Finding Algebraic Structure of Care in Time: A Deep Learning Approach

Phuoc Nguyen, Truyen Tran and Svetha Venkatesh
Centre for Pattern Recognition and Data Analytics
Deakin University, Geelong, Australia
{phuoc.nguyen,truyen.tran,svetha.venkatesh}@deakin.edu.au

Abstract

Understanding the latent processes from Electronic Medical Records could be a game changer in modern healthcare. However, the processes are complex due to the interaction between at least three dynamic components: the illness, the care and the recording practice. Existing methods are inadequate in capturing the dynamic structure of care. We propose an end-to-end model that reads medical record and predicts future risk. The model adopts the algebraic view in that discrete medical objects are embedded into continuous vectors lying in the same space. The bag of disease and comorbidities recorded at each hospital visit are modeled as function of sets. The same holds for the bag of treatments. The interaction between diseases and treatments at a visit is modeled as the residual of the diseases minus the treatments. Finally, the health trajectory, which is a sequence of visits, is modeled using a recurrent neural network. We report preliminary results on chronic diseases – diabetes and mental health – for predicting unplanned readmission.

1 Introduction

An ultimate goal of AI in healthcare is to reason about the past (historical illness), present (diagnosis) and future (prognosis). The learning path through recorded medical data necessitates the modeling of the dynamic interaction between the three processes: the illness, the care and the data recording [15]. For this paper, we assume that a hospital visit at a time manifests through a set of discrete diseases and treatments. A healthcare trajectory is therefore a sequence of time-stamped records.

Here we adopt the notion of reasoning as "algebraically manipulating previously acquired knowledge in order to answer a new question" [1]. For that we learn to embed discrete medical objects into continuous vectors, which lend themselves to a wide host of powerful algebraic and statistical operators [3, 17]. For example, with diseases represented as vectors, computing disease-disease correlation is simply a cosine similarity between the two corresponding vectors. Illness – recorded as a bag of discrete diseases – can then be a function of set of vectors. The same holds for care. Importantly, if diseases, treatments (or even doctors) are embedded in the same space then recommendation of treatments (or doctors) for a given disease will be as simple as finding the nearest vectors.

The algebraic view makes it easily to adapt powerful tools from the recent deep learning advances [12] for healthcare. In particular, we can build end-to-end models for risk prediction without manual feature engineering [2, 14, 15, 16]. As the models are fully differentiable, credit assignment to distant risk factors can be carried out [14], making the models more transparent than commonly thought.

One important aspect still remains, however, that is the dynamic interaction between illness and care. Although care is supposed to lessen the illness, it is often designed through highly controlled trials where one treatment is targeted at one disease, on a specific cohort, at a specific time. Much less is known for the effect of multiple treatments on multiple diseases, in general hospitalized patients, *over time*. A recent model known as DeepCare [15] partly addresses this problem by considering the moderation effect of treatments on illness state transition *between visits*, but not multi-disease–multi-treatment interaction *within visits*.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

This paper reports preliminary results on an investigation into finding the algebraic structure of care hidden in the Electronic Medical Records. The task is to predict future risk such as unplanned readmission or death at discharge. We focus on chronic diseases (diabetes and mental health) as they are highly complex with multiple causes, often associated with multiple cormobidities, and the treatments are not always effective.

2 Methods

We consider the problem of modeling Electronic Medical Records (EMR) for predicting future risk at the time of hospital discharge. Each medical record is a sequence of hospital visits by a patient. For simplicity, we consider a visit as consisting of diseases and treatments. While we might expect that the disease subset together with the treatment subset reflect the illness state at the time of discharge, it is not necessarily the case. This is because of several reasons. First, the coding of those diseases and treatments is often optimized for billing purposes, not all diseases are included. Second, errors do occur sometimes. And third, the treatments usually take time to get the full intended effect.
For this reason, we include historical visits to assess the current state as well as to predict future risk. An efficient way is to model the visit sequences as a Recurrent Neural Network (RNN) [6]. In this paper, we choose Long Short-Term Memory (LSTM) since it can remember distant events [9].

2.1 Visit Representation

A visit consists of two variable-size bags of discrete elements: a bag of diseases and a bag of treatments. Among older cohorts, non-singleton bags are prevalent, reflecting the comorbidity picture of modern healthcare. As a result, treatments must be carefully administered to work with, or at least not to cancel out, each other. This also calls for a sensible way to model the complexity of *multi-disease–multi-treatment interaction*. Most existing biostatistics methods, however, are designed for simplified treatment effect against just one condition.
We use vector representation of diseases and treatments, following the recent practice in NLP (e.g., see [5]). Let $\boldsymbol{e}_d$ be the vector representation of disease d, $\boldsymbol{e}_p$ the representation of the treatment p, and the vectors are embedded in a common space. The bags of diseases and bags of treatments are also represented as vectors in the same space. The representation of a bag is computed using a differentiable set function $f_e(S)$ that receives a bag of vectors S and returns another vector of the same dimensions. We use the following set function:

$$f_e(S) \leftarrow \frac{\bar{\boldsymbol{e}}_S}{\epsilon + \|\bar{\boldsymbol{e}}_S\|} \quad \text{where} \quad \bar{\boldsymbol{e}}_S = \max\left(\mathbf{0}, \sum_{i \in S} \boldsymbol{e}_i\right) \tag{1}$$

where $\epsilon > 0$ is a smoothing factor. This is essentially a linear rectifier [13] of the sum, approximately normalized to unit vector. The factor ϵ lets $\|f_e(S)\| \to 0$ when$\|\bar{\boldsymbol{e}}_S\| \to 0$, but $\|f_e(S)\| \to 1$ when$\|\bar{\boldsymbol{e}}_S\| \gg 0$.
Denote by D_t the bag of diseases and I_t the bag of treatments recorded for the visit at time t. Let $\boldsymbol{d}_t = f_e(D_t)$ be the representation of the disease bag, and $\boldsymbol{p}_t = f_e(I_t)$ the representation of the treatment bag. We compute the visit vector as:

$$\boldsymbol{v}_t = \rho(\Delta) \quad \text{where} \quad \Delta = \boldsymbol{d}_t - \boldsymbol{p}_t \tag{2}$$

where ρ is an element-wise transformation. The difference Δ reflects the intuition that treatments are supposed to lessen the illness. We found $\rho(\Delta) = (1 + \Delta)^2$ works well, suggesting that the disease-treatment interaction is highly nonlinear, and thus warranting a deeper investigation.

2.2 Illness Memory with LSTM

Given a sequence of input vectors, one per visit, the LSTM reads an input $\boldsymbol{v}_t$ at a time and estimates the illness state $\boldsymbol{h}_t$. To connect to the past, LSTM maintains an internal short-term memory $\boldsymbol{c}_t$, which is updated after seeing the input. Let $\tilde{\boldsymbol{c}}_t$ be the new candidate memory update after seeing $\boldsymbol{v}_t$, the memory is updated over time as $\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{c}}_t$, where $\boldsymbol{f}_t \in (\mathbf{0}, \mathbf{1})$ is forget gate determining how much of past memory to keep; $\boldsymbol{i}_t \in (\mathbf{0}, \mathbf{1})$ is the input gate controlling the amount of new information to add into the present memory. The input gate is particularly useful when some recorded information is irrelevant to the final prediction tasks.
The memory gives rise to the state as $\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{c}_t)$, where $\boldsymbol{o}_t \in (\mathbf{0}, \mathbf{1})$ is the output gate, determining how much external information can be extracted from the internal memory. The candidate memory and the three gates are parameterized functions of $(\boldsymbol{v}_t, \boldsymbol{h}_{t-1})$.

(a) Worsening progression ($P = 0.70$) (b) Improving progression ($P = 0.23$)

Figure 1: Illness state progression over time, measured as $\boldsymbol{h}_t$ for the last 10 visits. **Left figure:** a high-risk case with 70% chance of readmission at time T=10. **Right figure**: a low-risk case with 23% chance at the end of the sequence. Best viewed in color.

With this long short-term memory system in place, information of the far past is not entirely forgotten, and credit can be assigned to it. Second, partially recorded information can be integrated to offer a better picture of current illness.

2.3 Risk Prediction

Once the LSTM is specified, its states are pooled for risk prediction at each discharge, i.e., $\bar{\boldsymbol{h}}_t = \text{pool}(\boldsymbol{h}_{1:t})$. The pooling function can be as simple as the `mean()` (i.e., $\bar{\boldsymbol{h}}_t = \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j$) or `last()` (i.e., $\bar{\boldsymbol{h}}_t = \boldsymbol{h}_t$). We also tried exponential smoothing, $\bar{\boldsymbol{h}}_t = \alpha\bar{\boldsymbol{h}}_{t-1} + (1-\alpha)\boldsymbol{h}_t$, for $\bar{\boldsymbol{h}}_1 = \boldsymbol{h}_1$ and $\alpha \in [0, 1]$. A small α would mean the recent visits have more influence in future risk. Finally, a differentiable classifier (e.g., a feedforward neural net) is placed on top of the pooled state to classify the medical records (e.g., those in population stratification) or to predict the future risk (e.g., unplanned readmissions). The loss function is typically the negative log-likelihood of outcome given the historical observations, e.g., $-\sum_t \log P\left(y_t \mid \boldsymbol{D}_{1:t}, \boldsymbol{I}_{1:t}\right)$. We emphasize here is the system is end-to-end differentiable, starting from the disease and treatment lookup table at the bottom to the final classifier at the top. No feature engineering is needed.

2.4 Regularizing state transitions

For chronic diseases, it might be beneficial to regularize the state transition. We consider adding the regularizers $\frac{\beta}{T}\sum_{t=2}^{T}\left(\|\boldsymbol{h}_t\|_2 - \|\boldsymbol{h}_{t-1}\|_2\right)^2$ to the loss function as suggested in [11]. This asks the amount of information available at each time step, encapsulated in the norm $\|\boldsymbol{h}_t\|_2$, to be stable over time. This is less aggressive than maintaining state coherence, i.e., by minimizing $\frac{\beta}{T}\sum_{t=2}^{T}\|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\|_2^2$.

3 Results

Cohorts Data is previously studied in [15], which consists of two chronic cohorts: diabetes and mental health. There are over 7000 diabetes patients with a median age of 73 making over 53,000 visits. For mental health cohort, the figures are 6,100, 37 and 52,000 respectively. The number of disease and treatment codes for each cohort are around 240 and 1100 respectively. Diseases and treatments are coded using the ICD-10 coding scheme. For diseases, the first two-level in the ICD-10 tree is used. The data was collected between 2002-2013 from a large regional Australian hospital. Each record contains at least 2 hospital visits.

Implementation Models are implemented in Julia using the Knet.jl package [18]. Optimizer is Adam [10] with learning rate of 0.01 and other default parameters. Two baselines are implemented. One is bag-of-words trained using regularized logistic regression (BoW+LR), where diseases and treatments are considered as words, and the medical history as document. No temporal information is modeled. Although this is a simplistic treatment, prior research has indicated that BoW works surprisingly well [14, 15]. The other is a recent model known as Deepr [14], which is based on convolutional net for sequence classification. Unlike the BoW, which are unordered, in Deepr words are sequenced by their temporal order. Words of the same visit are randomly sequenced. However, the Deepr does not model the temporal transition between illness states.

Visualization The progression of illness states and probability of readmission over time is visualized in Fig. 1 for two typical patients. The high-risk case is shown in Fig. 1(a) – it seems that the illness gets worse over time. In contrast, the low-risk case is depicted in Fig. 1(b), where the illness is rather stable over time.

Prediction accuracy Table 1 reports the Area Under the ROC Curve (AUC) for all methods in predicting unplanned readmission. The proposed methods shows a competitive performance against the baselines. The Multi-Disease–Multi-Treatment method shows better prediction rate in the mental health data while the Multi-Disease–Multi-Treatment with Progression method information seems better in the diabetes data. It suggests that a proper modeling of care over time is needed, not only for understanding the underlying processes, but also to achieve a competitive predictive performance.

Method	*Diabetes*	*Mental health*
BoW+LR	0.673	0.705
Deepr [14]	0.680	0.714
MDMTP+LTSM	**0.718**	**0.726**
MDMT+LSTM	**0.701**	**0.730**

Table 1: Area Under the ROC Curve averaged over 5 folds in predicting unplanned readmission. BoW = bag-of-words, LR = logistic regression, MDMT = Multi-Disease–Multi-Treatment, MDMTP = Multi-Disease–Multi-Treatment with Progression information.

4 Related Work

The past few years have witnessed an intense interest in applying recent deep learning advances to healthcare [16]. The most ready area is perhaps medical imaging [8]. Thanks to the record-breaking successes in convolutional nets in computer vision, we now can achieve diagnosis accuracy comparable with experts in certain sub-areas such as skin-cancer [7]. However, it is largely open to see if deep learning succeeds in other areas where data are less well-structured and of lower quality such as electronic medical records (EMR) [15].

Within EMRs, three set of techniques have been employed. The first is finding distributed representation of medical objects such as diseases, treatments and visits [3, 17]. The techniques are not strictly deep but they offer a compelling algebraic view of healthcare. The second group of techniques involve 1D convolutional nets, which are designed for detecting short translation invariant motifs over time [2, 14]. The third group, to which this paper belongs, employs recurrent neural nets to capture the temporal structure of care [4, 15].

Predictive healthcare begs the question of modeling treatment effects *over time*. This has traditionally been in the realm of randomized controlled trials. Our work here is, on the other hand, based entirely on observational administrative data stored in Electronic Medical Records.

5 Discussion

The continuous representation of diseases make it easy to study the disease space, that is, which diseases are related and may be interacting. The same holds for the treatments. The interaction between diseases and treatments can be modelled easily.

5.1 Conclusion

In summary, we have introduced a temporal model of care, where the emphasis is to build a continuous representation of discrete medical entities such as diseases, treatments and hospital visits. Representing disease comorbidity is via simple algebraic manipulation of disease vectors. The same holds for the care package, which is a bag of treatments targeted at multiple diseases present in the patient. Multi-disease–multi-treatment interaction is a function of difference between the comorbidity vector and the care package vector. A healthcare trajectory is then modelled using a recurrent neural network known as Long Short-Term Memory, which is capable of memorizing distant events. Importantly, the entire system is *end-to-end*: the model reads the Electronic Medical Record and predicts future risks *without* any manual feature engineering. Initial results on two chronic cohorts, the diabetes and the metal health, demonstrate the usefulness of the model.

Future work will include investigation into (a) interaction models between diseases and treatments, (b) disease progression, where we conjecture that simple linear albegra operators such as matrix multiplications can be used.

Acknowledgments The paper is partly supported by the Telstra-Deakin CoE in Big Data and Machine Learning.

References

[1] Léon Bottou. From machine learning to machine reasoning. *Machine Learning*, 94(2):133–149, 2014.

[2] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records: A deep learning approach. In *SIAM International Conference on Data Mining (SDM 2016)*. SIAM, 2016.

[3] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. Multi-layer representation learning for medical concepts. *KDD*, 2016.

[4] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[6] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[8] Hayit Greenspan, Bram van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] David Krueger and Roland Memisevic. Regularizing RNNs by Stabilizing Activations. *arXiv preprint arXiv:1511.08400*, 2015.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[13] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. 27th International Conference on Machine Learning*, 2010.

[14] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A Convolutional Net for Medical Records. *Journal of Biomedical and Health Informatics*, 21(1), 2017.

[15] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69:218–229, May 2017.

[16] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[17] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of biomedical informatics*, 54:96–105, 2015.

[18] Deniz Yuret. Knet: beginning deep learning with 100 lines of julia. In *Machine Learning Systems Workshop at NIPS 2016*, 2016.