Attention based deep neural networks in patient level prediction

Egill Fridgeirsson, David Sontag, Peter Rijnbeek

Background

Patient-Level prediction (PLP) models are being used more in recent years for research and to assist for clinical decision support. The Observational Health Data Science and Informatics (OHDSI) community has developed a framework for developing and validating predictions models on observational data standardized to the OMOP Common Data Model (CDM) [1]. To date, although the framework supports deep learning models, the focus of research using the PLP framework has mostly been on traditional machine learning models. In the meantime, there have been rapid advances in the deep learning field which might work well on the type of observational data that is in the OMOP CDM.

A relevant advance in the field of deep learning is that of attention based models [2]. Attention is a mechanism where the relations between the input features of a sequence are learnt. These relations are then used to build representations which are used for the task of interest. These types of models have improved the state-of-the art in diverse fields such as natural language processing and computer vision. Recently there has been work done in translating these models to electronic health record data [3,4]. While it has been hard to improve upon a strong linear baseline on electronic health record (EHR) data, one approach in particular shows promising result where a model learns, using reverse distillation, from a strong linear baseline and then subsequently outperforms it.

Methods

We implemented a recent attention-based model and integrated it into the patient level prediction framework. The model is the Self-Attention with Reverse Distillation (SARD) model [4] and the we used a windowed L1 regularized linear regression model (LASSO) as baseline. We predicted mortality in the following year after a first occurrence of a general practitioners visit after the patient reached the age of 60. We used condition, drug and procedure codes as features from 3 years before the index date. The data used was the IPCI[1] database from the Netherlands. We split the data into 60% training set, 20% validation and 20% test set with a stratified split. Loss on validation set was used for early stopping and to select the best hyperparameters. Results are reported for the test set. For LASSO a full grid search was used to select hyperparameters. C is the parameter controlling the sparsity, we searched over a grid of C from 0.0001 to 10.000 spaced evenly on a log scale with 20 values. For SARD a randomized search of hyperparameters was used with 100 samples, then with the resulting model a full grid search was done over the values of alpha=[0, 0.05, 0.1, 0.15, 0.2]. Alpha is the parameter that controls the influence of distillation loss vs cross entropy loss during fine-tuning of the SARD model. More tasks and models will be added before the OHDSI symposium.

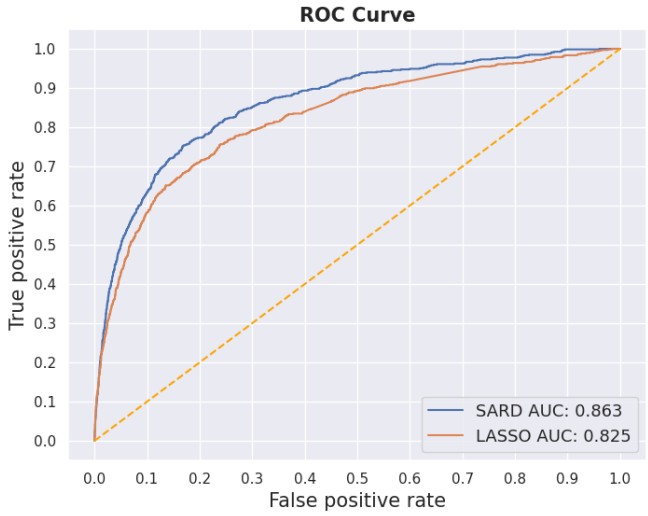
Results

For our task, the population at risk includes 150.277 patients with 3329 outcomes. The hyperparameters selected for the linear model are a C of 0.01 and windows of [30, 90, 180, 365 and 1095] before index. For SARD the selected hyperparameters were 2 attention heads with embedding dimension of 32 per head. The number of attention layers were 6 with no dropout and alpha was 0.15. For SARD all concepts in visits

[1] www.ipci.nl

over 365 days before index were considered to be a part of the same visit.

The linear baseline performs well with an area under the curve (AUC) of 82.5% while SARD reached 86.3% AUC (Fig 1). These results are in line with those in [4].



Conclusion

Attention based deep learning models are promising although a simple linear baseline is still competitive. When fully integrated into the PLP framework these kinds of models can be used in a fast and straightforward way on various OMOP CDM databases. Future work will explore other classes of attention-based models, look at their performance with external validation as well as explore the interpretability of the attention weights.

References/Citations

[1] Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Informatics Assoc 2018;25:969–75. https://doi.org/10.1093/jamia/ocy032.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv. Neural Inf. Process. Syst., vol. 2017- Decem, 2017, p. 5999–6009. https://doi.org/10.5555/3295222.3295349.

[3] Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. Adv Neural Inf Process Syst 2016:3512–20.

[4] Kodialam RS, Boiarsky R, Lim J, Dixit N, Sai A, Sontag D. Deep Contextual Clinical Prediction with Reverse Distillation. Proc AAAI Conf Artif Intell 2020;35:249–58.