CSE8803 Project Proposal: Unsupervised Learning for Computational Phenotyping

Chris Hodapp (chodapp3@gatech.edu)

I. Introduction

With large volumes of health care data comes the research area of *computational phenotyping*[1], making use of techniques such as machine learning to describe illnesses and other clinical concepts from the data itself. Lasko *et al.*[5] notes the "traditional" approach of using supervised learning over specific information that a domain expert must supply, and notes two main limitations: requiring skilled humans to supply correct labels limits its scalability and accuracy, and relying on existing clinical descriptions limits the sorts of patterns that can be found (and, for instance, may fail to acknowledge that a disease treated as a single condition may really have several subtypes with different phenotypes; [5] cites asthma and heart disease as examples of this.)

Some recent papers ([6], [5], [3]) cite successes with approaches that instead use unsupervised learning on time-series data. This shows potential for finding patterns in Electronic Health Records that would otherwise be hidden and that can lead to greater understanding of conditions and treatments.

In [5], such an approach applied to serum uric acid measurements was able to distinguish gout and acute leukemia with no prior classifications given in training. Marlin *et al.*[6] examines 13 physiological measures from a pediatric ICU (such as pulse oximetric saturation, heart rate, and respiratory rate). Che *et al.*[1] likewise uses ICU data, but focuses on certain ICD-9 codes rather than mortality.

This approach still has technical barriers. Time-series in healthcare data frequently are noisy, spare, heterogeneous, or irregularly sampled[2], and commonly Gaussian processes are employed here in order to condition the data into a more useful form. In [5], Gaussian process regression is used as a transformation step prior to a series of autoencoders used for deep learning. While [1] notes that "shallow" clustering models such as Gaussian mixture models (as in [6]) are used too, deep learning succeeds more here at finding latent factors and extracting concepts. Rather than transforming input data with Gaussian process regression, this work uses an unsupervised stacked denoising autoencoder (SDAE).

II. Goals & Deliverables

My goal in this project is to apply similar methology as in [5], [1], but apply it more generally to various laboratory measurements (i.e. `LABEVENTS` table) available in MIMIC-III[4]. As [6] handles ICU data, its treatment of data may be more relevant to MIMIC-III, so I may make use of techniques from there as well.

From these laboratory measurements and that methodology, I aim to produce a transformation to a reduced feature space using a combination of Gaussian process regression to condition irregularly-sampled input data and autoencoders to learn a reduced-dimension form of them. Using labels such as certain ICD-9 codes, I will train classifiers from the reduced form, and compute their AUCs for these codes. I will also examine them for apparent clusterings that may suggest multiple different phenotypes for some label, as seen in [5, figure 5].

The deliverables of this should include source code of the above, some form of trained models that may be applicable outside of MIMIC-III, and visualizations (perhaps with t-SNE) of the reduced feature space representation.

III. Tools & Infrastructure

I am aiming to implement this work in Apache Spark or Hadoop, perhaps running on clusters with Amazon EC2 and AWS. To implement the autoencoders, I will probably make use of some deep learning framework such as TensorFlow or caffe via CaffeOnSpark. If I use Gaussian process regression, I may make use of `GaussianProcessRegressor` in the scikit-learn library, or the SparkGP library.

Acknowledgment

Thank you to Dr. Jimeng Sun and the TAs for their time and effort in this course.

References

[1] Z. Che, D. Kale, W. Li, M. Taha Bahadori, and Y. Liu. Deep Computational Phenotyping. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516, 2015.

[2] M. Ghassemi, T. Naumann, T. Brennan, D. a. Clifton, and P. Szolovits. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse , Heterogeneous Clinical Data. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453, 2015.

[3] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. Clifton, and G. D. Clifford. Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, 104(2):444–466, 2016.

[4] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[5] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE*, 8(6), 2013.

[6] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models.