Coupling Heterogeneous Graph Embeddings with Convolution Neural Networks Improves Mortality Prediction

Tingyi Wanyan
tingyi.wanyan@mssm.edu
Hasso Plattner Institute for Digital Health at Mount Sinai,
Icahn School of Medicine at Mount Sinai
New York, NY

Ying Ding
ying.ding@ischool.utexas.edu
School Of Information, University of Texas Austin
Austin, TX

Ariful Azad
azad@iu.edu
Intelligent System Engineering, Indiana University
Bloomington, IN

Benjamin S Glicksberg
benjamin.glicksberg@mssm.edu
Hasso Plattner Institute for Digital Health at Mount Sinai,
Icahn School of Medicine at Mount Sinai
New York, NY

ABSTRACT

Computational prediction of in-hospital mortality in the setting of an intensive care unit can help clinical practitioners guide care and make early decisions for interventions. In this work, we train a Heterogeneous Graph Relational Model on Electronic Health Record data and use the resulting embedding vector as additional information added to an Convolution Neural Network model for predicting in-hospital mortality. We show that the additional information provided from including time as a vector in the embedding captures the relationships between medical concepts, lab tests, and diagnosis and aids in predictive performance. We find that adding HGM to a CNN model can increase the mortality prediction accuracy to a certain extent. This framework can serve for a foundation for future experiments involving more EHR data types on other important healthcare prediction tasks .

CCS CONCEPTS

• **Mathematics of computing** → *Graph algorithms*; • **Applied computing** → **Health informatics**.

KEYWORDS

Electronic Health Records, Heterogeneous Graph Model, Convolution Neural Network, Morality Prediction

ACM Reference Format:
Tingyi Wanyan, Ying Ding, Ariful Azad, and Benjamin S Glicksberg. 2018. Coupling Heterogeneous Graph Embeddings with Convolution Neural Networks Improves Mortality Prediction. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Predicting in-Hospital mortality in the hospital Intensive Care Units is crucial for patient care [7, 12], as it can help practitioners tailor care and allow for earlier interventions . Electronic Health Record (EHR) consists of information relating to patient encounters with a health system, such as demographics, disease diagnoses, vital signs, and medications, among others which are often used for machine learning (ML)-based tasks in the biomedical space, such as mortality prediction [6, 13]. The inherent complexity of EHR data often require complex modeling frameworks for robust performance for these tasks, such as convolution neural networks (CNN), which treats the time as horizontal dimensions and medical concepts as vertical dimensions. [2, 8, 14]. For the vertical medical concept features of many such CNN models directly concatenate these unordered sets of raw data and use them as direct inputs[11]. This scenario of providing features to CNN model is simple, straightforward, and often performs well in these health-related prediction tasks. This strategy of inputting raw features, however, also disregards the graphical structure and inner connectivity between medical concepts[3, 4]. Furthermore, the medical events recorded in EHR data are often sparse, as a result of missing or incomplete data, which results in a dearth of information for CNN, which could thereby affect performance model[2].

In this work, we propose to use and extend a Heterogeneous Graph Model (HGM) to provide a patient embedding vector to fill in missing gaps of information for training CNN model in EHR data. The HGM model can captures the relationships between different medical concept types (e.g., diagnoses and lab tests) due to its graphical structure. Integrating the context between concept types in a model can further facilitate capturing more complex patient patterns and encode similarities. In the procedure of building the HGM model, we also add an edge connection representing time that reflects states of patient across their progression in the hospital (i.e., on admission and in the next hour). Therefore, after the HGM model is trained, we can provide an additional patient embedding vector of the next hour based on the patient embedding vector of the current hour. This addition of time could provide extra information to the embedding vectors from the HGM model of time points where data are missing.

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

In this work, we show that by concatenating these additional time-based embedding vectors to the raw features as the final feature input to CNN model, we can increase performance of in-hospital mortality prediction to a certain degree compared to another advanced modeling strategy, specifically a CNN model with pure raw data as the input feature.

2 METHODOLOGY

2.1 Data Set

We conduct our experiment on de-identified EHR data from MIMIC-III. This data set contains various clinical data relating to patient admission to a hospital ICU, such as demographics, lab test results, and disease diagnoses. We collected data for 5956 patients, extracting lab tests within every hour from admission. There were a total of 409 unique lab tests and 3387 unique disease diagnoses observed. We cropped the lab test events into six, 12, 24, and 48 hours prior to patient death or discharge from ICU. From these data, we sought to predict mortality using two modeling strategies and diving the data into 70% for training and 30% for testing.

2.2 CNN model

CNNs are best known by achieving tremendous success on image processing tasks [9] due to its ability to extract distinct groups of features in two dimensional data, which increases the accuracy for classification tasks. In this work, we use CNN model as the baseline method for mortality prediction.

As a CNN model requires two dimensional inputs, we treated time as the horizontal dimension and medical events as the vertical dimension. In the dimension of time, we recorded every event that happened within every hour increments counting down from the patient death or discharge time. In the baseline model, the vertical dimension was constructed by concatenating two medical event vectors: lab tests and diagnoses. Every entry of the lab test vector records the value of a specific lab test for that hour; for the diagnosis vector, the i-th entry is 1 if the i-th diagnosis is observed, and we concatenate these two vectors to form a medical event happened in one hour.

The prediction is a binary classification task on predicting mortality. We used a softmax layer with two dimension as the prediction layer.

2.3 Building the Heterogeneous Graph Model

The features used in baseline CNN model were purely raw data, which lacks considration of the inner relations between medical concepts. We used an HGM to capture these inherent relationships by creating three different type of nodes: patient, lab test, and diagnosis. These different type of nodes are connected by three relation types: tested, diagnosed, and time. These could be represented with two triples:

$$Lab \xrightarrow{tested} patient : \{Labtest, tested, patient\}$$

$$Patient \xrightarrow{diagnosed} Diagnosis : \{patient, diagnosed, diagnosis\}$$

$$Patient \xrightarrow{time} patient\_next\_hour : \{patient, time, patient\_next\_hour\}$$

the testing relationship shows whether a specific lab test was given to a patient at a specific time, the diagnosed relationship shows whether a patient was diagnosed with a disease, the time relationship captures the patient condition at a specific time, and his/her condition in the following hour.

To represent the lab test and diagnosis node types, we separately use multi-hot encoding vector: $X_l \in \{0,1\}^{409}$ and $X_d \in \{0,1\}^{3387}$, the i-th entry is 1 indication of the whether a specific lat test was performed or a specific diagnosis was given. The patient node is also represented as a vector $X_p \in \mathbb{R}^{477}$ containing the numerical values measured from lab tests at that time.

2.4 Embedding Different Type of Nodes Into the Same Latent Space

For capturing the inner relations between different medical events related to a patient, we utilized the TransE model[1] to project different type of nodes into a same latent space, then classified those nodes that are connected as a similar group. Meanwhile, we classified the disconnected nodes into a dissimilar group.

The TransE model uses a set of 1) projection matrices and 2) relation vectors. After initialization, projections and translations can be optimized end-to-end. Heterogeneous nodes X_p, X_l, X_d are projected into a shared latent space with trainable projection matrices W_p, W_i, W_d using the nonlinear mappings:

$$c_p = \sigma(W_p \cdot X_p)$$
$$c_i = \sigma(W_i \cdot X_i)$$
$$c_d = \sigma(W_d \cdot X_d)$$

Where σ is a non-linear activation function and c_p, c_i, c_d are the latent representations of each type of node. Despite the fact that the EHR-space uses different dimensions for different data types X_p, X_i, X_d, all nodes types were projected into the same latent space. Then we apply translation operations to link these different types of nodes:

$$c_p = c_i + r_{ip}$$
$$c_d = c_p + r_{pd}$$
$$c_p^{'} = c_p + r_{time}$$

Where r_{ip} and r_{pd} are the relation vectors connecting patients to lab testss and diagnoses, respectively. Note that $c_p^{'}$ is the patient latent representation in the next hour corresponding to c_p, r_{time} captures this relationship. Both $c_p^{'}$ and c_p use the same projection matrix W_p.

2.5 Optimization Model

For learning the HGM, we apply a heterogeneous skip-gram optimization model [5],which increases the proximity between those embedding points whose corresponding graph nodes are often connected after the projection and translation operations:

$$\max \sum_{u \in V} \sum_{t \in T_V} logPr(N_t(u)|f(u)) \tag{1}$$

Where $N_t(u)$ are the heterogeneous neighborhood vertices of center node u, and $t \in T_V$ is the node type. Here, we learn effective

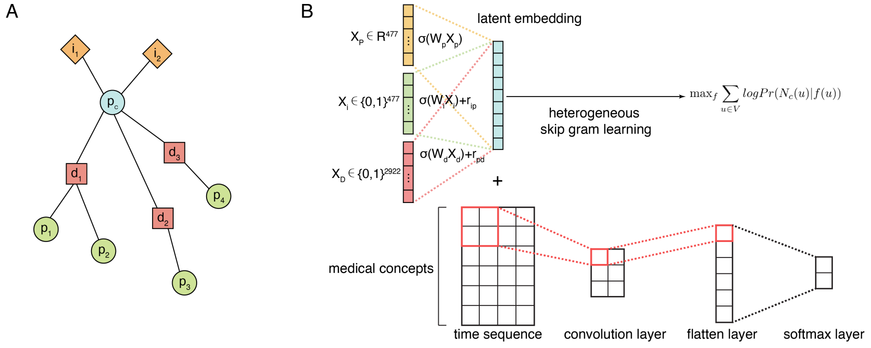

Figure 1: (A) A graphical representation of the HGM for patient, lab test, and diagnosis data. (B) All graph nodes in (A) have a corresponding vector like those shown in (B). The vector representations can be projected into a shared space with the TransE method, and this projection is optimized for retaining relations in the original data in the embedding via skip-gram optimization. Finally, these vectors are concatenated into the CNN model for mortality prediction.

node embeddings by maximizing the probability of correctly predicting the a patient node's associated lab tests and diagnoses. The prediction probability is modeled as a softmax function:

$$Pr(c_t|f(u)) = \frac{e^{\vec{c}_t \cdot \vec{u}}}{Z_u} \tag{2}$$

Where $\vec{u}$ is the latent representation of patient u, $\vec{c}_t$ is the latent representation of lab and diagnosis neighbors of node of u, and $\vec{c}_t \cdot \vec{u}$ is the inner product of the two embedding vectors representing their similarity. Z_u is the normalization term $Z_u = \sum_{v \in V} e^{\vec{v}_t \cdot \vec{u}}$. Where Z_u integrates over all vertices. Therefore, equation 1 could be simplified to:

$$\mathcal{L}_s = -\sum_{t \in T}\sum_{u \in V}\Big[\sum_{c_t \in N_t(u)} \vec{c}_t \cdot \vec{u} - logZ_u\Big] \tag{3}$$

Numerical computation of Z_u is intractable for very large graphs with millions of nodes. So we adopt negative sampling strategy [10] to approximate the normalization factor, and the optimization function becomes:

$$\mathcal{L}_s = -\sum_{t \in T}\sum_{u \in V}\Big[\sum_{c_t \in N_t(u)} log\sigma(\vec{c}_t \cdot \vec{u}) + \sum_{j=1}^{\mathbb{K}} E_{c_j \sim P_v(c_j)} log\sigma(-\vec{c}_j \cdot \vec{u})\Big] \tag{4}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$, $\mathbb{K}$ is the number of negative samples. $P_v(c_j)$ is the negative sampling distribution. Equation 4 is the final objective function we are using for heterogeneous graph learning.

2.6 Details of the Training Process

For training our HGM, we performed heterogeneous neighborhood sampling by its one-hop connectivity, and picked *Patient* node as the center node, since it has one-hop connections to both *Diagnoses* and *Lab_test* nodes. Specifically, for one training center *Patient* node, we uniformly sampled 10 *Diagnoses* one-hop direct connected nodes, and 10 *Lab_test* one-hop direct connected nodes. From these sampled 10 *Diagnoses* nodes, we sampled another 10 *Patient* nodes, each having connections with each of the prior 10 *Diagnoses* nodes. In this way, we connected the center patient node with similar other *Patient* nodes by their common diagnoses. We also sample the patient node which belongs to the next hour corresponding to the center *Patient* node. For negative sampling [10], we performed uniform sampling through all *Diagnoses* node and *Lab_test* nodes that do not have one-hop connections with the center training patient node. We then projected these different nodes into same latent space through TransE model. After unifying the embeddings for different node types, each concept is represented as a point in a Euclidean space. In this space, we can measure the similarity between any two points by the angle of the vectors between them and the origin.

2.7 Incorporating the HGM Embedding Vector into CNN Model

With the Heterogeneous Graph Embedding model, feeding the model with a raw patient data will result in a patient embedding vector output. This embedding vector encodes not only a patient's current lab test results, but also their relation with different kind of diagnoses, lab tests, and subsequent items in time.

We concatenated this outcome embedding vector on the baseline CNN vertical feature dimension to form a final feature vector within every hour, and use these new features as the CNN input to predict mortality. In addition, since we also encode time as a relation type, we can infer the embedding vector of time points with missing data based on information from the previous hour. We visualize this procedure in the schematic Fig 1.

3 EXPERIMENTS

We aimed to predict mortality in 6, 12, 24, and 48 hours prior to death and/or discharge. The CNN model wass used as the prediction model which was introduced in section 2.2. We compared three different scenarios on testing the impact of adding HGM embedding vectors as additional features to the framework:

- HGM: Concatenate the patient embedding feature from HGM model with raw diagnosis vector.
- CNN: Concatenate the raw lab test feature with diagnosis vector.

Table 1: Mortality Prediction Accuracy

Hours	Models		
	HGM	CNN	HGM+CNN
6	0.776	0.766	**0.803**
12	0.775	0.773	**0.807**
24	0.771	0.775	**0.806**
48	0.775	0.772	**0.801**

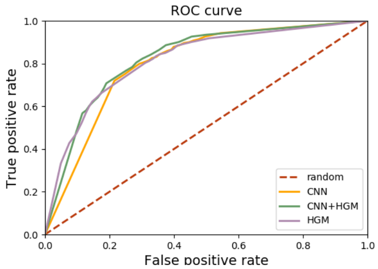

Figure 2: The ROC curve derived experiments on the testing data

- HGM+CNN: Concatenate the HGM patient embedding vector, the raw lab test feature vector, and the diagnosis vector.

We split the data into 70% for training and 30% percent for testing. The primary output metric was prediction accuracy, specifically the number of patients with correctly detected predictions divided by the total patents tested. We show the results of these experiments in Table1. We show the receiver operator characteristic (ROC) curve of the performance of these tasks on the test data in Fig.2

The testing results shows that the HGM+CNN outperforms both the basic HGM and CNN models, indicating the additional information added from the HGM patient embedding of time increases the accuracy of predicting in-patient mortality. The prediction accuracy of using different hours prior to death&discharge did not largely vary, indicating that different time windows did not have a big impact on the result on this particular task. The prediction accuracy in the CNN model which uses the concatenation of raw lab test and diagnosis data drops in the case of six hours prior to death and/or discharge, but not in the other two models, indicating that using the embedding features from HGM model was more robust than the raw data.

4 DISCUSSION AND CONCLUSION

In this work, we propose a method to incorporate patient embedding vector from HGM model into the raw data for CNN model in an attempt to provide more information to CNN model. We assess the value of this implementation on a task of predicting mortality in EHR data. The results of our experiment shows the superior performance of adding the additional patient embedding vector, which was pretrained from the HGM model, compared to pure raw features as the input to CNN model. In one aspect, this is due to the fact that the HGM embedding vector captures additional relational information between different medical concepts, thus providing additional information to CNN model. In another aspect, we fill in the missing EHR data in a specific hour by deriving the embedding vector inferred from the patient embedding of the previous hour. Therefore, more information was provided in a time sequence to the CNN model, resulting in an increased accuracy in predicting mortality rates within this modeling framework.

Furthermore, we observed that purely concatenating the HGM embedding vector with diagnosis feature vectors did not increase the accuracy against using the concatenation between raw lab test and diagnosis feature vectors. This finding indicates that the raw lab test feature vector could provide unique information for CNN to utilize. At the same time, this finding indicates that the embedded patient vector from HGM model could lose some information from the raw lab test feature along the process of projecting these data into a low dimensional latent space. By concatenating all feature vectors, we aim to preserve the information from different data points, which helped achieve higher mortality prediction accuracy. We hope the findings from this work can be expanded in future directions that may add more EHR node types and time components on a variety of other important health-related predictive tasks.

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. 2787–2795.

[2] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 432–440.

[3] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in neural information processing systems*. 4547–4557.

[4] Edward Choi, Zhen Xu, Yujia Li, Michael W Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M Dai. 2019. Graph convolutional transformer: Learning the graphical structure of electronic health records. *arXiv preprint arXiv:1906.04716* (2019).

[5] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 135–144.

[6] Benjamin S Glicksberg, Riccardo Miotto, Kipp W Johnson, Khader Shameer, Li Li, Rong Chen, and Joel T Dudley. [n.d.]. Automated disease cohort selection using word embeddings from Electronic Health Records. World Scientific.

[7] Alistair EW Johnson and Roger G Mark. 2017. Real-time mortality prediction in the Intensive Care Unit. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association, 994.

[8] Soo Yeon Kim, Saehoon Kim, Joongbum Cho, Young Suh Kim, In Suk Sol, Youngchul Sung, Inhyeok Cho, Minseop Park, Haerin Jang, Yoon Hee Kim, et al. 2019. A deep learning model for real-time mortality prediction in critically ill children. *Critical Care* 23, 1 (2019), 279.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[11] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* 6, 1 (2016), 1–10.

[12] Alok Sharma, Anupam Shukla, Ritu Tiwari, and Apoorva Mishra. 2017. Mortality Prediction of ICU patients using Machine Leaning: A survey. In *Proceedings of the International Conference on Compute and Data Analysis*. 49–53.

[13] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1589–1604.

[14] Jinghe Zhang, Jiaqi Gong, and Laura Barnes. 2017. HCNN: heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 214–221.