Towards Better Drug Repositioning Using Joint Learning

Apoorva Vikram Singh
Department of
Electrical Engineering
National Institute of Technology, Silchar
Silchar, Assam 788010, India
Email: singhapoorva388@gmail.com

Atul Negi
School of Computer and
Information Sciences
University of Hyderabad
Hyderabad, Telengana 500046, India
Email: atulcs@uohyd.ernet.in

***Abstract*—Drug repositioning offers an economical and efficient alternative to traditional drug discovery. It means that, a drug approved for effect against a particular disease is considered and its applications for novel pharmaceutical purposes are explored in shorter development timelines. Unlike conventional approaches, this work attempts to explore the network of existing drugs and its unmapped indications by treating drug repositioning as a classification problem. The proposed classification model attempts estimation of the relevance of a drug with an unmapped indication. An enhanced word representation model is used for this purpose by integrating knowledge obtained from a structured biological knowledge graph and medical literature. To harvest the structured biological data, we have leveraged multiple biological ontologies to achieve a formal framework in the form of a semantic knowledge graph. Our novelty lies in that we have exploited knowledge from biological knowledge graph and medical corpora to complement each other. This makes our method competent with well established drug repositioning techniques.**

***Index Terms*—Drug Repositioning, Machine Learning, Word Embedding, Knowledge Graph, Ontology.**

I. Introduction

The pathway involved in *de novo* drug discovery and development is quite draining in terms of financial commitment and elapsed time along with the substantial risk bracketed with the same. Eastern Research Group (ERG) [1] reports that time window of 10-15 years is essential for the development of a drug including expenditure of US$ 2-3 billion with an average success rate of 2.01% [2]. In order to surpass these hindrances, drug repositioning [3] has been a significant notion in pharmaceutical research and development in recent years.

Ontology, in most universal terms, is defined as “specifications of conceptualization in a domain” [4]. The recent developments in Semantic Web technologies enabled us to integrate multiple ontologies formatted in Web Ontology Language (OWL) [5] into graph structured knowledge particularly using Resource Development Framework (RDF) [6].

We have used a knowledge graph [7] constructed by automated reasoning while leveraging Disease Ontology [8], Gene Ontology [9] and Human Phenotype Ontology [10]. We have further utilized neural network-based approaches to generate the representations of the nodes present in the graph to encode the semantic knowledge present within the knowledge graph. This representation (node embeddings) can be harnessed to predict edges (relations) between these nodes containing biological significance.

Combinations of different forms of data fed into a single predictive model has the potential of enhancing the efficiency of computational drug repositioning techniques. In a nutshell, our method involves harvesting of semantic, structured knowledge extracted from a knowledge graph, combined with the extracted features from unstructured medical corpus. Exploiting the procured features by means of supervised learning algorithms, we established that our model outperforms the utilization of individual features and adds a new dimension to computational drug repositioning techniques.

II. Related work

In past, the task of drug repositioning has been successfully carried out with scientific legitimacy. Thalidomide, Raloxifene and Sildenafil (Viagra) are some classical examples of repositioned drugs [11]. Computational approaches to drug repurposing can be predominantly grouped into chemoinformatics based approaches, side-effects based approaches, network based approaches, and text-mining based approaches.

One may consider the first order of configuration which states drugs are lipophilic molecules. Therefore, it is intuitive to study the chemical structure of drug molecules and claim that similar structures lead to similar biological outcomes. The identification of off-targets of drugs with adequate validation was implemented in 2009 [12].

Network based approaches intend to utilize the data sources containing biological entities (drugs, proteins, disease, etc.) in the same module of structured biological framework. DBSCAN [13] and OPTICS [14] presented cluster-based network approaches.

Text mining approaches are based on the presence of patterns in natural language for prediction of novel therapeutic drug applications. Tools like DrugQuest [15] has been developed to detect drug-drug relationships.

Concept combination approaches like PREDICT [16] combine similarity between drug on the basis of side-effects, and genes based on its sequence and functions. PREDICT

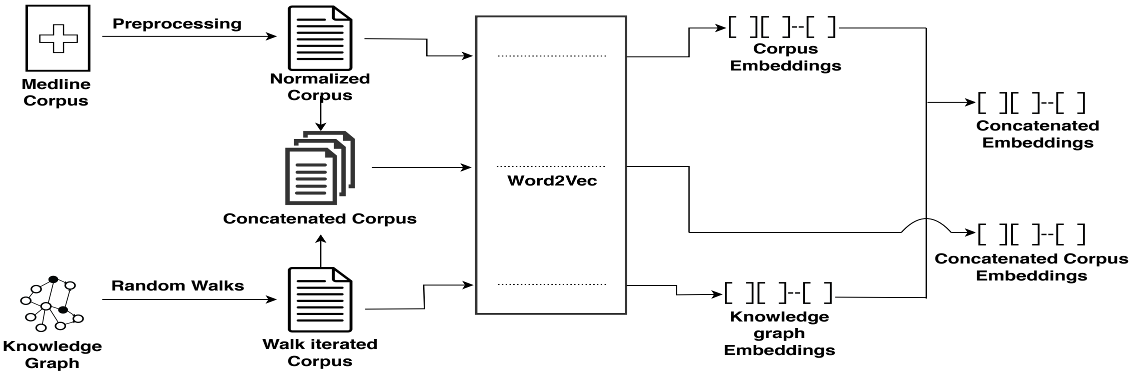

Fig. 1. Workflow for embedding creation

displayed promising results using these gold standard relations to predict real associations from the fake ones.

III. Our Approach: The proposed model

The proposed approach is novel in the aspect that it treats drug repositioning as a process of segregating drug-indication pairs of entities into positive and negative associations. For a better representation of concerned biological entities and their associations, we propose a joint learning technique. The technique [17] combines semantics from knowledge graphs with features from text data as illustrated in Fig. 1.

Joint learning based techniques have already been successfully applied in domains like neural machine translation, semantic parsing, relation extraction, etc. The drawback of using a knowledge graph for knowledge representation is the deprivation of rich contextual content while that of textual corpus is the exclusion of semantic wealth data has to offer. Joint learning based approaches assist in overcoming this anomaly by complementing and delivering on the drawbacks of these two techniques.

A binary classification model is used for supervising the problem that calculates a confidence score for the predictions made. Further, we arranged all the associations predicted based on their ranks of confidence scores. Our model was able to successfully classify the pairs into positive and negative ones. Predicted negative pairs with low confidence scores and predicted positive pairs with high confidence scores can be considered for drug repositioning by undertaking pharmaceutical research on them. This paves the way for risk-free and efficient drug repositioning with lesser investments.

IV. Dataset Description

A. Knowledge Graph Data

The approach uses a knowledge graph comprising of information about drugs, proteins/genes, diseases, and their relational characteristics. To construct this structured framework, three different ontologies have been used-Disease Ontology (DO) [8], Human Phenotype Ontology (HPO) [10] and Gene Ontology (GO) [9]. The biological entities present in the concerned knowledge graph can be classified into-Drugs, diseases or indications and genes. From the SIDER database [18], drug and its indications relationships are labeled. STITCH database [19] was used to label human chemical-protein interactions. Human proteins/genes interactions were obtained from STRING database [20].

A knowledge graph using RDF [6] was created from symbolic representation learning method [21]. Drug and target/indications as nodes and the association as edge were used. Also, all the information present in the knowledge graph is in the form of IRIs (International Resource Identifiers).

B. Corpus Data

For extracting features from the text, we have used Medline corpus made available by PubTator project [22]. The aforementioned corpus comprises of 27,599,238 abstracts of diseases, proteins/genes and drugs/chemicals. As a part of preprocessing, we have replaced the entity names (drugs/diseases/genes) with its corresponding IRIs for ease in integration of this data with knowledge extracted from graph.

C. Prediction

To find the relevance of the relationship between entities (drug and indication), we treat it like a classification problem. For this purpose, a dataset containing positive and negative relations is generated with the help of above-mentioned knowledge graph. The relations present in the knowledge graph are considered as positive. An equal number of weak or negative relations are selected i.e. when the relation between the two entities is not present in the knowledge graph.

V. Experimentation

A. Embedding Creation

We undertook comparative studies among three different state-of-the-art techniques to create node embeddings (as discussed in Table II) from the knowledge graph. The techniques included-FAIR Poincare embedding [23], Translating embeddings (TransE) [24] and modified DeepWalk [25]. By analyzing experimental results, we determined that the modified Deepwalk technique outperforms the remaining two methods.

For modified DeepWalk [25], the input is a knowledge graph with all drug-indication removed from the graph. The output

TABLE I
PERFORMANCE RESULTS FOR PREDICTION OF DRUG-INDICATIONS FOR DIFFERENT EMBEDDINGS

Classification Model	Knowledge Graph	Medline Corpus	Concatenated Corpus	Concatenated Embeddings
Logistic Regression	0.845	0.857	0.862	0.859
Random Forest	0.895	0.902	0.918	0.908
Arificial Neural Networks	0.881	0.887	0.921	0.917

is a corpus containing a set of edge-labeled iterated random walks. We generate a sentence for each node in the graph using a brief random walk. The corpus has been generated by keeping the length of walk 15 and iterating 40 walks per node. Now, to generate embeddings through this corpus, skip-gram Word2Vec model [26] has been used. The embeddings created had embedding size 128, window size 10 and negative sampling using 5 words.

The textual features of medical literature were inculcated in knowledge representation by the creation of vector representations of IRIs present in corpora [17]. For this, skip-gram based Word2Vec model and GloVe [27] were applied to the corpus. Word2Vec was able to marginally outperform GloVe achieving 90.2% over 89.9%. For the creation of vectors, a window size of 30, negative sampling 5 and embedding size of 128 was used for 15 epochs over the Medline literature.

The two standard methods used to this end, are concatenation of embeddings and concatenation of corpus to generate embeddings [28].

For concatenation of embeddings, as inferred by comparative studies conducted, we have utilized top performers on knowledge graph and Medline corpus for embedding creation. Modified DeepWalk [25] for knowledge graph and skip-gram Word2Vec for corpus have been used with that motivation.

On the other hand, to generate embeddings from the concatenated corpus, we have combined corpus obtained from the textual corpus and random iterated walks on knowledge graph. We have then employed skip-gram Word2Vec model [26] on the concatenated corpus to create embeddings for the entities.

B. Evaluation Details

We have compared the quality of embeddings produced using Knowledge Graph, Medline corpus, and joint learning techniques. This has been carried out by using three classification based supervised models: Logistic regression, Random Forest and Artificial Neural Network. Based on supervised learning methods, we trained the model on randomly selected 80% drug-indication associations and tested by predicting remaining associations between 20% data.

SIDER database [18] has been used as the evaluation set for drug indications. 2552 diseases were ranked for 754 drugs each (where each drug can have one or more indications). This provides a hint for indicated disease for each drug. It was ensured that all the entities used (drugs and diseases) had an overlapping presence in literature, Knowledge graph, and SIDER [18].

Each machine learning model consists of two embedding vectors as input out of which one represents a drug and the other one a disease. The model predicts if a valid association between drug and indication exists. A confidence score has been calculated for each prediction which has been utilized to rank predict the associations. Further, we calculated the area under the ROC (receiver operating characteristics) curve (ROCAUC) [29]. Table I summarizes the results obtained for three different classification models for different embeddings.

The random forest classifier having 50 trees is used with minimum of one training sample present in each leaf node. To measure the split quality, Gini Impurity index was used. The Logistic regression based classification model was used with the standard parameter settings mentioned in Python Scikit-learn. The Artificial Neural Network used contains a single hidden layer containing twice the size of input feature vector. Rectified Linear Unit (ReLU) has been used as activation function throughout the networks and Rmsprop [30] as optimizer.

TABLE II
PERFORMANCE RESULTS FOR KNOWLEDGE GRAPH EMBEDDINGS

Poincare	TransE	Modified DeepWalk
0.681	0.633	0.895

VI. Results and Discussions

We successfully achieved 0.921, 0.918 and 0.862 ROCAUC scores for concatenated corpus embeddings which were marginally higher than the scores for individual embeddings (Knowledge Graph/Corpus). Therefore, the joint learning technique had an upper hand for data representation as compared to individual methods.

Through this work, we were able to competently develop a technique through which we have projected data of different types into a vector space by means of feature learning. The work also establishes that combination of different modes of information serves as an effective technique for data representation

VII. Conclusion

The presented work proposes a unique workflow to comprehensively incorporate multifaceted data from medical literature and knowledge graph into predictive machine learning simultaneously. A comparison of three graph embedding methods was done for hypothesising drug repositioning.

Acknowledgment

We thank Ms. Divyansha for assistance with domain expertise and comments that greatly improved the manuscript.

References

[1] A. Sertkaya, A. Birkenbach, A. Berlind, and J. Eyraud, "Examination of clinical trial costs and barriers for drug development," *Report, US Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, Washington, DC*, pp. 1–92, 2014.

[2] H. Xue, J. Li, H. Xie, and Y. Wang, "Review of drug repositioning approaches and resources," *International journal of biological sciences*, vol. 14, no. 10, pp. 1232–1244, 2018.

[3] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, p. 673, 2004.

[4] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?" in *Handbook on ontologies*. Springer, 2009, pp. 1–17.

[5] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[6] K. S. Candan, H. Liu, and R. Suvarna, "Resource description framework: metadata and its applications," *ACM SIGKDD Explorations Newsletter*, vol. 3, no. 1, pp. 6–19, 2001.

[7] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs." *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 48, 2016.

[8] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein *et al.*, "Human disease ontology 2018 update: classification, content and workflow expansion," *Nucleic acids research*, vol. 47, no. D1, pp. D955–D962, 2018.

[9] D. P. Hill, B. Smith, M. S. McAndrews-Hill, and J. A. Blake, "Gene ontology annotations: what they mean and where they come from," in *BMC bioinformatics*, vol. 9, no. 5. BioMed Central, 2008, p. S2.

[10] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *The American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[11] J. S. Shim and J. O. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *International journal of biological sciences*, vol. 10, no. 7, p. 654, 2014.

[12] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, p. 175, 2009.

[13] I. Babur, J. Ahmad, B. Ahmad, and M. Habib, "Analysis of dbscan clustering technique on different datasets using weka tool," *Science International*, vol. 27, pp. 5087–5090, 12 2015.

[14] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *ACM Sigmod record*, vol. 28, no. 2. ACM, 1999, pp. 49–60.

[15] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, I. S. Vizirianakis, and I. Iliopoulos, "Drugquest-a text mining workflow for drug association discovery," *BMC bioinformatics*, vol. 17, no. 5, p. 182, 2016.

[16] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "Predict: a method for inferring novel drug indications with application to personalized medicine," *Molecular systems biology*, vol. 7, no. 1, 2011.

[17] M. Alsuhaibani, D. Bollegala, T. Maehara, and K.-i. Kawarabayashi, "Jointly learning word embeddings using a corpus and a knowledge base," *PloS one*, vol. 13, no. 3, p. e0193094, 2018.

[18] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular systems biology*, vol. 6, no. 1, 2010.

[19] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "Stitch: interaction networks of chemicals and proteins," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D684–D688, 2007.

[20] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Research*, vol. 45, no. D1, pp. D362–D368, oct 2016. [Online]. Available: https://doi.org/10.1093%2Fnar%2Fgkw937

[21] M. Alshahrani, M. A. Khan, O. Maddouri, A. R. Kinjo, N. Queralt-Rosinach, and R. Hoehndorf, "Neuro-symbolic representation learning on biological knowledge graphs," *Bioinformatics*, vol. 33, no. 17, pp. 2723–2730, 2017.

[22] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," *Nucleic acids research*, vol. 41, no. W1, pp. W518–W522, 2013.

[23] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, 2017, pp. 6338–6347.

[24] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, ser. AAAI'14. AAAI Press, 2014, pp. 1112–1119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2893873.2894046

[25] 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.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[27] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. [Online]. Available: https://doi.org/10.3115%2Fv1%2Fd14-1162

[28] J. Goikoetxea, E. Agirre, and A. Soroa, "Single or multiple? combining word representations independently learned from text and wordnet," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 2608–2614. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016266

[29] P. F. Pinsky, "Scaling of true and apparent ROC AUC with number of observations and number of variables," *Communications in Statistics - Simulation and Computation*, vol. 34, no. 3, pp. 771–781, jul 2005. [Online]. Available: https://doi.org/10.1081%2Fsac-200068366

[30] S. Ruder, "An overview of gradient descent optimization algorithms." 2016, cite arxiv:1609.04747Comment: Added derivations of AdaMax and Nadam. [Online]. Available: http://arxiv.org/abs/1609.04747