Towards Supervised Biomedical Semantic Similarity

Rita T. Sousa[1,*], Sara Silva[1] and Catia Pesquita[1]

[1]*LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal*

Abstract

Ontology-based semantic similarity between entities in knowledge graphs is essential for several bioinformatics applications, including the prediction of protein-protein interactions and the discovery of associations between diseases and genes. Knowledge graphs typically describe entities according to different aspects modeled in ontologies, but both classical and graph embeddings-based semantic similarity measures consider the graph as a whole. This can be a limitation since different use cases may require different similarity perspectives and ultimately depend on expert knowledge for manual fine-tuning.

We present a new approach that uses supervised machine learning to tailor aspect-oriented semantic similarity measures to fit a particular view on biological similarity. This results in a supervised semantic similarity that is independent of the downstream application. We implement and evaluate it using different combinations of representative semantic similarity measures and machine learning methods with three biological similarity views: protein function family similarity, protein sequence similarity and phenotype-based gene similarity.

The results demonstrate that our approach outperforms non-supervised methods, producing semantic similarity models that fit different biological perspectives significantly better than the commonly used manual combinations of semantic aspects.

Keywords

Ontology, Knowledge Graph, Graph Embedding, Semantic Similarity, Machine Learning

1. Introduction

Measuring the similarity or distance between two entities is fundamental to many research fields, including biomedical informatics and artificial intelligence. When data is described according to an ontology and structured as a knowledge graph (KG), it can be explored to produce a semantic similarity score between two represented entities. Therefore, several semantic similarity measures (SSMs) for ontologies and KGs have been proposed over the years. Classical SSMs were initially based on ontologies and compute similarity between classes structured in a hierarchical taxonomy [1]. KG embeddings, a more recent research direction, can also be used to compute semantic similarity through vector similarity [2].

Ontologies express knowledge about a domain and allow the description of complex biological phenomena that are not easily captured in mathematical form [3]. As such, they provide the scaffolding for comparing biological entities at a higher level of complexity by comparing the ontology classes with which they are annotated. There are a wide variety of bioinformatics

[*] Corresponding author.

risousa@ciencias.ulisboa.pt (R. T. Sousa); sgsilva@ciencias.ulisboa.pt (S. Silva); clpesquita@ciencias.ulisboa.pt (C. Pesquita)

0000-0002-7241-8970 (R. T. Sousa); 0000-0001-8223-4799 (S. Silva); 0000-0002-1847-9393 (C. Pesquita)

CEUR Workshop Proceedings (CEUR-WS.org)

applications that benefit from using semantic similarity over biomedical ontologies, namely protein-protein interaction (PPI) prediction [4], disease-associated genes identification [5], and drug-drug interaction prediction [6].

However, the specificity of these data mining tasks contrasts with the broad domains covered by many biomedical ontologies. Large and successful biomedical ontologies often afford multiple perspectives (or semantic aspects) over the entities it describes. For instance, the Gene Ontology (GO) [7] describes protein function according to three semantic aspects: the *molecular functions* they perform, the *biological processes* they intervene in and the *cellular components* where they are active. In the same way, ChEBI [8] provides information about small chemical entities (e.g., atoms, molecules, ion pairs, radicals, radical ions, complexes, conformers) from three perspectives: the *molecular structure*, the *role* within a biological context or based on the intended use by humans, and the *subatomic particle*. Human Phenotype Ontology (HP) [9] is another example of a biomedical ontology that contains about terms describing phenotypic abnormalities found in human hereditary diseases according to five categories: *phenotypic abnormality*, *mode of inheritance*, *clinical course*, *clinical modifier* and *frequency*. Moreover, it can also be the case that multiple ontologies describe the same real-world entities, each covering different semantic aspects.

Depending on our viewpoint of the domain or the analytical task for which we want to use semantic similarity, some semantic aspects may be irrelevant for a specific definition of similarity. Consider the following example on comparing proteins according to their function. From a biochemist's point of view, two proteins playing the same molecular functions are very similar. However, these proteins can be very different from a physiological perspective if they participate in different biological processes at the whole-organism level. Therefore, depending on our goal, different semantic aspects should be taken into consideration in similarity computation. Selecting which semantic aspects to use and how they should be taken into account usually falls to the domain expert, rendering semantic similarity applications dependent on fine-tuning. This brings us to the challenge of tailoring SSMs to fit a specific application and biological perspective on similarity. In previous work, we developed a method to predict protein-protein interactions that uses genetic programming to evolve combinations of aspect-oriented semantic similarities that are tailored for PPI prediction [10]. However, this method has a tight connection between the tailoring of the similarity and the task it is used in.

In this work, we uncouple the tailoring of the similarity from the application task and develop a novel approach that learns semantic similarity models tailored to better capture particular biological similarity views, in effect producing a supervised semantic similarity. Since there is no gold standard for the similarity between complex biomedical entities, we take advantage of biological similarity proxies to train the models and evaluate them. These proxies of similarity rely on objective representations of entities (e.g., gene sequence, domains) and calculate similarity using mathematical expressions or other algorithms (e.g., BLAST-based similarity for sequences). The proposed approach was implemented using different KG-based SSMs, based on embeddings or taxonomic semantic similarity, coupled with different machine learning (ML) methods. This way, we compare the behaviour of different combinations of SSMs and ML methods in capturing different similarity perspectives.

We evaluate the proposed approach in a set of 11 benchmark datasets [11] that have varying sizes with different semantic annotation characteristics and include data from two biomedical ontologies, GO and HP. These datasets contain three proxies for biomedical entity similarity calculated based on protein sequence similarity, protein function family similarity, and phenotype-based gene similarity that are known to relate to relevant characteristics of the underlying entities. Our approach is compared with combinations of semantic aspects that emulate expert choices to understand how well the approach captures entity similarity. The results achieved on the benchmark datasets demonstrate the ability of our approach to significantly improve the estimation of similarity between biomedical entities.

2. Related Work

A SSM can be defined as a function that estimates the closeness in meaning between two entities. Several SSMs have been proposed with most measures falling in the category of taxonomic semantic similarity (also referred to as ontology-based semantic similarity, or only semantic similarity) [1]. Taxonomic SSMs are generally designed by an expert based on assumptions about how an ontology is used and what should constitute a similarity. They make extensive use of the taxonomical aspect of an ontology, comparing classes based on subclass/superclass relations.

KG embeddings can also be used to compute semantic similarity [2]. While some graph embedding methods focus on exploring the graph facts solely (like translational models or distMult [12]), others also include additional information, such as entity types, relation paths, axioms and rules, or textual information. More recently, path-based approaches, such as RDF2Vec [13], have been proposed by transforming the ontology graph into node sequences.

Approaches that combine taxonomic semantic similarity with ML have also been proposed. GARUM [14] is based on a supervised regression algorithm that receives several similarity measures of hierarchy, neighborhood, shared information, and attributes, and then predicts a final similarity score. In evoKGsim [10], we have used genetic programming over aspect-oriented semantic similarities to predict protein-protein interactions. However, the majority of the work that combines ontologies and ML is focused on embeddings. Kulmanov *et al.* [15] provide an overview of methods that incorporate SSMs and ontology embeddings into ML methods.

3. Methodology

We have developed a novel approach[1] to learn the similarity between entities represented in KGs (Definition 3.1) optimized towards a specific similarity proxy. This tailoring is achieved by considering the similarities for different semantic aspects (Definition 3.2), as opposed to the static SSMs (Definition 3.3).

Definition 3.1. A **KG** is created to describe real-world entities using links to ontology classes, represented in a graph. The nodes of the KGs represent ontology classes and entities, and edges are employed in representing ontology classes' relations and semantic annotations for entities.

[1] https://github.com/liseda-lab/Supervised-SS

Definition 3.2. A **semantic aspect** represents a perspective of the representation of KG entities. It can correspond to portions of the graph (e.g., describing a protein only through the biological process subgraph of the GO) or a given set of property types (e.g., describing a person only through properties having geographical locations as a range).

Definition 3.3. A **static SSM** calculates values of similarity by processing the KG without additional external input or tailoring to a specific similarity proxy.

An overview of the approach is shown in Figure 1. The first step consists of identifying the semantic aspects that describe the KG entities. Our approach takes as pre-defined semantic aspects the subgraphs when the KGs have multiple roots (such as GO) or the subgraphs rooted in the classes at a distance of one from the KG root class. Semantic aspects can also be manually defined by selecting the root classes which will anchor the aspects. The second step is representing each instance (i.e., a pair of KG entities) by computing the static KG-based similarities computed for each semantic aspect. The last step is to train a supervised semantic similarity according to the similarity proxy for which we want to tailor the similarity. The ML algorithms are used for regression where the expected outputs are the proxy similarity values.

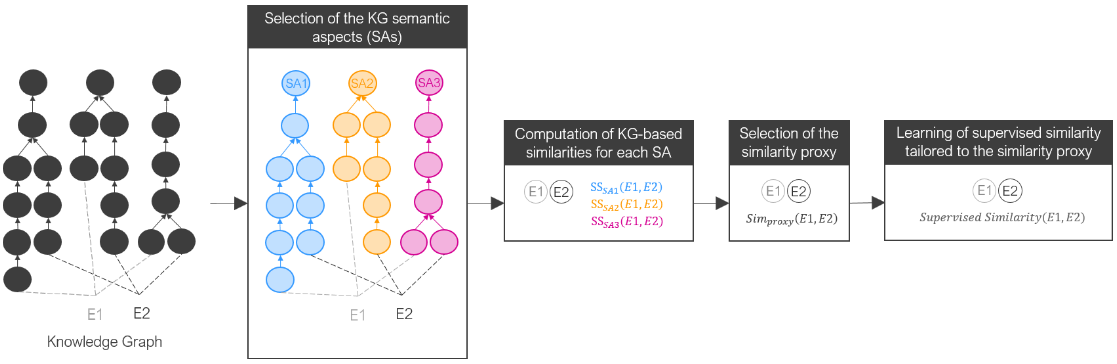

Selection of the KG semantic aspects (SAs)
SA1
SA2
SA3
E1
E2
Knowledge Graph
Computation of KG-based similarities for each SA
$SS_{SA1}(E1, E2)$
$SS_{SA2}(E1, E2)$
$SS_{SA3}(E1, E2)$
Selection of the similarity proxy
$Sim_{proxy}(E1, E2)$
Learning of supervised similarity tailored to the similarity proxy
$Supervised\ Similarity(E1, E2)$

Figure 1: Overview of the proposed approach.

This approach is independent of the semantic aspects, the specific implementation of KG-based similarity and the ML algorithm employed in regression. The following sections present the specific details of the implementation that currently supports four different SSMs and eight targeted supervised learning approaches.

3.1. Static Similarity Computation

Currently, our approach supports four different KG-based SSMs: two based on taxonomic similarity and two based on embeddings. The taxonomic semantic similarity is calculated using two state-of-the-art measures, derived by combining one IC approach (IC_{Seco} [16]) with one of two set similarity measures (ResnikBMA [17], SimGIC [18]). These were selected for their representativeness and good performance in the biomedical domain [19].

Regarding the embedding similarity, we employ two graph embedding approaches, namely RDF2Vec [13] and distMult [12], using an RDF2Vec python implementation[2] and the OpenKE

[2] https://github.com/IBCNServices/pyRDF2Vec

library[3]. These approaches were selected because they are representative of different types of graph embedding techniques. We generate protein or gene graph embeddings for each semantic aspect using these approaches and then, to compute the graph embeddings similarities, we employ cosine similarity between the vectors representing each entity in the pair.

3.2. Supervised Similarity Computation

Our approach combines the semantic similarities computed for each semantic aspect and returns a supervised similarity. The supervised semantic similarity model is computed by a supervised regression algorithm. Therefore, each regressor receives the similarity values for each semantic aspect as input features (independent variables) and a similarity proxy value as the expected output (dependent variable), and returns a single similarity score as the predicted output. We employ eight well-known classes of ML models, representative of different types of ML methods, to train regressors: linear regression (LR) [20], bayesian ridge (BR) [21], K-nearest neighbors (KNN) [22], genetic programming (GP) [23], decision tree (DT) [24], random forest (RF) [25], extreme gradient boosting, better known as XGBoost (XGB) [26], and multi-layer perception (MLP) [27]. Except for GP and XGB, we used the scikit-learn 21.3 library [28]. For running GP and XGB, we use gplearn 3.0[4] and the XGBoost 1.1.1 package[5], respectively.

4. Evaluation

The novel approach is evaluated using 10 protein benchmark datasets, one gene benchmark dataset [11] and two different KGs. These datasets, described in Table 1, explore three proxy similarities based on protein and gene properties. In the protein datasets, two proxies of protein similarity based on their biological properties were employed: sequence similarity and PFAM similarity (computed as the ratio of shared PFAM annotations). These datasets cover multiple species (*Drosophila melanogaster*, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*) and present two levels of annotation completion: the datasets ending in "1" include proteins with annotations in, at least, one aspect; the datasets ending in "3" include proteins with at least one annotation in each aspect. Concerning the gene benchmark dataset, the proxy similarity is based on the ratio of shared OMIM phenotypic series annotations.

Regarding the used KGs, we consider the GO KG and its three semantic aspects for the protein datasets. GO [7] defines the universe of classes associated with gene product (proteins or RNA) functions and how these functions are related with each other with respect to these three aspects: (i) molecular function (MF), the activities that occur at the molecular level performed by the gene product; (ii) biological process (BP), the larger process in which the gene product is active;; (iii) cellular component (CC), the cellular compartments in which the gene product performs a function. We built the GO KG with GO, gene product as instances, and GO annotations. Therefore, the nodes of the GO KG represent gene product or GO classes. The KG edges represent relationships between the GO classes or links between gene products annotated with GO classes.

[3] https://github.com/thunlp/OpenKE/tree/OpenKE-Tensorflow1.0

[4] https://gplearn.readthedocs.io/en/stable/

[5] https://xgboost.readthedocs.io

Table 1
Number of proteins/genes and pairs for all datasets.

Dataset	Ents	Pairs
PFAM		
DM1	7494	53797
DM3	5810	52460
EC1	1250	4623
EC3	748	1813
SC1	4783	42192
SC3	3660	30747
HS1	13604	60176
HS3	12487	60163
ALL1	27131	158512
ALL3	22705	142736
Gene	2026	12000

For the gene dataset, we also used the HP KG to compute the semantic similarity between two genes based on the phenotypes that describe them. The HP [9] contains about terms describing phenotypic abnormalities found in human hereditary diseases. The HP, genes and associated HP annotations compose the HP KG. Therefore, in addition to the three GO aspects, we also consider the HP phenotypic abnormality subgraph as a semantic aspect.

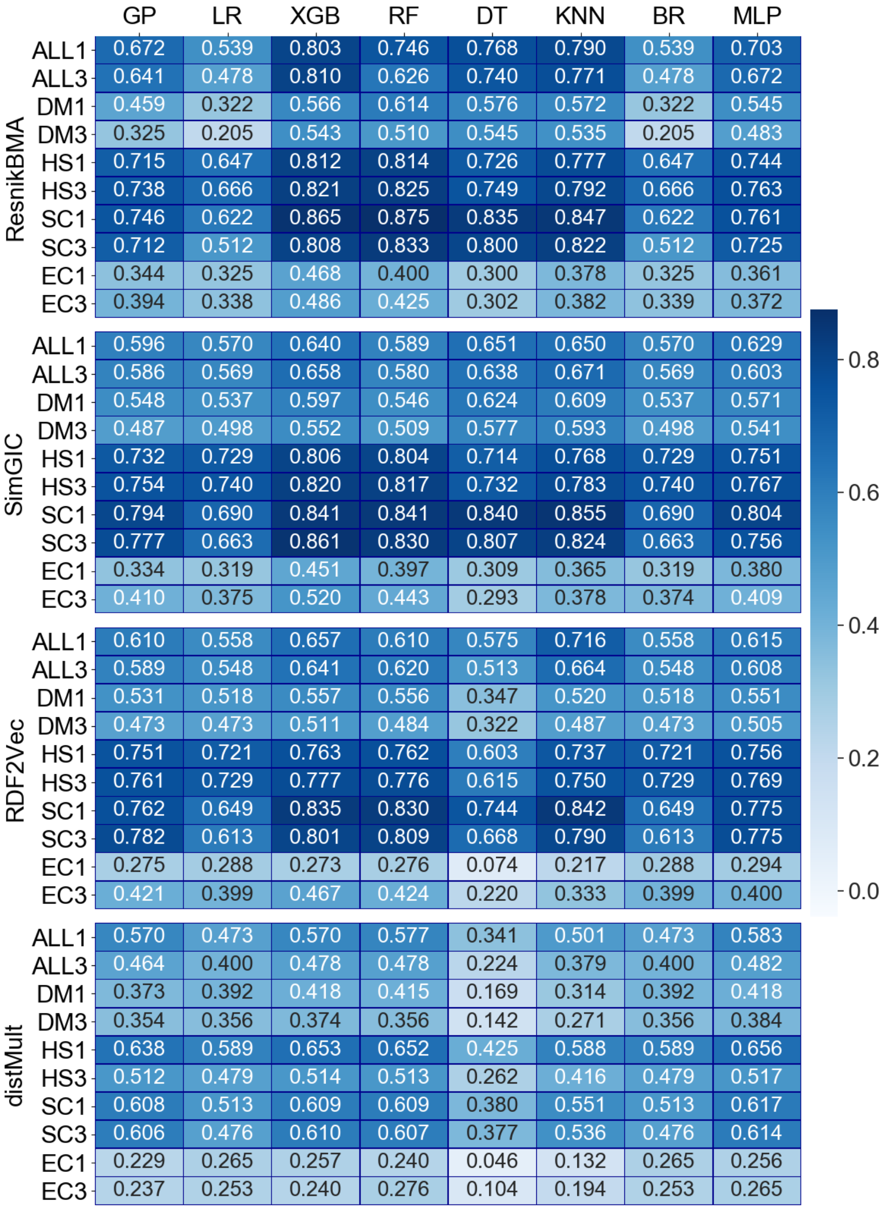
After semantic similarity computations, each instance of the dataset, that represents a protein or gene pair, is represented by several features corresponding to the semantic similarity for each semantic aspect, labeled with a proxy similarity value. The learned models correspond to a supervised semantic similarity tailored to a specific biological similarity.

For cross-validation, each dataset is split into ten folds. The same ten folds are used throughout all the experiments. The regression models are evaluated with the Pearson's correlation coefficient between the respective similarity proxies (expected values) and the obtained supervised similarity (predicted values). Since we use 10-fold cross-validation, the results we report are the median and the interquartile range (IQR) of the 10 Pearson's correlation coefficients calculated on the 10 folds.

5. Results and Discussion

5.1. Supervised Similarity

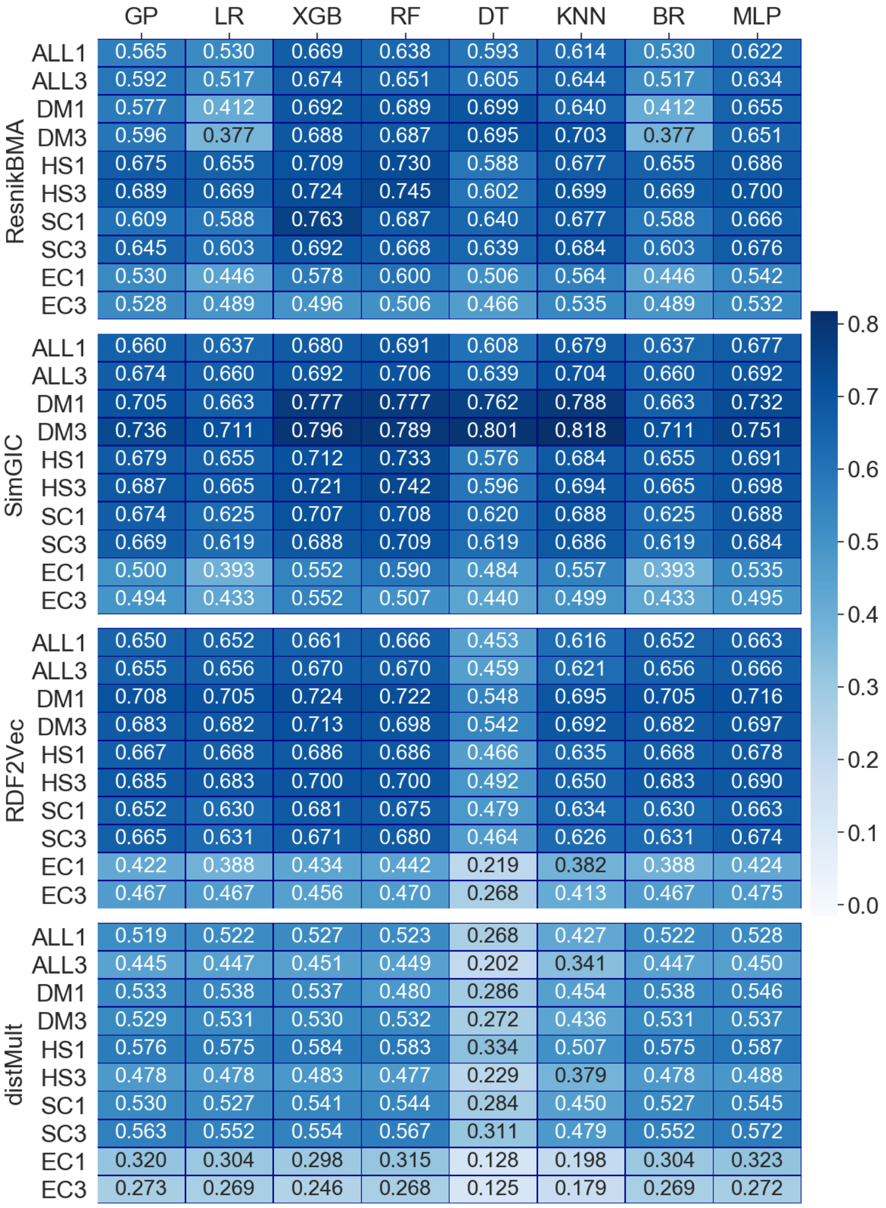
Figures 2, 3, and 4 contain the heat maps depicting the median Pearson's correlation coefficients between the similarity proxy (expected output) and the supervised similarity obtained with different ML methods and SSMs (predicted output), for each similarity proxy. To better compare the eight ML algorithms, we also generated radar charts (Figure 5) showing the median Pearson's correlation coefficients between similarity proxy and supervised similarity. Radar charts reveal which ML algorithms combined with different SSMs are scoring high or low within a dataset. In each radar plot, the ML algorithms are represented by different colors, and the SSMs are represented on different axes. For the sake of brevity, these radar plots only show the results for the protein datasets combining all species' protein pairs in the same group proxy.



		GP	LR	XGB	RF	DT	KNN	BR	MLP
ResnikBMA	ALL1	0.672	0.539	0.803	0.746	0.768	0.790	0.539	0.703
	ALL3	0.641	0.478	0.810	0.626	0.740	0.771	0.478	0.672
	DM1	0.459	0.322	0.566	0.614	0.576	0.572	0.322	0.545
	DM3	0.325	0.205	0.543	0.510	0.545	0.535	0.205	0.483
	HS1	0.715	0.647	0.812	0.814	0.726	0.777	0.647	0.744
	HS3	0.738	0.666	0.821	0.825	0.749	0.792	0.666	0.763
	SC1	0.746	0.622	0.865	0.875	0.835	0.847	0.622	0.761
	SC3	0.712	0.512	0.808	0.833	0.800	0.822	0.512	0.725
	EC1	0.344	0.325	0.468	0.400	0.300	0.378	0.325	0.361
	EC3	0.394	0.338	0.486	0.425	0.302	0.382	0.339	0.372
SimGIC	ALL1	0.596	0.570	0.640	0.589	0.651	0.650	0.570	0.629
	ALL3	0.586	0.569	0.658	0.580	0.638	0.671	0.569	0.603
	DM1	0.548	0.537	0.597	0.546	0.624	0.609	0.537	0.571
	DM3	0.487	0.498	0.552	0.509	0.577	0.593	0.498	0.541
	HS1	0.732	0.729	0.806	0.804	0.714	0.768	0.729	0.751
	HS3	0.754	0.740	0.820	0.817	0.732	0.783	0.740	0.767
	SC1	0.794	0.690	0.841	0.841	0.840	0.855	0.690	0.804
	SC3	0.777	0.663	0.861	0.830	0.807	0.824	0.663	0.756
	EC1	0.334	0.319	0.451	0.397	0.309	0.365	0.319	0.380
	EC3	0.410	0.375	0.520	0.443	0.293	0.378	0.374	0.409
RDF2Vec	ALL1	0.610	0.558	0.657	0.610	0.575	0.716	0.558	0.615
	ALL3	0.589	0.548	0.641	0.620	0.513	0.664	0.548	0.608
	DM1	0.531	0.518	0.557	0.556	0.347	0.520	0.518	0.551
	DM3	0.473	0.473	0.511	0.484	0.322	0.487	0.473	0.505
	HS1	0.751	0.721	0.763	0.762	0.603	0.737	0.721	0.756
	HS3	0.761	0.729	0.777	0.776	0.615	0.750	0.729	0.769
	SC1	0.762	0.649	0.835	0.830	0.744	0.842	0.649	0.775
	SC3	0.782	0.613	0.801	0.809	0.668	0.790	0.613	0.775
	EC1	0.275	0.288	0.273	0.276	0.074	0.217	0.288	0.294
	EC3	0.421	0.399	0.467	0.424	0.220	0.333	0.399	0.400
distMult	ALL1	0.570	0.473	0.570	0.577	0.341	0.501	0.473	0.583
	ALL3	0.464	0.400	0.478	0.478	0.224	0.379	0.400	0.482
	DM1	0.373	0.392	0.418	0.415	0.169	0.314	0.392	0.418
	DM3	0.354	0.356	0.374	0.356	0.142	0.271	0.356	0.384
	HS1	0.638	0.589	0.653	0.652	0.425	0.588	0.589	0.656
	HS3	0.512	0.479	0.514	0.513	0.262	0.416	0.479	0.517
	SC1	0.608	0.513	0.609	0.609	0.380	0.551	0.513	0.617
	SC3	0.606	0.476	0.610	0.607	0.377	0.536	0.476	0.614
	EC1	0.229	0.265	0.257	0.240	0.046	0.132	0.265	0.256
	EC3	0.237	0.253	0.240	0.276	0.104	0.194	0.253	0.265

Figure 2: Heat map representing the median Pearson's correlation coefficient using sequence proxy for each PFAM dataset.

The performance of regression models obtained by DT is globally lower compared to the other ML algorithms. DT is one of the most commonly used approaches for supervised learning. However, since it is based on recursive binary splitting, DT may not be suitable for the current regression problem of finding the best combination of semantic aspects. LR and BR also show lower correlations in many cases. The Pearson's correlation coefficients obtained by LR and



		GP	LR	XGB	RF	DT	KNN	BR	MLP
ResnikBMA	ALL1	0.565	0.530	0.669	0.638	0.593	0.614	0.530	0.622
	ALL3	0.592	0.517	0.674	0.651	0.605	0.644	0.517	0.634
	DM1	0.577	0.412	0.692	0.689	0.699	0.640	0.412	0.655
	DM3	0.596	0.377	0.688	0.687	0.695	0.703	0.377	0.651
	HS1	0.675	0.655	0.709	0.730	0.588	0.677	0.655	0.686
	HS3	0.689	0.669	0.724	0.745	0.602	0.699	0.669	0.700
	SC1	0.609	0.588	0.763	0.687	0.640	0.677	0.588	0.666
	SC3	0.645	0.603	0.692	0.668	0.639	0.684	0.603	0.676
	EC1	0.530	0.446	0.578	0.600	0.506	0.564	0.446	0.542
	EC3	0.528	0.489	0.496	0.506	0.466	0.535	0.489	0.532
SimGIC	ALL1	0.660	0.637	0.680	0.691	0.608	0.679	0.637	0.677
	ALL3	0.674	0.660	0.692	0.706	0.639	0.704	0.660	0.692
	DM1	0.705	0.663	0.777	0.777	0.762	0.788	0.663	0.732
	DM3	0.736	0.711	0.796	0.789	0.801	0.818	0.711	0.751
	HS1	0.679	0.655	0.712	0.733	0.576	0.684	0.655	0.691
	HS3	0.687	0.665	0.721	0.742	0.596	0.694	0.665	0.698
	SC1	0.674	0.625	0.707	0.708	0.620	0.688	0.625	0.688
	SC3	0.669	0.619	0.688	0.709	0.619	0.686	0.619	0.684
	EC1	0.500	0.393	0.552	0.590	0.484	0.557	0.393	0.535
	EC3	0.494	0.433	0.552	0.507	0.440	0.499	0.433	0.495
RDF2Vec	ALL1	0.650	0.652	0.661	0.666	0.453	0.616	0.652	0.663
	ALL3	0.655	0.656	0.670	0.670	0.459	0.621	0.656	0.666
	DM1	0.708	0.705	0.724	0.722	0.548	0.695	0.705	0.716
	DM3	0.683	0.682	0.713	0.698	0.542	0.692	0.682	0.697
	HS1	0.667	0.668	0.686	0.686	0.466	0.635	0.668	0.678
	HS3	0.685	0.683	0.700	0.700	0.492	0.650	0.683	0.690
	SC1	0.652	0.630	0.681	0.675	0.479	0.634	0.630	0.663
	SC3	0.665	0.631	0.671	0.680	0.464	0.626	0.631	0.674
	EC1	0.422	0.388	0.434	0.442	0.219	0.382	0.388	0.424
	EC3	0.467	0.467	0.456	0.470	0.268	0.413	0.467	0.475
distMult	ALL1	0.519	0.522	0.527	0.523	0.268	0.427	0.522	0.528
	ALL3	0.445	0.447	0.451	0.449	0.202	0.341	0.447	0.450
	DM1	0.533	0.538	0.537	0.480	0.286	0.454	0.538	0.546
	DM3	0.529	0.531	0.530	0.532	0.272	0.436	0.531	0.537
	HS1	0.576	0.575	0.584	0.583	0.334	0.507	0.575	0.587
	HS3	0.478	0.478	0.483	0.477	0.229	0.379	0.478	0.488
	SC1	0.530	0.527	0.541	0.544	0.284	0.450	0.527	0.545
	SC3	0.563	0.552	0.554	0.567	0.311	0.479	0.552	0.572
	EC1	0.320	0.304	0.298	0.315	0.128	0.198	0.304	0.323
	EC3	0.273	0.269	0.246	0.268	0.125	0.179	0.269	0.272

Figure 3: Heat map representing the median Pearson's correlation coefficient using using PFAM proxy for PFAM datasets.

BR are identical in most of the datasets. LR and BR assume a linear relationship between the independent and dependent variables, which is not true for many cases. This characteristic may explain why these ML methods were not capable of learning suitable combinations of semantic aspects.

The very tight lines in the radar plots show that KNN, GP, and MLP achieve comparable results. Ensemble methods, like XGB and RF, achieve better results in most experiments. These

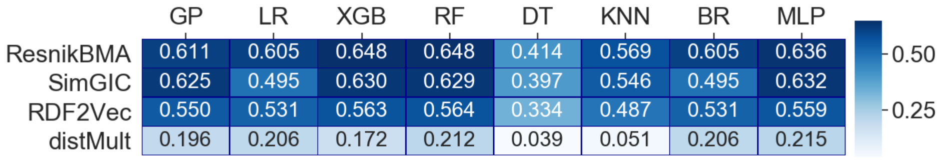


Figure 4: Heat map representing the median Pearson's correlation coefficient using using phenotype series proxy for gene dataset.

results were expected, since the ensemble methods combine the decisions from multiple models to improve the overall performance, and these methods have been successfully applied to different domains [29].

Comparing the SSMs, taxonomic similarity performs well across many evaluations and, in the majority of the datasets, has better performance than embedding similarity. The initial assumption was that embedding similarity could potentially outperform taxonomic similarity since semantic similarity is limited to the taxonomic relations within the ontology. However, the ability of taxonomic similarity to take into account class specificity may give it the advantage over embedding similarity to estimate similarity more accurately. Besides, taxonomic similarity measures are usually hand-crafted, providing human interpretable results for further analysis. Comparing the two taxonomic semantic similarity approaches, we verify that, in most cases, the maximum correlation is achieved when the ResnikBMA approach is used. Regarding the graph embedding approaches, RDF2Vec achieves the maximum correlation in the majority of datasets.

In order to assess whether a particular combination of an ML method and a specific SSM increases performance, for each proxy similarity we ranked the possible combinations of SSMs with ML algorithms within each dataset. Then, we calculated the average ranking of each SSM-ML combination. Table 2 shows the best combination for each proxy similarity. Although it is not straightforward to identify the best combination of SSM with ML algorithm that will work for all datasets and use cases, the results seem to indicate that combining a taxonomic SSM with an ensemble method is a good choice.

Table 2
Best SSM-ML combination for each proxy similarity.

Proxy	SSM	ML Algorithm
$\mathbf{Sim_{seq}}$	ResnikBMA	RF
$\mathbf{Sim_{PFAM}}$	SimGIC	RF
$\mathbf{Sim_{PS}}$	ResnikBMA	XGB

5.2. Static versus Supervised Similarity

Tables 3, 4, and 5 compare the results obtained using static similarity and supervised similarity for sequence, PFAM and phenotypic series proxies, respectively. The static similarity was obtained using taxonomic SSMs (SimGIC or ResnikBMA) and embedding-based SSMs (RDF2Vec

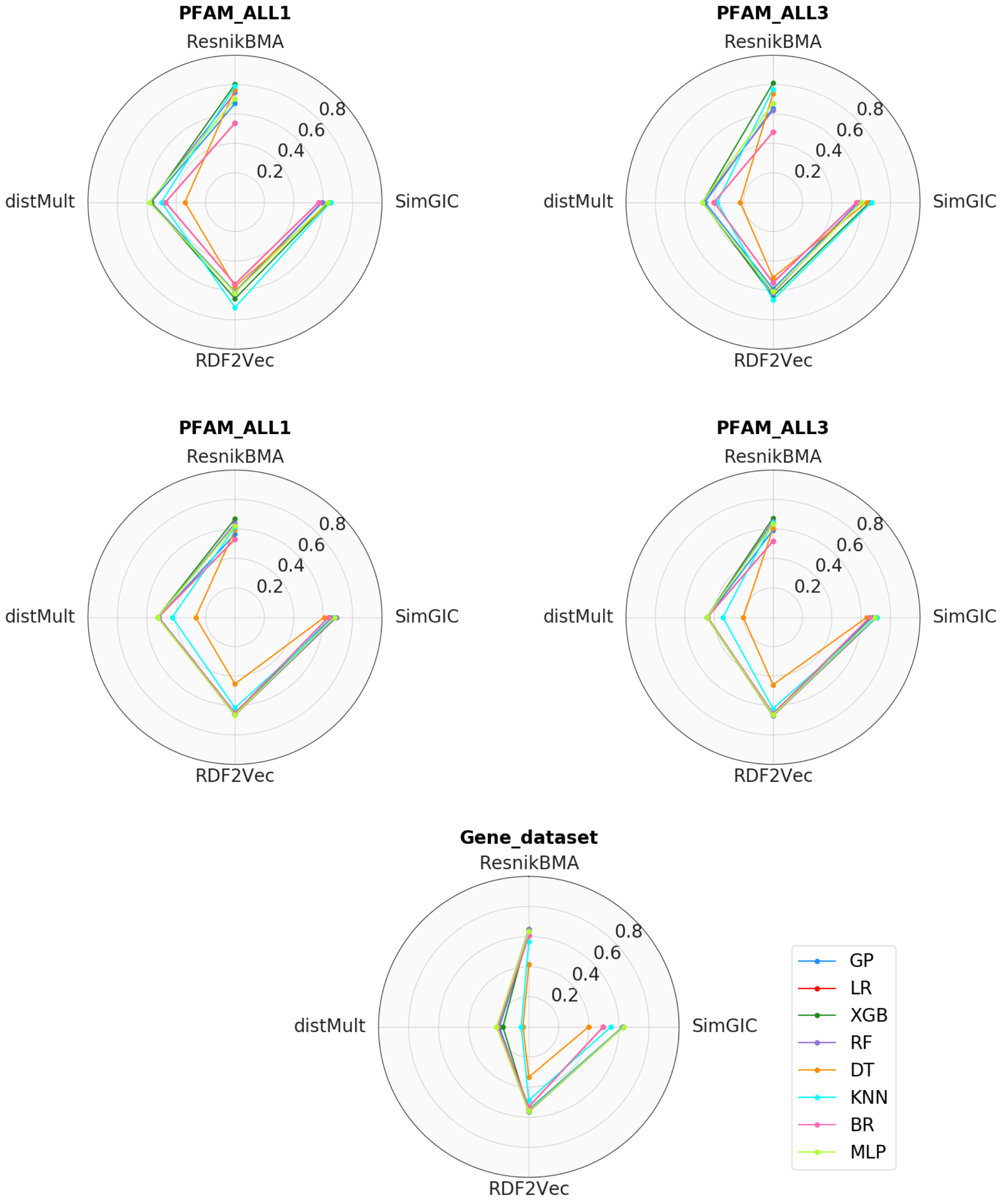


Figure 5: Radar charts using the sequence proxy (top) and PFAM proxy (middle) for the PFAM datasets and phenotype series proxy for the gene dataset (bottom).[a]

[a] the line for BR overlaps the line for LR.

and distMult), and computed for the whole graph, each single semantic aspect, and the average and maximum combinations of single semantic aspects. The Pearson's correlation coefficient was computed for each proxy. Regarding supervised similarity, the median and inter-quartile range (IQR) of Pearson's correlation values were calculated for the proposed approach using a SSM with an ensemble method (XGB or RF) for each proxy, the combinations previously shown to produce the best results. Once again, for the sake of brevity, these tables only show the results for the protein datasets combining all species' protein pairs in the same group proxy.

Table 3
Pearson's correlation coefficient between Sim_{seq} and different SSMs for the baselines and the median and IQR of Pearson's correlation coefficient between Sim_{seq} and supervised similarity obtained using XGB or RF. In bold, the best result for each dataset-SSM.

Dataset	SSM	Static						Supervised			
		ALL	BP	CC	MF	AVG	MAX	XGB		RF	
								Median	IQR	Median	IQR
ALL1	ResnikBMA	0.510	0.528	0.373	0.291	0.481	0.399	**0.803**	0.013	0.746	0.015
	SimGIC	0.568	0.552	0.406	0.415	0.547	0.406	**0.640**	0.033	0.589	0.004
	RDF2Vec	0.501	0.540	0.437	0.419	0.544	0.457	**0.657**	0.014	0.610	0.014
	distMult	0.435	0.398	0.236	0.322	0.467	0.429	0.570	0.009	**0.577**	0.009
ALL3	ResnikBMA	0.472	0.466	0.334	0.325	0.445	0.349	**0.810**	0.012	0.626	0.009
	SimGIC	0.564	0.544	0.374	0.451	0.539	0.411	**0.658**	0.037	0.580	0.009
	RDF2Vec	0.485	0.520	0.394	0.469	0.533	0.442	**0.641**	0.008	0.620	0.008
	distMult	0.445	0.382	0.184	0.011	0.341	0.380	**0.478**	0.018	**0.478**	0.018

Table 4
Pearson's correlation coefficient between Sim_{PFAM} and different SSMs for the baselines and the median and IQR of Pearson's correlation coefficient between Sim_{PFAM} and supervised similarity obtained using XGB or RF. In bold, the best result for each dataset-SSM.

Dataset	SSM	Static						Supervised			
		ALL	BP	CC	MF	AVG	MAX	XGB		RF	
								Median	IQR	Median	IQR
ALL1	ResnikBMA	0.534	0.448	0.370	0.456	0.525	0.500	**0.669**	0.008	0.638	0.005
	SimGIC	0.577	0.494	0.451	0.591	0.621	0.604	0.680	0.015	**0.691**	0.003
	RDF2Vec	0.636	0.524	0.466	0.619	0.627	0.623	0.661	0.007	**0.666**	0.009
	distMult	0.396	0.414	0.254	0.388	0.516	0.457	**0.527**	0.007	0.523	0.007
ALL3	ResnikBMA	0.521	0.431	0.387	0.463	0.514	0.480	**0.674**	0.015	0.651	0.005
	SimGIC	0.596	0.506	0.498	0.608	0.644	0.622	0.692	0.009	**0.706**	0.005
	RDF2Vec	0.648	0.535	0.514	0.612	0.640	0.627	**0.670**	0.009	**0.670**	0.006
	distMult	0.406	0.413	0.242	0.036	0.400	0.378	**0.451**	0.009	0.449	0.009

The results in Tables 3 to 5 show that whatever the ensemble method and SSM, supervised similarity always achieves higher values of correlation than static similarity. Improvements over the whole graph similarity and the single aspect similarities are consistent for all datasets and also clear when considering the combination of single aspects. However, there are some differences between the similarity proxies. For the sequence proxy, it is known that the relationship between

Table 5
Pearson's correlation coefficient between Sim_{PS} and different SSMs for the baselines and the median and IQR of Pearson's correlation coefficient between Sim_{PS} and supervised similarity obtained using XGB or RF. In bold, the best result for each SSM.

SSM	Static							Supervised			
	ALL	HP	BP	CC	MF	AVG	MAX	XGB		RF	
								Median	IQR	Median	IQR
ResnikBMA	0.524	0.601	0.210	0.142	0.055	0.413	0.552	**0.648**	0.022	0.648	0.023
SimGIC	0.459	0.489	0.205	0.158	0.095	0.399	0.429	**0.630**	0.011	0.629	0.013
RDF2Vec	0.554	0.526	0.230	0.182	0.123	0.396	0.351	0.563	0.014	**0.564**	0.010
distMult	0.155	0.015	0.184	0.105	0.041	0.179	0.182	0.172	0.018	**0.212**	0.052

sequence similarity and semantic similarity is non-linear [11], so improvements over the best static similarity are very pronounced. Regarding the PFAM proxy, we verify that MF is a relevant semantic aspect. The more functional (or PFAM) domains two proteins share, the more likely it will be to share molecular functions since these domains are usually responsible by assigning functions to proteins. Supervised similarity outperforms the GO, the GO single aspects and static combinations (average and maximum), although the improvements are more relevant for single aspects. In the gene dataset, the differences between static and supervised similarity are much more accentuated for the GO single aspects. These results were also expected, since the more phenotypic series two genes are associated with, the more likely it is that they share HP classes.

Finally, the comparison of results using protein datasets with different levels of annotation completion can be interesting. It is known that the annotation completeness of biological entities impacts semantic similarity [30]. Analyzing our results, we conclude that in the PFAM datasets, lower correlations were generally found for the incomplete annotation datasets, but the opposite happens in the PPI datasets. These results are in agreement with conclusions in [11].

6. Conclusion

Measuring the similarity between two genes or two gene products is a fundamental aspect of today's biomedical informatics research. Biomedical ontologies and KGs provide meaningful context to data and support the comparison of biomedical entities through semantic similarity. Many KGs afford different perspectives over the data, however, existing SSMs are general-purpose and either use the whole KG indiscriminately or depend on expert knowledge to select and combine the relevant KG semantic aspects for each use case.

This work presented a novel approach to tailor SSMs to better capture specific biological similarities by using semantic similarity features derived from different semantic aspects with ML methods. We tested our approach with four KG-based similarity measures based on embeddings or taxonomic semantic similarity, and eight ML methods. However, our approach is independent of the SSM and the chosen ML method. A comparative evaluation of the five SSMs combined with the eight ML algorithms was conducted using 11 benchmark datasets covering different species, levels of annotation completion, KGs describing them, and similarity proxies employed in them. The biological similarity proxies include protein family function similarity, protein sequence similarity and phenotype-based gene similarity - and were used to train and evaluate the supervised models. The results showed that our approach is able to learn a supervised semantic similarity that outperforms static semantic similarity in capturing biological similarity both using KG embeddings and standard taxonomic SSMs.

Currently, we have used SSMs that take into consideration semantic and structural information. Recently, KG embedding methods that also consider lexical information to generate embeddings, such as OPA2Vec [2], have been proposed, so there is a potential for these embeddings to improve the overall performance. However, we expect the main conclusion that the tailoring of SSM using semantic aspects increases the ability of SSMs to capture specific biological similarities to remain.

This work applied supervised ML algorithms to tailor semantic similarity to different similarity proxies and evaluated the correlation for supervised and static similarity. In future work, we will apply these supervised semantic similarities to bioinformatics tasks such as predicting protein-protein interactions, drug-target interactions or gene-disease associations. Our expectation is that a supervised similarity tailored to relevant biological similarities can transfer to these predictive tasks, outperforming static similarity and moreover performing competitively with supervised learning approaches without requiring specific training.

Acknowledgments

CP, SS, RTS are funded by the FCT through LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020), the projects GADgET (ref. DSAIPA/DS/0022/2018) and BINDER (ref. PTDC/CCI-INF/29168/2017), and the 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] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic Similarity from Natural Language and Ontology Analysis, Morgan Claypool Publishers, Williston, VT, USA, 2015.

[2] F. Z. Smaili, X. Gao, R. Hoehndorf, OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction, Bioinformatics 35 (2018) 2133–2140.

[3] S. Staab, R. Studer, Handbook on ontologies, Springer-Verlag, Berlin Heidelberg, 2010.

[4] K.-H. Chen, T.-F. Wang, Y.-J. Hu, Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme, BMC Bioinformatics 20 (2019) 308.

[5] M. Asif, H. F. M. C. M. Martiniano, A. M. Vicente, F. M. Couto, Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology, PLOS ONE 13 (2018) 1–15.

[6] S. Nunes, R. T. Sousa, C. Pesquita, Predicting gene-disease associations with knowledge graph embeddings over multiple ontologies, 2021.

[7] The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong, Nucleic Acids Research 47 (2018) D330–D338.

[8] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, ChEBI in 2016: Improved services and an expanding collection of metabolites, Nucleic Acids Research 44 (2015) D1214–D1219.

[9] S. e. a. Köhler, Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources, Nucleic Acids Research 47 (2018) D1018–D1027.

[10] R. T. Sousa, S. Silva, C. Pesquita, Evolving knowledge graph similarity for supervised learning in complex biomedical domains, BMC Bioinformatics 21 (2020) 6.

[11] C. Cardoso, R. T. Sousa, S. Köhler, C. Pesquita, A collection of benchmark data sets for knowledge graph-based similarity in the biomedical domain, in: Proceedings of Extended Semantic Web Conference 2020, Springer International Publishing, Cham, 2020, pp. 50–55.

[12] B. Yang, W. tau Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, 2015. arXiv:1412.6575.

[13] P. Ristoski, H. Paulheim, RDF2Vec: RDF graph embeddings for data mining, in: P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, Y. Gil (Eds.), Proceedings of International Semantic Web Conference 2016, Springer International Publishing, Cham, 2016, pp. 498–514.

[14] I. Traverso-Ribón, M.-E. Vidal, GARUM: A semantic similarity measure based on machine learning and entity characteristics, in: S. Hartmann, H. Ma, A. Hameurlain, G. Pernul, R. R. Wagner (Eds.), Database and Expert Systems Applications, volume 11029, Springer International Publishing, Cham, 2018, pp. 169–183.

[15] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics (2020) bbaa199.

[16] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in wordnet, in: Proceedings of the 16th European Conference on Artificial Intelligence, IOS Press, NLD, 2004, p. 1089–1090.

[17] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, p. 448–453.

[18] C. Pesquita, D. Faria, H. Bastos, A. Falcao, F. Couto, Evaluating GO-based semantic similarity measures, in: Proceedings of the 10th Annual Bio-Ontologies Meeting, Vienna, Austria, 2007, pp. 37–40.

[19] C. Pesquita, Semantic similarity in the Gene Ontology, in: C. Dessimoz, N. Škunca (Eds.), The Gene Ontology Handbook, Springer New York, New York, NY, 2017, pp. 161–173.

[20] M. A. Poole, P. N. O'Farrell, The assumptions of the linear regression model, Transactions of the Institute of British Geographers (1971) 145–158.

[21] P. J. Brown, J. V. Zidek, Adaptive multivariate ridge regression, Annals of Statistics 8 (1980) 64–74.

[22] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1967) 21–27.

[23] J. R. Koza, J. R. Koza, Genetic Programming: on the programming of computers by means

of natural selection, volume 1, MIT press, Cambridge, 1992.

[24] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1986) 81–106.

[25] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[26] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794.

[27] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.

[28] F. e. a. Pedregosa, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[29] O. Sagi, L. Rokach, Ensemble learning: A survey, WIREs Data Mining and Knowledge Discovery 8 (2018) e1249.

[30] P. H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Briefings in Bioinformatics 13 (2011) 569–585.