Programa de Pós-graduação em Sistemas de Informação

Representação do conhecimento, grafos de conhecimento, ontologias e suas aplicações

Programa de Verão 2023

Leonardo C. Santos

São Paulo / 2023

DataSelection

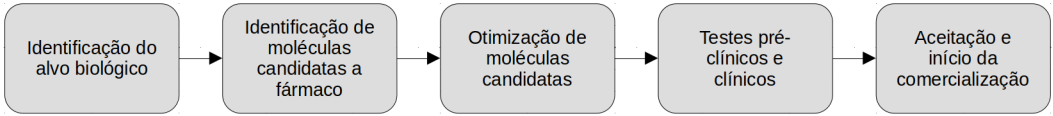
Uma abordagem semântica para seleção de conjuntos de dados em experimentos de transferência de aprendizado

Objetivo

Construir uma ferramenta que tome como entrada um conjunto de dados em um domínio-alvo e retorne um ranqueamento de conjuntos de dados candidatos a exercer o papel de domínio-fonte

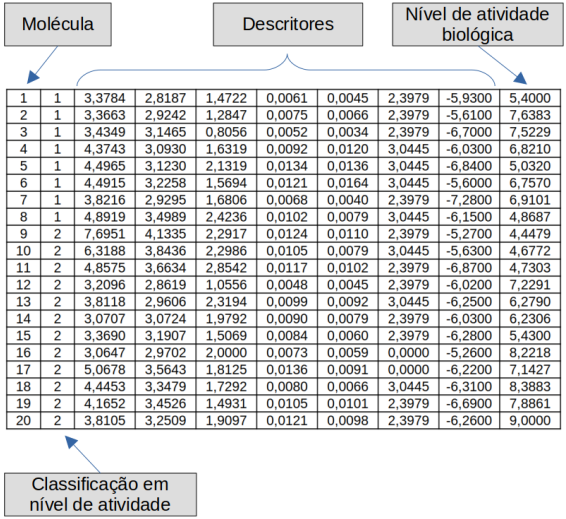
DataSelection

Drug discovery



DataSelection

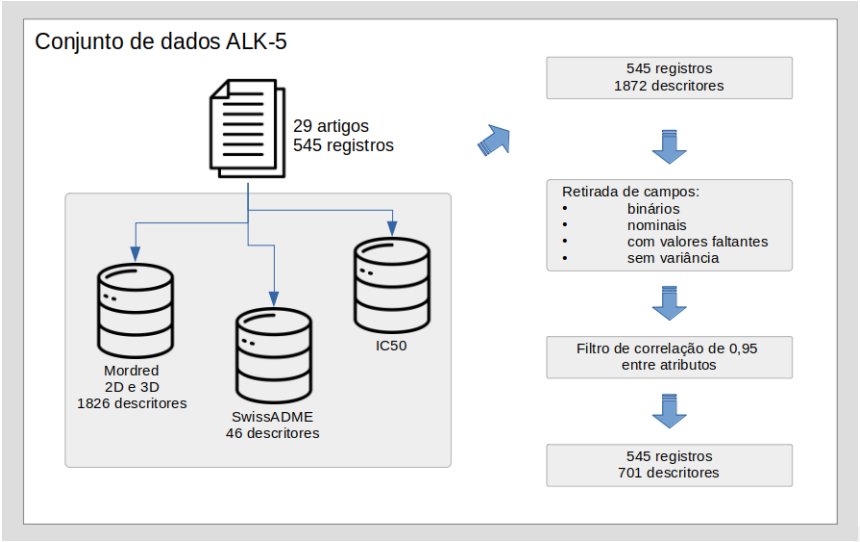
Descritores Químicos



1	1	3,3784	2,8187	1,4722	0,0061	0,0045	2,3979	-5,9300	5,4000
2	1	3,3663	2,9242	1,2847	0,0075	0,0066	2,3979	-5,6100	7,6383
3	1	3,4349	3,1465	0,8056	0,0052	0,0034	2,3979	-6,7000	7,5229
4	1	4,3743	3,0930	1,6319	0,0092	0,0120	3,0445	-6,0300	6,8210
5	1	4,4965	3,1230	2,1319	0,0134	0,0136	3,0445	-6,8400	5,0320
6	1	4,4915	3,2258	1,5694	0,0121	0,0164	3,0445	-5,6000	6,7570
7	1	3,8216	2,9295	1,6806	0,0068	0,0040	2,3979	-7,2800	6,9101
8	1	4,8919	3,4989	2,4236	0,0102	0,0079	3,0445	-6,1500	4,8687
9	2	7,6951	4,1335	2,2917	0,0124	0,0110	2,3979	-5,2700	4,4479
10	2	6,3188	3,8436	2,2986	0,0105	0,0079	3,0445	-5,6300	4,6772
11	2	4,8575	3,6634	2,8542	0,0117	0,0102	2,3979	-6,8700	4,7303
12	2	3,2096	2,8619	1,0556	0,0048	0,0045	2,3979	-6,0200	7,2291
13	2	3,8118	2,9606	2,3194	0,0099	0,0092	3,0445	-6,2500	6,2790
14	2	3,0707	3,0724	1,9792	0,0090	0,0079	2,3979	-6,0300	6,2306
15	2	3,3690	3,1907	1,5069	0,0084	0,0060	2,3979	-6,2800	5,4300
16	2	3,0647	2,9702	2,0000	0,0073	0,0059	0,0000	-5,2600	8,2218
17	2	5,0678	3,5643	1,8125	0,0136	0,0091	0,0000	-6,2200	7,1427
18	2	4,4453	3,3479	1,7292	0,0080	0,0066	3,0445	-6,3100	8,3883
19	2	4,1652	3,4526	1,4931	0,0105	0,0101	2,3979	-6,6900	7,8861
20	2	3,8105	3,2509	1,9097	0,0121	0,0098	2,3979	-6,2600	9,0000

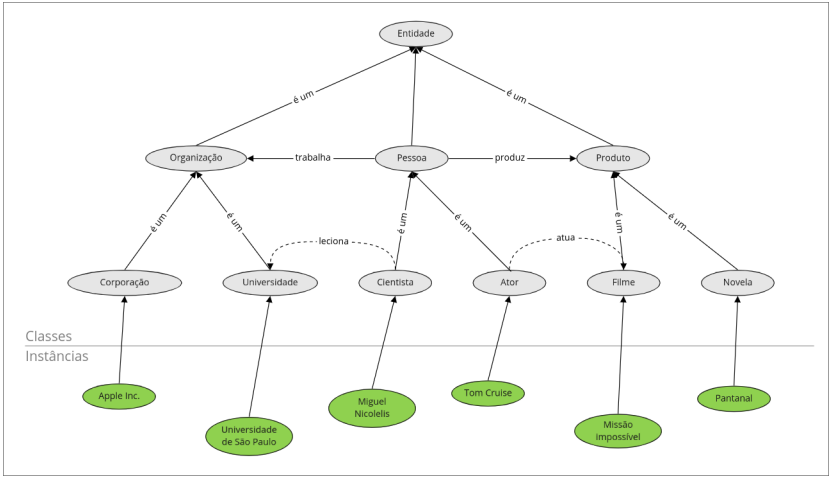
DataSelection

Descritores Químicos



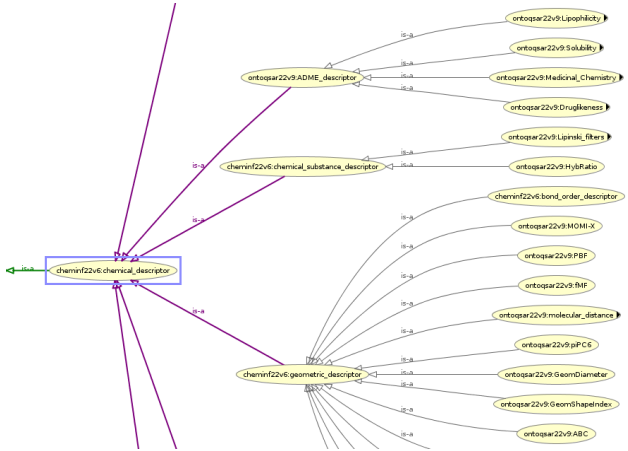
DataSelection

Ontologias



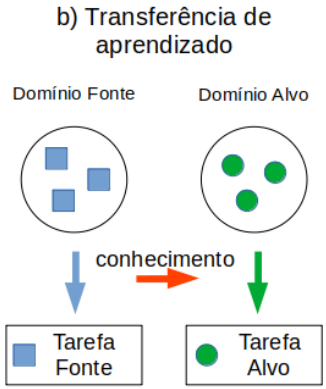
DataSelection

OntoQSAR + Cheminf + Novos conceitos

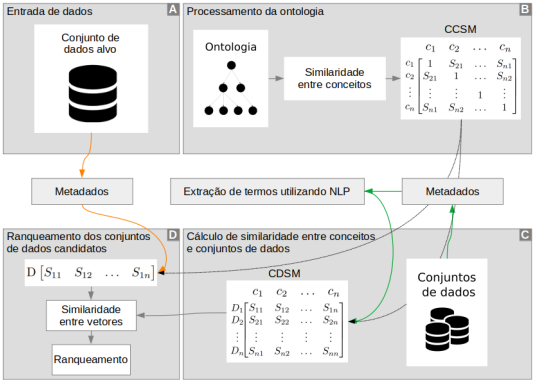


DataSelection

Transferência de aprendizado



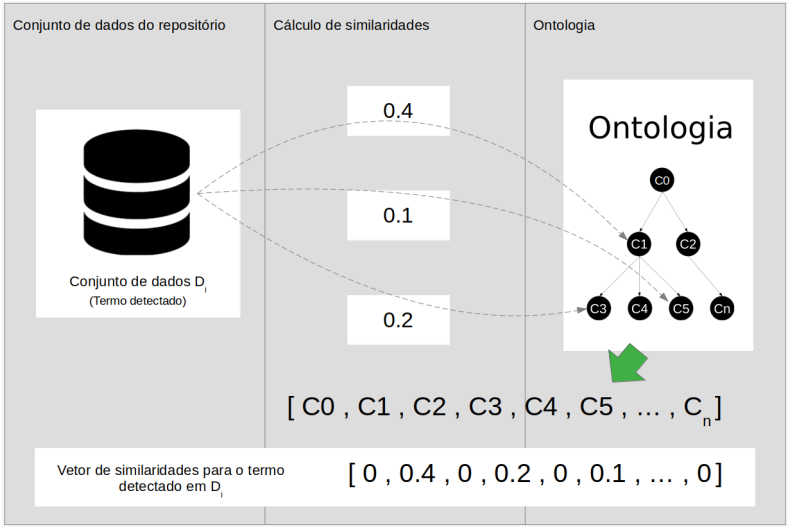
DataSelection



DataSelection

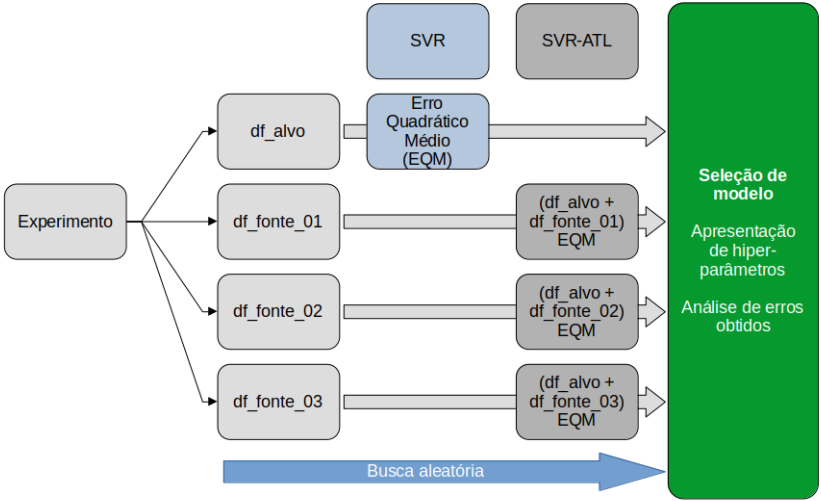
CCSM	ab initio quantum chemistry Computational method	abs máx(ev)	abs máx(nm)	absorption	acidic group count	alcohol	aldehyde	algorithm	alogp descriptor	amide	amine	aromatic atom count
ab initio quantum chemistry Computational method	1	0,5	0,5	0	0	0	0	0	0	0	0	0
abs máx(ev)	0,5	1	0,5	0	0	0	0	0	0	0	0	0
abs máx(nm)	0,5	0,5	1	0	0	0	0	0	0	0	0	0
absorption	0	0	0	1	0	0	0	0	0	0	0	0
acidic group count	0	0	0	0	1	0	0	0	0,5	0	0	0,5
alcohol	0	0	0	0	0	1	0,5	0	0	0,5	0,5	0
aldehyde	0	0	0	0	0	0,5	1	0	0	0,5	0,5	0
algorithm	0	0	0	0	0	0	0	1	0	0	0	0
alogp descriptor	0	0	0	0	0,5	0	0	0	1	0	0	0,5
amide	0	0	0	0	0	0,5	0,5	0	0	1	0,5	0
amine	0	0	0	0	0	0,5	0,5	0	0	0,5	1	0
aromatic atom count	0	0	0	0	0,5	0	0	0	0,5	0	0	1

DataSelection



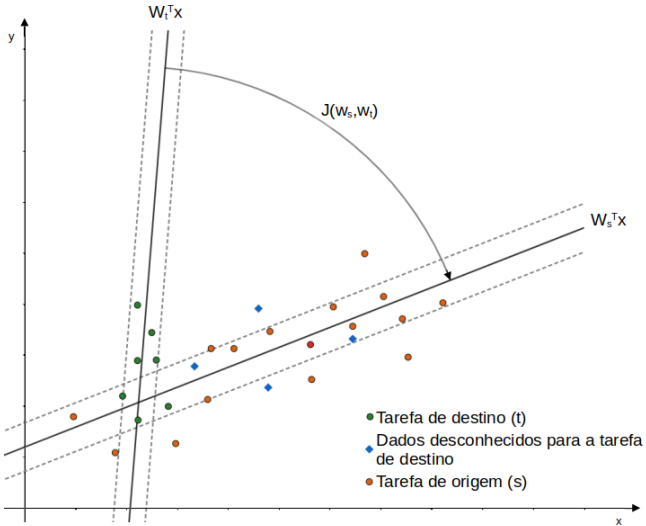
DataSelection

Configuração experimental



DataSelection

SVR-ATL



DataSelection

SVR-ATL

Tabela 2 – Resultados de similaridade obtidos pelo sistema *DataSelection* utilizando a medida de de caminho mais curto. Aqui, três conjuntos candidatos a fonte foram comparados com um conjunto alvo.

Atributos	Similaridade		
	Fonte 01	Fonte 02	Fonte 03
10	**0.87726**	0.71637	0.38634
20	0.81679	**0.87963**	0.75119
25	**0.90868**	0.72468	0.77018
35	**0.87145**	0.86534	0.79474
50	0.79109	**0.85111**	0.84249
55	**0.95701**	0.95380	0.86911
75	0.89444	**0.95989**	0.49030

Fonte: SANTOS, L. C., 2023

DataSelection

SVR-ATL

Tabela 3 – Resultados de EQM obtidos em experimentos com os métodos SVR e SVR-ATL.

Atributos	SVR	SVR-ATL		
		Fonte 01	Fonte 02	Fonte 03
10	0.65827	**0.61622**	0.62729	0.62555
20	0.70437	0.70009	0.70275	**0.69565**
25	0.64947	**0.62178**	0.63623	0.64115
35	0.60133	**0.59315**	0.62480	0.60365
50	0.59292	0.62392	0.60522	**0.59057**
55	0.40816	**0.39222**	0.43404	0.42459
75	0.91201	0.88875	**0.87656**	0.89725

Fonte: SANTOS, L. C., 2023

Conclusões

DataSelection

- Os resultados obtidos em experimentos de transferência de aprendizado, envolvendo cálculos de similaridade entre o conjunto alvo e outros conjuntos candidatos a fonte, mostraram que a transferência trouxe benefícios em todos os cenários considerados, diminuindo o erro quadrático médio com relação aos resultados obtidos pela abordagem de regressão tradicional.

Conclusões

DataSelection

- Além disso, o sistema *DataSelection* conseguiu indicar o conjunto fonte mais adequado na maioria dos cenários avaliados, demonstrando o seu potencial para pesquisas futuras, inclusive em outras áreas de aplicação, já que o único artefato dependente do domínio, nesta proposta, é a ontologia.

Obrigado!

Thanks! / ¡Gracias!

lattes.cnpq.br/5620610314140397
leonardo.cunha.santos@usp.br