MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231058

Graph Representation Learning-Based Fixed-Length Clinical Feature Vector Generation from Heterogeneous Medical Records

Tomohisa Seki[a,1], Yoshimasa Kawazoe[a,b] and Kazuhiko Ohe[a,c]
[a] *Department of Healthcare Information Management, The University of Tokyo Hospital, Japan*
[b] *Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo, Japan*
[c] *Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Japan*
ORCiD ID: Tomohisa Seki https://orcid.org/0000-0002-4281-135X,
Yoshimasa Kawazoe https://orcid.org/0000-0002-7277-0827,
Kazuhiko Ohe https://orcid.org/0000-0002-4296-9536

Abstract. Transformation of patient data extracted from a database into fixed-length numerical vectors requires expertise in topical medical knowledge as well as data manipulation—thus, manual feature design is labor-intensive. In this study, we propose a machine learning-based method to for this purpose applicable to electronic medical data recorded during hospitalization, which utilizes unsupervised feature extraction based on graph embedding. Unsupervised learning is performed on a heterogeneous graph using Graph2Vec, and the inclusion of clinically useful data in the obtained embedding representation is evaluated by predicting readmission within 30 days of discharge based on it. The embedded representations are observed to improve predictive performance significantly as the information contained in the graph increases, indicating the suitability of the proposed method for feature design corresponding to clinical information.

Keywords. Electronic health record, graph embedding, machine learning, unsupervised learning, feature extraction

1. Introduction

Medical records of individual patients are stored in medical information databases to facilitate consistent treatment based on their specific requirements over time. This is important, as the course of treatment can vary widely over different patients even for a single disease. Extraction of specific features from the vast amount of accumulated medical data requires clinical knowledge related to the target medical concept, data

[1] Corresponding Author: Tomohisa Seki, Department of Healthcare Imformation Management, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655 Japan, email: seki@m.u-tokyo.ac.jp.

conversion and processing techniques based on medical informatics, and expertise in processing and supplementing missing values reasonably [1-3]. Thus, manual feature design satisfying the aforementioned functions is difficult.

Machine learning, and neural networks in particular, have garnered significant attention in this regard owing to their ability to extract useful features for various tasks by learning a large amount of data [4]. One of the major novelties of this technology is the automation of complex feature design and extraction processes [5]. The application of automatic feature extraction technology to medical information databases is expected to facilitate the verification of similarities with past cases and identification of new patient clusters. Medical information stored as textual data in a database following information exchange standards can be represented using graphs [6], thereby integrating unformatted data into a single piece of information and enabling it to be considered as a single digital input. In particular, graphed patient information converted into fixed-length embedded representations via unsupervised learning, while retaining important features for case identification, can be used for automatic feature extraction. In this context, we propose a program to convert dated patient data recorded during hospitalization into graphs. In turn, the transformed graph data are converted into a fixed-length embedded representation of patient information using unsupervised graph representation learning. To verify the extraction of clinically meaningful patient characteristics, the clinical information contained in the obtained embedded representation is evaluated by examining the viability of predicting readmission within 30 days after discharge based on it.

2. Methods

2.1. Data preparation

Anonymized SS-MIX2 standardized storage [7] at the University of Tokyo Hospital was used in this study. The SS-MIX2 standardized storage standard has been approved as a standard for storing HL7 Ver.2.5 text files in Japan by the Ministry of Health, Labor and Welfare. For training, 52667 hospitalization data points corresponding to 31679 patients with hospitalization dates in 2015 and 2016 were used. For validation, 21763 hospitalization data corresponding to 15417 patients with hospitalization dates in 2017 were used.

2.2. Conversion to graphs

The proposed program included laboratory test results, prescriptions, and disease information recorded during hospitalization in HL7 messages stored in the SS-MIX2 standardized storage within its scope for conversion into graphs. It was created and executed using Python (version 3.8.5), and networkx (version 2.5) was used to handle the graphed data.

Within laboratory data, the analyte, identification, material code, and examination result of the Japan laboratory code version 10 (JLAC10 code) were included in the scope of the program. For items whose examination results were recorded as numerical values with accompanying normal ranges, abnormal directions were included in the graphing process. Corresponding to prescription information, standard drug master reference numbers (HOT9 codes) recorded during hospitalization were included within the scope of the graphing program after conversion into drug price standard list codes used by the national health insurance of Japan. For disease name information, International Classification of Diseases 10th Revision codes (ICD10 codes) for disease classification were used by the program. During the graphing process, the codes for disease names, prescriptions, and laboratory tests were partitioned and converted to graphs such that the agreement of higher-level concepts was expressed as node-label agreement, enabling facile evaluation of similarity. In addition, edges incident with date nodes corresponding to recorded dates were used. All graph edges were taken to be undirected. The patient data graph was utilized as input to the model as structured graph data by networkx.

2.3. Graph embedding

Graph2vec, which executes the Skip-gram algorithm using a neural network and learns an embedded representation of the entire graph via unsupervised learning, is a typical graph representation learning method used to obtain embedded representations of graphs, and is known for its widespread application to similarity determination in compound structures [8]. Graph2Vec uses the Weisfeiler-Lehman Graph Kernel to aggregate the label information of neighboring nodes within a pre-defined number of edges of each node. Subsequently, it obtains the embedded representation by utilizing unsupervised learning as an input to the Skip-gram algorithm. In this study, the single graph corresponding to information recorded during one hospitalization instance was used the input to Graph2Vec, and an embedded representation was obtained for each hospitalization via unsupervised learning.

2.4. Evaluation of embedded representations

The obtained embedded representations were plotted on a two-dimensional plane using t-distributed stochastic neighbor embedding (T-SNE) and color-coded with respect to the 20 ICD10 code block classification items. To verify the inclusion of clinically meaningful information within the embedded representations, they were first input into a multilayer perceptron. Then, a model was trained using supervised learning based on the label of readmission within 30 days after discharge, and its predictive performance was evaluated using a five-fold crossover test. The training and testing data contained 8,711 and 3,079 readmissions within 30 days of discharge, respectively. A schematic diagram of the validation experiment is depicted in Figure 1.

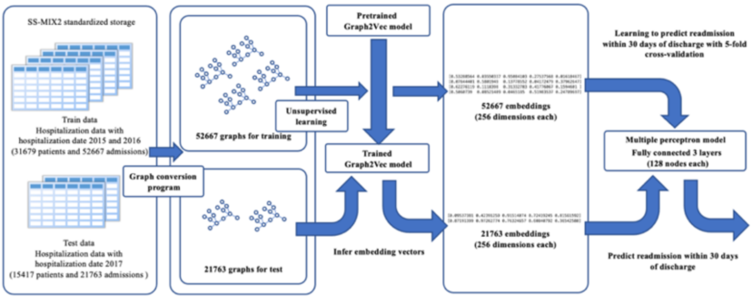
SS-MIX2 standardized storage
Train data
Hospitalization data with
hospitalization date 2015 and 2016
(31679 patients and 52667 admissions)
Test data
Hospitalization data with
hospitalization date 2017
(15417 patients and 21763 admissions)
Graph conversion program
52667 graphs for training
Unsupervised learning
21763 graphs for test
Pretrained Graph2Vec model
Trained Graph2Vec model
Infer embedding vectors
52667 embeddings
(256 dimensions each)
21763 embeddings
(256 dimensions each)
Learning to predict readmission within 30 days of discharge with 5-fold cross-validation
Multiple perceptron model
Fully connected 3 layers
(128 nodes each)
Predict readmission within 30 days of discharge

Figure 1. Schematic diagram of validation experiment used in this study.

3. Results

The two-dimensional plots obtained via T-SNE revealed an even distribution of embedded expressions with respect to block categories of ICD10 codes recorded as the primary diagnoses of admissions, as depicted in Figure 2. Some block category items formed denser clusters, suggesting that similar embedded expressions were obtained from medical data of hospitalizations with similar characteristics.

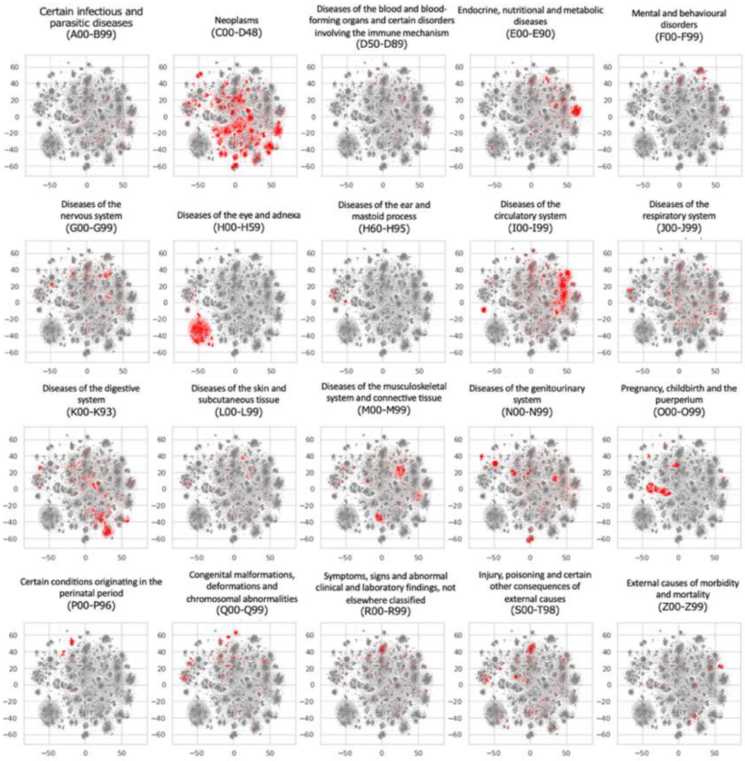

Figure 2. Two-dimensional plots of embedding vectors of test data using T-SNE. Red highlights indicate the inclusion of the corresponding disease name as the main disease name during hospitalization.

Readmission within 30 days of discharge was predicted on the test dataset for embedded representations. The prediction performance tended to improve as the information contained in the graph increased. These results suggest that clinical features were adequately extracted from the embedded expressions.

Table 1. Multi-perceptron Prediction of Readmission within 30 Days of Discharge. SD based on repetition during five-fold cross-validation.

Included information				**Validation results on test data**			
Date	**Drug**	**Lab**	**Diagnosi s**	**AUC (SD)**	**Precision (SD)**	**Recall (SD)**	**F1 score (SD)**
+	-	-	-	0.500 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
+	+	-	-	0.570 (0.009)	0.526 (0.021)	0.165 (0.022)	0.250 (0.024)
+	-	+	-	0.587 (0.004)	0.205 (0.010)	0.484 (0.038)	0.288 (0.004)
+	+	+	-	0.609 (0.017)	0.445 (0.031)	0.278 (0.055)	0.337 (0.027)
+	+	+	+	0.627 (0.021)	0.410 (0.027)	0.336 (0.063)	0.363 (0.036)

4. Discussion

In this study, we used unsupervised graph representation learning on SS-MIX2 standardized storage data recorded during hospitalization to convert it into a fixed-length embedded representation. We also confirmed that the obtained embedded representation partially contained information enabling successful prediction of readmission within 30 days of discharge.

An important limitation of this study is that the results obtained serve only as a proof of concept due to the limited amount of information transformed into graphs. In addition, the graph structure itself merits verification to obtain an embedded representation that is easier to characterize and more useful than the current one. Moreover, the design of model structures capable of extracting time-series features should be considered. In addition, in the prediction experiment, scheduled readmission and admission not anticipated in advance were not differentiated—this should be done in future works.

5. Conclusions

Using unsupervised graph representation learning, we developed a method to transform patient information recorded during hospitalization into a fixed-length embedded representation that preserves clinical transitional features. The obtained embedded representations were demonstrated to contain adequate clinical features of patients.

References

[1] Rehman A, Naz S, Razzak I. Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimed Syst. 2022 Aug;28(4):1339-71, doi: 10.1007/s00530-020-00736-8.

[2] Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, Zheng S, Xu A, Lyu J. Brief introduction of medical database and data mining technology in big data era. J Evid-Based Med. 2020 Feb;13(1):57-69, doi: 10.1111/jebm.12373.

[3] Sharafoddini A, Dubin JA, Lee J. Patient similarity in prediction models based on health data: a scoping review. JMIR Med Inform. 2017 Mar;5(1):e6730, doi:10.2196/medinform.6730.

[4] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436-44, doi: 10.1038/nature14539.

[5] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. InProceedings of the IEEE international conference on computer vision 2015. 1026-1034 p, doi: 10.48550/arXiv.1502.01852.

[6] Kawazoe Y, Imai T, Ohe K. A querying method over RDF-ized health level seven v2. 5 messages using life science knowledge resources. JMIR Med Inform. 2016 Apr;4(2):e5275, doi:10.2196/medinform.5275.

[7] Kimura M, Nakayasu K, Ohshima Y, Fujita N, Nakashima N, Jozaki H, Numano T, Shimizu T, Shimomura M, Sasaki F, Fujiki T. SS-MIX: a ministry project to promote standardized healthcare information exchange. Methods Inf Med. 2011;50(02):131-9, doi:10.3414/ME10-01-0015.

[8] Narayanan A, Chandramohan M, Venkatesan R, Chen L, Liu Y, Jaiswal S. graph2vec: Learning distributed representations of graphs. arXiv preprint arXiv:1707.05005. 2017 Jul, doi: 10.48550/arXiv.1707.05005.