2018 IEEE International Conference on Healthcare Informatics Workshop

Constructing Node Embeddings for Human Phenotype Ontology to Assist Phenotypic Similarity Measurement

Feichen Shen, Ph.D.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
shen.feichen@mayo.edu

Sijia Liu, M.S.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
liu.sijia@mayo.edu

Yanshan Wang, Ph.D.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
wang.yanshan@mayo.edu

Liwei Wang, M.D., Ph.D.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
wang.liwei@mayo.edu

Andrew Wen, M.S.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
wen.andrew@mayo.edu

Andrew H. Limper, M.D.
Division of Pulmonary and Critical Care Medicine
Mayo Clinic
Rochester, MN
limper.andrew@mayo.edu

Hongfang Liu, Ph.D.
Department of Health Sciences Research
Mayo Clinic
Rochester, MN
liu.hongfang@mayo.edu

Abstract— **The Human Phenotype Ontology (HPO) was developed to be a semantically computable vocabulary that captures the phenotypic abnormalities found in human diseases discovered through biomedical research. Usage of this ontology facilitates the translation between genotype and phenotype. Many studies have been conducted to accelerate the implementation of precision medicine into clinical practice by utilizing the informative contents provided in the HPO. No work, however, has been done in constructing a distributed representation for nodes in HPO to provide a deep insight of phenotypic similarities by analyzing its graph structure. Node2vec is a model for generating node embeddings based on large networks. In this study, we constructed node embeddings for the HPO leveraging node2vec to assist phenotypic similarity measurement. A downstream application on link prediction driven by HPO embedding achieved 0.81 ROAUC and 0.75 F-measure. A use case study was conducted on idiopathic pulmonary fibrosis (IPF) and we demonstrated the potential possibility of using HPO embeddings in assisting phenotypic similarity measurement.**

Keyword—Human Phenotype Ontology, node embeddings, phenotypic similarity

I. INTRODUCTION

As a tool for annotating human phenotypic abnormalities, the Human Phenotype Ontology (HPO) [1] was developed as a controlled vocabulary for phenotypes by mining and integrating phenotype knowledge from a large variety of sources, including medical literature, Orphanet [2], the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) [3], and Online Mendelian Inheritance in Man (OMIM) [4]. The HPO also provides associations with other biomedical resources such as the Gene Ontology [5].

In our previous work, we used the HPO to annotate a large collection of clinical narratives [6] and proposed a collaborative filtering model [7] using patients' phenotypic information to provide rare disease diagnosis recommendations [8]. However, one limitation of this work was that we only considered the HPO as a dictionary without fully utilizing its graph structure to calculate phenotypic similarity for automatic diagnosis recommendation. Another limitation was that we only treated patients' phenotypic information as binary data (i.e., patients either did or did not have a specific phenotype), which is an oversimplification without consideration of distributed semantics among different HPO terms.

Recently, deep learning has been widely adopted as a feature representation method due to its ability to transform raw data into a high level representation or feature vector [9]. Word embeddings are one of the most successful applications in deep learning for capturing the high level semantic properties of words [10, 11]. Specifically, the Skip-gram model introduced by Mikolov et al. [12] provided an efficient method for generating vector representations for large amounts of unstructured data for the purpose of natural language processing (NLP).

It is therefore natural to consider adopting a similar approach on graph data to learn feature representations for networks. The node2vec model was proposed by Grover et al. to address this consideration so as to provide a scalable semi-supervised algorithm to represent node embeddings for complex networks [13].

In this study, we applied the node2vec model on the HPO graph to build node embeddings based on homophily and structural analysis [14], with the end goal of generating

978-1-5386-6777-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI-W.2018.00011

distributed semantic representations for each node in the HPO and assisting phenotypic similarity measurement. To tune hyperparameters and provide optimal embeddings, we conducted a downstream experiment on link prediction with various metrics. We conducted a use case study of the generated embeddings on a rare disease, idiopathic pulmonary fibrosis (IPF), to evaluate the viability of using the generated HPO embedding to provide a differential diagnosis based on a distributed phenotypic similarity measurement.

The paper is organized as follows. We first describe the proposed methods for constructing HPO embeddings in Section II and then presented the experimental results in Section III, followed by a discussion on our results and a conclusion.

II. METHODS

We present the workflow of this study in two parts. We first describe data preprocessing and preparation tasks performed on the HPO data, including feature selection and link identification, and then briefly introduce how the node2vec model can be applied on prepared input data to generate embeddings.

A. Preprocessing

Feature Selection

In this study, we aimed to use genetic information as features to describe each phenotype. An annotation file provided by the HPO team recorded the specific genes (via entrez ID and gene symbol) known to be associated with most of phenotypic terms contained in the HPO. We used these identified gene-phenotype associations as gene features for phenotypes.

Link Identification

Since the only relationship between nodes maintained by the HPO ontology is represented by “is a”, we treated each relationship as a directed relationship from subclass to superclass, and we denoted subclass and superclass as source node and target node, respectively.

B. Node2vec

Similar to the word2vec model, the node2vec model accepts as input each node's feature representation (analogue to single word in word2vec) and produces output as a group of neighbor nodes (analogue to context in word2vec). However, unlike unstructured text data, graph data is represented in a non-linear manner, and as such a richer notion for neighborhood (context) selection is needed. Specifically, node2vec considers two types of equivalences in a graph: homophily and structural equivalence [14]. In homophily equivalence, node2vec adopts a Breadth-first Sampling (BFS) searching strategy [15] to find similar node embeddings based on how similar they belonged to the same cluster or communities [16, 17], while in structural equivalence [18], node2vec implements a Depth-first Sampling (DFS) algorithm [19] to detect similar node embeddings based on the presence of shared structural roles within the same community. Most graphs maintain a mixture of homophily and structural relationships, whereas node2vec provides a flexible way to switch between the BFS and DFS to balance the graph searching.

Random Walk

This step was used to generate neighborhoods of each specific node in order to deliver context information to the following Skip-gram model. As defined in node2vec [13], for any source node s in the HPO, node2vec simulates a random walk from s to any target node t within a fixed length. The normalized transition probability TP(s, t) of random walk is defined as:

$$TP(s,t) = \begin{cases} \frac{\alpha(s,t)\cdot w(s,t)}{Z} & if\ a\ link < s,t > exists \\ 0 & otherwise \end{cases} \quad \text{Eq (1)}$$

where α(s,t) is a search bias term, w(s,t) is the weightage assigned for each link, and Z is the normalizing constant. Specifically in this study, we consider w(s,t) as 1 for all links.

The search bias term α plays an important role to switch between the BFS and DFS strategies, in order to make a good balance between homophily and structural equivalence. Node2vec applies two parameter p and q in a 2nd order random walk to control bias:

$$\alpha(s',t) = \begin{cases} \frac{1}{p} & if\ sd(s',t) = 0 \\ 1 & if\ sd(s',t) = 1 \\ \frac{1}{q} & if\ sd(s',t) = 2 \end{cases} \quad \text{Eq (2)}$$

where s’ is the node that on the path between node s and t. <s,s’> forms a link and <s’,t> forms another link. sd(s’,t) denotes the shortest path between node s’ and t. The return parameter p is used to control re-visitations of nodes and the in-out parameter q is responsible for searching between BFS and DFS. The combination of p and q determines the set of neighborhood nodes given any input.

Embeddings Generation

Node2vec extended the Skip-gram architecture to graph data for the generation of node embeddings [13]. The objective function used for this step is shown in Eq (3)

$$\max_f \sum_{n\in V} \log P(L|f(n)) \quad \text{Eq (3)}$$

where n indicates any node in the graph that belongs to vertices set V, L denotes all neighborhood nodes selected for n, and f is the function for feature representation.

As shown in Eq (4), a softmax function was applied to output a vector of normalized probabilities for each neighbor li and input node feature f(n):

$$P\big(l_i\big|f(n)\big) = \frac{\exp\left(f(l_i)\cdot f(n)\right)}{\sum_{v\in V}\exp\left(f(v)\cdot f(n)\right)} \quad \text{Eq (4)}$$

Combining Eq (3) and (4), the objective function was thus simplified as shown in Eq (5), where T_n=$\sum_{v\in V}\exp\left(f(v)\cdot f(n)\right)$. We then used Stochastic gradient descent to optimize it.

$$\max_f \sum_{n\in V}[-\log T_n + \sum_{l_i\in L(n)} f(l_i)\cdot f(n)] \quad \text{Eq (5)}$$

III. EXPERIMENTS

The HPO contains about 11,000 nodes in total and the gene annotation file involves 7,280 HPO nodes. In this study, we therefore only selected a subset of the entire HPO nodes and resulted in 7,280 nodes with annotated gene information. We deleted any isolated nodes and formed a subgraph with 7,258 nodes and 8,999 edges. In addition, 3,430 genes were defined within the gene annotation file to prepare a feature matrix for each HPO node.

We used link prediction as a downstream application to select the optimal hyperparameters for constructing node embeddings. We prepared positive and negative links to conduct this experiment. For positive examples, we randomly removed 50% of the links from the HPO subgraph while still verifying that the remaining subgraph was still connected. For negative examples, we randomly created the same number of fake links between any two nodes from the HPO subgraph that have no edge connecting them. Specifically, we created 9,750 training edges, 1,625 validation edges, and 4,875 testing edges for both positive and negative examples, respectively. We used four different methods to generate edge embeddings (Table 1) and applied 10-fold cross-validation with logistic regression on the validation edges and fitted the model on testing edges for evaluation purpose. In addition to plotting a ROC curve for displaying link prediction performance, we also used precision, recall, and F-measure to quantify prediction performance as defined in Eq (6)-(8).

TABLE 1 Four methods for edge embeddings generation. f(u) and f(v) denote feature representation for node u and v.

Methods	Definition
Average	$\frac{f(u)+f(v)}{2}$
Hadamard	$f(u) * f(v)$
Weighted-L1	$\lvert f(u)-f(v)\rvert$
Weighted-L2	$\lvert f(u)-f(v)\rvert^{2}$

$$Precision = \frac{|\{Correct\ Results\} \cap \{Predicted\ Results\}|}{|\{Predicted\ Results\}|} \quad \text{Eq (6)}$$

$$Recall = \frac{|\{Correct\ Results\} \cap \{Predicted\ Results\}|}{|\{Correct\ Results\}|} \quad \text{Eq (7)}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad \text{Eq (8)}$$

After constructing node embeddings for the HPO, we conducted a use case study on idiopathic pulmonary fibrosis (IPF) to demonstrate phenotypic similarity measurement using embeddings. We treated embedding for each node as a vector and applied cosine similarity on pair-wise of nodes to compute similarity between phenotypes.

IV. RESULTS

A. Link Prediction

The optimal hyperparameters we found for link prediction with the HPO node embeddings are shown in Table 2. We found that less revisits with a larger p and more DFS with smaller q achieved the best performance. In addition, we found that a longer walk length hampered the performance and thus distance 5 was determined as the optimal option. For 3,430 gene feature dimensions, we found that setting embedding dimension as 128 produced better prediction outputs.

TABLE 2 Optimal hyperparameters for generating node embeddings for the HPO

Hyperparameters	Optimal Value
p	1
q	0.05
neighbor size	10
# of walks	10
walk length	5
dimensions	128

Figure 1 shows the ROC curve for each of the different edge embedding strategies on the performance of link prediction. We found that Hadamard yielded the lowest AUC as 0.76. The weighted-L1 strategy achieved a slightly better AUC of 0.81 compared to all other methods.

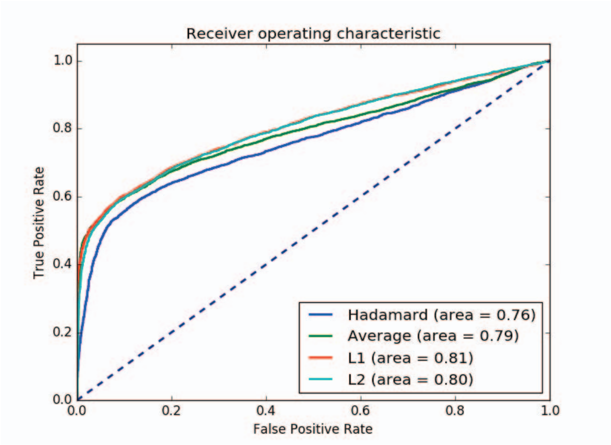

Fig 1 ROC curve for link prediction

Precision, recall and F-measure among four edge embedding methods are described in Table 3. We drew a consistent conclusion between ROAUC and F-measure, indicating that Weighted-L1 was the optimal method found for edge embedding. The hyperparameters selected using this strategy were thus also considered as optimal for embedding generation.

TABLE 3 Precision, recall and F-measure for different edge embedding methods

Methods	Precision	Recall	F-measure
Average	0.76	0.75	0.74
Hadamard	0.74	0.73	0.72

Weighted-L1	0.76	0.75	0.75
Weighted-L2	0.78	0.75	0.74

B. Use Case Study on IPF

IPF is a type of lung disease that results in fibrosis of the lungs for unknown reasons. IPF is considered a rare disease due to its low prevalence among the worldwide population [20]. In the HPO, 11 phenotypes are identified as relevant observable characteristics for IPF, but only 7 out of 11 are annotated with gene information, as shown in Table 4. For each relevant phenotype, we listed the top 5 most similar phenotypes ranked by cosine similarity. For gastroesophageal reflux, we found that even the top similar phenotype only had a shared similarity of 0.56, indicating that gastroesophageal reflux plays only a moderate role in the graph. Bronchiectasis had a higher similarity with abnormal bronchus morphology and bronchomalacia, which are all are associated with bronchiectasis related symptoms. For pulmonary fibrosis, cough and clubbing of fingers, all five similar phenotypes were considered to be relevant. We also observed some similar phenotypes that were not considered as regularly associated symptoms for the given phenotype. For example, based on our embeddings, exertional dyspnea is thought to be related to hypoplastic distal radial epiphyses, and pulmonary insufficiency is associated with calf muscle hypertrophy. These cases could be worth further exploration as they could provide potential evidence for differential diagnosis.

TABLE 4 Phenotypic similarity for seven IPF related phenotypes

IPF-related Phenotype	**Similar Phenotypes**	**Similarity Score**
gastroesophageal reflux	Prominent interphalangeal joints	0.56
	Abnormality of body height	0.53
	Depigmented fundus	0.52
	Wrist swelling	0.52
	Facial wrinkling	0.51
bronchiectasis	Abnormal bronchus morphology	0.96
	Bronchomalacia	0.93
	Recurrent bronchitis	0.89
	Recurrent upper respiratory tract infections	0.79
	Recurrent lower respiratory tract infections	0.77
pulmonary fibrosis	Atelectasis	0.9
	Pulmonary hypoplasia	0.89
	Abnormal lung morphology	0.89
	Alveolar proteinosis	0.88
	Unilateral primary pulmonary dysgenesis	0.88
exertional dyspnea	Dyspnea	0.98
	Respiratory distress	0.94
	Hypoplastic distal radial epiphyses	0.47
	Spastic dysarthria	0.46
	Nonketotic hyperglycinemia	0.45
pulmonary insufficiency	Abnormal pulmonary valve physiology	0.97
	Pulmonic stenosis	0.93
	Renal steatosis	0.5
	Calf muscle hypertrophy	0.48
	Prominent coccyx	0.47
cough	Neonatal breathing dysregulation	0.93
	Abnormal breath sound	0.92
	Snoring	0.92
	Abnormal blood gas level	0.92
	Restrictive ventilatory defect	0.91
clubbing of fingers	Abnormality of the fingertips	0.93
	Broad fingertip	0.86
	Abnormality of finger	0.81
	Finger swelling	0.78
	Swan neck-like deformities of the fingers	0.76

V. DISCUSSION AND FUTURE WORK

We used 3,430 genes as features to generate HPO embeddings. Too many features, however, may lead to a sparse matrix to build the distributed representations. In future work, we will leverage the Gene Ontology to merge some genes with fine granularity at a higher level, in order to shrink the feature space.

One limitation for the HPO graph was that it only contained “is a” relationships to record hierarchical associations. As such, there were no links between any two sibling nodes. Some informative predicates such as SameAs, hasSynonym, seeAlso could be introduced to create a more complex network with semantic meaning. In the future, we will enrich the HPO graph with more such sibling relationships by literature mining and incorporating heterogeneous knowledge bases.

The node2vec model only considered a 2nd order random walk ranging in {0, 1, 2}. In the future, we will include a more general number of steps (n) in one random walk and optimize n using a dynamic programming approach [21, 22].

VI. CONCLUSION

In this study, we investigated applying the node2vec model on the HPO to generate node embeddings. Experiments in our study showed using link prediction to generate the optimal embeddings and demonstrated the viability of using the HPO embeddings to assist phenotypic similarity measurement.

ACKNOWLEDGEMENT

This manuscript was made possible by joint funding from National Institute of Health grants U01TR02062 and NIH R01 EB19403-1.

REFERENCE

[1] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics. 2008;83:610-5.

[2] Weinreich SS, Mangon R, Sikkens J, Teeuw M, Cornel M. Orphanet: a European database for rare diseases. Nederlands tijdschrift voor geneeskunde. 2008;152:518-9.

[3] Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. The American Journal of Human Genetics. 2009;84:524-33.

[4] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005;33:D514-D7.

[5] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000;25:25.

[6] Shen F, Wang L, Liu H. Phenotypic Analysis of Clinical Narratives Using Human Phenotype Ontology. Studies in health technology and informatics. 2017;245:581-5.

[7] Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence: Morgan Kaufmann Publishers Inc.; 1998. p. 43-52.

[8] Shen F, Liu S, Wang Y, Wang L, Afzal N, Liu H. Leveraging Collaborative Filtering to Accelerate Rare Disease Diagnosis. American Medical Informatics Association, Washington DC. 2017.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521:436.

[10] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems2013. p. 3111-9.

[11] Wang Y, Liu S, Rastegar-Mojarad M, Wang L, Shen F, Liu F, et al. Dependency and AMR Embeddings for Drug-Drug Interaction Extraction from Biomedical Literature. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics: ACM; 2017. p. 36-43.

[12] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

[13] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining: ACM; 2016. p. 855-64.

[14] Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. Journal of the american Statistical association. 2002;97:1090-8.

[15] Bundy A, Wallen L. Breadth-first search. Catalogue of Artificial Intelligence Tools: Springer; 1984. p. 13-.

[16] Fortunato S. Community detection in graphs. Physics reports. 2010;486:75-174.

[17] Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1: Association for Computational Linguistics; 2003. p. 173-80.

[18] Henderson K, Gallagher B, Eliassi-Rad T, Tong H, Basu S, Akoglu L, et al. Rolx: structural role extraction & mining in large graphs. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining: ACM; 2012. p. 1231-9.

[19] Tarjan R. Depth-first search and linear graph algorithms. SIAM journal on computing. 1972;1:146-60.

[20] Nalysnyk L, Cid-Ruzafa J, Rotella P, Esser D. Incidence and prevalence of idiopathic pulmonary fibrosis: review of the literature. European Respiratory Review. 2012;21:355-61.

[21] Shen F, Lee Y. Knowledge discovery from biomedical ontologies in cross domains. PloS one. 2016;11:e0160005.

[22] Shen F, Liu H, Sohn S, Larson DW, Lee Y. Predicate Oriented Pattern Analysis for Biomedical Knowledge Discovery. Intelligent information management. 2016;8:66.