Review for "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records"

Farzan Farnia, Haitong Li, Ruishan Liu, Wanyi Qian, Fei Xia

I. Overview

Conventionally, what has been impeding the clinic predictive decision systems using electronic health record (EHRs) is the high cost and low efficiency of extracting features manually out of the high-dimensional and noisy EHR data. The authors' work was motivated by the observation that traditional feature selection and data representations are unreliable and hard to generalize due to the need of supervised feature space exploration/definition. For the medial records, this can be even more challenging because there might always be "unknown" patterns and features escaping from the feature selection. The alternative data-driven methods are slightly better but still run into the data representation shortcomings given the inherent hierarchical nature of EHR data. With these observations, Miotto and other co-authors propose and demonstrate utilizing unsupervised deep learning to extract the hierarchical features and patterns from aggregated EHR data as generalized "deep patient" data representation for classification. Their "deep patient" framework essentially uses stacked denoising autoencoders (SDA) to learn the data representations from multi-domain clinical data. Finally, the proposed "deep patient" representation is evaluated based on disease prediction results, showing greatly improved results over shallow models and conventional methods due to representative and compact features learned through SDA. The following sections will cover detailed methods (architecture, dataset, processing etc.), various evaluation results of the "deep patient", and the discussions about this work and its future applications.

II. Methods

EHR data can be highly complicated and dataset selection and pre-processing can have large impact on the training and evaluation results. The work uses the Mount Sinai data warehouse (records including inpatient, outpatient and emergency room visits). The authors retain all patients with at least one diagnosed disease expressed as numerical ICD-9 between 1980 and 2014, which leads to a dataset of ~1.2 million patients, with every patient having an average of 88.9 records. 704,587 patients who had at least 5 records by December 31, 2013 are considered in the training data and

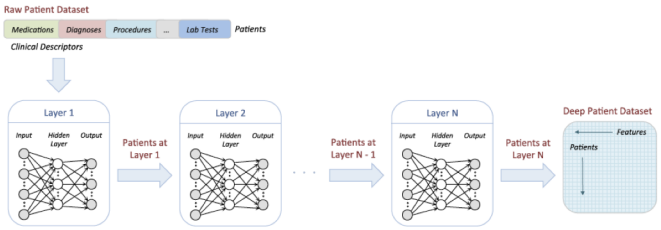

Fig. 1. Deep patient: unsupervised feature learning on EHR dataset through multi-stack denoising autoencoders. High-level representation is generated at the output of the network.

76,214 patients having at least one new ICD-9 diagnosis assigned in 2014 and at least ten records before that are considered in the testing data. Another 5,000 patients' data are used as the validation data for hyper-parameter searching. All the clinical records are pre-processed using the Open Biomedical Annotator to obtain harmonized codes for procedures and lab tests, normalized medications based on brand name and dosages, and to extract clinical concepts from the free-text notes. The parsed notes are further processed with topic modeling to reduce the sparseness of the representation and to obtain a semantic abstraction of the embedded clinical information.

As illustrated in Fig. 1, the pre-processed dataset is fed into a 3-layer SDA. All the feature values in the dataset are first normalized to reduce the variance of the data. All the autoencoders share the same set of parameters. In particular, there are 500 hidden units per layer with sigmoid activation function. Noise corruption factor is set to be 5% through tuning the model with the validation data (5,000 patients). For training, the reconstruction cross-entropy function is used as the loss function, and mini-batch stochastic gradient descent is performed. The learned encoding function is then applied to the clean inputs and to obtain the distributed representation.

III. Results

After constructing the "deep patient" model, the authors then evaluate the disease predictions in two different clinical tasks: disease classification (i.e., evaluation by disease) and patient disease tagging (i.e., evaluation by patient). For the first disease classification task, it aims to use the trained classifier to determine if test patients are likely to be diagnosed with a certain disease within a one-year interval. The authors also include PCA, GMM, K-Means, ICA, and RawFeat (original

Time Interval = 1 year (76,214 patients)			
		Classification Threshold = 0.6	
Patient Representation	AUC-ROC	Accuracy	F-Score
RawFeat	0.659	0.805	0.084
PCA	0.696	0.879	0.104
GMM	0.632	0.891	0.072
K-Means	0.672	0.887	0.093
ICA	0.695	0.882	0.101
DeepPatient	**0.773***	**0.929***	**0.181***

Time Interval	Metrics	UppBnd	Patient Representation			
			RawFeat	PCA	ICA	DeepPatient
30 days *(16,374 patients)*	Prec@1	1.000	0.319	0.343	0.345	**0.392***
	Prec@3	0.492	0.217	0.251	0.255	**0.277***
	Prec@5	0.319	0.191	0.214	0.215	**0.226***
60 days *(21,924 patients)*	Prec@1	1.000	0.329	0.349	0.353	**0.402***
	Prec@3	0.511	0.221	0.254	0.259	**0.282***
	Prec@5	0.335	0.199	0.216	0.219	**0.230***
90 days *(25,220 patients)*	Prec@1	1.000	0.332	0.353	0.360	**0.404***
	Prec@3	0.521	0.243	0.257	0.262	**0.285***
	Prec@5	0.345	0.201	0.219	0.220	**0.232***
180 days *(33,607 patients)*	Prec@1	1.000	0.331	0.361	0.363	**0.418***
	Prec@3	0.549	0.246	0.261	0.265	**0.290***
	Prec@5	0.370	0.207	0.221	0.224	**0.236***

Table. I. Left: disease classification results. Right: patient disease tagging results for diagnoses assigned during different time intervals in terms of precision-at-k, with k = 1, 3, 5. UppBnd shows the best results achievable.

descriptors) for comparison. For evaluation metrics, the AUC-ROC is the area under a curve of true positive rate versus false positive rate found over the set of predictions. Accuracy (both true positives and true negative) and F-score (harmonic mean of classification precision and recall) use a threshold of 0.6 to discriminate positive/negative predictions. Table I (left) summarizes the evaluation results. DeepPatient achieves an average AUC-ROC of 0.773, with a 15% improvement over RawFeat. Accuracy and F-score are improved by 15% and 54% respectively, showing that the quality of the positive predictions is indeed improved by pre-processing EHRs with a deep architecture. For 10 different diseases, DeepPatient outperforms all other feature learning methods. The authors also find that PCA does not lead to any improvement for several diseases.

Furthermore, the authors investigate how DeepPatient performs in patient-specific level, as summarized in Table I (right). The test measures the quality of disease annotations over different temporal windows (from 30 days to 180 days) for all the patients having true diagnoses in that period. It's clear that DeepPatient consistently outperformed other methods across all time intervals examined, which provides more support in addition to the disease-level evaluation.

IV. Discussions & Outlook

As claimed by the authors, deep learning has not been employed in large-scale EHR dataset for general-purpose feature learning until this work. Previous work, as cited by the authors, mostly focused on the domain-specific or disease-specific evaluations using machine learning approaches, while the DeepPatient can be applied to various tasks. The major contribution of this work is showing unsupervised deep learning is indeed suitable for complicated datasets like EHR to derive the underlying hierarchical features for a better representation, which was a bottleneck for previous conventional approaches. Analogous to the computer vision domain, various descriptors in a patient clinical records draw a complicated "picture" of history-dependent health condition, where certain underlying patterns that can be extremely hard to reveal or define in the sense of traditional medical data analytics. This is the niche where deep architecture fits, with the ability of producing higher and higher level of representation, especially when large-scale datasets are available. The authors mention that better raw EHR data representation together with better pre-processing may improve the learned features and prediction accuracy, which is reasonable given the complexity of EHR data. For example, in raw EHR data, there could be intrinsically "noisy" and biased records. PCA is used in baselines, but could also be used to pre-process the EHR data before SDA to remove irrelevant factors. However, this can be tricky for certain tasks since in the paper the authors show that PCA doesn't provide improvement over RawFeat for certain diseases, which maybe was part of the reason why the authors didn't perform PCA pre-processing. Besides, more work about sensitivity tests and hyper-parameter selection will be beneficial to improve the paper.

There is a wide array of applications such as personalized prescriptions/recommendation can be beneficial from this research, and the authors envision more collaboration between clinic/hospital systems to improve the patient representation. To achieve this, on the one hand, there may be other deep architectures other than the SDA used in this work more suitable for future exploration. On the other hand, since new EHR data are being collected every day, in order to update the models for improved performance, it would be more desirable to have one-shot learning capability rather than brute-force iterative training on the increasing large-scale datasets. One possibility is to explore hyperdimensional computing, a neural-inspired computational framework with holographic data representation (eg., computing with 10,000-bit-length random vectors) and capability of one-shot learning. Anyway, this paper indeed shows interesting and timely exploration for this filed and would be helpful for future studies.