (2013-06-19) https://doi.org/gb8v2x
DOI: 10.1186/1471-2105-14-198 · PMID: 23777239 · PMCID: PMC3718705

16. **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

17. **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

18. **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

19. **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

20. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

21. **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

22. **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

23. **Automatic detection of rare pathologies in fundus photographs using few-shot learning**
Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, Béatrice Cochener
Medical Image Analysis (2020-04) https://doi.org/ggsrc7
DOI: 10.1016/j.media.2020.101660 · PMID: 32028213

24. **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

25. **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

26. **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

27. **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