Graph-Based Data Representation and Prediction in Medical Domain Tasks Using Graph Neural Networks

Vdovkina Sofiia [1[0009-0000-5589-9533]], Derevitskii Ilya[1[0000-0002-8624-5046]], Abramyan Levon[2[0000-0002-6830-5172]], Vatian Aleksandra[1[0000-0002-5483-716X]]

[1] ITMO University, Saint-Petersburg, Russia
{sophia.vdovkina, ilyaderevitskii}@niuitmo.ru,
alexvatyan@gmail.com

[2] World-Class Research Centre for Personalized Medicine, Almazov National Medical Research Centre, 197341 St. Petersburg, Russia

Abstract. Medical data often presents as a time series, reflecting the disease's progression. This can be captured through longitudinal health records or hospital treatment notes, encompassing diagnoses, health states, medications, and procedures. Understanding disease evolution is critical for effective treatment. Graph embedding of such data is advantageous, as it inherently captures entity relationships, offering significant utility in medicine. Hence, this study aims to develop a graph representation of Electronic Health Records (EHRs) and combine it with a method for predictive analysis of COVID-19 using network-based embedding. Evaluation of Graph Neural Networks (GNNs) against Recurrent Neural Networks (RNNs) reveals superior performance of GNNs, underscoring their potential in medical data analysis and forecasting.

Keywords: Graph neural networks, data representation, electronic health records.

1 Introduction

Recent advancements in deep learning have yet to fully integrate into clinical decision-support systems, but the digitization of health records has spurred machine learning research in medicine, particularly in EHRs. While this surge underscores the growing relevance of the field, the utilization of EHRs spans from database establishment to embedding methodologies [1]. Our focus lies specifically on leveraging graph embeddings within medical records, representing a novel approach in this domain.

Medical data often manifests as time series, portraying disease progression over time within longitudinal health records or clinical notes documenting hospital treatments. EHRs serve as rich repositories for various electronic health tasks, including predictive modeling of clinical risks such as in-hospital mortality and readmission rates, disease correlations exploration, classification, and medical decision support. Predictive modeling of future clinical events is thus a vital objective in medical practice. Predicting sequences of clinical events typically involves latent entity and event embedding coupled with neural network models like RNNs. However, medical records pose unique challenges due to their sparsity, irregularity, and heterogeneity, hampering effective

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_32

predictive model creation. Graph embedding offers inherent advantages by capturing entity relationships, particularly beneficial in the medical domain. Therefore, our study aims to develop a predictive analysis tool for disease progression using network-based embedding methodologies, specifically exploring methods for transforming longitudinal medical record data into graphs to elucidate disease trajectories, followed by the construction and comprehensive analysis of GNNs.

The remainder of this paper is organized as follows. Section II reviews related work in the field. Section III provides details on the creation of patient graphs. In Section IV, the experimental setups with the proposed embedding approach and GNN are explained, along with the presentation of results. Finally, Section V concludes this paper, summarizing the findings and suggesting avenues for future research.

2 Related Work

2.1 Transformation Methods

The temporal dynamics inherent in time series data can be effectively captured through a graph model representation, akin to a discrete model of the underlying dynamic system. This model, guided by Takens theorem, facilitates the restoration of the system's state space. In [2], various transformation methods for time series data into complex network representations are categorized into three distinct classes: proximity networks, visibility graphs, and transition networks. In the context of personal medicine, prediction, and classification tasks utilizing EHRs, the conversion of temporal EHR data into graph structures has emerged as a pivotal approach for leveraging inherent graph properties in subsequent analyses. Given the heterogeneous nature of EHRs encompassing medications, diagnoses, clinical notes, and lab results, diverse modeling techniques are imperative to accommodate this diversity. Despite the prevalence of articles mentioning “medical records” and “graphs” a significant portion does not delve into graph theory or utilize graphs to represent individual patient data [2]. However, a subset of studies effectively leverages graph structures to encode temporal relationships within time series data.

2.2 Methods of Health Records Transformation to Graph

Patient medical records can be represented in a knowledge model in two primary ways: as a scope depicting common patient features and their interrelations or as temporal graphs for individual patients. Notably, Khademi M. and Nedialkov N. S. [3] constructed a probabilistic graphical model using clinical data, integrating it with deep belief networks for breast cancer prognosis. Chen et al. [4] proposed a graph-based semi-supervised learning algorithm for risk prediction, utilizing Cause of Death information as labels and temporal relationships between examination items as edges. Liu et al. [5] developed a temporal phenotyping approach based on graph representations of EHR events, extracting significant graph bases for interpretability. Esteban et al. [6] applied latent embedding models to clinical data for event sequence prediction, while Zhang et

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_32

al. [7, 8] proposed integrative medical temporal graph-based prediction approaches, albeit with varied success rates. Notably, Tong et al. [9] introduced an LSTM-GNN model for patient outcome prediction, effectively combining temporal and graph-based features.

Similarly, graph methods for patient clustering were developed [10], emphasizing the construction of network-like structures based on patient similarities. Additionally, Hanzlicek et al. [11] and Kaur et al. [12] described graph-based models for storing medical records, facilitating efficient data retrieval and decision support functionalities. These studies underscore the diverse applications and potential of graph representations in healthcare, from predictive modeling to decision support systems.

From the review, it can be inferred that graphs are not yet widely adopted as a form of health data representation. However, this area of research shows promise and warrants further exploration. For sequence prediction tasks, methods based on individual patient data graphs are recommended, as they effectively capture temporal relationships between health states. Conversely, for classification tasks, scope representation may be more beneficial, as it encapsulates common features and their interrelations. Nevertheless, the most promising avenue for future research lies in the discovery of methods that represent inter-patient connections. Such approaches hold the potential to enhance prognostic capabilities and inform clinical decision-making by leveraging similarities among medical cases. Thus, further investigation into graph-based representations in healthcare is imperative for advancing predictive analytics and decision support systems in medicine.

3 Creating EHR representation through graphs

EHRs serve as comprehensive repositories of patient health information, capturing vital aspects of their medical journey. Transforming this rich temporal data into a structured graph representation holds immense potential for facilitating various healthcare tasks, ranging from predictive modeling to clinical decision support. In this chapter, we present a novel pipeline for creating graph representations of EHRs. Our approach involves gathering patient state data during hospitalization periods, embedding each state, and subsequently constructing a graph where nodes represent patient states and edges encode temporal and proximity-based relationships.

3.1 Data description

The study utilized a dataset comprising 6188 medical records encompassing 1992 distinct patients who received treatment for COVID-19 at Almazov National Medical Research Centre in St. Petersburg, Russia, spanning from June 2020 to March 2021. Each treatment case is characterized by a comprehensive set of indicators, as detailed in Table 1.

These medical indicators comprise a spectrum of information, spanning past medical conditions, laboratory analysis results, physical measurements, lifestyle factors, and medication types. They collectively provide a comprehensive overview of patients'

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_32

health profiles, supporting clinical analysis and decision-making. The dataset comprises 40 integer features, 11 floating-point features, and 3 date features.

Table 1. Medical indicators of treatment cases.

Group	Features
Controlled Medications	Omeprazole, nadroparin calcium, esomeprazole, amlodipine, ambroxol, domperidone, mebrofenin, technetium, mometasone, bisoprolol, dexamethasone, hydrochlorothiazide, hydroxychloroquine, rabeprazole, enoxaparin sodium, perindopril, acetylcysteine, azithromycin, valsartan, methylprednisolone, loratadine, chloroquine, sodium chloride, indapamide, prednisolone, atorvastatin, dextran, lisinopril, losartan
Dynamic Factors	Temperature, lymphocytes count, aspartate aminotransferase, heart rate, respiratory rate, total bilirubin, mean platelet volume, platelet crit, lymphocytes percentage, decreased consciousness, severity grade on CT scan, lactate dehydrogenase, platelet distribution width
Static Factors	Age, sex
Controlled Procedures	Blood transfusion, oxygen therapy, non-invasive ventilation, invasive ventilation
Process Variables	Process stages, current process duration
Dates and length of stay	Admission date, end episode, length of observation
Target Variable	Outcome

3.2 Pipeline description

The pipeline is initiated by collecting data pertaining to patients' states during hospitalization periods. Each state is characterized by a set of features capturing relevant clinical parameters and temporal information. Preprocessing steps involve cleaning the data, handling missing values, and standardizing features to ensure uniformity across the dataset.

To capture the intricate relationships within patient states, TabNet, a state-of-the-art tabular data embedding technique, is employed [13]. This network leverages sequential attention mechanisms to learn informative embeddings from tabular data, effectively capturing both local and global dependencies. By embedding each patient state using TabNet, the multidimensional characteristics are encoded into a compact representation, facilitating downstream graph construction.

The crux of our pipeline lies in the construction of a graph representation that encapsulates the temporal and proximity-based relationships among patient states. A two-pronged approach to edge creation is adopted: temporal connections and proximity-based connections. Incorporating both features is crucial for capturing the sequential order of events and the contextual relationships between patient states in EHRs, our

DOI: 10.1007/978-3-031-63772-8_32

model gains a deeper understanding of disease dynamics and patient interactions. This enriched context improves the model's ability to generalize to new data and adapt to changes in healthcare practices.

Temporal Connections: Patient states are sequentially connected in time, reflecting the temporal progression of their medical journey. Each state is linked to its subsequent state within the same patient's trajectory, forming a temporal sequence of nodes. Proximity-based Connections: To capture inter-patient relationships, the Euclidean distance between each patient state is computed to identify the closest states of other patients. Two approaches are explored: (1) connecting each state to the closest three states of any other patient (see Fig. 1), or (2) imposing a constraint wherein a state can only be connected to a state of another patient once, thereby fostering diverse inter-patient connections (see Fig. 2). The choice between these approaches can be determined experimentally or tailored to specific healthcare objectives. For visualization purposes, graphs featuring 10 patients are provided. The complete graph comprises 6182 nodes, each with 24 features, wherein each node is labeled with either 0 or 1 based on the outcome, and 22737 edges representing temporal and interpersonal connections.

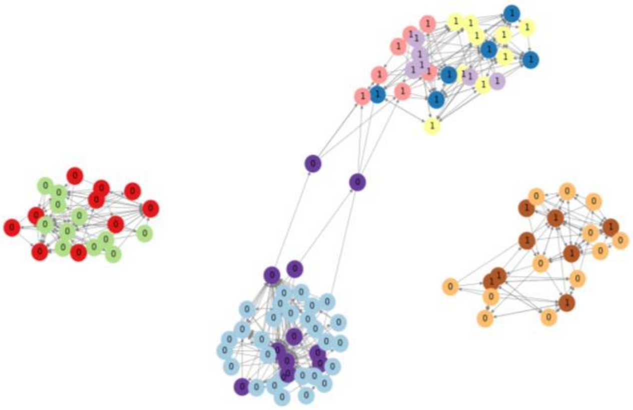

Fig. 1. Graph without constraints on the connections of patients. Colors represent different patients, and node labels are selected by the outcome.

Constructing a graph without connection constraints reveals discernible clusters within the data. Notably, patients experiencing negative outcomes tend to cluster together, suggesting shared characteristics or medical trajectories. Additionally, the observed clustering patterns hint at the potential grouping of patients with close medical histories, indicating the influence of shared anamnesis factors on cluster formation.

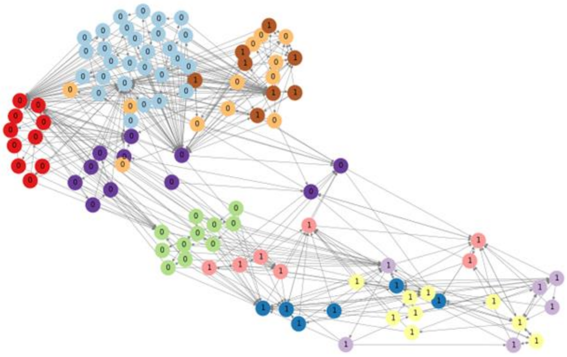

Fig. 2. A graph constructed with diverse inter-patient connections. Colors represent different patients, and node labels are selected by the outcome.

In a graph with connection constraints, distinct groups of patients are still discernible, albeit with more blurred boundaries. However, this method holds promise for classification tasks as it captures more ambiguous connections, allowing for the identification of subtle relationships between patient clusters.

Further experimentation and validation are needed to assess its efficacy and generalizability across healthcare contexts.

4 Experiments

The experimental results provide valuable insights into the performance of the suggested representation method used as a source for Graph Convolutional Network (GCN) and Graph Isomorphism Network (GIN) models for mortality prediction tasks based on patient state graphs, performing node classification. This choice is due to their unique ability to effectively capture relational information and structural dependencies inherent in graph data. By aggregating information from neighboring nodes, GCNs can effectively model the complex interactions between patient states, capturing crucial temporal and contextual dependencies. GINs are invariant under permutations of the nodes, so they learn and generalize patterns across different patient trajectories, regardless of their specific ordering or representation.

The proposed GCN architecture consists of three hidden layers of GCNConv, each followed by ReLU activation function. The input to the model is the feature vector representing each patient state, which is passed through the GCN layers along with the edge indices representing the connectivity of the graph. Dropout with a probability of 0.5 is applied after the first GCN layer to prevent overfitting. The output of the final GCN layer is passed through a linear layer with an output dimension of 1, followed by a sigmoid function to produce the final classification probability indicating the likelihood of patient mortality.

The GIN architecture consists of two GINConv layers, each comprising a sequence of linear transformations followed by batch normalization and ReLU activation

DOI: 10.1007/978-3-031-63772-8_32

functions. Similar to GCN, the input to the model is the feature vector representing each patient state, which is passed through the GINConv layers along with the edge indices representing graph connectivity. The aggregated features are then passed through a linear layer with an output dimension of 1, followed by a sigmoid function as well. And for the sake of experiment, along with graph networks RNN model tested with 3 hidden layers was.

The choice of metrics was driven by their relevance to mortality prediction. While accuracy measures overall correctness, precision assesses the proportion of true positive cases among predicted positives, and recall gauges the model's ability to identify all true positive cases. Emphasizing recall ensures the accurate identification of patients at risk of mortality, vital for early intervention.

The observed metrics of the models are presented in Table 2.

Table 2. Observed metrics of the models.

Model	Accuracy	Precision	Recall
GCN without connection constraints	0.8208	0.5341	0.7743
GIN without connection constraints	**0.8722**	0.5771	**0.8329**
GCN with connection constraints	0.8023	0.2875	0.6145
GIN with connection constraints	0.8458	0.3298	0.6875
RNN	0.7832	**0.6680**	0.7352

The first type of graph, without connection constraints, allowed for the formation of clusters, leading to improved performance of both GCN and GIN models. This suggests that the graph structure contributed to more accurate representations and enhanced predictive performance. The absence of constraints likely facilitated the models' ability to capture underlying patterns and relationships within the data. While both models performed well, GIN showed superior ability in capturing true positive cases while minimizing false positives.

In perspective of outcome prediction, these results suggest that incorporating graph-based representations of patient states can enrich predictive modeling.

5 Conclusion

In conclusion, our study introduces a novel approach for Electronic Health Record (EHR) representation using graphs, promising advancements in healthcare analytics and predictive modeling. By employing graph-based representations of patient states, temporal and relational dependencies crucial for accurate predictions were captured. Leveraging advanced machine learning techniques like Graph Convolutional Networks and Graph Isomorphism Networks further enhances predictive capabilities.

Our experiments confirm the effectiveness of our approach. While the incorporation of connection constraints in graph construction didn't improve efficiency, its potential applications warrant further investigation. Analyzing specific cluster characteristics could shed light on their impact on model performance.

In summary, adopting graph-based representations for EHRs revolutionizes healthcare analytics, enabling comprehensive analysis while preserving temporal and relational contexts. Our approach lays the groundwork for future advancements in EHR representation and predictive modeling, driving towards more effective healthcare solutions.

Acknowledgements

This work was supported by Russian Science Foundation, Grant № 23-11-00346.

References

1. Solares J. R. A. et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures //Journal of biomedical informatics. – 2020. – T. 101. – p. 103337.
2. Schrodt J. et al. Graph-representation of patient data: a systematic literature review //Journal of medical systems. – 2020. – T. 44. – №. 4. – pp. 1-7.
3. Khademi M., Nedialkov N. S. Probabilistic graphical models and deep belief networks for prognosis of breast cancer //2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). – IEEE, 2015. – pp. 727-732.
4. Chen L. et al. Mining health examination records—a graph-based approach //IEEE Transactions on Knowledge and Data Engineering. – 2016. – T. 28. – №. 9. – pp. 2423-2437.
5. Liu C. et al. Temporal phenotyping from longitudinal electronic health records: A graph based framework //Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. – 2015. – pp. 705-714.
6. Esteban C. et al. Predicting sequences of clinical events by using a personalized temporal latent embedding model //2015 International Conference on Healthcare Informatics. – IEEE, 2015. – pp. 130-139.
7. Zhang S. et al. MTPGraph: A data-driven approach to predict medical risk based on temporal profile graph //2016 IEEE Trustcom/BigDataSE/ISPA. – IEEE, 2016. – pp. 1174-1181.
8. Zhang J., Gong J., Barnes L. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records //2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). – IEEE, 2017. – pp. 214-221.
9. Tong C. et al. Predicting patient outcomes with graph representation learning //International Workshop on Health Intelligence. – Springer, Cham, 2021. – pp. 281-293. 3
10. Ochoa J. G. D., Mustafa F. E. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses //Artificial Intelligence in Medicine. – 2022. – T. 131. – p. 102359.
11. Hanzlicek P. et al. User interface of MUDR electronic health record //International journal of medical informatics. – 2005. – T. 74. – №. 2-4. – pp. 221- 227.
12. Kaur K., Rani R. Managing data in healthcare information systems: many models, one solution //Computer. – 2015. – T. 48. – №. 3. – pp. 52-59.
13. Arik S. Ö., Pfister T. Tabnet: Attentive interpretable tabular learning //Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – T. 35. – №. 8. – pp. 6679-6687.