Generate the concept representation using OMOP ontology graph

Junghwan Lee[1], Cong Liu, PhD[1], Ning Shang, PhD[1], Xinzhuo Jiang, MS[1], Kai Chen[1],
Krishna Sai Dheeraj Kalluri, MS[1], Chao Pang, PhD[1], Karthik Natarajan, PhD*[1],
Patrick Ryan, PhD*[1*], Chunhua Weng, PhD*[1]

[1]Columbia University Medical Center, New York, NY

Abstract

The OMOP Common Data Model (CDM) allows for the transformation of various medical data types into a common data format and for the sharing of source algorithms and analysis pipelines. Many of these analysis pipelines are related to observational studies, but more recently, newly developed machine learning algorithms such as deep neural networks have been applied to the OMOP CDM. The goal of this study is to generate a robust distributed representation for any concept defined in OMOP CDM and allow this representation to be used as an input for the deep learning tasks based on the OMOP CDM. We applied the node2vec algorithm to learn distributed representations of concepts based on the graph structure as defined in the concept_relationship table in the OMOP CDM. We then evaluated the learned representation using three different metrics: phenotype concept-set, ontology distance, and lexical similarity. For a given pair of concepts, cosine similarity of their learned representations are higher if they are from a defined concept-set, close to each other in the ontology graph, or share a similar string.

1. INTRODUCTION

Learning feature representations have been shown to be useful in EHR-based machine learning tasks. For example, Med2Vec showed strong performance in the prediction of clinical applications by considering the hierarchical structure of Electronic Health Record (EHR) while learning feature representations of medical concepts[3]. Other related works, such as cui2vec or hpo embeddings, were also significantly improved after learning proper feature representations [1,7]. While the OMOP Common Data Model (CDM) provides a unifying data format for applying various analysis pipelines[5], we expect that proper feature representations can be helpful when applying OMOP concepts to various machine learning tasks particularly to deep learning[2]. Further, these learned representation could be easily and crucially shared among OMOP community.

2. METHOD

Node2vec is an algorithm that learns continuous representation for nodes in a graph[4]. Similar to the word2vec[6], it learns distributed representations of the nodes in a graph taking both homophily and structural equivalence into account based on random walk. In this research, we generated distributed representations for concepts using node2vec. We constructed a graph structure (i.e. concept network) of the concepts by using the concept_relationship table defined in the OMOP CDM. Figure 1 depicts the structure of concept graph. Based on the constructed graph that consists of 2,349,171 concept nodes and 15,705,202 edges, distributed representations of 128 dimensions for the concept nodes were learned. We set d=128, r=10, l=80, k=10, and the optimization is run for a single epoch. Cosine similarities for each pair of node vectors were then calculated.

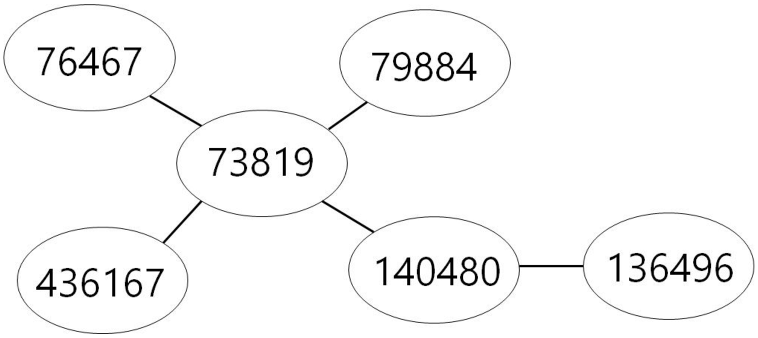
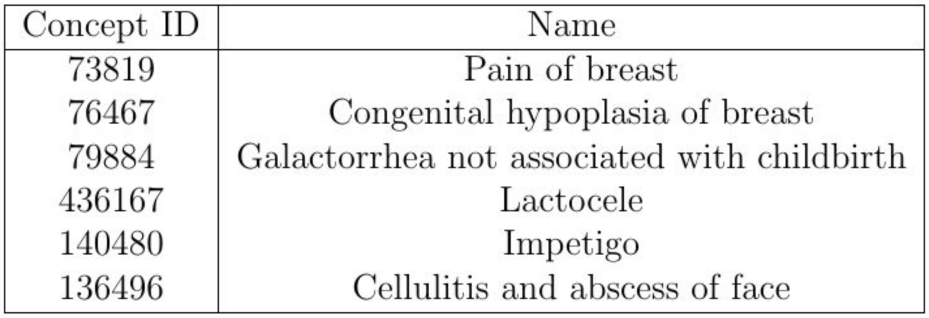


Figure 1. Example of graph constructed by using six concepts. Two concepts are connected with a line if there exists a relationship in concept_relationship table in OMOP CDM.

3. RESULTS

Concept codes from 53 validated phenotype algorithms containing defined concept-sets from the eMERGE Network were extracted to create positive and negative pairs. Positive pairs were defined as any two concepts belonging to the same concept-set, and negative pairs were defined as any two concepts not belonging to any of the phenotypes. Figure 2 showed that positive pairs have higher cosine similarities than negative node pairs. The median cosine similarity of the negative pairs and positive pairs is 0.712 and 0.844, respectively.

Table 1. The six concepts used to construct the graph in Figure 1.



Concept ID	Name
73819	Pain of breast
76467	Congenital hypoplasia of breast
79884	Galactorrhea not associated with childbirth
436167	Lactocele
140480	Impetigo
136496	Cellulitis and abscess of face

We then calculated the distance of any pair of concepts according to min_distance defined in concept_relationship table. In addition, we calculated the lexical similarity ratio of two concept strings. In Figure 3 (left), it is shown that the cosine similarity between two concepts decreases if the distance between the two nodes increases. The median of cosine similarity for each node distance group is 0.942, 0.937, 0.859, and 0.810 respectively. In Figure 3 (right), it is shown that cosine similarity increases as lexical similarity between two concepts increases. For simplicity and visualization, the lexical similarity ratios were broken into deciles.

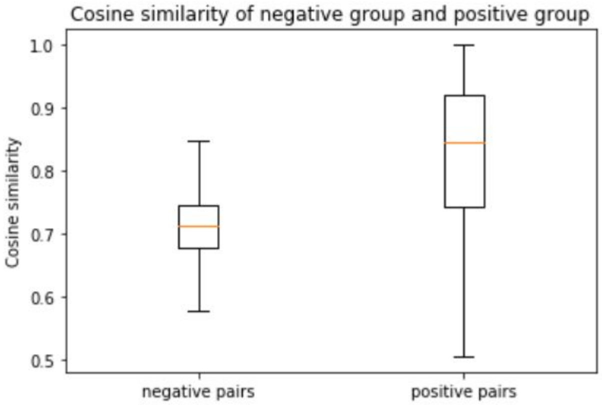


Figure 2. Cosine similarity in two different groups.

4. CONCLUSION

Based on the evaluation results, the learned representations of OMOP concepts reflect the graph structure and semantic properties of concepts in OMOP CDM well. The representation of the concepts can be shared among the OMOP community for various machine learning tasks particularly for deep learning.

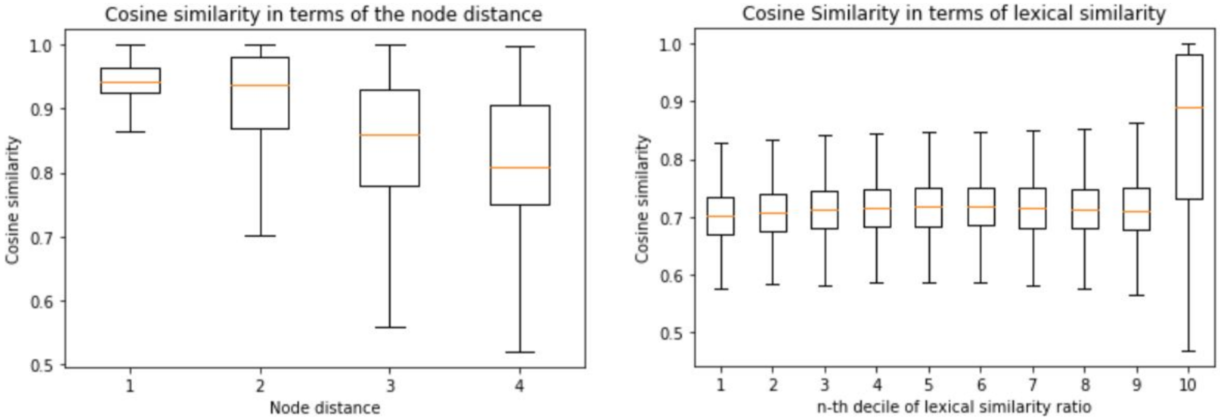


Figure 3. Cosine similarity across different node distances and lexical similarities.

REFERENCES

1. Beam A.L, et al. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. arXiv:1804.01486. 2018.
2. Bengio Y, Courvile A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(8):1798-1828
3. Choi E, et al. Multi-layer Representation Learning for Medical Concepts. KDD. 2016;1495-1504
4. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
5. Hripcsak G, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–578.
6. Mikolov T, et al. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013;26:3111–3119

7. Shen F, et al. Constructing Node Embeddings for Human Phenotype Ontology to Assist Phenotypic Similarity Measurement. IEEE International Conference on Healthcare Informatics Workshop (ICHI-W). 2018.