A Collection of Benchmark Datasets for Systematic Evaluations of Machine Learning on the Semantic Web

Petar Ristoski[1], Gerben Klaas Dirk de Vries[2], and Heiko Paulheim[1]

[1] University of Mannheim, Germany
Research Group Data and Web Science
{petar.ristoski,heiko}@informatik.uni-mannheim.de
[2] WizeNoze, Amsterdam, Netherlands
g.k.d.devries@outlook.com

Abstract.
Resource type: Datasets
Permanent URL: http://w3id.org/sw4ml-datasets
In the recent years, several approaches for machine learning on the Semantic Web have been proposed. However, no extensive comparisons between those approaches have been undertaken, in particular due to a lack of publicly available, acknowledged benchmark datasets. In this paper, we present a collection of 22 benchmark datasets of different sizes.Such a collection of datasets can be used to conduct quantitative performance testing and systematic comparisons of approaches.

Keywords: Linked Open Data, Machine Learning, Datasets, Benchmarking

1 Introduction

In the recent years, applying machine learning to Semantic Web data has drawn a lot of attention. Many approaches have been proposed for different tasks at hand, ranging from reformulating machine learning problems on the Semantic Web as traditional, propositional machine learning tasks to developing entirely novel algorithms. However, systematic comparative evaluations of different approaches are scarce; approaches are rather evaluated on a handful of often project-specific datasets, and compared to a baseline and/or one or two other systems.

In contrast, evaluations in the machine learning area are often more rigorous. Approaches are usually compared using a larger number of standard datasets, most often from the UCI repository[3]. With a larger set of datasets used in the evaluation, statements about statistical significance are possible as well [3].

At the same time, collections of benchmark datasets have become quite well accepted in other areas of Semantic Web research. Notable examples include the Ontology Alignment Evaluation Initiative (OAEI) for ontology matching[4], the

[3] http://archive.ics.uci.edu/ml/
[4] http://oaei.ontologymatching.org/

Berlin SPARQL Benchmark[5] for triple store performance, the Lehigh University Benchmark (LUBM)[6] for reasoning, or the Question Answering over Linked Data (QALD) dataset[7] for natural language query systems.

In this paper, we introduce a collection of datasets for benchmarking machine learning approaches for the Semantic Web. Those datasets are either existing RDF datasets, or external classification or regression problems, for which the instances have been enriched with links to the Linked Open Data cloud [14]. Furthermore, by varying the number of instances for a dataset, scalability evaluations are also made possible.

2 Related Work

Recent surveys on the use of Semantic Web for machine learning organize the proposed approaches in several categories, i.e., approaches that use Semantic Web data for machine learning [16], approaches that perform machine learning on the Semantic Web [11], and approaches that use machine learning techniques to create and improve Semantic Web data [8, 16]. Furthermore, there are some challenges, like the *Linked Data Mining Challenge*[8] or the *Semantic-Web enabled Recommender Systems Challenge*[9], which usually focus on only a few datasets and a very specific problem setting.

3 Datasets

Our dataset collection has three categories: (i) existing datasets that are commonly used in machine learning experiments, (ii) datasets that were generated from official observations, and (iii) datasets generated from existing RDF datasets. Each of the datasets in the first two categories are initially linked to DBpedia[10]. This has two main reasons, (1) DBpedia being a cross-domain knowledge base usable in datasets from very different topical domains, and (2) tools like DBpedia Lookup and DBpedia Spotlight making it easy to link external datasets to DBpedia. However, DBpedia can be seen as an entry point to the Web of Linked Data, with many datasets linking to and from DBpedia. In fact, we use the RapidMiner Linked Open Data extension [9], to retrieve external links for each entity to YAGO[11] and Wikidata[12]. Such links could be exploited for systematic evaluation of the relevance of the data of different LOD dataset in different learning tasks.

In the dataset collection, there are four datasets that are commonly used for machine learning. For these datasets, we first enrich the instances with links to LOD datasets, and reuse the already defined target variable to perform machine learning experiments:

[5] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/
[6] http://swat.cse.lehigh.edu/projects/lubm/
[7] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/
[8] http://knowalod2016.informatik.uni-mannheim.de/en/linked-data-mining-challenge/
[9] http://challenges.2014.eswc-conferences.org/index.php/RecSys
[10] http://dbpedia.org
[11] http://yago-knowledge.org/
[12] http://www.wikidata.org

– The *Auto MPG* dataset[13] captures different characteristics of cars, and the target is to predict the fuel consumption (MPG) as a regression task.
– The *AAUP* (American Association of University Professors) dataset contains a list of universities, including eight target variables describing the salary of different staff at the universities[14]. We use the average salary as a target variable both for regression and classification, discretizing the target variable into "high", "medium" and "low", using equal frequency binning.
– The *Auto 93* dataset[15] captures different characteristics of cars, and the target is to predict the price of the vehicles as a regression task.
– The *Zoo* dataset captures different characteristics of animals, and the target is to predict the type of the animals as a classification task.

For those datasets, cars, universities, and animals are linked to DBpedia based on their name.

The second category of datasets contains a list of datasets where the target variable is an observation from different real-world domains, as captured by official sources. Again, the instances were enriched with links to LOD datasets. There are thirteen datasets in this category:

– The *Forbes* dataset contains a list of companies including several features of the companies, which was generated from the Forbes list of leading companies 2015[16]. The target is to predict the company's market value as a classification and regression task. To use it for the task of classification we discretize the target variable into "high", "medium", and "low", using equal frequency binning.
– The *Cities* dataset contains a list of cities and their quality of living, as captured by Mercer [7]. We use the dataset both for regression and classification.
– The *Endangered Species* dataset classifies animals into endangered species[17].
– The *Facebook Movies* dataset contains a list of movies and the number of Facebook likes for each movie[18]. We first selected $10,000$ movies from DBpedia, which were then linked to the corresponding Facebook page, based on the movie's name and the director. The final dataset contains $1,600$ movies, which was created by first ordering the list of movies based on the number of Facebook likes, and then selecting the top 800 movies and the bottom 800 movies. We use the dataset for regression and classification.
– Similarly, the *Facebook Books* dataset contains a list of books and the number of Facebook likes. Each book was linked to the corresponding Facebook page using the book's title and the book's author. Again, we selected the top 800 books and the bottom 800 books, based on the number of Facebook likes.
– The *Metacritic Movies* dataset is retrieved from Metacritic.com[19], which contains an average rating of all time reviews for a list of movies [12]. The initial

[13] http://archive.ics.uci.edu/ml/datasets/Auto+MPG
[14] http://www.amstat.org/publications/jse/jse_data_archive.htm
[15] http://www.amstat.org/publications/jse/v1n1/datasets.lock.html
[16] http://www.forbes.com/global2000/list/
[17] http://a-z-animals.com/
[18] We use the Facebook Graph API: https://developers.facebook.com/docs/graph-api
[19] http://www.metacritic.com/browse/movies/score/metascore/all

dataset contained around 10,000 movies, from which we selected 1,000 movies from the top of the list, and 1,000 movies from the bottom of the list. We use the dataset both for regression and classification.
- Similarly, the *Metacritic Albums* dataset is retrieved from Metacritic.com[20], which contains an average rating of all time reviews for a list of albums [13].
- The *HIV Deaths Country* dataset contains a list of countries with the number of deaths caused by HIV, as captured by the World Health Organization[21]. We use the dataset both for regression and classification.
- Similarly, the *Traffic Accidents Deaths Country* dataset contains a list of countries with the number of deaths caused by traffic accidents[22].
- The *Energy Savings Country* dataset contains a list of countries with the total amount of energy savings of primary energy in 2010[23], which was downloaded from WorldBank[24]. We use the dataset both for regression and classification.
- Similarly, the *Inflation Country* dataset contains a list of countries with the inflation rate for 2011[25].
- The *Scientific Journals Country* dataset contains a list of countries with a number of scientific and technical journal articles published in 2011[26].
- The *Unemployment French Region* dataset contains a list of regions in France with the unemployment rate, used in the SemStats 2013 challenge [10].

Again, for those datasets, the instances (cities, countries, etc.) are linked to DBpedia. For datasets which are used for classification and regression, the regression target was discretized using equal frequency binning, usually into a *high* and a *low* class.

The third, and final, category contains datasets that were generated from existing RDF datasets, where the value of a certain property is used as a classification target. There are five datasets in this category:

- The *Drug-Food Interaction* dataset contains a list of drug-recipe pairs and their interaction, i.e., "negative" and "neutral" [6]. The dataset was retrieved from FinkiLOD[27]. Furthermore, each drug is linked to DrugBank[28]. We drew a stratified random sample of 2,000 instances from the complete dataset. When generating the features, we ignore the `foodInteraction` property in DrugBank, since it highly correlates with the target variable.
- The *AIFB* dataset describes the AIFB research institute in terms of its staff, research group, and publications. In [1] the dataset was first used to predict the affiliation (i.e., research group) for people in the dataset. The dataset contains 178 members of a research group, however the smallest group contains only 4

[20] http://www.metacritic.com/browse/albums/score/metascore/all
[21] http://apps.who.int/gho/data/view.main.HIV1510
[22] http://apps.who.int/gho/data/view.main.51310
[23] http://data.worldbank.org/indicator/10.1_ENERGY.SAVINGS
[24] http://www.worldbank.org/
[25] http://data.worldbank.org/indicator/NY.GDP.DEFL.KD.ZG
[26] http://data.worldbank.org/indicator/IP.JRN.ARTC.SC
[27] http://linkeddata.finki.ukim.mk/
[28] http://wifo5-03.informatik.uni-mannheim.de/drugbank/

people, which is removed from the dataset, leaving 4 classes. Also, we remove the `employs` relation, which is the inverse of the *affiliation* relation.
- The *AM* dataset contains information about artifacts in the Amsterdam Museum [2]. Each artifact in the dataset is linked to other artifacts and details about its production, material, and content. It also has an artifact category, which serves as a prediction target. We have drawn a stratified random sample of 1,000 instances from the complete dataset. We also removed the `material` relation, since it highly correlates with the artifact category.
- The *MUTAG* dataset is distributed as an example dataset for the DL-Learner toolkit[29]. It contains information about complex molecules that are potentially carcinogenic, which is given by the `isMutagenic` property.
- The *BGS* dataset was created by the British Geological Survey and describes geological measurements in Great Britain[30]. It was used in [17] to predict the lithogenesis property of named rock units. The dataset contains 146 named rock units with a lithogenesis, from which we use the two largest classes.

An overview of the datasets is given in Tables 1, 2, and 3. For each dataset, we depict the number of instances, the machine learning tasks in which the dataset is used (C stands for classification and R stands for regression), the source of the dataset, and the LOD datasets to which the dataset is linked. For each dataset, we depict basic statistics of the properties of the LOD datasets, i.e., average, median, maximum and minimum number of *types*, *categories*, *outgoing relations* (rel out), *incoming relations* (rel in), outgoing relations including values (rel-vals out) and incoming relations including values (rel-vals in). The datasets, as well as a detailed description, a link quality evaluation, and licensing information, can be found online[31].

From the given statistics, we can infer the following observations: (i) DBpedia contains significantly less *owl:sameAs* links to YAGO, compared to Wikidata; (ii) DBpedia provides the highest number of types and categories on average per entity; (iii) Wikidata contains the highest number of outgoing and incoming relations for most of the datasets; (iv) YAGO contains the highest number of outgoing and incoming relations values for most of the datasets.

4 Conclusion and Outlook

In this paper, we have introduced a collection of 22 benchmark datasets for machine learning on the Semantic Web. So far, we have concentrated on classification and regression tasks. There are methods to derive clustering and outlier detection benchmarks from classification and regression datasets [4, 5], so that extending the dataset collection for such unsupervised tasks is possible as well. Furthermore, as many datasets on the Semantic Web use extensive hierarchies in the form of ontologies, building benchmark datasets for tasks like *hierarchical multi-label classification* [15] would also be an interesting extension.

[29] http://dl-learner.org
[30] http://data.bgs.ac.uk/
[31] http://w3id.org/sw4ml-datasets

Table 1: Datasets statistics

Dataset					types				categories				rel out				rel in				rel-vals out				rel-vals in			
Name	Source	Task	LOD	#links	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min
Auto MPG	UCI ML	R	DBpedia	371	29.70	31	46	5	11.20	10	25	2	13.48	13	27	3	5.62	5	25	1	16.50	15	70	0	36.65	23	509	0
			YAGO	331	13.99	16	21	0	9.26	9	23	0	8.76	9	18	0	16.96	2	138	1	77.08	70	278	0	3,236.24	60	28,418	0
			Wikidata	371	1.05	1	3	0	0.29	0	3	0	20.20	18	61	9	5.32	5	31	1	13.92	12	54	4	59.33	21	755	3
AAUP	JSE	R/C (c=3)	DBpedia	960	24.40	28	41	0	9.38	9	20	0	12.68	15	28	0	8.20	7	36	0	11.74	11	66	0	62.18	23	2,488	0
			YAGO	889	10.49	11	17	0	3.31	3	11	0	11.37	12	15	0	13.61	3	138	1	85.83	68	446	0	2,455.27	110	28,418	1
			Wikidata	959	2.13	2	5	0	0.88	1	2	0	30.71	29	83	0	8.51	7	44	0	22.38	21	97	0	296.92	20	31,777	0
Auto 93	JSE	R	DBpedia	93	28.76	31	43	5	11.13	10	25	3	12.69	12	22	8	4.92	5	7	2	14.35	11	64	4	22.60	18	64	2
			YAGO	80	13.80	16	19	0	9.09	10	18	0	8.37	10	11	0	21.09	2	138	2	59.33	59	129	0	4,025.90	46	28,418	4
			Wikidata	93	1.00	1	2	0	0.12	0	1	0	17.31	17	26	9	3.56	3	8	1	11.23	11	25	4	19.91	19	57	3
Zoo	UCI ML	C (c=3)	DBpedia	101	8.61	11	26	0	4.67	3	34	0	8.22	9	15	3	3.54	3	8	1	13.26	11	87	1	146.28	24	3,686	2
			YAGO	8	0.74	0	13	0	0.15	0	6	0	0.63	1	8	0	127.23	138	138	2	5.39	1	156	0	26,173.23	28,418	28,418	3
			Wikidata	101	1.00	1	2	0	0.67	1	2	0	29.69	35	57	3	8.28	7	27	0	18.20	21	45	1	125.82	92	785	0
Forbes	Forbes	R/C (c=2)	DBpedia	1,585	14.77	19	62	0	4.87	4	52	0	10.15	11	27	0	2.76	2	27	0	10.44	10	136	0	14.30	4	1,925	0
			YAGO	1,003	7.28	10	33	0	2.35	2	42	0	7.57	11	21	0	52.07	2	138	1	34.42	27	510	0	10,531.37	107	28,418	1
			Wikidata	1,189	0.82	1	4	0	0.22	0	3	0	16.59	16	137	0	5.00	5	52	0	12.69	10	207	0	30.14	8	2,881	0
Cities	Mercer	R/C (c=3)	DBpedia	212	31.28	35	53	0	6.98	7	26	0	18.08	19	38	0	25.66	25	68	0	16.26	13	131	0	1,474.57	678	19,810	0
			YAGO	187	16.66	19	30	0	4.46	4	15	0	13.75	15	32	0	23.56	9	138	2	222.54	214	681	0	8,087.34	3,555	72,320	5
			Wikidata	212	2.11	2	9	1	3.40	4	6	0	69.08	67	153	6	39.99	37	108	1	105.29	89	390	2	5,298.23	1,599	99,865	1
FB Books	Facebook	R/C (c=2)	DBpedia	1,600	19.08	20	42	0	5.15	5	23	0	11.15	11	20	0	1.64	2	7	0	7.04	7	60	0	2.80	2	42	0
			YAGO	1,334	8.37	10	24	0	2.03	2	15	0	8.41	10	13	0	25.32	3	138	1	24.37	22	149	0	4,735.50	8	28,418	1
			Wikidata	1,578	1.00	1	3	0	0.01	0	1	0	21.19	22	55	0	3.15	3	17	0	16.41	16	69	0	7.47	4	165	0
FB Movies	Facebook	R/C (c=2)	DBpedia	1,600	24.90	27	55	0	12.50	11	60	0	12.43	13	21	0	1.46	1	12	0	11.65	12	51	0	4.96	2	110	0
			YAGO	1,339	12.08	14	32	0	6.51	6	27	0	8.39	10	17	0	26.89	6	138	1	55.01	47	280	0	4,682.42	43	28,418	1
			Wikidata	1,585	1.01	1	4	0	0.04	0	1	0	48.75	48	107	0	2.22	1	22	0	56.37	53	372	0	20.75	12	230	0
Metacritic Albums	Metacritic	R/C (c=2)	DBpedia	1,600	17.92	19	36	0	4.27	4	26	0	10.85	12	17	2	2.63	3	7	0	8.92	9	63	0	5.28	3	50	0
			YAGO	1,444	7.22	8	19	0	3.22	3	20	0	8.05	9	10	0	16.02	3	138	1	40.27	32	361	0	2,749.90	10	28,418	1
			Wikidata	1,576	0.99	1	2	0	0.00	0	1	0	17.64	18	45	0	4.00	5	9	1	11.73	12	49	0	8.77	7	54	1

Table 2: Datasets statistics

Dataset					types				categories				rel out				rel in				rel-vals out				rel-vals in			
Name	Source	Task	LOD	#links	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min
Metacritic Movies	Metacritic	R/C (c=2)	DBpedia	2,000	24.38	27	45	0	11.87	11	42	0	12.54	14	19	3	1.35	1	7	0	11.42	12	30	0	3.56	2	31	0
			YAGO	1,588	11.79	14	19	0	6.43	6	28	0	8.34	10	11	0	28.22	6	138	1	48.84	43	216	0	4,960.84	37	28,418	1
			Wikidata	1,981	0.98	1	1	0	0.03	0	1	0	47.86	49	99	0	1.98	1	13	0	52.70	53	237	0	15.77	11	117	0
HIV Deaths Country	WHO	R/C (c=2)	DBpedia	114	35.69	37	52	0	12.61	13	23	0	23.59	24	28	3	34.26	31	89	6	27.75	25	162	10	4,828.36	1,065	70,426	24
			YAGO	108	13.90	15	24	0	9.28	9	18	0	28.41	31	35	0	15.18	9	138	5	302.34	244	1,267	0	12,464.42	4,879	112,032	550
			Wikidata	114	4.12	4	8	1	4.83	5	6	0	120.87	119	173	7	55.68	51	148	2	229.46	210	595	2	45,671.15	4,971	669,273	66
Trafic Accidents Country	WHO	R/C (c=2)	DBpedia	146	36.40	38	53	0	13.12	13	23	0	23.40	24	28	1	37.87	36	94	8	27.44	24	162	0	7,528.18	1,587	218,957	77
			YAGO	139	14.29	15	27	0	9.62	10	16	0	28.44	31	35	0	14.61	9	138	5	345.03	290	2,104	0	17,882.47	6,126	423,559	693
			Wikidata	146	4.42	4	10	1	4.94	5	6	0	124.31	121	191	7	61.68	55	148	2	242.38	213	713	2	85,575.10	7,369	1,557,157	66
Energy Savingss Country	WorldBank	R/C (c=2)	DBpedia	162	36.07	38	53	0	13.12	13	23	0	23.46	24	28	1	36.64	33	94	8	26.72	23	162	0	6,876.80	1,440	218,957	77
			YAGO	152	14.09	15	27	0	9.52	10	16	0	27.82	31	35	0	16.40	9	138	5	329.28	279	2,104	0	16,969.96	5,821	423,559	757
			Wikidata	162	4.41	4	10	1	4.92	5	6	0	123.36	119	191	7	60.02	55	148	2	238.69	210	713	2	77,485.01	5,810	1,557,157	66
Inflation Country	WorldBank	R/C (c=2)	DBpedia	160	36.00	38	53	0	13.11	13	23	0	23.46	24	28	1	36.74	33	94	8	26.85	24	162	0	6,947.59	1,440	218,957	77
			YAGO	150	14.09	15	27	0	9.44	10	16	0	27.80	31	35	0	16.48	9	138	5	331.16	279	2,104	0	17,114.88	5,821	423,559	693
			Wikidata	160	4.39	4	10	1	4.88	5	6	0	123.23	119	191	7	60.12	55	148	2	237.94	210	713	2	78,453.16	5,810	1,557,157	66
Scientific Journals Country	WorldBank	R/C (c=2)	DBpedia	160	36.00	38	53	0	13.11	13	23	0	23.46	24	28	1	36.74	33	94	8	26.85	24	162	0	6,947.59	1,440	218,957	77
			YAGO	150	14.09	15	27	0	9.44	10	16	0	27.80	31	35	0	16.48	9	138	5	331.16	279	2,104	0	17,114.88	5,821	423,559	693
			Wikidata	160	4.39	4	10	1	4.88	5	6	0	123.23	119	191	7	60.12	55	148	2	237.94	210	713	2	78,453.16	5,810	1,557,157	66
Unemployment French Region	SemStats	R/C (c=2)	DBpedia	26	16.38	21	32	0	3.73	3	15	0	7.81	9	10	3	14.19	13	24	7	7.19	7	19	1	975.88	969	2,292	37
			YAGO	26	8.92	8	14	8	2.77	2	8	1	12.42	12	14	12	3.73	4	6	3	81.12	60	299	28	1,793.19	1,424	4,527	88
			Wikidata	26	1.35	1	3	1	2.58	3	4	1	86.23	84	119	74	34.00	33	51	21	83.12	79	157	58	332.69	193	1,464	137
Endangered Species	a-z-animals	R/C (c=2)	DBpedia	301	11.84	12	33	0	6.32	5	34	0	10.77	11	25	0	2.96	3	55	1	12.65	11	87	0	566.25	15	114,742	1
			YAGO	65	2.48	0	16	0	0.76	0	12	0	1.78	0	9	0	108.62	138	138	1	9.53	0	136	0	22,286.36	28,418	28,418	1
			Wikidata	301	1.05	1	4	0	0.44	0	6	0	34.32	37	137	3	6.94	6	78	0	22.04	22	400	1	21,909.94	70	6,460,930	0
Drug-Food Interaction	FinkiLOD	C (c=2)	DBpedia	1,989	8.83	4	38	0	5.46	5	18	0	12.65	14	15	0	1.40	1	5	0	3.63	3	12	0	34.71	24	158	0
			YAGO	588	4.46	0	31	0	0.68	0	6	0	2.15	0	8	0	99.69	138	138	1	7.28	0	61	0	20,427.08	28,418	28,418	1
			Wikidata	1,908	1.96	2	3	0	0.01	0	1	0	45.92	47	79	0	2.78	2	17	1	34.99	27	159	0	32.25	26	487	4
			DrugBank	2,000	2.00	2	2	2	\	\	\	\	61.68	64	71	41	1.70	2	2	0	41.96	41	132	14	62.49	50	211	0
			FinkiLOD	2,000	1.00	1	1	1	\	\	\	\	3.00	3	3	3	0.00	0	0	0	1.00	1	1	1	0.00	0	0	0

Table 3: Datasets statistics

Dataset			types				rel out				rel in				rel-vals out				rel-vals in			
Name	Task	#links	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min	avg	med	max	min
AIFB	C (c=4)	176	1.4	1	2	1	7.1	7	9	5	2.0	2	5	0	18.2	7	219	2	19.8	9	246	0
AM	C (c=11)	1,000	1.0	1	1	1	19.8	20	29	9	0.6	1	3	0	21.9	20	283	7	3.2	1	273	0
MUTAG	C (c=2)	340	1.0	1	1	1	9.8	10	14	5	\	\	\	\	65.8	56	465	4	\	\	\	\
BGS	C (c=2)	146	1.0	1	1	1	29.7	31	36	21	1.4	2	4	0	25.2	24	54	15	2.7	2	12	0

Acknowledgements The work presented in this paper has been partly funded by the German Research Foundation (DFG) under grant number PA 2373/1-1 (Mine@LOD), and the Dutch national program COMMIT.

References

1. Bloehdorn, S., Sure, Y.: Kernel Methods for Mining Instance Data in Ontologies. The Semantic Web pp. 58–71 (2007)
2. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: The Semantic Web: Research and Applications, pp. 733–747. Springer (2012)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)
4. Emmott, A.F., Das, S., Dietterich, T., Fern, A., Wong, W.K.: Systematic construction of anomaly detection benchmarks from real data. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description. pp. 16–21. ACM (2013)
5. Färber, I., Günnemann, S., Kriegel, H.P., Kröger, P., Müller, E., Schubert, E., Seidl, T., Zimek, A.: On using class-labels in evaluation of clusterings. In: MultiClust: Workshop on Discovering, Summarizing and Using Multiple Clusterings (2010)
6. Jovanovik, M., Bogojeska, A., Trajanov, D., Kocarev, L.: Inferring cuisine-drug interactions using the linked data approach. Scientific reports 5 (2015)
7. Paulheim, H.: Generating possible interpretations for statistics from linked open data. In: 9th Extended Semantic Web Conference (ESWC) (2012)
8. Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., Fanizzi, N.: Mining the semantic web. Data Mining and Knowledge Discovery pp. 613–662 (2012)
9. Ristoski, P., Bizer, C., Paulheim, H.: Mining the web of linked data with rapidminer. Web Semantics: Science, Services and Agents on the WWW (2015)
10. Ristoski, P., Paulheim, H.: Analyzing statistics with background knowledge from linked open data. In: Workshop on Semantic Statistics (2013)
11. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: A comprehensive survey. Web Semantics 36, 1–22 (2016)
12. Ristoski, P., Paulheim, H., Svátek, V., Zeman, V.: The linked data mining challenge 2015. In: KNOW@ LOD (2015)
13. Ristoski, P., Paulheim, H., Svátek, V., Zeman, V.: The linked data mining challenge 2016. In: KNOW@LOD (2016)
14. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: The Semantic Web–ISWC (2014)
15. Silla, Jr., C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. pp. 31–72 (2011)
16. Tresp, V., Bundschus, M., Rettinger, A., Huang, Y.: Towards machine learning on the semantic web. In: Uncertainty Reasoning for the Semantic Web I (2008)
17. de Vries, G.K.D.: A fast approximation of the Weisfeiler-Lehman graph kernel for RDF data. In: ECML/PKDD (1). pp. 606–621 (2013)