Is there Data Leakage in Protein-Protein Interaction Prediction using Knowledge Graphs?

Rita T. Sousa ✉, Sara Silva, Catia Pesquita

LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
{risousa,sgsilva,clpesquita}@ciencias.ulisboa.pt

Abstract. There is a high potential for data leakage in biomedical machine learning applications since biomedical data resources share, reuse and import data from each other routinely. We have investigated potential data leakage in the prediction of protein-protein interactions using the Gene Ontology knowledge graph, by comparing the performance of models trained and tested on the same versions of data versus training on archived data and predicting only for newly discovered protein interactions. Our results were not able to detect an influence of data leakage, indicating that if this problem exists, its magnitude is not affecting the performance of knowledge graph-based protein interaction predictions.

1 Introduction

Machine learning methods have become a significant trend in several research fields in recent years, and the semantic web is no exception. As machine learning is increasingly being used, concerns about data leakage have been raised [1]. Leakage occurs when information about the target of a data mining problem that should not be legitimately available to mine from is introduced [3], and it can lead to overestimation of the model's performance.

In biomedical applications, such as protein-protein interaction (PPI) prediction, data leakage can also be an issue. It is not uncommon that multiple databases and resources reuse the same sources of information. The majority of PPI prediction methods that are based on knowledge graphs (KGs) [7,11] explore the Gene Ontology (GO) KG that defines the universe of classes associated with proteins functions. The GO KG, composed of the GO [9] and GO annotations [2] that link proteins to GO classes, is continuously evolving as more data become available [10]. The majority of GO annotations are inferred by electronic annotation (IEA), which means they are based on the automated processing of other data sources. This could result in the same information that is used to support a PPI in a database (e.g. STRING [8]) to also be used to establish a GO annotation for the proteins.

We hypothesize that if this type of data leakage is common, then the performance of GO-based PPI prediction methods would be artificially increased. To test this hypothesis, we compare PPI prediction models trained on older GO data and PPI interactions and tested on previously unknown interactions captured in more recent versions of STRING with same version training and testing. Furthermore, by training the models on labeled examples from the past, we more closely simulate real-world applications.

2 Methods

PPI prediction is cast as a classification task that takes as input the GO KG and a set of protein pairs. The first step of our approach is using historical data to build the PPI datasets. Then we use the GO KG and the protein pairs to predict interactions using several machine learning algorithms.

2.1 Data

The PPI datasets were obtained from the STRING Database[1] which is one of the largest available PPI databases that integrates both physical interactions as well as functional associations between proteins collected from several sources. We considered the following criteria to select protein pairs from STRING: (i) each protein must be annotated with the GO; (ii) protein interactions must be experimentally determined or from curated databases (as opposed to computationally determined); (iii) interactions must have a confidence score above 950 to retain only high confidence interactions. We employed random sampling to create negative pairs composed of the human proteins present in the positive pairs but without any STRING interactions between them, building a balanced dataset.

We built several PPI datasets using three archived versions of the STRING database (v9.1, v10, and v10.5) and the current version (v11). For the current version, we created three datasets each excluding protein pairs present in each of the older versions (see Table 1). Regarding the GO KG, we obtained archived versions of the GO and GO annotations in 2015, 2017 and 2019 from the Gene Ontology Data Archive[2].

2.2 Protein-Protein Interaction Prediction

We follow the setup in [7] that predicts relations between KG entity pairs that are not encoded in the graph using similarity-based semantic representations. We employed three KG-based semantic similarity measures to compute semantic similarity: two taxonomic measures (ResnikMax [5], SimGIC [4]) and one based on graph embedding methods (RDF2Vec [6]). We applied six well-known classes

[1] https://string-db.org
[2] http://release.geneontology.org/

STRING Version	Date	Number of positive pairs
v9.1	04/2015	12 681
v10	05/2017	26 863
v10.5	01/2019	31 384
v11 (excluding pairs in v9.1)	10/2020	41 227
v11 (excluding pairs in v10)	10/2020	31 642
v11 (excluding pairs in v10.5)	10/2020	23 571

Table 1. Number of positive pairs in each version of the STRING database.

of machine learning models to train classifiers using the scikit-learn library: K-nearest neighbor (KNN), genetic programming (GP), decision tree (DT), XGBoost (XGB), random forest (RF), and multi-layer perceptron (MLP). The classification performance was evaluated using the weighted average of F-measures (WAF).

3 Results and Discussion

We conducted two types of experiments: (i) ***Same version***, where we train the model with randomly chosen 10 000 protein interacting pairs from the archived STRING version and test it with the remaining pairs; (ii) ***Future version***, where we train the model with randomly chosen 10 000 protein pairs from the archived STRING version and test it on data from the current STRING version (excluding interactions present in the archived version). The same randomly chosen 10 000 protein pairs are used in both settings.

Since we used three archived versions, the *Future version* experiments also allow us to measure the impact of using increasingly older versions of STRING and GO in training. Table 2 shows no substantial differences between *Same version* and *Future version* experiments.

The results do not support a clear indication for data bias. While for the 2019 version, it is always slightly easier to predict future PPIs, this is reversed in the 2017 version, and varies between methods for the 2015 version, so no clear trend is discernible. The median weighted F-measure for the *Same version* experiments is 0.844, while it is 0.845 for the *Future version* (see Figure 1).

In addition to not detecting data leakage, the results also indicate that the relation between the functions of a protein and its interactions do not fundamentally change over time. Even for more recently discovered interactions that can be biologically different, protein functions are still a good predictor of PPIs.

4 Conclusion

Biomedical data resources share, reuse and import data from each other routinely. This can be a potential source of data leakage for machine learning applications. We investigated potential data leakage between the GO KG and the STRING database in the task of PPI prediction, by comparing performance on unseen interactions using archived data. Our results were not able to detect an

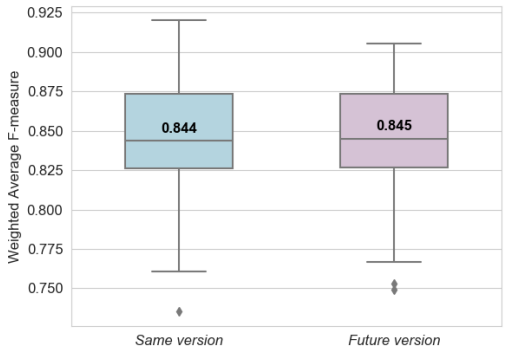

Fig. 1. Weighted Average F-measure Boxplot using the *Same version* and the *Future version* to test.

influence of data leakage, indicating that if this problem exists, its magnitude is not affecting the performance of KG-based PPI predictions.

Acknowledgements

CP, SS, RTS are funded by the FCT through LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020. CP and RTS are funded by project SMILAX (ref. PTDC/EEI-ESS/4633/2014), SS by projects BINDER (ref. PTDC/CCI-INF/29168/2017) and PREDICT (ref. PTDC/CCI-CIF/29877/2017), and RTS

ML	SSM	KG Version 01/19		05/17		04/15	
		Same	*Future*	*Same*	*Future*	*Same*	*Future*
KNN	ResnikMax	0.905	0.892	0.877	0.891	0.853	0.861
	SimGIC	0.858	0.843	0.821	0.842	0.825	0.826
	RDF2Vec	0.832	0.813	0.788	0.806	0.815	0.800
GP	ResnikMax	0.896	0.893	0.877	0.888	0.856	0.843
	SimGIC	0.873	0.855	0.835	0.859	0.835	0.842
	RDF2Vec	0.848	0.829	0.813	0.830	0.836	0.823
DT	ResnikMax	0.900	0.880	0.863	0.875	0.855	0.852
	SimGIC	0.815	0.800	0.768	0.788	0.767	0.772
	RDF2Vec	0.784	0.767	0.735	0.753	0.761	0.749
XGB	ResnikMax	0.920	0.903	0.887	0.906	0.874	0.880
	simGIC	0.873	0.858	0.839	0.860	0.834	0.846
	RDF2Vec	0.851	0.830	0.812	0.831	0.837	0.821
RF	ResnikMax	0.912	0.902	0.885	0.902	0.867	0.880
	SimGIC	0.874	0.858	0.837	0.859	0.832	0.845
	RDF2Vec	0.851	0.830	0.813	0.831	0.838	0.822
MLP	ResnikMax	0.902	0.894	0.880	0.896	0.860	0.869
	SimGIC	0.871	0.857	0.838	0.861	0.835	0.845
	RDF2Vec	0.851	0.829	0.813	0.831	0.839	0.823

Table 2. Weighted Average of F-Measures for each combination of semantic similarity measure (SSM) and machine learning (ML) algorithm for different GO KG version.

by FCT PhD grant (ref. SFRH/BD/145377/2019). It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453.

References

1. Akrami, F., Saeef, M.S., Zhang, Q., Hu, W., Li, C.: Realistic re-evaluation of knowledge graph completion methods: An experimental study. In: Proc. of the 2020 ACM SIGMOD Int. Conference on Management of Data. pp. 1995–2010 (2020)
2. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., O'Donovan, C.: The GOA database: gene ontology annotation updates for 2015. Nucleic acids research **43**(D1), D1057–D1063 (2015)
3. Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(4), 1–21 (2012)
4. Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O., Couto, F.M.: Metrics for GO based protein semantic similarity: a systematic evaluation. In: BMC Bioinformatics. vol. 9, pp. 1–16. Springer (2008)
5. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the 14th Int. Joint Conference on Artificial Intelligence - Volume 1. p. 448–453. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995)
6. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: The Semantic Web. pp. 498–514 (2016)
7. Sousa, R.T., Silva, S., Pesquita, C.: evoKGsim+: a framework for tailoring knowledge graph-based similarity for supervised learning. In: ESWC 2021 Poster and Demo Track (2021)
8. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., et al.: The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research **49**(D1), D605–D612 (2021)
9. The Gene Ontology Consortium: The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research **47**(D1), D330–D338 (11 2018)
10. Tomczak, A., Mortensen, J.M., Winnenburg, R., Liu, C., Alessi, D.T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N.H., et al.: Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. Scientific reports **8**(1), 1–10 (2018)
11. Zhong, X., Rajapakse, J.C.: Graph embeddings on gene ontology annotations for protein–protein interaction prediction. BMC bioinformatics **21**(16), 1–17 (2020)