Patient representation learning from EHR data

Guillaume Desmarais-de Grandpré
Department of Computer Science and Software Engineering,
Université Laval, Québec, Canada
guded1@ulaval.ca

Keywords: Representation learning, Electronic Health Records, Natural Language Processing, Transformers, Deep Learning

1. Introduction

This research is part of a multi-disciplinary project that aims to define polypharmacy usage (the use of a combination of drugs) as well as detect and predict dangerous cases. The particular goal of the research presented herein is to learn patient representations from a massive number of Electronic Health Records (EHR) with the intent to discover socio-demographic biases in the outcomes of polypharmacy usage.

1.1. Representation learning

Electronic Health Records are, by nature, high dimensional and sparse. They contain temporal, sequential data about a patient's medical visits. Each medical visit is comprised of one or many medical codes representing either a medical procedure or a medication. With the excessively large number of medical codes available, every patient's data contains only a small subset of all possible values. This means that each patient is represented by a large vector containing mostly empty values. The features of this data make it challenging to use it in its raw form in machine learning tasks [1]. With the intent to compare patients with each other to form clusters of records with similar outcomes based on socio-demographic features, a dense vector representation must be learned. This representation can then be used in vector similarity measurements such as Euclidian distance, which calculates the distance between two vectors **p** and **q**:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)} \tag{1.1}$$

Vector distances such as this one cannot be used on vectors that contain mostly zeros; the resulting distance between any two vectors would always be close to zero.

1.2. Natural language processing and patient representation learning

The concept of learning a dense representation of a high dimensional sparse dataset is one that has been explored in the field of Natural Language Processing (NLP), where algorithms like Word2Vec are used to transform one-hot-vectors into dense representations of words such that two words used in similar contexts have similar representations. Clinical data can be linked to textual data because they both contain sequential information with semantic relationships among data points. For example, a parallel can be drawn between the two types of data by treating a medical visit as a sentence and the medical codes contained within as words. Because of this parallel, NLP techniques have proved to be successful in the clinical field as well [2]. In 2018, Zhang et al. used the Word2Vec approach in Patient2Vec [3] to learn a personalized representation for each patient to use it in future hospitalization prediction. In a recent research, Steinberg et al. [4] suggest using word embeddings to

represent medical codes and train their model to predict the medical codes contained in a medical visit using the codes contained in previous visits.

2. Planned research

2.1. Data generation

Due to the highly sensitive nature of the data involved, access to it is limited. The first part of the research consists in building a dataset that closely resembles the real data in order to train models on it. Such data should include detectable patterns of outcomes, and be high dimensional enough to evaluate the ability of the trained models to compact it into dense representations. The data generation process will be done in collaboration with a team of researchers in pharmaceutical sciences. This team has had access to the data and will help understand its important features. In particular, the different distributions of outcomes and medical codes will be extracted from the actual data in order to generate a simulated dataset with the correct distributions encapsulated.

2.2. Learning patient representations

Most research [2] based on patient representation learning using EHR uses Recurrent Neural Networks (RNN) with Gated Reccurent Units (GRU) or bi-directional Long Short Term Memory (bi-LSTM) networks. Given that NLP techniques have shown good results in the particular task of patient representation learning, the plan of this research is to explore several NLP models in order to apply them to the available data. Most notably, the state-of-the-art in NLP is to use transformers [5] which have been introduced with great success both in terms of performance and reduction of computational time. Transformers are based on an encoder-decoder architecture, and use self-attention to attribute similarity scores between words (in our case, medical codes). Such use of transformers is presented by Song et al. in 2018 and uses "positional encoding and dense interpolation strategies for incorporating temporal order" [6] in an architecture based solely on self-attention mechanisms. It is therefore expected that designing a transformer-based architecture should lead to interesting results. In addition to the use of transformers, socio-demographic data will have to be added to the representations in order to detect biases in outcomes based on these features.

There are two significant types of patient representation learning models found in literature. *End-to-end* models, like Patient2Vec [3], are trained on the specific task of predicting a defined medical outcome. Non *end-to-end* models, such as CLMBR [4], are models that first learn patient representations (in this case, by constructing a clinical language model) independently of any prediction task. The learned representations are then tested by using them as inputs in several clinical prediction models such as logistic regressions or gradient boosted trees. The apparent advantage of non *end-to-end* models is that they can capture semantic relationships between medical concepts and produce representations that can be used in all sorts of clinical tasks, instead of having to train a new model for each new task. The intuition for this research will therefore be to begin by focusing on building a non *end-to-end* model to learn representations, and test these representations on the prediction of clinical outcomes present in the dataset. This first model can then be used as a baseline for the comparison of more complex models.

This research is still in its embryonic state, but given the effectiveness of NLP methods in patient representation learning and the performance of transformers in NLP, we expect interesting results with the model in development.

Acknowledgements

This research is funded by a CIHR/NSERC collaborative grant and supported by the INSPQ. I am grateful to my supervisor Richard Khoury and co-supervisor Audrey Durand.

References

[1] J. Wu, J. Roy, and W. F. Stewart. "Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches". In: *Medical Care* 48.6 (2010), S106–S113. ISSN: 00257079. URL: http://www.jstor.org/stable/20720782.

[2] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, and K. Roberts. "Deep Representation Learning of Patient Data from Electronic Health Records (EHR): A Systematic Review". In: *arXiv preprint arXiv:2010.02809* (2020).

[3] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes. "Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record". In: *IEEE Access* 6 (2018), 65333–65346. ISSN: 2169-3536. DOI: 10.1109/access.2018.2875677. URL: http://dx.doi.org/10.1109/ACCESS.2018.2875677.

[4] E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, and N. H. Shah. *Language Models Are An Effective Patient Representation Learning Technique For Electronic Health Record Data*. 2020. arXiv: 2001.05295 [cs.CL].

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[6] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias. "Attend and diagnose: Clinical time series analysis using attention models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.