Journal of Biomedical Informatics 64 (2016) 168–178

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Semi-supervised learning of the electronic health record for phenotype stratification

Brett K. Beaulieu-Jones [a,b], Casey S. Greene [b,c,d,*], the Pooled Resource Open-Access ALS Clinical Trials Consortium [1]

[a] Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, United States
[b] Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, United States
[c] Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, United States
[d] Institute for Translational Medicine and Therapeutics, University of Pennsylvania, Perelman School of Medicine, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 18 February 2016
Revised 5 October 2016
Accepted 8 October 2016
Available online 12 October 2016

Keywords:
Electronic health record
Denoising autoencoder
Unsupervised
Electronic phenotyping
Patient stratification
Disease subtyping

ABSTRACT

Patient interactions with health care providers result in entries to electronic health records (EHRs). EHRs were built for clinical and billing purposes but contain many data points about an individual. Mining these records provides opportunities to extract electronic phenotypes, which can be paired with genetic data to identify genes underlying common human diseases. This task remains challenging: high quality phenotyping is costly and requires physician review; many fields in the records are sparsely filled; and our definitions of diseases are continuing to improve over time. Here we develop and evaluate a semi-supervised learning method for EHR phenotype extraction using denoising autoencoders for phenotype stratification. By combining denoising autoencoders with random forests we find classification improvements across multiple simulation models and improved survival prediction in ALS clinical trial data. This is particularly evident in cases where only a small number of patients have high quality phenotypes, a common scenario in EHR-based research. Denoising autoencoders perform dimensionality reduction enabling visualization and clustering for the discovery of new subtypes of disease. This method represents a promising approach to clarify disease subtypes and improve genotype-phenotype association studies that leverage EHRs.

1. Introduction

Biomedical research often considers diseases as fixed phenotypes, but many have evolving definitions and are difficult to classify. The electronic health record (EHR) is a popular source for electronic phenotyping to augment traditional genetic association studies, but there is a relative scarcity of research quality annotated patients [1]. Electronic phenotyping relies on either codes designed for billing or time intensive manual clinician review. This is an ideal environment for semi-supervised algorithms, performing unsupervised learning on many patients followed by supervised learning on a smaller, annotated, subset. Denoising autoencoders (DAs) are a powerful tool to perform unsupervised learning [2]. DAs are a type of artificial neural network trained to reconstruct an original input from an intentionally corrupted input. Through this training they learn higher-level representations modeling the structure of the underlying data. We sought to determine whether applying DAs to the EHR could reduce the number of annotated patients required, construct non-billing code based phenotypes and elucidate disease subtypes for fine-tuned genetic association.

The United States federal government mandated meaningful use of EHRs by 2014 to improve patient care quality, secure and communicate patient information, and clarify patient billing [3,4]. Despite not being designed specifically for research, EHRs have already proven an effective source of phenotypes in genetic association studies [5,6]. Initially, phenotypes were hand designed based on manual clinician review of patient records. These studies were limited by the time and cost inherent in manual review [7,8],

[*] Corresponding author at: Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, United States.

E-mail addresses: brettbe@med.upenn.edu (B.K. Beaulieu-Jones), csgreene@upenn.edu (C.S. Greene).

[1] Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH, Northeast ALS Consortium, Novartis, Prize4Life, Regeneron Pharmaceuticals, Inc., Sanofi, Teva Pharmaceutical Industries, Ltd.

http://dx.doi.org/10.1016/j.jbi.2016.10.007
1532-0464/© 2016 The Author(s). Published by Elsevier Inc.

but DAs can make use of unlabeled data. After unsupervised pretraining, the trained DA's hidden layer can be used as input to a traditional classifier to create a semi-supervised learner. This allows the DA to learn from all samples, even those without labels, and requires only a small subset to be annotated. Today, phenome-wide association studies (PheWAS) are the most prevalent example of EHR phenotyping, proving particularly effective at identifying pleiotropic genetic variants [9]. PheWASs often use algorithms based on the International Classification of Disease (ICD) codes to construct a phenotype. This coding system was designed for billing, not to capture research phenotypes. DA constructed features are combinations of many components of clinical data and may provide a more holistic view of a patient than billing codes alone.

Through extensive study, disease diagnoses can become more precise over time [10–14]. Cancers, for example, were historically typed by occurrence location and the efficacy of different treatments. As the mechanisms of cancer are better understood, they are further categorized by their physiological nature. The progression of subtypes in lung cancer illustrates this increased understanding over time [10]. Beginning with a single diagnosis based on occurrence in the lung, lung cancer has been divided into dozens of subtypes over several decades based on histological analysis, and genetic markers [11–14]. The unsupervised nature of DAs means that even if the definitions of a disease change, they would not need to be retrained. The ability to identify more homogenous phenotypes showed increased genotype to phenotype linkage in schizophrenia, bipolar disease [15], and Rett Syndrome [16–19]. Furthermore, type 2 diabetes subtypes have been discovered using topological analysis of EHR patient similarity [20]. The dimensionality reduction possible with a DA makes clustering and visualization more feasible. Subtyping exposes disease heterogeneity and may contribute to additional physiological understanding.

Previous work in semi-supervised learning of the EHR relies on closed source commercial software [20], and natural language processing of free text fields to match clinical diagnosis [21,22]. We are not aware of any previous work performing semi-supervised classification and clustering from quantitative structured patient data.

We evaluate DAs for phenotype construction using four simulation models of EHR data for complex phenotypes, modify DAs to effectively handle missingness in data and use the DA to create cluster visualizations that can aid in the discovery of subtypes of complex diseases. We apply these methods to predict ALS patient survival and to visualize ALS patient clusters. ALS is a progressive neurodegenerative disorder, which attacks the neurons responsible for controlling muscle function [23]. ALS patients typically die within 3–5 years, but some patients can survive more than 10 years, the disease is considered clinically heterogeneous and predicting the rate of progression can be challenging [24].

2. Methods

We developed an approach, entitled "Denoising Autoencoders for Phenotype Stratification (DAPS)," that constructs phenotypes through unsupervised learning. This generalized phenotype construction can be used to classify whether patients have a particular disease or to search for disease subtypes in patient populations. To evaluate DAPS, we created a simulation framework with multiple hidden factors influencing potentially overlapping observed variables. We evaluated the reduced DA models against feature-complete representations with popular supervised learning algorithms. These evaluations covered both complete datasets, as well as the more realistic cases of incompletely labeled and missing data. We developed a technique that uses the reduced feature-space of the DA to visualize potential subtypes. Finally, we evaluate DAs ability to predict ALS patient survival in both classification and clustering tasks. Each of these is fully described below and full parameters included in sweeps are available in the supplementary materials.

Source code to reproduce each analysis is included in our repository (https://github.com/greenelab/DAPS) [25] and is provided under a permissive open source license (3-clause BSD). A docker build is included with the repository to provide a common environment to easily reproduce results without installing dependencies [26]. In addition, Shippable, a continuous integration platform, is used to reanalyze results in a clean environment and generate figures after each commit [27].

2.1. Unsupervised training with denoising autoencoders

DAs were initially introduced as a component in constructing the deep networks used in deep learning [28]. Deep learning algorithms have become the dominant performers in many domains including image recognition, speech recognition and natural language processing [29–34]. Recently they have also been used to solve biological problems including tumor classification, predicting chromatin structure and protein binding [2,35,36]. DAs showed strong performance early in the deep learning revolution but have been surpassed in most domains by convolutional neural networks or recurrent neural networks [28]. While these complex deep networks have surpassed the performance of DAs in these areas, they rely on strictly structured relationships such as the relative positions of pixels within an image [31,37]. This structure is unlikely to exist in the EHR. In addition, complex deep networks are notoriously hard to interpret. DAs are easily generalizable, benefit from both linear and nonlinear correlation structure in the data, and contain accessible, interpretable, internal nodes [2]. Oftentimes the hidden layer is a "bottle-neck", a much smaller size than the input layer, in order to force the autoencoder to learn the most important patterns in the data [37].

We used the Theano library [38,39] to construct a DA consisting of three layers, an input layer x, a single hidden layer y, and a reconstructed layer z [28] (Fig. 1A). Noise was added to the input layer through a stochastic corruption process, which masks 20% of the input values, selected at random, to zero.

The hidden layer y was calculated by multiplying the input layer by a weight vector W, adding a bias vector b and computing the sigmoid (Formula 1). The reconstructed layer z was similarly computed using tied weights, the transpose of W and b (Formula 2). The cost function is the cross-entropy of the reconstruction, a measure of distance between the reconstructed layer and the input layer (Formula 3).

$$y = s(Wx + b) \qquad \text{(Formula1)}$$

$$z = s(W'y + b') \qquad \text{(Formula2)}$$

$$cost = -\sum_{k=1}^{d}[x_k \log(z_k) + (1 - x_k)\log(1 - z_k)] \qquad \text{(Formula3)}$$

Stochastic gradient descent was performed for 1000 training epochs, at a learning rate of 0.1. Hidden layers of two, four, eight and sixteen hidden nodes were included in the parameter sweep with a 20% input corruption level. Vincent et al. [28] provide a through explanation of training for DAs without missing data.

In the event of missing data, the cost calculation was modified to exclude missing data from contributing to the reconstruction cost. A missingness vector m was created for each input vector, with a value of 1 where the data is present and 0 when the data is missing. Both the input sample x and reconstruction z were multiplied by m and the cross entropy error was divided by the sum of