Name:	Mandev S. Gill
Affiliation:	Department of Statistics, Columbia University
Email:	msg2191@columbia.edu
Presentation type (select one):	Poster

Sparse Coding for Predictive Modeling of Observational Health Outcomes

Mandev S. Gill, PhD[1], Patrick B. Ryan, PhD[2,3], David Madigan, PhD[1,3]

[1]Department of Statistics, Columbia University, New York, NY, USA; [2]Janssen Research and Development, Titusville, NJ, USA; [3]Observational Health Data Sciences and Informatics, New York, NY, USA

Abstract

The Observational Health Data Sciences and Informatics (OHDSI) network aims to improve healthcare by leveraging large-scale observational health databases with detailed patient histories. A central goal is to improve medical decision-making, which requires accurate predictive models. The nature of massive longitudinal observational health data presents unique challenges, and it is crucial to understand the impact of data representation on prediction. To this end, we implement a sparse coding representation of medical records and interface it with OHDSI software tools for predictive modeling. We empirically evaluate the performance of traditional predictive models with and without sparse coding and demonstrate the importance of data representation as a step in building predictive models.

Introduction

The emergence of massive-scale patient-level databases of electronic health records and administrative claims provides promising new opportunites to test clinical hypotheses, inform clinical decision-making, and transformatively improve healthcare. To help realize this potential, the Observational Health Data Sciences and Informatics (OHDSI) network has been formed. OHDSI is a multi-stakeholder, interdisciplinary, international collaborative of observational health researchers that develops open-source software tools and provides expertise at multiple levels to ensure that the tools satisfy clinical research needs. The OHDSI network has grown to encompass over 50 databases totalling more than 600 million patient records[1].

A central goal of OHDSI is to improve patient-level predictive modeling. Building upon popular extant statistical models can be fruitful, but this line of attack typically assumes that the data are in "regression format" and neglects the challenges presented by the representation of the data. Clinical databases contain highly detailed, longitudinal patient histories, including information about medications, conditions diagnosed, and procedures performed. However, these databases are typically sparse and voluminous, and observational patient data are quite irregular, comprising events that occur at particular moments in time (e.g., procedures or certain acute conditions), events that occur over periods of time (e.g., in-patient visits), measurements at specific times (e.g., cholesterol), and non-temporal data (e.g., ethnicity, gender, and date of birth). For temporal data such as prescription drug records, for example, binary predictors provide a natural baseline representation of the data. For each patient, we posit a matrix with a row for each available drug and a column for each day in the patient history. The (i,j)-th entry of the matrix will be 1 if the patient consumed drug i on day j, and 0 otherwise. However, it is possible to construct more tailored, optimized representations of patient data. Here, we explore the role of patient data representation in the performance of predictive models.

As a motivating prediction problem, we consider a retrospective new-user cohort of patients who are prescribed celecoxib. The outcomes of interest are acute renal failure, myocardial infarction, gastrointestinal hemorrhage, and angiodema.

Sparse Coding

We propose an alternative representation of medical records via sparse coding[2,3]. We represent a medical record as a sparse linear combination of a data-derived "dictionary" or basis. Specifically, given a collection of data vectors $\{x_1, \ldots, x_n\}$, sparse coding algorithms learn a set of basis vectors $\{b_1, \ldots, b_m\}$ and coefficient vectors $a^i = (a_1^i, \ldots, a_m^i)$ such that each x_i

can be represented as $x_i = a_1^i b_1 + \ldots + a_m^i b_m$. L_1 regularization is employed to ensure that coefficient vectors are extremely sparse with most elements equal to zero. Sparse coding in essence reveals patterns and structure inherent in the data. The basis vectors can be thought of as exemplar patterns of healthcare usage, and real individual patients are then represented as sparse linear combinations of the exemplars.

We integrate this sparse coding representation with the OHDSI network patient-level prediction software tools[4]. To accommodate massive data sets, we utilize efficient online optimization algorithms for dictionary learning[5,6]. We empirically evaluate the performance of widely-used predictive models, such as high-dimensional logistic regression, under different representations of patient data, including standard "regression format" and various sparse coding representations. To compare the performance of the different approaches, we examine dicriminatory power as quantified by the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. We also examine the Brier score (mean-squared-error) and calibration.

Conclusion

Appropriate representation of medical records is an important step in building accurate predictive models. Sparse coding provides an elegant way of summarizing patient data yet preserving individual patient characteristics, and we examine its utility in predicting health outcomes at the individual patient level. We hope that our work sets the stage for further exploration of the affects of combining predictive models with different methods of representing patient data, such as phenotyping.

References

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Noren GN, Li Y, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015;216:574-578.
2. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381:607-609.
3. Raina R, Battle A, Lee H, Ng AY. Self-taught learning: transfer learning from un-labeled data. ICML. 2007;749-756.
4. Schuemie MJ, Suchard MA, Ryan PB, Reps J. PatientLevelPrediction: package for patient level prediction using data in the OMOP Commond Data Model. R package version 1.1.0. 2015.
5. Mairal J, Bach F, Ponce J, Sapiro G. Online dictionary learning for sparse coding. ICML. 2009;689-696.
6. Mairal J, Bach F, Ponce J, Sapiro G. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research. 2010;11:19-60.