Graph Learning for Prediction of Drug-Disease Interactions: Preliminary Results

Andrej Kastrin* and Dimitar Hristovski†
Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana
Ljubljana, Slovenia
Email: *andrej.kastrin@mf.uni-lj.si, †dimitar.hristovski@mf.uni-lj.si

***Abstract*—One of the fundamental problems to complex network research is understanding of link formation. We study the problem of representation learning in a bipartite drug-disease network of semantic predications extracted from biomedical literature. We employ DeepWalk and node2vec node embedding methods with deep learning link predictor, as well as standard baseline predictors including common neighbors, Jaccard coefficient, and Adamic/Adar. Experimental results show that both network embedding algorithms outperform traditional link predictors.**

Keywords–Complex networks; Network analysis; Network learning; MEDLINE.

I. Introduction

The corpus of biomedical papers is growing at an exponential rate. For instance, MEDLINE [1], the largest bibliographic database in biomedicine, at the time of this writing, aggregates more than 28 million citations to life science papers. However, a significant amount of potentially useful knowledge still remains undiscovered. It is hard to synthesize divergent research evidence into coherently interpretable knowledge. Here, we tackle the problem of unravelling hidden relations between drugs and diseases using link prediction methodology from biomedical literature.

An elementary problem in graph research is the analysis of the connections between the nodes. In computer science and statistics, this is known as link prediction problem. Although novel research area, link prediction attracted numerous researchers in the last decade. Link prediction refers to the discovery of relations between nodes that are not connected in the current snapshot of a given network but will be connected in the future. The aim of link prediction in general is to estimate the probability that a link exists among a pair of nodes, based on the topology of existing nodes, edges, and their attributes.

In the case of link prediction, we need to encode pairwise properties between nodes, such as the number of common neighbors or relationship strength. Traditional approaches rely on summary network statistics (e.g., centrality measures) to extract structural information from networks. However, recently, approaches have emerged that seek to learn representations that encode structural information about the network and present a powerful alternative to traditional feature engineering. The general idea behind representation learning is to learn a mapping that embeds nodes as points in a low-dimensional vector space (i.e., embedding space). The goal is to optimize this mapping so that geometric relationships in this space reflect the structure of the original network. The learned embeddings can then be used as feature inputs for machine learning tasks. Embedding methods can be generally categorized into three groups [2]: (i) factorization methods (e.g., locally linear embedding [3]), (ii) random walk techniques (e.g., DeepWalk [4], node2vec [5]), and (iii) deep learning (e.g., structural deep network embedding [6]).

In this work, we investigate the performance of two network embedding algorithms, namely DeepWalk and node2vec. More formally, we examine, how neural network predictor, using computed embeddings from DeepWalk and node2vec, behaves on the task of link prediction in a large-scale network of semantic predications extracted from biomedical literature. In addition, we examine these methods in contrast to traditional baseline predictors such as common neighbors, Jaccard coefficient, and Adamic/Adar.

The rest of the paper is structured as follows. In Section II, we present dataset and methodology used in this study. Results are presented in Section III. Finally, we conclude in Section IV.

II. Methods

A. Dataset

SemRep is a symbolic natural language processing tool that extracts semantic predications from MEDLINE citations [7]. A predication is a formal representation of textual content that consists of a subject, predicate, and object. Subject and object arguments are concepts from the Unified Medical Language System (UMLS) Metathesaurus [8] as available through MetaMap [9]. The predicate is from UMLS Semantic Network [10]. These predications provide a normalized representation of the meaning of the source text in a machine-readable form for automatic processing. SemRep extracts about 30 predicate types, related to clinical medicine (e.g., `TREATS`, `DIAGNOSES`), substance interactions (e.g., `INTERACTS_WITH`, `STIMULATES`), genetic etiology of disease (e.g., `ASSOCIATED_WITH`), and pharmacogenomics (e.g., `AFFECTS`). In this paper, we focus only on `TREATS` relation. We extract all `TREATS` relations that connect drugs (Metathesaurus concepts with semantic type “Pharmacologic Substance”) and diseases (concepts with semantic type “Disease or Syndrome”).

B. Baseline Predictors

We implemented a link prediction baseline using various proximity measures, which are used to find similarity among a pair of nodes. Our assumption is that similar nodes are more likely to form a link in the future. For each non-observed link (u, v) in a testing network, a link prediction computes a score $s(u, v)$, which can be considered as an estimate of the presence of edge creation between nodes u and v. In our initial settings we used common neighbors, Jaccard coefficient, and Adamic/Adar.

ISBN: 978-1-61208-715-3

C. Network Embeddings

For network representation learning models, we used two state-of-the-art algorithms, namely DeepWalk [4] and node2vec [5]. Both methods employ random walk algorithm to sample topological properties and node representations are learned to preserve pairwise similarities of nodes.

DeepWalk learns an embedding by sampling random walks from each node and applying skip-gram learning on those walks. We use the default parameters described in the seminal paper, i.e., walk length $t = 80$, number of walks per node $\gamma = 80$, and window size $w = 10$. node2vec improves the random walk step of DeepWalk by defining hyperparameters p and q that control the depth and breadth of random walks, respectively. The special case with parameters $p = 1$ and $q = 1$ corresponds to DeepWalk. In our settings we used the same values for parameters as for DeepWalk; the remaining parameters were set to $p = 2$ and $q = 4$.

After network embedding, we need to set up a statistical classification framework for link prediction. Both algorithms, DeepWalk and node2vec, described above, are designed to learn feature representations for nodes in a network. However, in our study we are interested in prediction involving pairs of nodes and not individual nodes. To this end, we need to define a binary operator $\circ$ over the corresponding feature vectors $f(u)$ and $f(v)$ in order to generate a composite representation $g(u, v)$. We consider three different alternatives for the $\circ$ operator:

1) concatenation: $u_i + v_i$,
2) average: $(u_i + v_i)/2$, and
3) Hadamard product: $(u_i * v_i)$,

where u and v are two vectors and u_i and v_i are i-th element of u and v, respectively.

D. Machine Learning

For the classification task, we use deep learning model implemented in TensorFlow [11] as feed-forward neural network with a single hidden layer. Input to the model is a vector representation with the binary operator defined above. The output of the model is a probability of a link formation between the input nodes. We draw a fixed proportion of the existing edges for training, and use the rest of edges for the testing. Training was defined for a time period from 1843 to 2003 and testing for a time range from 2004 to 2018.

In this study, we used the following five measures to compare the performance of the statistical learning: area under a receiver operating characteristic curve (AUROC), area under a precision-recall curve (AUPR), mean average precision (mAP), and precision at k (Prec@k).

III. Results and Discussion

In our experiment we used the knowledge network, constructed as a subset of SemMedDB network [12], as defined previously in the Methods section. The network comprises 13,182 unique vertices that refer to drugs and 8856 vertices that refer to diseases. In total, there were 170,707 relations between both sets of nodes. The mean degree of the bipartite network was 12.95 links and the average path length was 1.76 hops.

The results in terms of classification performances of the performed experiment are summarized in Table I. The best performer across all four performance measures is node2vec with average merge type. If we consider only AUROC and AUPR measures, the baseline predictors are slightly better than DeepWalk and node2vec. Common neighbors measure performs best, followed by Jaccard coefficient, and Adamic/Adar. mAP and Prec@k scores for DeepWalk and node2vec are an order of magnitude higher in comparison to baseline predictors.

TABLE I. Performance Measures of Link Prediction Algorithms

Method	Binary operator	AUROC	AUPR	Pred@k	mAP
CN	–	0.86	0.86	0.86	0.64
JC	–	0.85	0.84	0.86	0.62
AA	–	0.81	0.74	0.82	0.54
DeepWalk	Co	0.83	0.86	0.96	0.79
	Av	0.83	0.86	0.97	0.80
	Ha	0.72	0.72	0.82	0.65
node2vec	Co	0.83	0.86	0.96	0.80
	Av	0.83	0.86	0.97	0.81
	Ha	0.72	0.73	0.83	0.65

Note: CN = Common Neighbors, JC = Jaccard Coefficient, AA = Adamic/Adar; (Co)ncatenate, (Av)erage, and (Ha)damard merge type; further details are provided in text

As far as we know, this is the first work discussing knowledge network, network embeddings as well as deep learning approach to discover new drug-disease interactions from literature. Results of this study show that both network embedding algorithms, DeepWalk and node2vec, outperform traditional link predictors such as common neighbors or Jaccard coefficient.

IV. Conclusion

We investigate the representation learning in bipartite drug-disease network of semantic predications. We design a deep learning model that includes the network structure into the embedding. Experimental results demonstrated that performance measures in terms of AUROC and AUPR are comparable. However, we found evidence that DeepWalk and node2vec outperformed baseline predictors in terms of Pred@k and mAP measures.

References

[1] "PubMed," https://www.ncbi.nlm.nih.gov/pubmed, accessed: 2019-06-27.

[2] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," Knowledge-Based Systems, vol. 151, 2018, pp. 78–94.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, 2000, pp. 2323–2326.

[4] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014, pp. 701–710.

[5] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 855–864.

[6] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016, pp. 1225–1234.

[7] T. C. Rindflesch and M. Fiszman, “The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text,” Journal of Biomedical Informatics, vol. 36, no. 6, 2003, pp. 462–477.

[8] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” Nucleic Acids Research, vol. 32, no. Database issue, 2004, pp. D267–D270.

[9] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” Journal of the American Medical Informatics Association, vol. 17, no. 3, 2010, pp. 229–236.

[10] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, “The Unified Medical Language System,” Methods of Information in Medicine, vol. 32, Aug. 1993, pp. 281–291.

[11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., “Tensorflow: A system for large-scale machine learning,” in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16), 2016, pp. 265–283.

[12] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, “SemMedDB: A PubMed-scale repository of biomedical semantic predications,” Bioinformatics, vol. 28, no. 23, 2012, pp. 3158–3160.