Predicting Treatment Initiation from Clinical Time Series Data via Graph-Augmented Time-Sensitive Model

Fan Zhang [1] **Tong Wu** [2] **Yunlong Wang** [1] **Yong Cai** [1] **Cao Xiao** [1] **Emily Zhao** [1] **Lucas Glass** [1] **Jimeng Sun** [3]

Abstract

Many computational models were proposed to extract temporal patterns from clinical time series for each patient and among patient group for predictive healthcare. However, the common relations among patients (e.g., share the same doctor) were rarely considered. In this paper, we represent patients and clinicians relations by bipartite graphs addressing for example from whom a patient get a diagnosis. We then solve for the top eigenvectors of the graph Laplacian, and include the eigenvectors as latent representations of the similarity between patient-clinician pairs into a time-sensitive prediction model. We conducted experiments using real-world data to predict the initiation of first-line treatment for Chronic Lymphocytic Leukemia (CLL) patients. Results show that relational similarity can improve prediction over multiple baselines, for example a 5% incremental over long-short term memory baseline in terms of area under precision-recall curve.

1. Introduction

Recent years there has been an explosion in the amount of digital information growth in electronic health records, which provides great opportunities for applications such as health analytics and clinical informatics (Xiao et al., 2018; Topol, 2019). Among the electronic health data, clinical time series is one major type. Over the years, many computational models have been proposed for disease detection (Choi et al., 2016c; 2018), disease progression (Bai et al., 2018), and patient subtyping (Baytas et al., 2017; Che et al., 2017) based on clinical time series data.

[1]IQVIA Inc., Plymouth Meeting, PA, USA [2]Department of Biomedical Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA [3]Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Yunlong Wang <Yunlong.Wang@iqvia.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

A disease progression model is designed to predict the development of potential treatments for many slowly progressing diseases, e.g. Alzheimer's disease, by detecting more granular stages as compared to those defined in clinical diagnosis (Sukkar et al., 2012). Longitudinal clinical time series along with related patient or physician data is important in informing disease progression patterns, and motivate many deep learning based disease progression models including recurrent neural network (RNN), attention model, and graph embedding (Choi et al., 2016a;b; Bai et al., 2018; Che et al., 2017; Suresh et al., 2017).

Despite these initial success, the relational structure between patients and clinicians has been overlooked and is of great potentials to enhance prediction accuracy. The assumption is that patients with the same disease who visit the same clinicians tend to receive similar treatments, which can be arguably attributed to that clinicians follow a set of common medical knowledge, and more likely make similar decisions for patients with the same disease.

We propose to model the patient-clinician relational structure as a bipartite graph, in which the two disjoint sets of vertices represent the patients and the clinicians, respectively; the set of edges records the number of visits made by each patient with each clinician. We note that many patients visit different clinicians for diagnosis and follow-up treatments. Thus we create two graphs, with one for patients and their diagnosis clinicians, and another for patients and their follow-up clinicians. An example of the graphs using patients' clinical data from IQVIA's prescription and claim database is given in Figure 1. Here we show six patients diagnosed with Chronic Lymphocytic Leukemia (CLL), and their visited clinicians for both diagnosis and follow-up treatments. We then apply spectral graph analysis to solve for the top eigenvectors of the graph Laplacian, and take the eigenvectors as latent representations of the similarity between patient-clinician pairs. The extracted features capture the latent proximity of patients who visit the same clinicians, and can enhance the prediction accuracy of a variety of disease progression models.

As a key contribution, to our best knowledge, this work is the first disease progression/detection model that leverages graph theory to exploit the relational similarity of clinician-

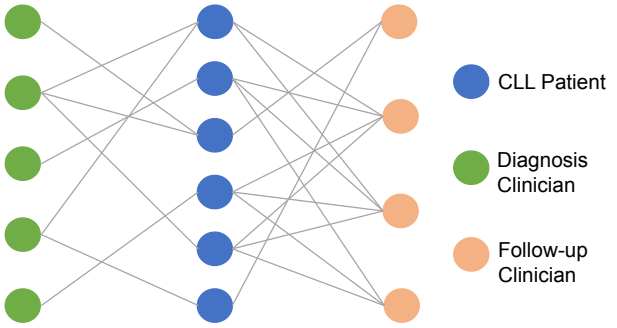

Figure 1. Illustration of the bipartite graph between CLL patients and diagnosis/follow-up clinicians. Follow-up clinicians are restricted to oncologists and hematologists. Diagnosis and follow-up clinicians can overlap.

visiting from patients' clinical time series data. Using the patients' clinical medical data from IQVIA's database, we show that the proposed similarity features can improve the performance of a wide choice of machine learning models, from XGBoost (Chen & Guestrin, 2016) to more recent deep learning based models, e.g. convolutional neural network (CNN) and long short-term memory (LSTM).

2. Related Works

Medical treatment prediction is a core research task of disease progression modeling. Recently, many deep learning models have made rapid advancements on this topic. In (Choi et al., 2016b), a two-level attention model was designed to detect influential past visits and significant clinical variables for better prediction accuracy and interpretability. In (Choi et al., 2017), a graph-based attention model was proposed to extract hierarchical information from medical oncologies and improve RNN-based rare disease prediction. In (Ma et al., 2017), a bi-directional RNN was designed to remember information of both the past and future visits based on three attention mechanism to measure the relationship of different visits for prediction. In (Che et al., 2017), a RNN architecture through dynamically matching temporal patterns was proposed to learn the similarity between two longitudinal patient record sequences for personalized prediction of Parkinson's Disease. Other similar approaches have also been proposed (Ma et al., 2018b;a).

3. Methods

3.1. Problem Formulation

We let $\mathcal{D}$ denote patients' clinical medical records. For a patient p, $\mathcal{D}^{(p)} = \{\mathbf{S}^{(p)}, \mathcal{I}^{(p)}\}$, where $\mathbf{S}^{(p)} = [\mathbf{s}_1^{(p)}, \mathbf{s}_2^{(p)}, \ldots, \mathbf{s}_T^{(p)}]$ is the sequence of visits and $\mathcal{I}^{(p)}$ includes demographics and related medical features. Each visit $\mathbf{s}_t^{(p)}$ consists of information such as diagnoses, procedures, prescriptions, visited clinicians, etc. The goal of prediction is to learn $f : \mathcal{D}^{(p)} \mapsto y^{(p)}$, where $y^{(p)}$ is the label indicating if the patient will start treatment in the next time window.

Table 1. List of notations defined in this paper.

Notation	Description
$\mathcal{D}$	Clinical medical records
$\mathbf{S}$	Patients' sequences of visits
$\mathcal{I}$	Patients' demographics
$\mathcal{U}, \mathcal{V}$	Sets of clinicians and patients.
$M = \lvert\mathcal{U}\rvert, N = \lvert\mathcal{V}\rvert$	Sizes of $\mathcal{U}$ and $\mathcal{V}$.
$\mathcal{E} = \{\{w_{ij}\}_{i=1}^{M}\}_{j=1}^{N}$	Set of edges connecting $\mathcal{U}$ and $\mathcal{V}$; w_{ij} is the count of patient j visiting clinician i.
$\mathcal{G} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}\}$	Bipartite graph for patients and their visited clinicians
$\mathbf{A}$	Adjacency matrix of $\mathcal{G}$
$\mathbf{D}$	Degree matrix of $\mathbf{A}$
$\mathbf{L}$	Laplacian matrix of $\mathcal{G}$
$\{\mathbf{X}_i\}_{i=1}^{K}$	Relational similarity features

We model the patient-clinician relation as a bipartite graph $\mathcal{G} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}\}$ over $\mathcal{D}$, where $\mathcal{U}$, $\mathcal{V}$ are the sets of clinicians and patients, $\mathcal{E}$ is the set of edges connecting $\mathcal{U}$ and $\mathcal{V}$, with each weight w_{ij} denoting the count of patient j visiting clinician i. Table 1 lists the notations we used in the paper.

3.2. Extracting Relational Similarities from Patients' Clinical Time Series via Graph Laplacian

The algorithm for extracting relational similarity is motivated by spectral clustering (Ng et al., 2002), in which data points are considered as nodes of a similarity graph and mapped to a low-dimensional space where they can be segregated to form clusters.

Firstly, we construct K bipartite graphs $\{\mathcal{G}_i\}_{i=1}^{K}$ from $\mathcal{D}$, where each $\mathcal{G}_i = \{\mathcal{U}_i, \mathcal{V}, \mathcal{E}_i\}$ and $\mathcal{E}_i = \{\{w_{ij}\}_{i=1}^{|\mathcal{U}_i|}\}_{j=1}^{|\mathcal{V}|}$. We let $M = |\mathcal{U}_i|$ and $N = |\mathcal{V}|$ for simplicity. $\mathcal{V}$ represents patients thus remains the same across all graphs.

Secondly, for each graph $\mathcal{G}_i$ we construct the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(M+N)\times(M+N)}$. Since in $\mathcal{G}_i$ we ignore interaction within patients (or clinicians), $\mathbf{A}$ is block sparse, and takes the form as $\begin{bmatrix} \mathbf{0}_{M,M} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{0}_{N,N} \end{bmatrix}$, where $\mathbf{B} \in \mathbb{R}^{M\times N}$ is the matrix representation of $\mathcal{E}_i$.

Thirdly, we compute the Laplacian matrix of $\mathcal{G}_i$ as $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix whose (i,i)-th element is the sum of the i-th row of $\mathbf{A}$.

Lastly, we compute the top k eigenvectors of $\mathbf{L}$, i.e. $\mathbf{X}_i = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k] \in \mathbb{R}^{(M+N)\times k}$, using algorithms such as

Algorithm 1 Graph-based Similarity Feature Extraction

Input: Bipartite graphs $\{G_i\}_{i=1}^K$
Output: $\{\mathbf{X}_i\}_{i=1}^K$
for $i = 1$ **to** K **do**
  $\mathcal{G}_i = \{\mathcal{U}_i, \mathcal{V}, \mathcal{E}_i\}, M = |\mathcal{U}_i|, N = |\mathcal{V}|$
  $\mathbf{A} = \begin{bmatrix} \mathbf{0}_{M,M} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{0}_{N,N} \end{bmatrix}$ (where $\mathbf{B}[i,j] = w_{ij}$)
  $\mathbf{D} = \mathrm{diag}(\{\sum_{j=1}^{M+N} \mathbf{A}[i,j]\}_{i=1}^{M+N})$
  $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$
  $\mathbf{X}_i = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k] = \mathrm{eigendecomp}(\mathbf{L})$
  for $\alpha = 1$ **to** $M + N$ **do**
    $\mathbf{X}_i[\alpha, :] = \mathbf{X}_i[\alpha, :]/(\sum_\beta \mathbf{X}_i[\alpha, \beta]^2)^{1/2}$
  end for
end for

Implicitly Restarted Arnoldi Method (Lehoucq & Sorensen, 1996) that is suitable to process large sparse matrices. $\mathbf{X}_i$ is normalized such that each row of $\mathbf{X}_i$ has an L_2-norm of 1.

For a patient p, we extract the corresponding row from $\{\mathbf{X}_i\}_{i=1}^K$ as the relational similarity features. The complete algorithm is summarized in Algorithm 1.

4. Experiments and Results

4.1. Cohort

We extract data from IQVIA longitudinal prescription (Rx) and medical claims (Dx) database, including hundreds of millions of patient clinical records. In this study, we selected all the patients diagnosed with CLL from 01-2017 to 12-2018 from the IQVIA database and kept only the patients with complete Rx/Dx information. We split the time period from 07-2017 to 12-2018 into 3 equal intervals. Within each interval, we defined the positive cohort as patients who were diagnosed with CLL before the interval and started treatment within the interval, and the negative cohort as patients who were diagnosed with CLL before the interval, but did not start treatment during the interval. The final positive and negative cohorts have 11,259 and 109,563 patient profiles, each of which contains a medical sequence and a feature vector including demographics and relational similarity features, respectively.

In time series forecasting, it is common practice to reserve the last part of each time series for testing, and use the rest of the series for training (Bergmeir & Benítez, 2012). This is to avoid using information from the future to predict past events. As a result, we cannot apply conventional train/test split of data such as k-fold cross-validation that would shuffle data randomly. Therefore, the 07-2018 to 12-2018 interval is hold out for model testing, and the rest intervals are used for model training.

4.2. Feature preparation

BOW features: For each interval, we pulled one-year patient clinical records (diagnosis, procedure, etc.) before the interval. We divided the one-year look-back into four quarters, and used Bag-of-Words (BOW) (Mikolov et al., 2013) to extract features (counts of occurrences of services) from the records within each quarter. Medical services with less than 1% occurrences in the entire dataset were ignored. Eventually, we obtained 329 BOW features.

Demographics and medical features: Demographic features include age and gender. Medical features are 11 clinical services from patients' medical history, which are chosen by clinical experts and have been shown relevance to treatment initiation. Both demographics and medical features are converted to categorical variables.

Vectorized clinical time series: The medical service codes in the patients' clinical time series were converted to 300-dimensional dense vectors through a pre-trained embedding.

Relational similarity: We constructed two bipartite graphs, one for patients and their diagnosis clinicians, and another for patients and their follow-up clinicians. We took the top 5 eigenvectors from each of the two Laplacian matrices derived from the two graphs using Algorithm 1, resulting in a 10-dimensional vector as the relational similarity feature for one patient.

4.3. Evaluation

Because of the imbalance of positive and negative samples in the cohort, we use precision-recall area-under-curve (PR-AUC) and precision@k as the metrics to evaluate model performance.

All models are built using the training data set and evaluated using the testing data set. We use Adam optimizer (Kingma & Ba, 2014) with a default learning rate of 1e-4. The number of training epochs for each model is 50 and an early stopping criterion is invoked if the performance does not improve in 10 consecutive epochs. All models are implemented in Keras with Tensorflow backend and tested on a system equipped with 128GB RAM, 16 Intel(R) Core Xeon(R) E5-2683 v4 2.10GHz CPUs, and Nvidia Tesla P100-PCIE-16GB.

4.4. Models and Performance

To demonstrate the robustness of the proposed method, we have applied it on three baseline models for comparison: XGB, CNN, and LSTM. We evaluate the performance of these models with and without the relational similarity feature. For fairly comparison, we also introduce the diagnosis clinicians and follow-up clinicians IDs to the baseline models, to guarantee that the proposed method and baseline

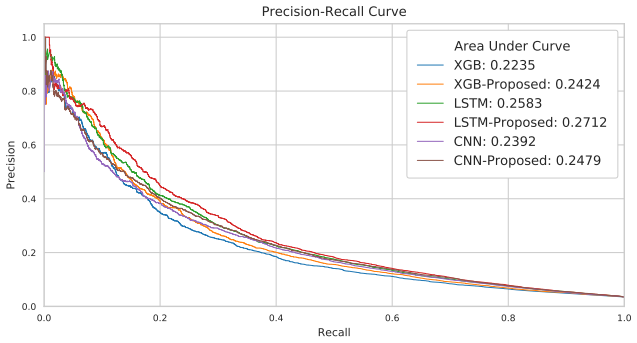

Figure 2. Comparison of baseline models with and without relational similarity feature measured in PR-AUC.

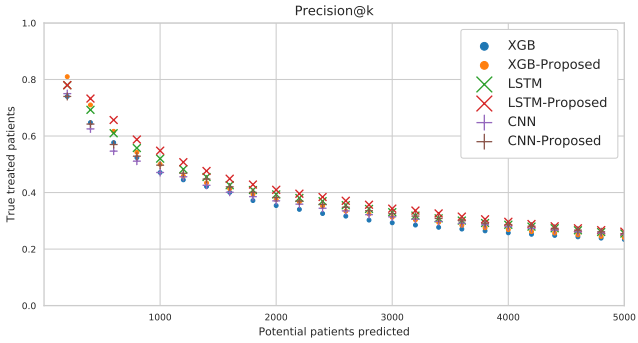

Figure 3. Comparison of baseline models with and without relational similarity feature measured in precision@k.

model leverage the same amount of information.

XGB: A tree boosting regression model implemented in XGBoost with 500 estimators. Input features include BOW features, demographics, medical features, and relational similarity features.

CNN: A 4-layer CNN with a structure of $1 \times 1(128) - 2 \times 2(128) - 3 \times 3(128) - 5 \times 5(128)$ (kernel size/number of kernels), followed by two dense layers. Clinical time series are fed into the first CNN layer. Features of relational similarity, demographics, and medical are concatenated with flattened CNN features at the first dense layer.

LSTM: A bi-directional LSTM handles the clinical time series. The hidden dimension of the LSTM is 256. The maximum hidden states at each time step of the top-level LSTM are concatenated with relational similarity features, demographics, and medical features, and processed by two dense layers.

Performance of the models with and without the relational similarity feature is evaluated in PR-AUC and precision@k, and shown in Figure 2 and 3. Among all the models, the feature number of proposed and baseline are on comparable level, e.g., XGB-baseline has 1,331, and XGB-proposed has 1,339 features.

Table 2. The number of successfully predicted treatments from 07-2018 to 12-2018 among the K selected patients.

K	600	1,200	1,800	3,600
XGB	346	534	669	975
XGB-Proposed	370	550	709	1,023
Improvement	6.9%	3.0%	6.0%	4.9%
CNN	328	547	694	1,048
CNN-Proposed	342	562	721	1,080
Improvement	4.3%	2.7%	3.9%	3.1%
LSTM	366	579	736	1,083
LSTM-Proposed	394	609	771	1,133
Improvement	7.7%	5.2%	4.8%	4.6%

A sampling of the precision@k curve is provided in Table 2. The results further confirm that by capturing the relational similarity feature, we can improve the prediction accuracy of all baseline models, suggesting the universal effectiveness of relational similarity feature for a wide range of model structures. When K goes large, the proposed method consistently outperforms the baseline model, although the incremental gradually become marginal, e.g., in LSTM the improvement is 2.9%, 1.6%, 0.6%, when K equals 5,000, 10,000, 20,000, respectively.

5. Conclusions

In this paper, we proposed a graph-based algorithm to extract latent relational similarity from patients' clinical time series. Experimental results using real-world data show that the proposed feature can improve the prediction accuracy of a wide range of model structures. We envision that the relational similarity can also enhance model performance on other tasks, such as rare disease detection or patient subtyping. In its current form, the algorithm is operating on a two-dimensional feature space, i.e. patients and clinicians. In the future, we will add medical services as another dimension into the feature space, and thus enable the application of more advanced signal processing techniques such as tensor decomposition to uncover more useful information from the multidimensional feature inputs.

References

Bai, T., Zhang, S., Egleston, B. L., and Vucetic, S. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 43–51. ACM, 2018.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74. ACM, 2017.

Bergmeir, C. and Benítez, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

Che, C., Xiao, C., Liang, J., Jin, B., Zho, J., and Wang, F. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 198–206. SIAM, 2017.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pp. 301–318, 2016a.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512, 2016b.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2016c.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795. ACM, 2017.

Choi, E., Xiao, C., Stewart, W., and Sun, J. MiME: Multilevel medical embedding of electronic health records for predictive healthcare. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4547–4557. Curran Associates, Inc., 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lehoucq, R. B. and Sorensen, D. C. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, 1996.

Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., and Gao, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911. ACM, 2017.

Ma, F., Wang, Y., Xiao, H., Yuan, Y., Chitta, R., Zhou, J., and Gao, J. A general framework for diagnosis prediction via incorporating medical code descriptions. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1070–1075. IEEE, 2018a.

Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 743–752. ACM, 2018b.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849–856, 2002.

Sukkar, R., Katz, E., Zhang, Y., Raunig, D., and Wyman, B. T. Disease progression modeling using hidden markov models. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2845–2848. IEEE, 2012.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., and Ghassemi, M. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pp. 322–337, 2017.

Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 2019.

Xiao, C., Choi, E., and Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10): 1419–1428, 2018.