Grenze International Journal of Engineering and Technology, June Issue

Estimating the Association of Gene Disease using Knowledge Graph Embedding

Pooja Lingan[1], Golden Nancy R[2], T. Jemima Jebaseeli[3] and R Venkatesan[4]
[1]Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbatore, India.
Email: poojal@karunya.edu.in
[2-3]Division of AIML, Karunya Institute of Technology and Sciences, Coimbatore, India.
Email: goldennancy@karunya.edu, jemima_jeba@karunya.edu
[4]Division of CSE, Karunya Institute of Technology and Sciences, Coimbatore, India.
Email: rlvenkei_2000@karunya.edu

***Abstract*—Strategies for predicting gene-disease connections based on ontology cover a wide range of techniques, from conventional semantic similarity algorithms to more contemporary developments such as knowledge graph embeddings. While semantic similarity often delves within the hierarchical relations embedded in the ontology, knowledge graph embeddings take a broader view, considering a multitude of connections. Nevertheless, these embeddings are usually constructed across a single network, and additional ontologies would be necessary for more complicated tasks like gene-disease correlation prediction. The proposed research is delved into the impact of utilizing richer semantic representations derived from multiple ontologies. These representations aim to capture a comprehensive understanding of both genes and diseases, taking into account diverse types of relationships within the ontologies. The utility of using random walk-based knowledge graph embeddings, emphasizes the need for close integration of different ontologies to improve prediction performance. The multi-ontology approach contributes to a more nuanced and holistic interpretation of gene-disease associations, paving the way for improved biomedical insights.**

***Index Terms*— Ontologies, semantic similarity, knowledge graph, knowledge graph embedding, machine learning.**

I. INTRODUCTION

Extensive lists of potential candidate genes are often generated by genomic studies and high-throughput experiments, but only a few of these genes are truly relevant to the targeted disease, biological process, and phenotype. One of the most challenging tasks in the biomedical domain is to unravel the intricate connections between genes and diseases. The genotypes of complex diseases exhibit significant heterogeneity, posing challenges for identifying biological markers. Computational approaches designed to predict gene-disease associations have the potential to expedite this process. Overcoming several obstacles, such as the requirement to create connections on a broad scale without sacrificing the organization's quality. Recent developments that make use of the scientific information stored in ontologies have demonstrated encouraging outcomes in this regard [1].

Innovative approaches in this domain frequently depend on Semantic Similarity Measures (SSM). Phenotype-phenotype similarity was employed by Zeng et al. (2017) to rank new gene-phenotype relationships [2]. Asif et

Grenze ID: 01.GIJET.10.2.546

al. [3] claim that the identification of genes associated with complex illnesses is improved by machine learning classifiers built on Gene Ontology's database of gene functional similarities. When evaluating the similarity between things defined using ontology concepts, classic SSM techniques only consider hierarchical linkages [4]. Biomedical entities create a Knowledge Graph (KG) when they are annotated using ontology concepts [5]. Knowledge Graph Embeddings (KGEs) allow each item to be represented by a vector approximating the similarity qualities of the graph. In essence, by taking into account various sorts of linkages, KGEs offer more robust representations [6]. Vector representations of biological entities in ontologies are produced by the recent development demonstrated by Onto2Vec [7] and OPA2Vec [8], which combine formal ontology axioms with annotation axioms from the ontology metadata. Even though Knowledge Graph Embeddings (KGEs) typically function on a single aggregated graph, multi-domain applications such as gene-disease connections may require more than limiting KGEs to a single ontology.

The semantic representations are more comprehensive as they originate from several ontologies. Human Phenotype Ontology (HP) explores adept genes and illnesses, and includes a variety of relationship types between the ontologies [9]. Gene Ontology (GO) [10] in particular shows the features of genes and diseases in a common semantic space, encompassing phenotypes as well as gene activity. For phenotypic anomalies seen in human illnesses, HP offers a uniform language, whereas GO clarifies the role of individual genes and gene products. According to this hypothesis, using several ontologies can enhance gene-disease association predictions, and a more thorough integration can provide positive results.

The proposed Logical Definitions (LD) establish a complex semantic link between things from disparate ontologies, enabling the bridging of domains and contextualizing interactions between entities like genes and illnesses. A restriction involving the GO phrase "sensory perception of sound" (GO:0007605) is equal to the HP term "hearing impairment" (HP:0000365). Furthermore, the proposed methodology examines several Knowledge Graph Embedding (KGE) techniques for representing genes and illnesses and examines the efficient combination of different vectors.

II. Related Works

Kulmanov et al., [4] computed the similarity for items labeled with a set of HP classes in the context of Human Phenotype Ontology and HP annotations, enabling comparisons such as illnesses based on phenotypes, gene pairs, or gene-disease correlations. These metrics are flexible; they may be used as parts of clustering algorithms, features in supervised machine learning models, and unsupervised methods of association prediction. The prediction of gene-disease connections, and protein-protein interactions, and to aid in the diagnosis of patients and the assessment of computer techniques for ontology class annotation prediction. State-of-the-art methods often fall into taxonomic semantic similarity, comparing entities based on ontology graph relations. Approaches for quantifying semantic similarity differ based on whether they compare two classes or two entities annotated with their own set of classes [11].

In knowledge representation, embeddings are vectors resulting from semantic information mapping techniques [12]. The goal of knowledge graph embeddings is to maintain the graph structures while representing concepts and interactions in a graph as low-dimensional vectors. These embeddings serve as features for machine learning and support similarity computation through operations such as cosine similarity. Various methods exist for building knowledge graph embeddings, with some focusing solely on knowledge graph facts, while others incorporate additional information like entity types, relation paths, axioms, rules, or textual information. Recently, there has been an increase in interest in path-based techniques, which convert the knowledge graph into node sequences. Categorization of knowledge graph embedding methods varies, with Wang et al., [5] recognizing matrix factorization, translational distance, semantic matching, and deep learning, while considering matrix factorization, deep learning, and random walks.

III. Methodology

The proposed approach is to build Knowledge Graphs (KGs) by merging several ontologies and annotating data. This work creates embeddings representing the gene and illness from their annotations in different KGs. Using a variety of vector operations, these embeddings are combined to create a single, common semantic space representation of genes and disorders. Finally, the combined embeddings are trained over using supervised learning algorithms to anticipate gene-disease relationships. The effectiveness of the approach is evaluated against expert rule systems techniques that depend on Knowledge Graph Embedding (KGE) similarity measures and standard Semantic Similarity Measures (SSM).

A. Knowledge Graphs

Three possible forms of Knowledge Graphs (KGs) are identified from the ones used in this study:

1. *Human Phenotype Ontology:* Involves Human Phenotype Ontology (HP) without Logical Definitions (LDs) and includes Human Phenotype annotations for both diseases and genes.
2. *Human Phenotype + Gene Ontology:* Encompasses Human Phenotype without Logical Definitions, Human Phenotype annotations (for genes and diseases), Gene Ontology (GO), and Gene Ontology annotations (for genes). Gene Ontology and Human Phenotype are linked via the same virtual root.
3. *Human Phenotype + Logical Definitions + Gene Ontology:* Integrates HP+GO by including Logical Definitions (LDs). Logical Definitions were reduced to direct equivalency statements between the Human Phenotype class and the Gene Ontology class mentioned in the Logical Definition, to improve graph embedding techniques.

B. Knowledge Graph Embeddings

Using the current graph structure as a guide, knowledge graph embedding is the process of filling in the gaps in the knowledge graphs by using probabilistic assumptions. This work generates 200 feature Knowledge Graph Embeddings (KGEs). It uses five different methodologies to cover a variety of KGE approaches: Translational distance is measured using random walk-based methods like RDF2Vec, OPA2Vec, and OWL2Vec, translational distance (TransE), and semantic matching (Distance Multi).

C. Vector Representation and Prediction

Two vectors, g and d, respectively, stand in for each pair of genes and diseases. For these vectors, an operator was developed to provide a composite representation, called r(g, d). Several operator options were taken into consideration, such as Hadamard, Average, Concatenation, Weighted-L1, and Weighted-L2. The resulting vectors served as input for different distinct machine learning (ML) algorithms: SSM, RF, MLP, etc. It employs grid search to identify optimal parameters for MLP. Additionally, this paper computed the cosine similarity (CS) between the vectors using the same approach as the baseline, employing a similarity threshold.

D. Baseline

The baseline serves the purpose of determining the performance of the methods that utilize a single ontology and classical Semantic Similarity Measures (SSM). This paper opted for the Human Phenotype (HP) Knowledge Graph (KG) as it provides annotations for both the genes and the diseases. Semantic similarity was computed for all gene-disease pairs using six different SSMs: BMA_{ICSeco}, $BMA_{ICResnik}$, $SimGIC_{ICSeco}$, $SimGIC_{ICResnik}$, Max_{ICSeco}, and $MAX_{ICResnik}$. The anticipation of associations is presented as a classification challenge, wherein a Semantic Similarity (SS) score exceeding a designated threshold signifies a positive association. In the case of Semantic Similarity Matrices (SSM), this research determined an SS threshold by assessing the weighted average of F-measures (for both positive and negative predictions) at different thresholds and selecting the maximum value (ranging from 0 to 1 with a step of 0.01). These benchmarks signify the best achievable outcomes with Semantic Similarity Matrices and the Human Phenotype (HP) ontology.

E. Vector Representation and Prediction

Each pair of genes and diseases is represented by two vectors, denoted as g and d, respectively. An operator was defined to manipulate these vectors, generating a representation denoted as r(g,d). Various operators, such as Concatenation, Average, Hadamard, Weighted-L1, and Weighted-L, were considered. The resultant vectors served as input for four distinct Machine Learning (ML) algorithms: Random Forest, Multi-Layer Perceptron, Naïve Bayesm and eXtreme Gradient Boosting. Optimal parameters for RF, XGB, and MLP were determined through grid search. Furthermore, this work evaluated the Cosine Similarity (CS) between vectors by using a Semantic Similarity threshold in conjunction with the same methodology as the baseline.

IV. IMPLEMENTATION

A. Dataset Description

The gathered curated gene-disease connections are from DisGeNET, an extensive database that is linked to human illness [13]. The screened relationships guarantee objectivity in annotations, keeping only those on which the source dependent on databases like Uniprot [14], OMIM [15], or Orphanet [16]. The dataset was then further refined using the following criteria.

i. Genes must match a protein annotated in Uniprot with Gene Ontology (GO) keywords.

ii. Genes must be annotated with terms related to human phenotypes (HP).
iii. Diseases must be annotated with terms related to HP. 2,716 genes, 1,807 disorders, and 8,189 disease-gene correlations were found as a consequence of this filtering method.

The proposed approach used random selection to generate negative instances because DisGeNET did not have any negative examples. To achieve a balanced dataset, these cases included genes and disorders that were extracted from the positive cases but did not have any established correlations. Gene Ontology annotations for the human species were sourced from the Gene Ontology Annotation (GOA) database [17]. HP annotations, establish connections between illnesses and genes and Human Phenotype terms, which were obtained from the HP database [9].

B. Performance measure

Among the techniques employed to evaluate the effectiveness of the classification was the weighted mean of F-measures (WAF). In addition to using WAF, the area under the receiver operating characteristic (ROC) curve, or AUC, scores were used in this study to evaluate performance. Every experiment used a split technique of 70% training and 30% testing. The same split-up was used consistently in every trial, even the baseline assessment.

C. Results

Baseline Performance

Table I presents the results for predictions using six semantic similarity measures with the knowledge graph HP-full. The choice between simple or full versions of HP is inconsequential, as these measures exclusively consider hierarchical "is a" relations between ontology concepts. In terms of information content for the WAF score with a single ontology, the Max approach is notably sensitive to the employed IC measure, showing differences up to 5% in both WAF and AUC-ROC. This underscores the significance of how frequently pairs of classes are used to annotate entities in comparison, rather than their placement in the graph. For the combination approach, the pairwise approach followed by BMA achieves optimal results, with a top WAF of 0.684 and AUC-ROC of 0.725 when combined with ICSeco. This highlights the trade-offs between globally considering all phenotypes (as done by simGIC) and restricting comparisons to the most similar phenotype (as done by Max). Despite offering insights into optimal similarity thresholds, these baselines still yield performance scores below 0.70 in WAF. BMA_{Seco} emerges as the primary semantic similarity baseline due to its superior overall results.

TABLE I. WAF AND AUC-ROC SCORES REPRESENTING OPTIMAL SEMANTIC SIMILARITY MEASURE (SSM) PERFORMACE WITH THE HP ONTOLOGY

SSM	**BMA_{Resnik}**	**BMA_{Seco}**	**MAX_{Resnik}**	**MAX_{Seco}**	**$simG_{ICResnik}$**	**$simG_{ICSeco}$**
WAF	0.682	**0.684**	0.681	0.633	0.633	0.636
AUC-ROC	0.713	**0.725**	0.712	0.662	0.671	0.667

Rich Semantic Representations Performance

a). Comparison of Vector Combination Approaches

It is difficult to provide an efficient strategy for combining gene and illness vectors to get a full semantic representation of diseases and genes using knowledge graph embeddings. Figure 1 provides a summary comparing five selected vector operations based on AUC-ROC evaluations. These evaluations were conducted using three top knowledge graph embedding methods (RDF2Vec, OPA2Vec, and DistMult) in conjunction with the Random Forest classifier, a highly effective machine learning algorithm. The analysis utilized the wealth knowledge graph (Human Phenotype-simple + Logical Definition + Gene Ontology). For a more detailed perspective, Table II presents the WAF scores achieved for each operator, considering all machine learning and embedding approaches within the same knowledge graph.

The Hadamard operator demonstrates superior performance compared to other operators in conjunction with RDF2Vec, OPA2Vec, and TransE. On the other hand, Concatenation emerges as the most effective choice when paired with OWL2Vec and DistMult. All things considered, Concatenation and Hadamard are the finest combination techniques; Hadamard performs best when paired with OPA2Vec, Random Forests, or XG-Boost. It's crucial to remember that weighted-L1, weighted-L2, average, Hadamard, and weighted-L2 all produce vectors of identical size (200), whereas concatenation produces double-sized (400) vectors. Machine learning algorithms' training time is impacted by this variation in vector size. Despite small performance losses observed

with Hadamard when paired with OWL2Vec and DistMult, all subsequent experiments focus on utilizing the Hadamard operator.

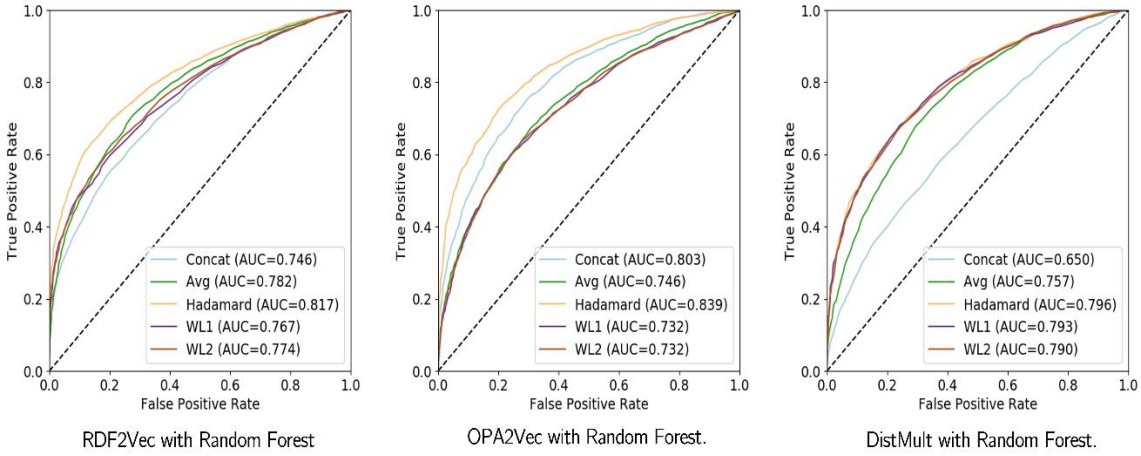

Fig 1: AUC values and ROC curves for different vector operators were produced by applying distinct algorithms to the HP+GO knowledge network

TABLE II. THE ASSESSMENT VECTOR COMBINATION OPERATORS (WAF SCORES) ARE CONDUCTED USING THE HPSIMPLE+LD+GO KNOWLEDGE GRAPH. THE OPTIMAL RESULT ACHIEVABLE IN EACH KNOWLEDGE GRAPH EMBEDDING IS HIGHLIGHTED IN BOLD

		RDF2Vec	OPA2Vec	OWL2Vec	DistMult	TransE
Concatenation	RF	0.672	0.728	0.617	0.603	0.484
	XGB	0.702	0.739	0.642	0.667	0.490
	NB	0.517	0.513	0.507	0.380	0.480
	MLP	0.732	0.743	**0.707**	**0.738**	0.487
Average	RF	0.714	0.683	0.652	0.683	0.509
	XGB	0.700	0.690	0.645	0.704	0.511
	NB	0.608	0.573	0.582	0.409	0.492
	MLP	0.711	0.717	0.679	0.700	0.366
Hadamard	RF	**0.743**	0.759	0.695	0.716	0.473
	XGB	0.739	**0.760**	0.694	0.724	**0.514**
	NB	0.617	0.530	0.582	0.467	0.500
	MLP	0.732	0.749	0.694	0.714	0.333
Weighted-L1	RF	0.693	0.677	0.623	0.714	0.508
	XGB	0.701	0.675	0.617	0.712	0.505
	NB	0.679	0.567	0.603	0.525	0.501
	MLP	0.699	0.695	0.620	0.696	0.487
Weighted-L2	RF	0.702	0.676	0.611	0.707	0.508
	XGB	0.701	0.675	0.630	0.715	0.505
	NB	0.669	0.540	0.593	0.629	0.498
	MLP	0.704	0.698	0.640	0.699	0.333

b). Comparison of Knowledge Graph Embedding Methods

Comparing embeddings to the semantic baseline, Hadamard consistently excels in machine learning with RDF2Vec, OPA2Vec, and DistMult, outperforming baseline and cosine. Overall, well-suited knowledge graph embeddings with effective ML algorithms consistently beat single-ontology predictions by up to 7.6% in WAF, surpassing cosine. OPA2Vec shines with Hadamard, reaching 0.760 WAF, while DistMult ranks third at 0.724. The optimal combo achieves 0.760 WAF with XGB, Hadamard, and OPA2Vec. Random Forest and MLP show promise, but NB struggles in linked data contexts due to its assumption of independent features.

c). Comparison of different Knowledge Graph

The Gene Ontology (GO) affects methods differently. RFD2Vec and OPA2Vec improve with HP-simple + GO, continuing with HP-full + GO. The new graph, (HP-simple + LD + GO), achieves the best results, indicating noise reduction. OPA2Vec and OWL2Vec exhibit a drop in performance from HP-simple to the full version, but

OPA2Vec improves with direct logical definitions. RDF2Vec's performance generally increases with a richer graph, boosted by noise reduction. Precision and recall values highlight that improved performance with semantic richness primarily stems from recall gains. In summary, the impact of adding GO is relatively small, and the added information may be limited by the number of logical definitions linking the two ontologies. The performance enhancement from GO is not significant beyond predictions derived from HP alone.

V. Conclusion

Associating genes with diseases represents a critical area of research, addressing a fundamental challenge in human health with significant implications for understanding disease etiologic and developing new approaches for prevention, diagnosis, and therapy. Using semantic representations based on Knowledge Graph Embeddings (KGEs) across several ontologies, this work presents a novel method for predicting gene-disease connections. We explore alternative ways to construct a shared semantic representation for genes and disorders by applying different KGE techniques and machine learning (ML) algorithms. With an AUC of 0.821, the proposed results show that the use of KGEs combined with the Hadamard operator outperforms traditional Semantic Similarity Measures (SSM). Regardless of a richer integration utilizing Logical Definitions (LDs), our investigations show that the differences between the usage of a single ontology and the combination of two ontologies (Human Phenotype Ontology, HP, and Gene Ontology, GO) are quite small. This research makes the hypothesis that the data supplied by LDs might not add much in the way of new information above what is currently included in HP. This may be due in part to the small number of LDs (351) that connect Gene Ontology and Human Phenotype. In the next phases, we will investigate ontology matching methods to provide more logical definitions and connections across ontologies. This will allow us to employ a wider variety of ontologies, especially ones that don't have LDs between them.

Reference

[1] Opap, K., & Mulder, N. (2017). Recent advances in predicting gene–disease associations. F1000Research, 6.

[2] Zeng, X., Ding, N., Rodríguez-Patón, A., & Zou, Q. (2017). Probability-based collaborative filtering model for predicting gene–disease associations. BMC medical genomics, 10, 45-53.

[3] Asif, M., Martiniano, H. F., Vicente, A. M., & Couto, F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PloS one, 13(12), e0208626.

[4] Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. Briefings in bioinformatics, 22(4), bbaa199.

[5] Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12), 2724-2743.

[6] Alshahrani, M., Khan, M. A., Maddouri, O., Kinjo, A. R., Queralt-Rosinach, N., & Hoehndorf, R. (2017). Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics, 33(17), 2723-2730.

[7] Smaili, F. Z., Gao, X., & Hoehndorf, R. (2018). Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics, 34(13), i52-i60.

[8] Smaili, F. Z., Gao, X., & Hoehndorf, R. (2019). OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics, 35(12), 2133-2140.

[9] Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J. P., ... & Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic acids research, 47(D1), D1018-D1027.

[10] Gene Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. Nucleic acids research, 47(D1), D330-D338.

[11] Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., ... & Furlong, L. I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015, bav028.

[12] UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1), D506-D515.

[13] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research, 43(D1), D789-D798.

[14] Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., ... & Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics, 28(2), 165-173.

[15] Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., & O'Donovan, C. (2015). The GOA database: gene ontology annotation updates for 2015. Nucleic acids research, 43(D1), D1057-D1063.

[16] Seco, N., Veale, T., & Hayes, J. (2004, August). An intrinsic information content metric for semantic similarity in WordNet. In Ecai (Vol. 16, p. 1089).
[17] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.
[18] Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., & Couto, F. M. (2008, April). Metrics for GO based protein semantic similarity: a systematic evaluation. In BMC bioinformatics (Vol. 9, No. 5, pp. 1-16). BioMed Central.
[19] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.
[20] Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.
[21] Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., & Paulheim, H. (2019). RDF2Vec: RDF graph embeddings and their applications. Semantic Web, 10(4), 721-752.
[22] Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O. M., Antonyrajah, D., & Horrocks, I. (2021). OWL2Vec*: Embedding of OWL ontologies. Machine Learning, 110(7), 1813-1845.
[23] Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2010). Modeling species distribution and change using random forest. In Predictive species and habitat modeling in landscape ecology: Concepts and applications (pp. 139-159). New York, NY: Springer New York.
[24] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[25] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29, 131-163.
[26] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.