FAIR Multi-label classification

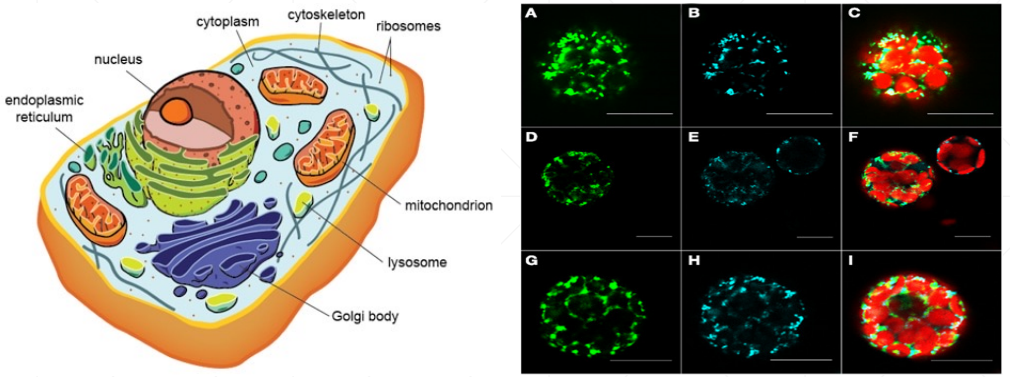
Comprehensive study of multi-label classification methods (part III)

Kocev Dragi, **Bogatinovski Jasmin**, Kostovska Ana, Panov Pance

ECML PKDD 2021Tutorial

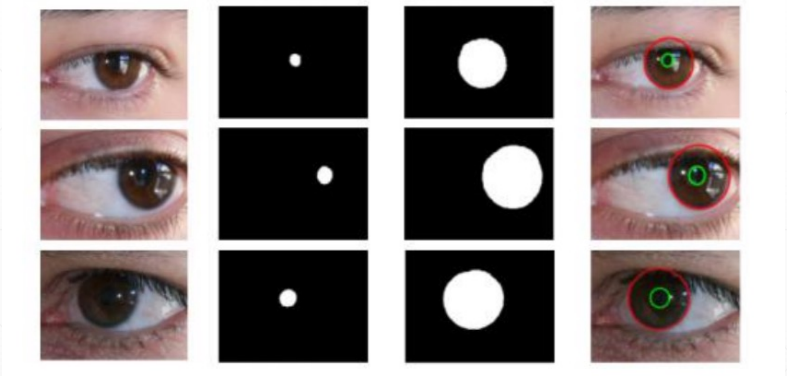
17.09.2021

Multi-label classification

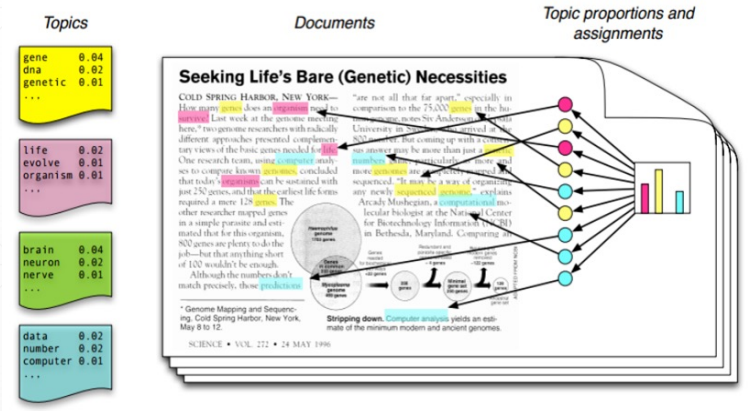


[a] subcellular localization

[b] video annotation



[c] mobile applications (high performing method for low resolution images)



[d] web page categorization

. . .

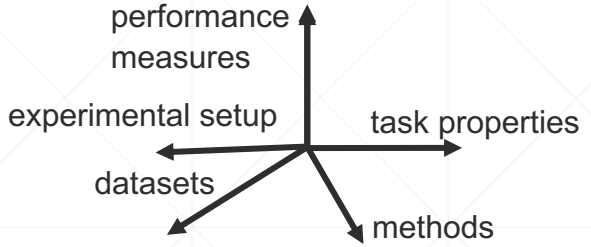
[a] Ryngajllo M, Childs L, Lohse M, Giorgi FM, Lude A, Selbig J and Usadel B (2011) SLocX: predicting subcellular localization of *Arabidopsis* proteins leveraging gene expression data. *Front. Plant Sci.* **2**:43. doi: 10.3389/fpls.2011.00043

[b] https://www.playment.io/blog/video-annotations-for-deep-learning-popular-applications-and-examples

[c] Xie, Z.; Chen, Y.; Lu, D.; Li, G.; Chen, E. Classification of Land Cover, Forest, and Tree Species Classes with ZiYuan-3 Multispectral and Stereo Data. *Remote Sens.* 2019

[d] https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc

Comprehensive study: Why?

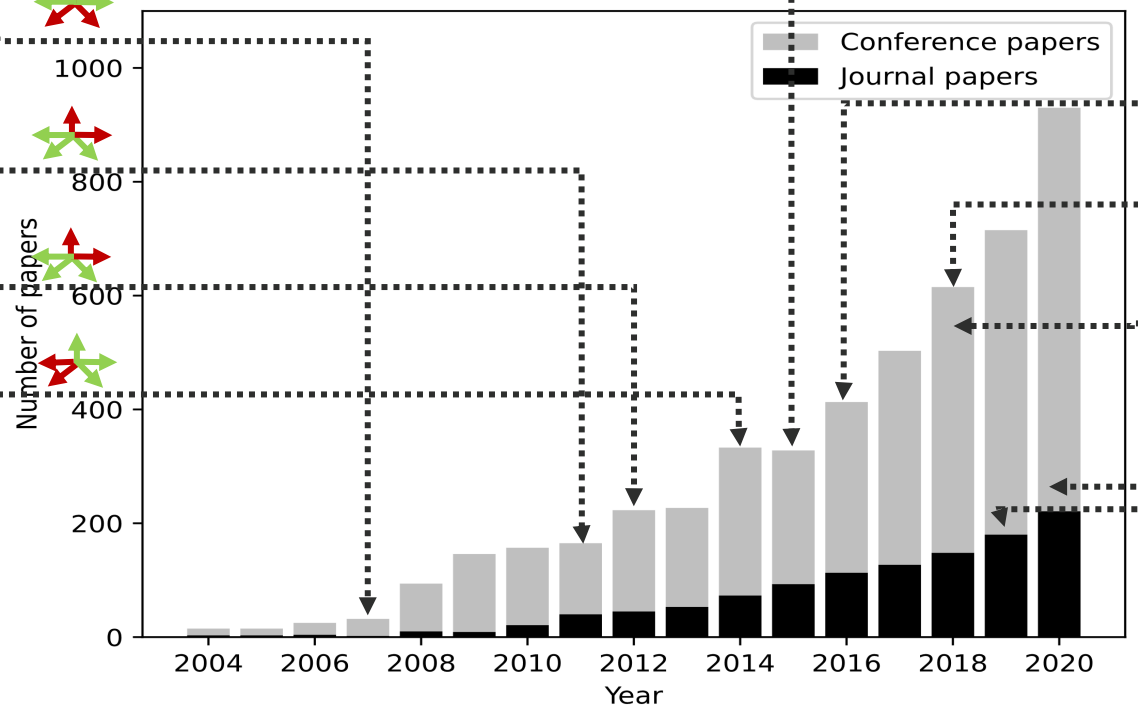


[Tsoumakas et al. 2007]

[Chekina et al. 2011]

[Madjarov et al. 2012]

[Zhang et al. 2014]



[Gibaja et al. 2015]

[Herrera et al. 2016]

[Moyano et al. 2018]

[Zhang et al. 2018]

[Rivoli et al. 2020]

[Bogatinovski. 2019]

[Tsoumakas et al. 2007] Tsoumakas G, Katakis I. 2007. Multi-label classification: an overview. International Journal of Data Warehousing and Mining 3(3):1–13 DOI 10.4018/jdwm.2007070101.

[Chekina et al. 2011] L. Chekina, L. Rokach and B. Shapira, "Meta-learning for Selecting a Multi-label Classification Algorithm," *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 220-227, doi: 10.1109/ICDMW.2011.118.

[Madjarov et al. 2012] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski, “An ˇ extensive experimental comparison of methods for multi-label learning,” Pattern Recognition, vol. 45, pp. 3084 – 3104, 2012

[Zhang et al. 2014] M. L. Zhang and Z. H. Zhou, “A review on multi-label learning algorithms,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, pp. 1819–1837, 2014

[Gibaja et al. 2015] E. Gibaja and S. Ventura, “A tutorial on multilabel learning,” ACM Computing Surveys, vol. 47, no. 3, pp. 52:1–52:38, 2015

[Herrera et al. 2016] F. Herrera, A. J. Rivera, M. J. del Jesus, and F. Charte, Multilabel Classification. Springer Cham, Switzerland: Springer, 2016

[Moyano et al. 2018] J. M. Moyano, E. L. G. Galindo, K. J. Cios, and S. Ventura, “Review of ensembles of multi-label classifiers: Models, experimental study and prospects,” Information Fusion, vol. 44, pp. 33–45, 2018

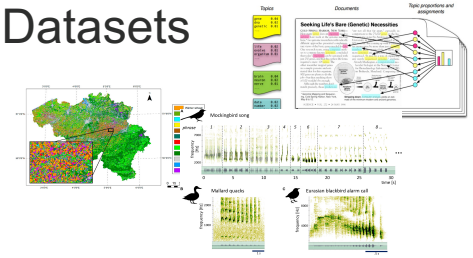
[Zhang et al. 2018] M.-L. Zhang, L. Yu-Kun, L. Xu-Ying, and X. Geng, “Binary relevance for multi-label learning: An overview,” Frontiers of Computer Science, vol. 12, pp. 191–202, 2018

[Bogatinovski 2019] Bogatinovski Jasmin, “A comprehensive study of multi-label classification methods”, Jozef Stefan International Postgraduate School, MSc, Ljubljana, 2019

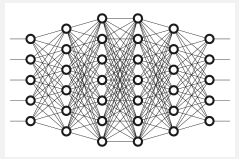
[Rivoli et al. 2020] A. Rivolli, J. Read, C. Soares, B. Pfahringer, and A. C. P. L. F. de Carvalho, “An empirical analysis of binary transformation strategies and base algorithms for multi-label learning,” Machine Learning, pp. 1–55, 2020

Overview

Datasets



Methods



Experimental setup

Software

Design

Model selection (SRM & HPO)

Thresholding

Results

Discussion

Recommendations

Multi-label classification

Multi-label classification: Task definition

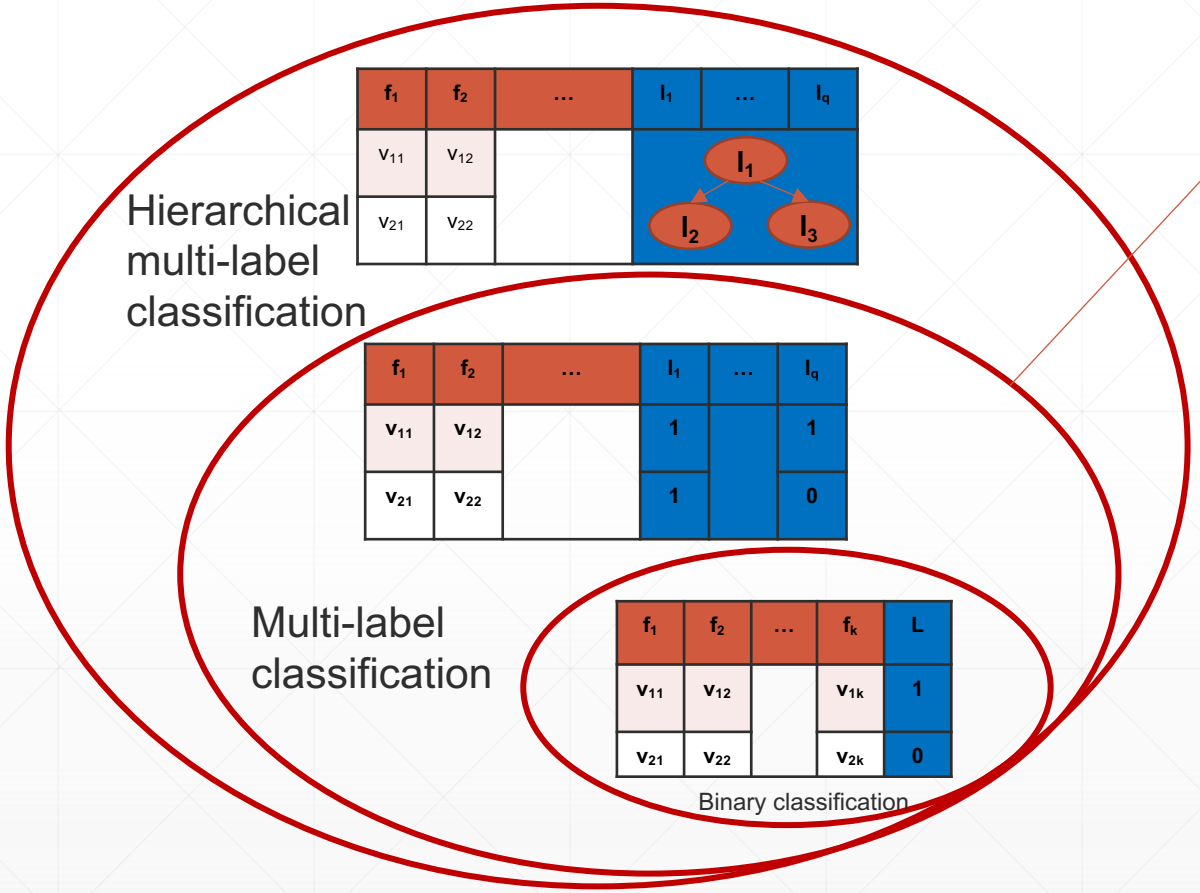
- An example space X consisting of values of primitive data types (boolean, discrete or continuous) $\forall \boldsymbol{x_i} \in X, \boldsymbol{x_i} = (x_1, x_2 \ldots x_d)$ where ***d*** is the number of descriptive features;
- A label space $L = \{\lambda_1, \lambda_2, \ldots \lambda_Q\}$ which is presented as a tuple of Q discrete variables;
- A set of examples E, where each example is a pair of tuples from the example and label space, respectively, i.e., $E = \{(x_i, Y_i) | x_i \in X, Y_i \in L, 1 \leq i \leq N\}$ and N is the number of examples;
- A quality criterion ***q***, which rewards models with high predictive accuracy and low complexity;

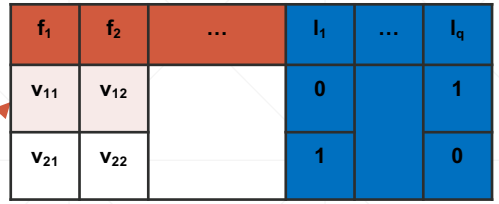
The task of MLC is to find:

- Function $f: X \rightarrow 2^L$, such that f optimizes ***q***

[Madjarov et al. 2012] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski, “An ˇ extensive experimental comparison of methods for multi-label learning,” Pattern Recognition, vol. 45, pp. 3084 – 3104, 2012

Multi-label classification: Related tasks





$\forall(\boldsymbol{x_i}, \boldsymbol{Y_i}) \in E,\ |Y_i|=1$ Multi-class classification

$\forall(\boldsymbol{x_i}, (\ldots, \lambda_j, \ldots)) \in E, \lambda_j \in \boldsymbol{R}$ Label ranking (*λ_j denotes ranking of the label *i*)

$\forall(\boldsymbol{x_i}, (\ldots, \lambda_j, \ldots)) \in E, |\lambda_j| > 2$ Multi-target multi-class classification

Multi-label classification can be seen as instance of:

- Multi-task learning
- Multiview learning

Multi-label classification: Subtasks in MLC

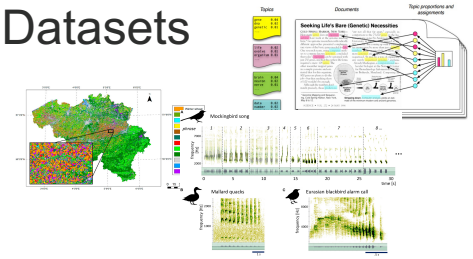
- Extreme multi-label classification
- Multi-label learning with limited supervision
  1. Multi-label learning with missing labels
  2. Semi-supervised multi-label classification
  3. Partial multi-label classification
  4. MLC with noisy labels
- Online multi-label learning
- Statistical multi-label learning

- Multi-instance multi-label learning ($h: 2^X \rightarrow 2^L$)
- Feature selection
- Labels selection
- Data augmentation
- MLC with Partial Abstention etc…

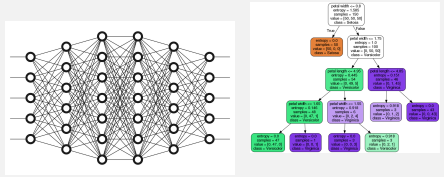
[Herrera et al. 2016] F. Herrera, A. J. Rivera, M. J. del Jesus, and F. Charte, Multilabel Classification. Springer Cham, Switzerland: Springer, 2016
[Liu et al. 2020] W. Liu, X. Shen, H. Wang, I. W. Tsang . The Emerging Trends of Multi-Label Learning, arxiv preprint, https://arxiv.org/pdf/2011.11197.pdf

Overview

Datasets



Methods



Experimental setup

Software

Design

Model selection (SRM & HPO)

Thresholding

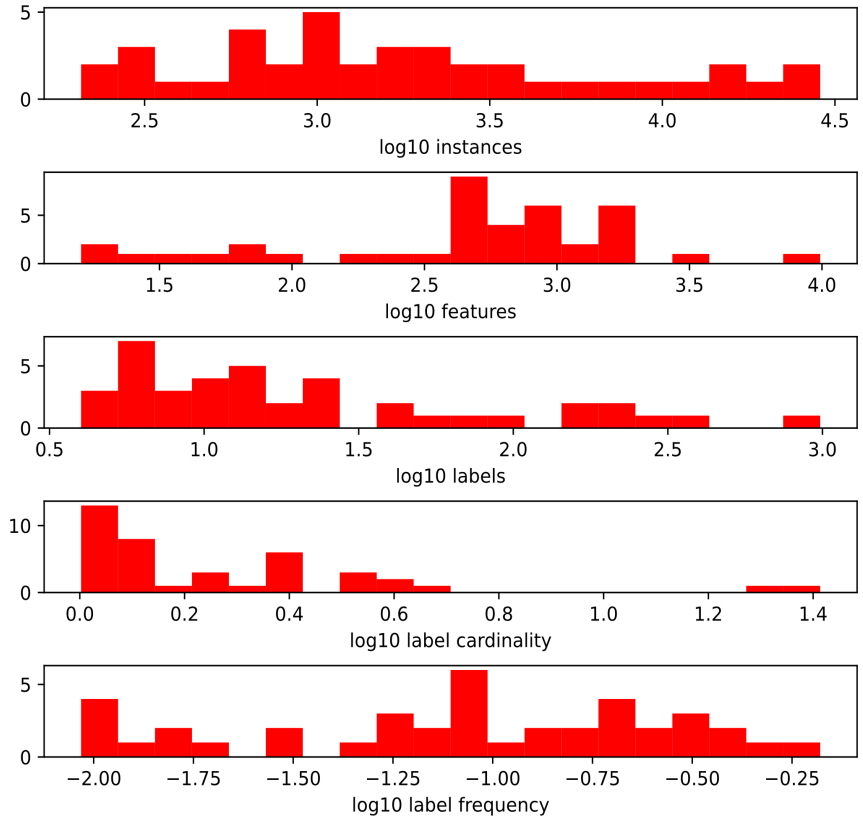
Results

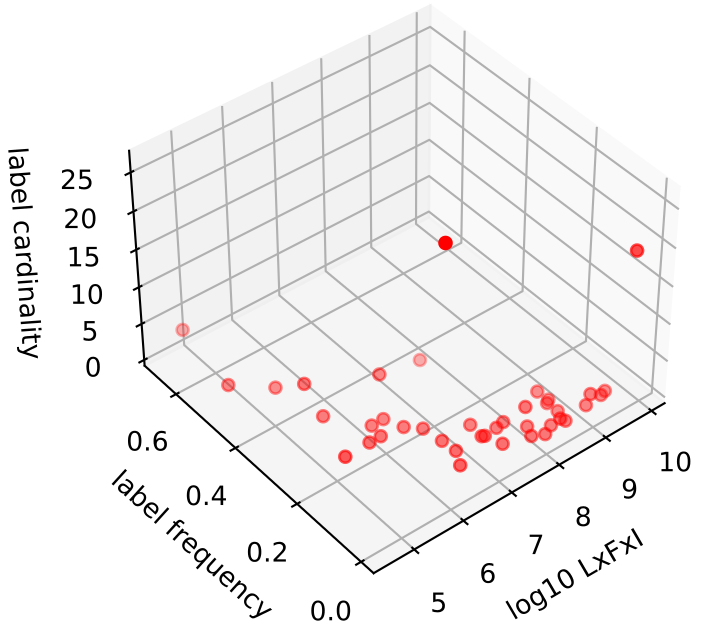
Discussion

Recommendations

Multi-label classification

MLC Datasets: Benchmark Landscape

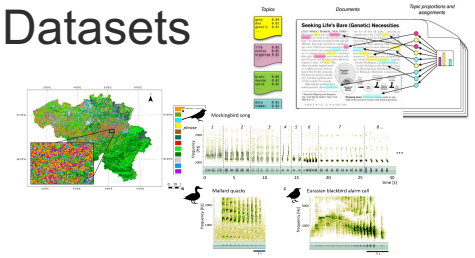




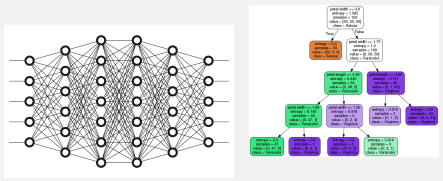
$$LC = \frac{1}{|D|}\sum_{i=1}^{|D|}|l_i| \qquad LD = \frac{1}{|D||L|}\sum_{i=1}^{|D|}|l_i| \qquad LxFxI = |L||F||I|$$

Overview

Datasets



Methods



Experimental setup

Software

Design

Model selection (SRM & HPO)

Thresholding

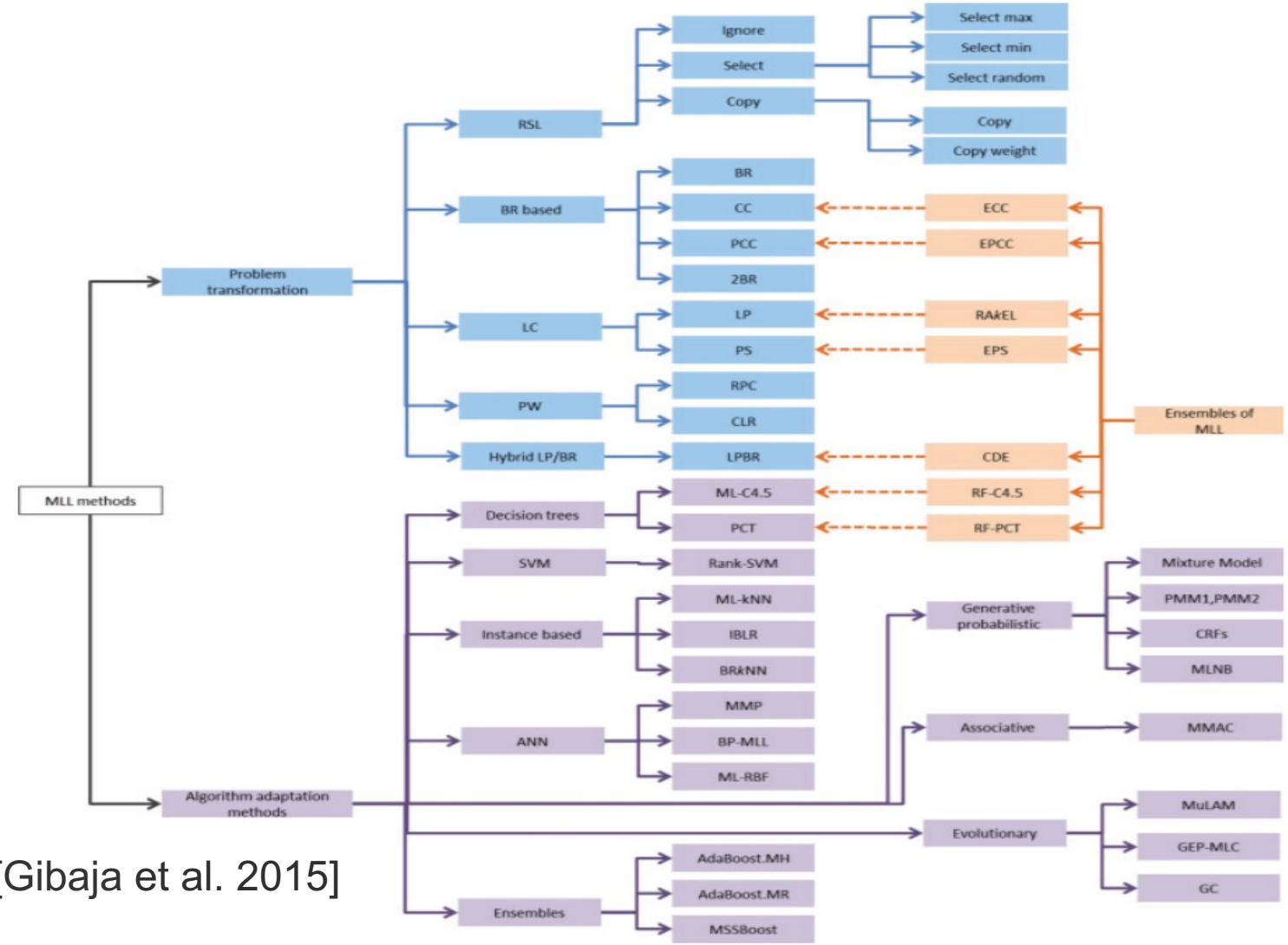
Results

Discussion

Recommendations

Multi-label classification

Mu hods



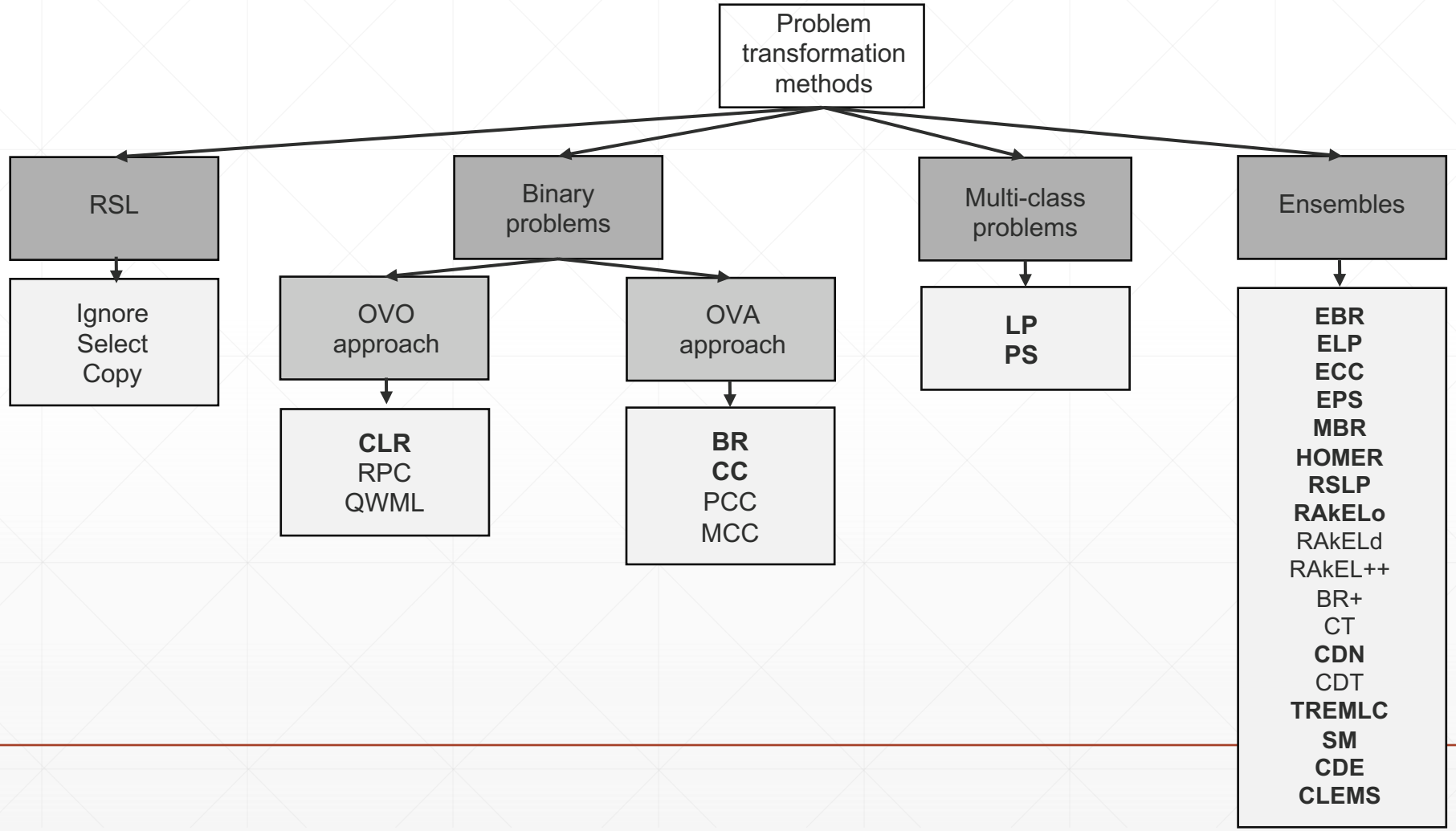
[Gibaja et al. 2015]

[Tsoumakas et al. 2007] Tsoumakas G, Katakis I. 2007. Multi-label classification: an overview. International Journal of Data Warehousing and Mining 3(3):1–13 DOI 10.4018/jdwm.2007070101.

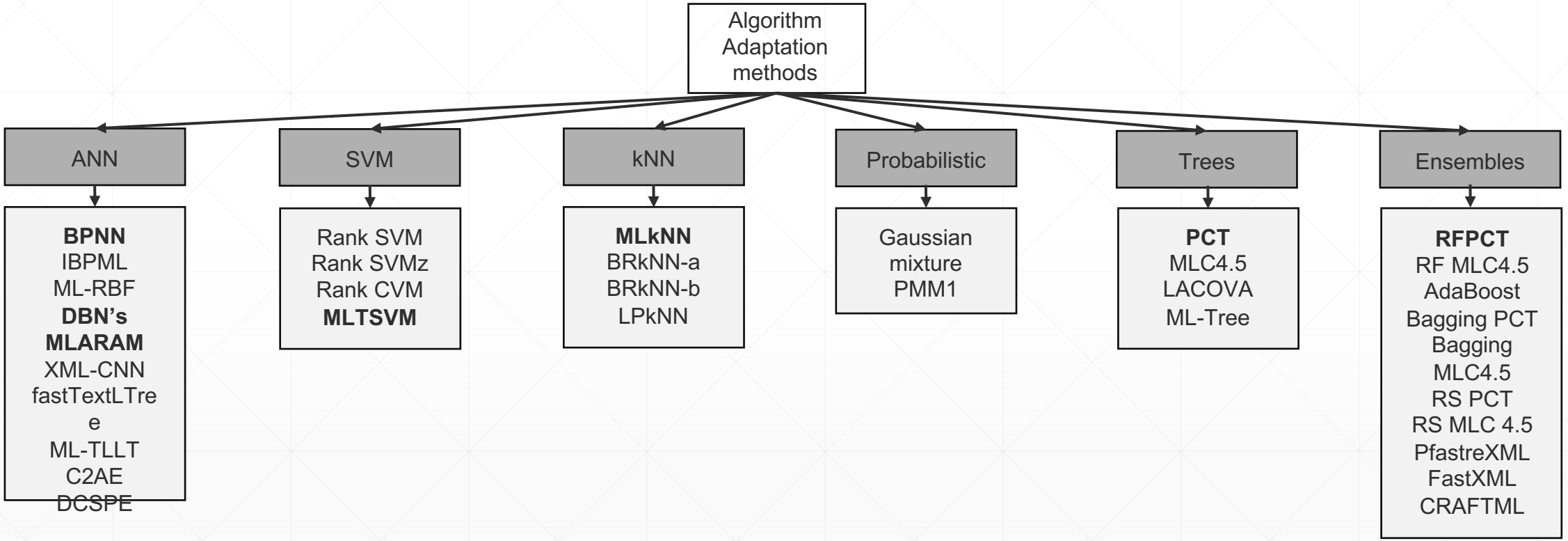
[Madjarov et al. 2012] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski, “An ˇ extensive experimental comparison of methods for multi-label learning,” Pattern Recognition, vol. 45, pp. 3084 – 3104, 2012

[Gibaja et al. 2015] E. Gibaja and S. Ventura, “A tutorial on multilabel learning,” ACM Computing Surveys, vol. 47, no. 3, pp. 52:1–52:38, 2015

MLC methods: Problem Transformation (PT) methods



MLC methods: Algorithm Adaptation (AA) methods



Datasets and methods

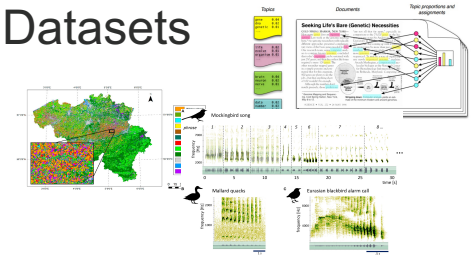
MLARAM, MLTSVM, BR, CLR, MLkNN, CC, PCT, DBN, BPNN, RFPCT, LP, PS, EBR, ELP, ECC, EPS, MBR, CLEMS, HOMER, RSLP, RAkELo, CDN,TREMLC, CDE, SM, CM,

42 datasets and 26 methods

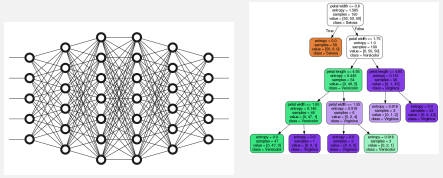
Dataset name	Domain	Training	Test	Features	Labels	LCardTr	LDeTr	LCardTs	LDeTs
ABPM	Medical	189	81	33	6	3.9683	0.6614	3.9877	0.6646
Foodtruck	Text	254	146	28	12	2.3268	0.1939	2.2671	0.1889
Flags	Multimedia	174	110	19	7	3.3793	0.4828	3.4273	0.4896
CHD 49	Medicine	372	183	49	6	2.5887	0.4315	2.5628	0.4271
Water quality	Chemistry	711	349	16	14	5.0689	0.3621	5.0802	0.3629
Emotions	Multimedia	413	270	72	6	1.9080	0.3180	1.8296	0.3049
Virus PseAAC	Bioinformatics	139	68	440	6	1.2734	0.2122	1.1029	0.1838
VirusGo	Bioinformatics	139	68	749	6	1.2734	0.2122	1.1029	0.1838
Gpositive PseAAC	Bioinformatics	448	171	440	4	1.0057	0.2514	1.0117	0.2529
Gpostive GO	Bioinformatics	348	171	912	4	1.0057	0.2514	1.0117	0.2529
Proteins Virus	Biology	206	62	1538	6	1.2292	0.2049	1.1774	0.1962
Yeast	Biology	1450	967	103	14	4.2566	0.3040	4.2079	0.3006
Birds	Multimedia	445	290	260	19	1.0404	0.0548	0.9655	0.0508
Scene	Multimedia	1447	960	294	6	1.0788	0.1798	1.0667	0.1778
Gnegative PseAAC	Bioinformatics	933	459	440	8	1.0407	0.1301	1.0566	0.1320
Plant PseACC	Bioinformatics	656	322	440	12	1.0854	0.0904	1.0652	0.0888
SPIRIT	Text	493	212	600	14	1.5781	0.1127	1.5613	0.1115
Cal500	Multimedia	362	230	68	174	25.9064	0.1489	26.2223	0.1507
20genres	Text	1075	461	600	20	1.3460	0.0673	1.3319	0.6656
Proteins Plants	Biology	673	288	1538	12	1.0847	0.0904	1.0625	0.0885
Gnegative GO	Bioinformatics	933	459	1717	8	1.0407	0.1301	1.0566	0.1321
Human PseAAC	Bioinformatics	2082	1024	440	14	1.1969	0.0855	1.1611	0.0829
Genbase	Biology	460	292	1185	27	1.2370	0.0458	1.2568	0.0466
Yelp	Text	7241	3565	671	5	1.6470	0.3294	1.6205	0.3241
Plant GO	Bioinformatics	656	322	3091	12	1.0854	0.0904	1.0652	0.0888
Proteins Humans	Biology	1663	713	1538	14	1.2790	0.0914	1.2637	0.0903
Medical	Text	648	420	1499	45	1.2253	0.0272	1.2714	0.0283
Slashdot	Text	2272	1513	1079	22	1.1787	0.0536	1.1844	0.0538
Enron	Text	1082	710	1001	53	3.3983	0.0641	3.3366	0.0630
Langlog	Text	940	610	1004	75	1.1489	0.0153	1.2295	0.0164
Arabic 200	Text	15000	8754	200	40	1.1073	0.0277	1.2536	0.0313
Stackexch. chess	Text	1005	670	585	227	2.4229	0.0107	2.3940	0.0106
Reutersk 500	Text	3607	2403	500	103	1.4649	0.0142	1.4580	0.0142
Tmc2007 500	Text	17190	11462	500	22	2.2253	0.1012	2.2107	0.1005
Ohsumed	Text	8392	5592	1002	23	1.6621	0.0723	1.6640	0.0724
Ng20	Text	11640	7750	1006	20	1.0288	0.0514	1.0291	0.0515
Bibtex	Text	4495	2990	1836	159	2.4220	0.0152	2.3625	0.0149
Human GO	Bioinformatics	2082	1024	9844	14	1.1969	0.0855	1.1611	0.0829
Stackexch. philsph	Text	2385	1593	842	233	2.2839	0.0098	2.2561	0.0097
Stackexch. cs	Text	5571	3717	635	274	2.5536	0.0093	2.5582	0.0093
Corel5k	Multimedia	3060	2030	499	374	3.5242	0.0094	3.5153	0.0094
Delicious	Text	9725	6470	500	983	19.0851	0.0194	18.9360	0.0193

Overview

Datasets



Methods



Experimental setup

Software

Design

Model selection (SRM & HPO)

Thresholding

Results

Discussion

Recommendations

Multi-label classification

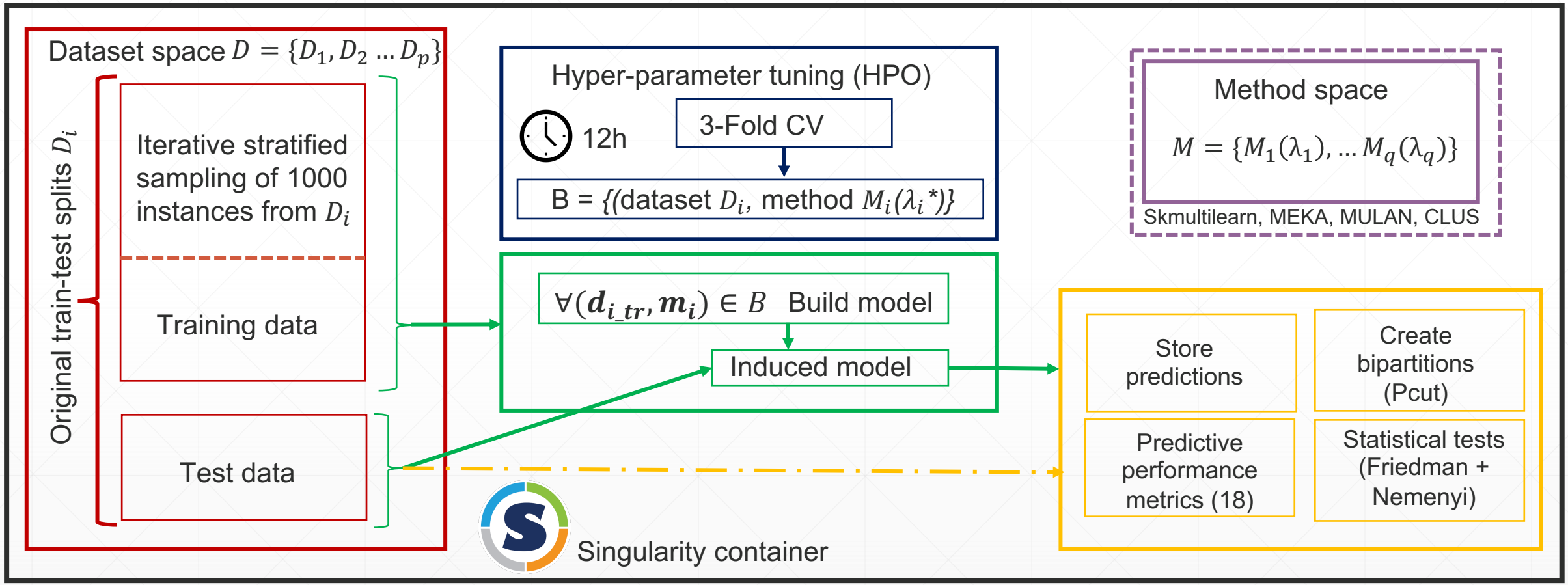
Experimental setup: Software tools

Properties	MULAN	MEKA	mldr	CLUS	Labiac	sckitmultilearn
Data format	.arff + XML	.arff	any	.arff	Labelset index:value	any
Require additional preprocessing on loading	No	No	Yes	No	No	Yes
type interface	programmatic	prog/GUI	prog/GUI/cmd	Prog/Cmd	cmd	programmatic
EDA tools	Yes	Yes	Yes	No	No	No
Partitioning (iterative/random/stra)	2	1	3	/	/	1
Programming lang.	Java	Java	R	Java	C++	Python

*For specific method implementation available check Table 4 [Charte 2020]

[Rivoli 2018] The utiml Package: Multi-label Classification in R, Adriano Rivolli and Andre C. P. L. F. de Carvalho, The R Journal , 24—37, vol. 10, 2018
[Charte 2020] A Comprehensive and Didactic Review on Multilabel Learning Software Tools, F. Charte IEEE Access 8, 50330-50354

Experimental setup: Design



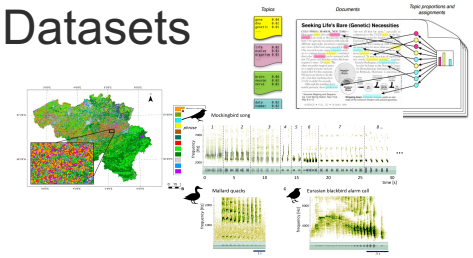
[Caruana 2006], R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006, pp. 161–168

[Dembczyński 2010] Dembczyński K., Waegeman W., Cheng W., Hüllermeier E. (2010) Regret Analysis for Performance Metrics in Multi-Label Classification: The Case of Hamming and Subset Zero-One Loss. In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol 6321. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15880-3_24

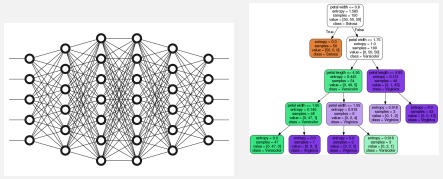
[Waegeman 2014] Willem Waegeman, Krzysztof Dembczyński, Arkadiusz Jachnik, Weiwei Cheng, and Eyke Hüllermeier. 2014. On the bayes-optimality of F-measure maximizers. J. Mach. Learn. Res. 15, 1 (January 2014), 3333–3388.

Overview

Datasets



Methods



Experimental setup

Software

Design

Model selection (SRM & HPO)

Thresholding

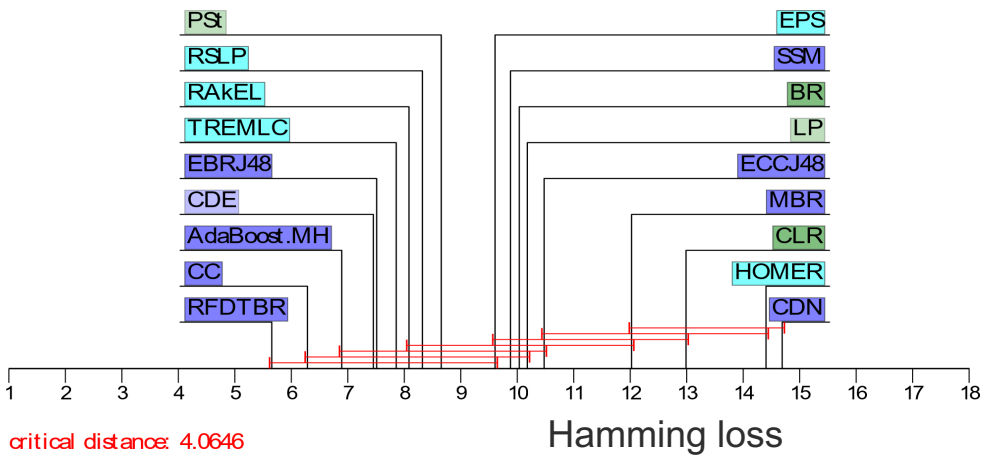
Results

Discussion

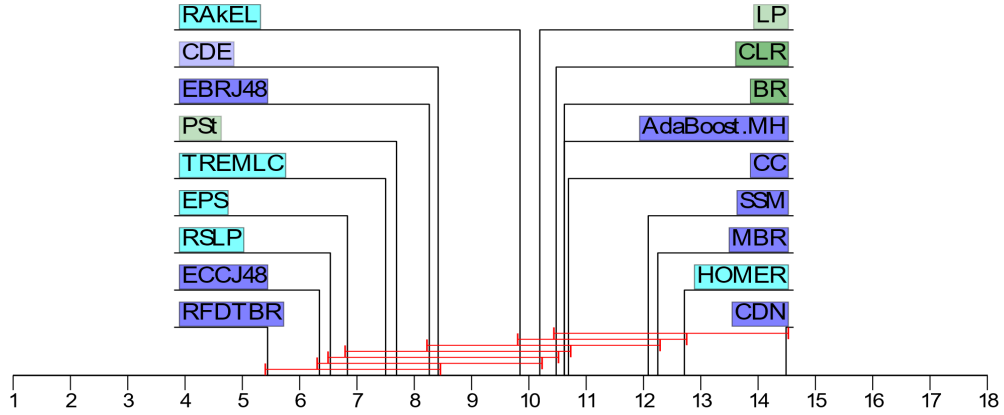
Recommendations

Multi-label classification

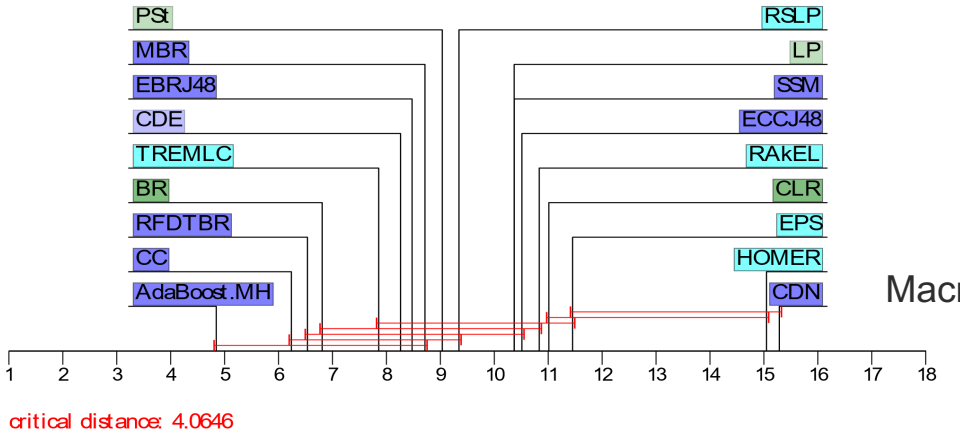
Comprehensive study: Results (PT methods)



Hamming loss

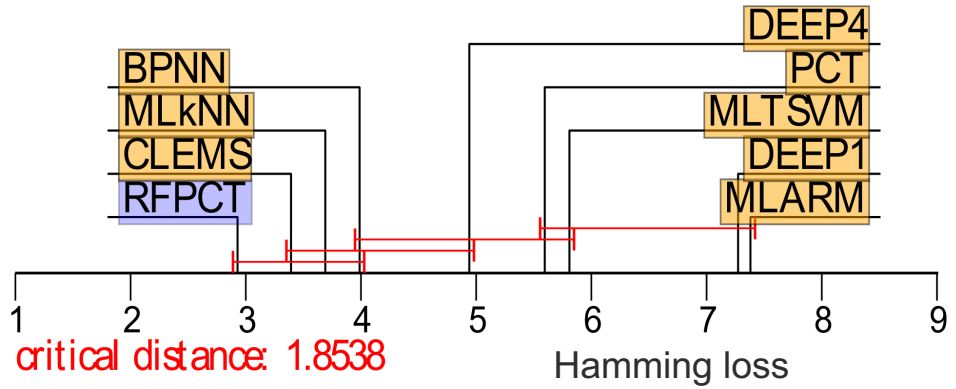


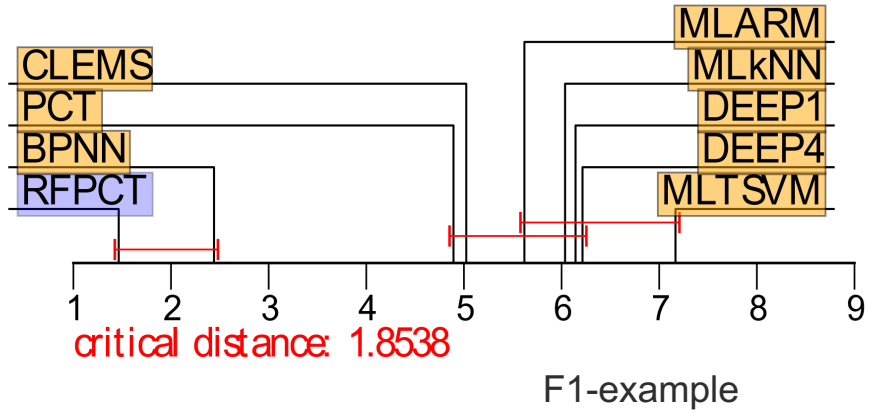
F1-example

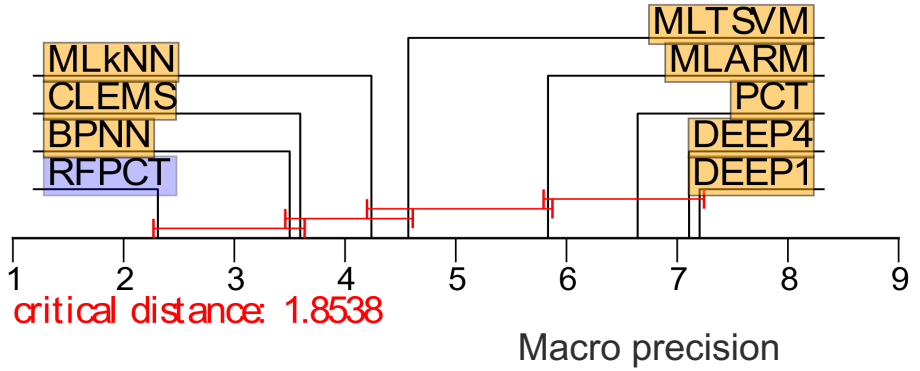


Macro precision

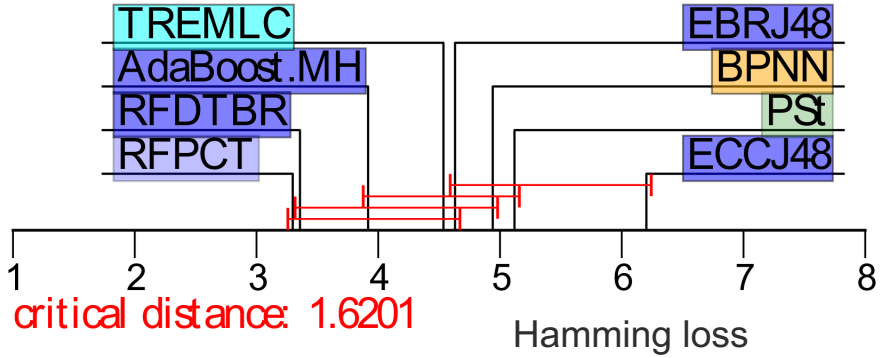
Comprehensive study: Results (AA methods)

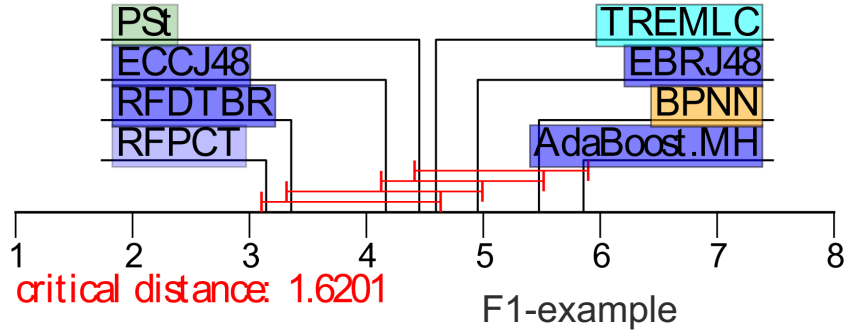


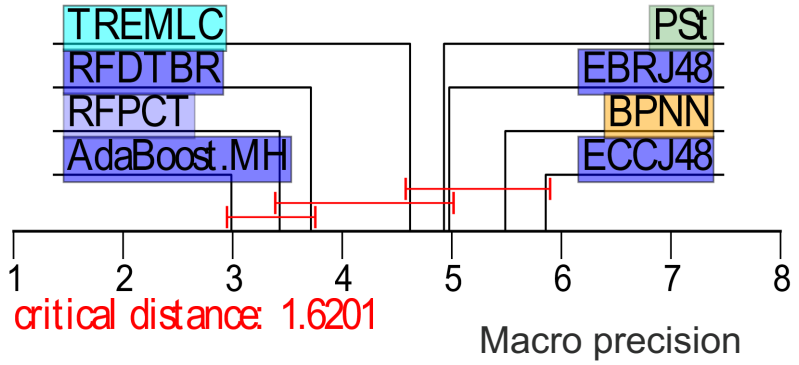




Comprehensive study: Results (best methods)







Comprehensive study: Recommendation

1. On example-based measures RFPCT and RFDTBR produce the best performance on 4 out of 6 measures, with RFPCT being in favour.
2. Methods that rely on LP achieve good performance on example-based recall, precision, F1, accuracy.
3. On label-based measures the BR-based methods (either single or ensemble) tend to perform the best. However, on micro measures, RFPCT has a similar competitive edge as these methods.
4. On ranking-based methods the competitive edge belongs to RFPCT, AdaBoost, EBRJ48 and RFDTBR.
5. Optimizing SVMs for the BR method is not worth it.

Comprehensive study: Recommendation

6. Compared by the family of methods, BR-based methods outperform LP and embedding-based methods in general.
7. Back-propagation Neural Networks is the best performing method from the algorithm adaptation methods.
8. Ensembles of problem transformation methods should be built with J48 instead of SVMs.
9. General values for the parameters of DBNs (Deep Belief Networks) should not be a practice.

Conclusion

- This is a large comprehensive study of MLC methods;
- It includes 26 methods, 42 datasets and 20 evaluation criteria;
- A description of the datasets and methods for MLC is given, with semantic and meta properties
- The descriptions allow for the inclusion of the methods and datasets into ontology and later providing various operations such as querying;
- We adhere to the literature recognized standards for large scale empirical studies with time budgeted constraint;
- We parametrized the methods that require parameterization according to literature recognized ranges for their values;
- An extensive empirical evaluation of the existing methods across the benchmark datasets outlining the landscape of methods for MLC.

Conclusion

- An extensive empirical evaluation of the existing methods across the benchmark datasets outlining the landscape of methods for MLC.

In a nutshell: The methods **RFDTBR**, **RFPCT**, **ECCJ48**, **EBRJ48** and **AdaBoost** have it all; providing premium predictive performance, while the first two are also being very efficient.

Personal information

Bogatinovski Jasmin

Personal web page: https://bogatinovskijasmin.github.io/bogatinovskijasmin/aboutme/
Google Scholar: https://scholar.google.si/citations?user=GgMAEc0AAAAJ&hl=en
LinkedIn: https://www.linkedin.com/in/jasmin-bogatinovski-6b697a116

Huawei-TUB Innovation Lab
https://huaweitubinovationlab.github.io/