MULTI-LABEL CLASSIFICATION

X_1	X_2	X_3	X_4	X_5	C_1	C_2	C_3	C_4
3.2	1.4	4.7	7.5	3.7	1	0	1	1
2.8	6.3	1.6	4.7	2.7	0	0	1	0
7.7	6.2	4.1	3.3	7.7	1	0	1	1
9.2	0.4	2.8	0.5	3.9	0	1	0	0
5.5	5.3	4.9	0.6	6.6	1	1	0	1

OUTLINE – MULTI-LABEL

- Basic vocabulary and framework
- Applications
- Evaluation metrics
- Overview of techniques
- Software and references

SINGLE versus MULTI-LABEL

X_1	X_2	X_3	X_4	X_5	C
3.2	1.4	4.7	7.5	3.7	1
2.8	6.3	1.6	4.7	2.7	0
7.7	6.2	4.1	3.3	7.7	1
9.2	0.4	2.8	0.5	3.9	0
5.5	5.3	4.9	0.6	6.6	1

X_1	X_2	X_3	X_4	X_5	C_1	C_2	C_3	C_4
3.2	1.4	4.7	7.5	3.7	1	0	1	1
2.8	6.3	1.6	4.7	2.7	0	0	1	0
7.7	6.2	4.1	3.3	7.7	1	0	1	1
9.2	0.4	2.8	0.5	3.9	0	1	0	0
5.5	5.3	4.9	0.6	6.6	1	1	0	1

MULTI-LABEL CLASSIFICATION

[PDF] **Multilabel text classification** for automated tag suggestion

I Katakis, G Tsoumakas, I Vlahavas - Proceedings of the ECML/PKDD, 2008 - Citeseer

The increased popularity of tagging during the last few years can be mainly attributed to its embracing by most of the recently thriving user-centric content publishing and management Web 2.0 applications. However, tagging systems have some limitations that have led …

Save Cite Cited by 387 Related articles All 10 versions

Deep learning for extreme **multi-label text classification**

J Liu, WC Chang, Y Wu, Y Yang - … of the 40th international ACM SIGIR …, 2017 - dl.acm.org

Extreme **multi-label text classification** (XMTC) refers to the problem of assigning to each document its most relevant subset of class labels from an extremely large label collection, where the number of labels could reach hundreds of thousands or millions. The huge label …

Save Cite Cited by 377 Related articles All 3 versions

Large-scale **multi-label text classification**—revisiting neural networks

J Nam, J Kim, EL Mencía, I Gurevych… - … european conference on …, 2014 - Springer

Neural networks have recently been proposed for **multi-label classification** because they are able to capture and model label dependencies in the output layer. In this work, we investigate limitations of BP-MLL, a neural network (NN) architecture that aims at minimizing …

Save Cite Cited by 356 Related articles All 14 versions

Multi-label text classification with a mixture model trained by EM

AK McCallum - AAAI 99 workshop on **text** learning, 1999 - Citeseer

In many important document **classification** tasks, documents may each be associated with multiple class labels. This paper describes a Bayesian **classification** approach in which the multiple classes that comprise a document are represented by a mixture model. While the …

Save Cite Cited by 703 Related articles All 4 versions

FRAMEWORK

$\mathcal{X}$: d-dimensional input space

$\mathcal{Y}$: output space of q labels $\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$

$\mathcal{S}$: multi-label training set of m samples, $\{(\mathbf{x}_i, \mathbf{y}_i) | 1 \leq i \leq m\}$

h: multi-label classifier, $h: \mathcal{X} \rightarrow 2^{\mathcal{Y}}$

or ranking the associated labels to a sample **x**,

e.g. $r_{\mathbf{x}}(\lambda_2) < r_{\mathbf{x}}(\lambda_4) < r_{\mathbf{x}}(\lambda_1) < r_{\mathbf{x}}(\lambda_3)$

FRAMEWORK

Given a set of initial labels $L=\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$

Given a new instance **x**

Multi-label classification → outputs a *bipartition* of the set of labels, i.e. positive and negative ones,

P_x: $\{\lambda_1, \lambda_4\}$ and

N_x: $\{\lambda_2, \lambda_3, \lambda_5\}$

X_1	X_2	X_3	X_4	X_5	$Y \subseteq \mathcal{Y}$
3.2	1.4	4.7	7.5	3.7	$\{\lambda_1, \lambda_4\}$
2.8	6.3	1.6	4.7	2.7	$\{\lambda_3, \lambda_4\}$
7.7	6.2	4.1	3.3	7.7	$\{\lambda_1, \lambda_4\}$
9.2	0.4	2.8	0.5	3.9	$\{\lambda_2\}$
5.5	5.3	4.9	0.6	6.6	$\{\lambda_1, \lambda_2, \lambda_3\}$

APPLICATIONS - TEXT

ACM COMPUTING CLASSIFICATION (Veloso et al. 2007)

- A *document* described by its title, abstract, citation, autorship: huge feature space
- First hierarchy level, 11 labels: general literature, hardware, software, information systems...
- Second hierarchy level with 81 labels
- 81,251 digital archives

acm

COMPUTING
CLASSIFICATION
SYSTEM

APPLICATIONS - TEXT

REUTERS CORPUS (Lewis et al. 2004)

- 804,414 newswire stories
- To be indexed in 103 topic codes
- Words: huge and sparse feature space
- A benchmark in multi-label learning

THOMSON REUTERS

APPLICATIONS – e-MAIL

ENRON COMPANY e-MAILS

- UC Berkeley Enron e-mail analysis project
- Company-professional e-mails of about 150 Enron senior managers
- 1,702 samples, 53 labels
- Public datasets during a Federal Energy Regulatory Commission investigation

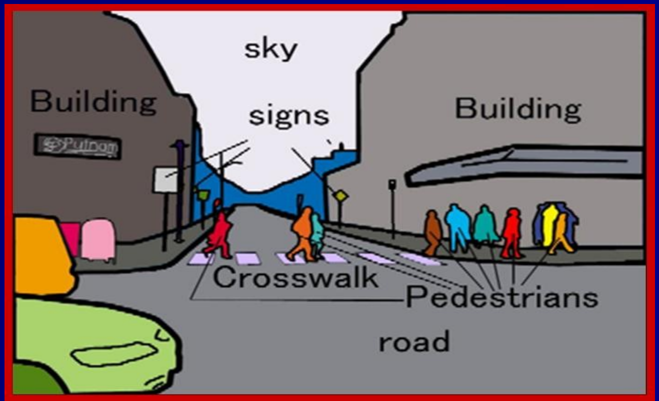
OTHER APPLICATIONS

- BIOLOGY:
  - Annotation of protein functions
  - Gene ontology annotations (e.g. of a gene)

- DIRECT MARKETING:
  - Product offers to customers

OTHER APPLICATIONS

- IMAGE AND AUDIO:
  - Simultaneous object class recognition
  - Demographic classification of facial images: sex, age, ethnicity...
  - Music categorization: instruments, country, rhythm...
  - Categorization of song emotions: happy, calm, amazed...



BENCHMARK DATASETS

- MULAN multi-label datasets
- MEKA multi-label datasets

			attributes					
name	domain	instances	nominal	numeric	labels	cardinality	density	distinct
bibtex	text	7395	1836	0	159	2.402	0.015	2856
birds NEW!	audio	645	2	258	19	1.014	0.053	133
bookmarks	text	87856	2150	0	208	2.028	0.010	18716
CAL500	music	502	0	68	174	26.044	0.150	502
corel5k	images	5000	499	0	374	3.522	0.009	3175
corel16k (10 samples)	images	13811±87	500	0	161±9	2.867±0.033	0.018±0.001	4937±158
delicious	text (web)	16105	500	0	983	19.020	0.019	15806
emotions	music	593	0	72	6	1.869	0.311	27
enron	text	1702	1001	0	53	3.378	0.064	753
EUR-Lex (directory codes)	text	19348	0	5000	412	1.292	0.003	1615
EUR-Lex (subject matters)	text	19348	0	5000	201	2.213	0.011	2504
EUR-Lex (eurovoc descriptors)	text	19348	0	5000	3993	5.310	0.001	16467

BENCHMARK DATASETS

Multi-Label Classification Dataset Repository

Dataset	Domain	m	d	q	Card	Dens	Div	avgIR	rDep	m×q×d	Full dataset Meka
20NG	Text	19300	1006	20	1.029	0.051	0.003	1.007	0.984	3.88E+08	
3s-bbc1000	Text	352	1000	6	1.125	0.188	0.234	1.718	0.733	2.11E+06	
3s-guardian1000	Text	302	1000	6	1.126	0.188	0.219	1.773	0.667	1.81E+06	
3s-inter3000	Text	169	3000	6	1.142	0.190	0.172	1.766	0.400	3.04E+06	
3s-reuters1000	Text	294	1000	6	1.126	0.188	0.219	1.789	0.667	1.76E+06	
Bibtex	Text	7395	1836	159	2.402	0.015	0.386	12.498	0.111	2.16E+09	
Birds	Audio	645	260	19	1.014	0.053	0.206	5.407	0.123	3.19E+06	
Bookmarks	Text	87860	2150	208	2.028	0.010	0.213	12.308	0.315	3.93E+10	

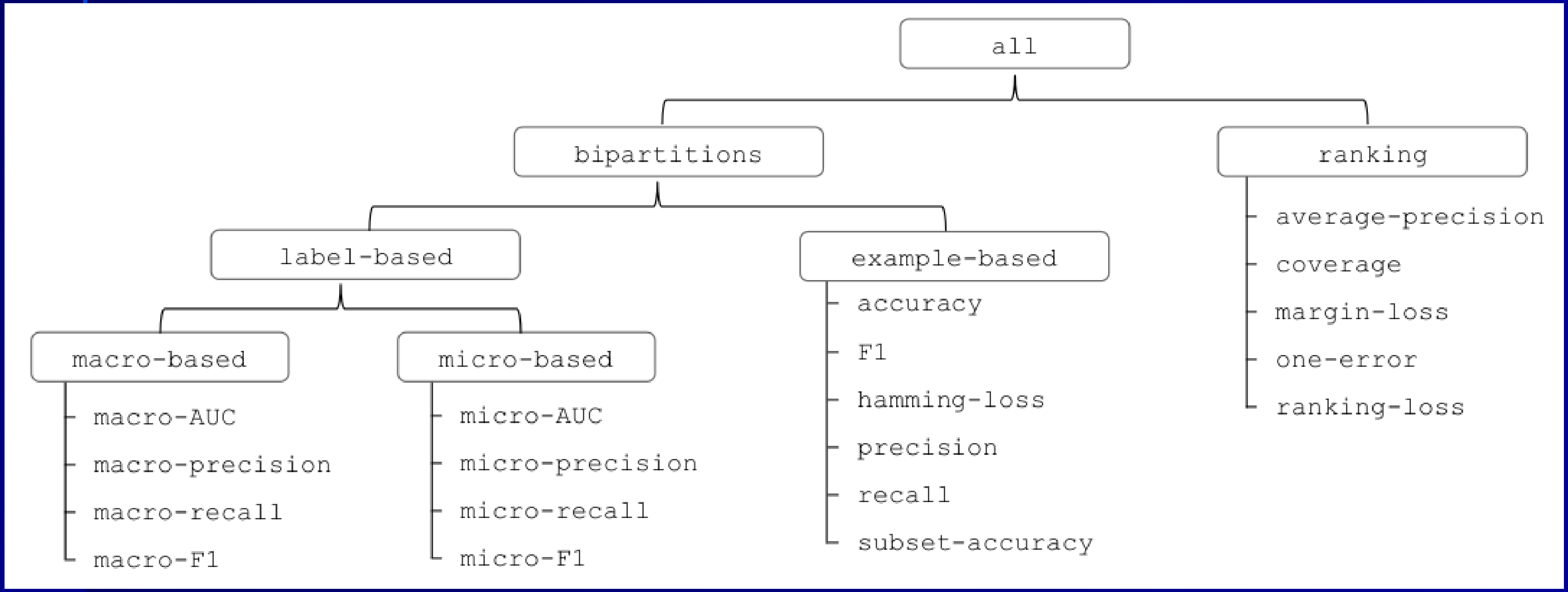
EVALUATION METRICS

- X : d-dimensional input space $\boldsymbol{X}=(X_1,\ldots X_m)$
- Y : output space of q labels $\{\lambda_1,\lambda_2,\ldots,\lambda_q\}$, real labels, $Y=\{\lambda_7,\lambda_9\}$
- S : multi-label training set of m samples, $\{(x_i,Y_i)|1\leq i\leq m\}$

- h: multi-label classifier, $h: X \rightarrow 2^y$
- *Predicted labels for a example*, $Pred_h(Y)=\{\lambda_6,\lambda_9,\lambda_{11},\lambda_{15}\}$

- Evaluation metrics:
  - Example-based:
    - → calculated separately **for each sample** and averaged
  - Label-based:
    - → calculated separately **for each label** and averaged

EVALUATION METRICS



0/1 SUBSET ACCURACY EXACT MATCH - PER SAMPLE

	$Y^{(i)}$	$\hat{Y}^{(i)}$
$\boldsymbol{x}^{(1)}$	$\{\lambda_1, \lambda_3\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(2)}$	$\{\lambda_2, \lambda_4\}$	$\{\lambda_2, \lambda_4\}$
$\boldsymbol{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(4)}$	$\{\lambda_2, \lambda_3\}$	$\{\lambda_2\}$
$\boldsymbol{x}^{(5)}$	$\{\lambda_1\}$	$\{\lambda_1, \lambda_4\}$

- EXACT MATCH = 1/5 x (0+1+1+0+0)
- For each sample → checks whether the predicted set of labels is an exact match of the true set of labels
- Very strict evaluation

ACCURACY – PER SAMPLE

	$Y^{(i)}$	$\hat{Y}^{(i)}$
$\boldsymbol{x}^{(1)}$	$\{\lambda_1, \lambda_3\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(2)}$	$\{\lambda_2, \lambda_4\}$	$\{\lambda_2, \lambda_4\}$
$\boldsymbol{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(4)}$	$\{\lambda_2, \lambda_3\}$	$\{\lambda_2\}$
$\boldsymbol{x}^{(5)}$	$\{\lambda_1\}$	$\{\lambda_1, \lambda_4\}$

$$\text{ACCURACY} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}^{(i)} \cap Y^{(i)}|}{|\hat{Y}^{(i)} \cup Y^{(i)}|} = \frac{1}{5}\left(\frac{1}{3} + \frac{2}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2}\right)$$

$$\text{PRECISION} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}^{(i)} \cap Y^{(i)}|}{|\hat{Y}^{(i)}|} = \frac{1}{5}\left(\frac{1}{2} + \frac{2}{2} + \frac{2}{2} + \frac{1}{1} + \frac{1}{2}\right)$$

$$\text{RECALL} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{Y}^{(i)} \cap Y^{(i)}|}{|Y^{(i)}|} = \frac{1}{5}\left(\frac{1}{2} + \frac{2}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{1}\right)$$

- Scores → from the “Information Retrieval” area
- Using AND and OR logical operations

HAMMING LOSS PER SAMPLE

$\mathbf{x}$	$Y^{(i)}$	$\hat{Y}^{(i)}$
$\mathbf{x}^{(1)}$	$\{\lambda_1, \lambda_3\}$	$\{\lambda_1, \lambda_4\}$
$\mathbf{x}^{(2)}$	$\{\lambda_2, \lambda_4\}$	$\{\lambda_2, \lambda_4\}$
$\mathbf{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$	$\{\lambda_1, \lambda_4\}$
$\mathbf{x}^{(4)}$	$\{\lambda_2, \lambda_3\}$	$\{\lambda_2\}$
$\mathbf{x}^{(5)}$	$\{\lambda_1\}$	$\{\lambda_1, \lambda_4\}$

$$\text{HAMMING LOSS} = \frac{1}{d} \cdot \frac{1}{N} \sum_{i=1}^{N} |\hat{Y}^{(i)} \Delta Y^{(i)}|$$

- HAMMING LOSS = 1/4 x 1/5 (2+0+0+1+1)
- Symmetric difference between both sets: XOR operation
- Average binary classification error

LABEL-BASED METRICS
METRICS PER LABEL

Contingency Table for λ_j		Actual Value	
		POS	NEG
Learner Output	POS	TP_j	FP_j
	NEG	FN_j	TN_j

$$\mathrm{ACC}_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$\mathrm{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\mathrm{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

$$\mathrm{F1}_i = \frac{2 * \mathrm{Precision}_i * \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

LABEL-BASED METRICS
METRICS PER LABEL

$$\mathrm{ACC}_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$\mathrm{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\mathrm{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

$$\mathrm{F1}_i = \frac{2 * \mathrm{Precision}_i * \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

	$Y^{(i)}$	$\hat{Y}^{(i)}$
$\boldsymbol{x}^{(1)}$	$\{\lambda_1, \lambda_3\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(2)}$	$\{\lambda_2, \lambda_4\}$	$\{\lambda_2, \lambda_4\}$
$\boldsymbol{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(4)}$	$\{\lambda_2, \lambda_3\}$	$\{\lambda_2\}$
$\boldsymbol{x}^{(5)}$	$\{\lambda_1\}$	$\{\lambda_1, \lambda_4\}$

- ACCURACY [LABEL_4] = (2+1)/(2+1+2+0)
- PRECISION [LABEL_4] = (2)/(2+2)
- RECALL-SENSITIVITY [LABEL_4] = (2)/(2+0)

LABEL-BASED METRICS
MACRO vs. MICRO

$$\mathrm{ACC}_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

$$\mathrm{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\mathrm{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

$$\mathrm{F1}_i = \frac{2 * \mathrm{Precision}_i * \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

Contingency Table for λ_j		Actual Value	
		POS	NEG
Learner Output	POS	TP_j	FP_j
	NEG	FN_j	TN_j

Macro-averaging

- Ordinary averaging of a binary measure
- $B_{macro} = \frac{1}{d}\sum_{j=1}^{d} B(TP_j, FP_j, TN_j, FN_j)$

Micro-averaging

- Labels as different instances of the same global label
- $B_{micro} = B(\sum_{j=1}^{d} TP_j, \sum_{j=1}^{d} FP_j, \sum_{j=1}^{d} TN_j, \sum_{j=1}^{d} FN_j)$

$$\mathrm{Precision}_{macro} = \frac{\sum_{i=1}^{q} \mathrm{Precision}_i}{q}$$

$$\mathrm{Precision}_{micro} = \frac{\sum_{i=1}^{q} TP_i}{\sum_{i=1}^{q}(TP_i + FP_i)}$$

TECHNIQUES - TAXONOMY

1. Problem transformation methods:
   - In several single-label tasks
   - Algorithm independent

2. Algorithm adaptation methods:
   - Extending supervised algorithms to multi-label data
   - Decision trees, SVM, Bayesian networks, K-NN...

TECHNIQUES - TAXONOMY

Strategy function	Description	Approach[a]	Reference
`baseline`	Baseline	-	Metz et al. (2012)
`br`	Binary Relevance	BR	Tsoumakas et al. (2010)
`brplus`	BR+	BR, STA	Cherman et al. (2012)
`cc`	Classifier Chains	BR, CC	Read et al. (2009)
`clr`	Calibrated Label Ranking	PW	Brinker et al. (2006)
`ctrl`	ConTRolled Label correlation	BR, ENS	Li and Zhang (2014)
`dbr`	Dependent Binary Relevance	BR, STA	Montañes et al. (2014)
`ebr`	Ensemble of Binary Relevance	BR, ENS	Read et al. (2009)
`ecc`	Ensemble of Classifier Chains	BR, CC, ENS	Read et al. (2009)
`eps`	Ensemble of Pruned Set	ENS, PS	Read et al. (2008)
`homer`	Hierarchy Of Multi-label classifiER	HIE	Tsoumakas et al. (2008)
`lift`	Learning with Label specIfic FeaTures	BR, CLU	Zhang and Wu (2015)
`lp`	Label Powerset	PS	Tsoumakas and Katakis (2007)
`mbr`	Meta-BR, 2BR or stacking	BR, STA	Tsoumakas et al. (2009)
`mlknn`	Multi-label kNN	AD	Zhang and Zhou (2007)
`ns`	Nested Stacking	BR, CC	Senge et al. (2013)
`ppt`	Pruned Problem Transformation	PS	Read et al. (2008)
`prudent`	PRUned and confiDENT Stacking	BR, STA	Alali and Kubat (2015)
`ps`	Pruned Set	PS	Read (2008)
`rakel`	Random k-labelsets	ENS, PS	Tsoumakas and Vlahavas (2007)
`rdbr`	Recursive Dependent Binary Relevance	BR, ENS, STA	Rauber et al. (2014)
`rpc`	Ranking by Pairwise Comparison	PW	Hüllermeier et al. (2008)

[a] AD = Adaptation; BR = Binary transformation; CC = Chain of classifiers; CLU = Clustering based; ENS = Ensemble; HIE = Hierarchy; PS = Powerset transformation; PW = Pairwise transformation; STA = Stacking

Table 3: Strategies available in the **utiml** package

PROBLEM TRANSFORMATION BINARY RELEVANCE - BR

$\boldsymbol{x}$	$Y \subseteq \mathcal{Y}$
$\boldsymbol{x}^{(1)}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(2)}$	$\{\lambda_3, \lambda_4\}$
$\boldsymbol{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$
$\boldsymbol{x}^{(4)}$	$\{\lambda_2\}$
$\boldsymbol{x}^{(5)}$	$\{\lambda_1, \lambda_2, \lambda_3\}$

- Learning one binary classifier per class
- Output the union of their predictions
- Not consider label relationships
- Ensemble of BR base classifiers → common

$\boldsymbol{x}$	λ_1
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(2)}$	false
$\boldsymbol{x}^{(3)}$	true
$\boldsymbol{x}^{(4)}$	false
$\boldsymbol{x}^{(5)}$	true

$\boldsymbol{x}$	λ_2
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(2)}$	false
$\boldsymbol{x}^{(3)}$	false
$\boldsymbol{x}^{(4)}$	true
$\boldsymbol{x}^{(5)}$	true

$\boldsymbol{x}$	λ_3
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(2)}$	true
$\boldsymbol{x}^{(3)}$	false
$\boldsymbol{x}^{(4)}$	false
$\boldsymbol{x}^{(5)}$	true

$\boldsymbol{x}$	λ_4
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(2)}$	true
$\boldsymbol{x}^{(3)}$	true
$\boldsymbol{x}^{(4)}$	false
$\boldsymbol{x}^{(5)}$	false

PROBLEM TRANSFORMATION LABEL POWERSET – LC-LP

$\boldsymbol{x}$	$Y \subseteq \mathcal{Y}$	Label
$\boldsymbol{x}^{(1)}$	$\{\lambda_1, \lambda_4\}$	1001
$\boldsymbol{x}^{(2)}$	$\{\lambda_3, \lambda_4\}$	0011
$\boldsymbol{x}^{(3)}$	$\{\lambda_1, \lambda_4\}$	1001
$\boldsymbol{x}^{(4)}$	$\{\lambda_2\}$	0100
$\boldsymbol{x}^{(5)}$	$\{\lambda_1, \lambda_2, \lambda_3\}$	1110

- Each set of labels → recodify as a different class value
- e.g. "1001" → classA, "0011" → classB, etc...
- → a new single-class-variable classification task
- Limited training samples for many new labelsets
- High complexity
- Can not predict unseen labelsets

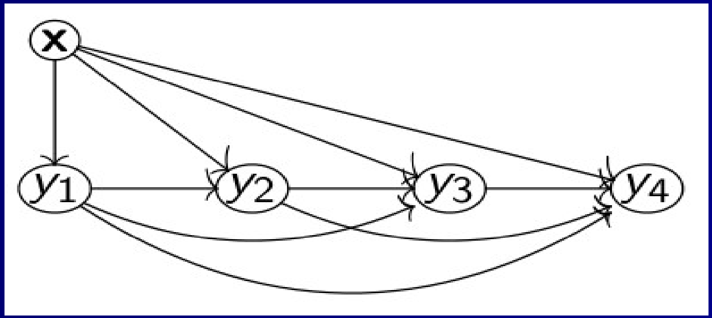
PROBLEM TRANSFORMATION
PRUNED SETS – PS

Labelset	Count
$\{\lambda_1\}$	12
$\{\lambda_2\}$	10
$\{\lambda_2, \lambda_3\}$	9
$\{\lambda_4\}$	7
$\{\lambda_3, \lambda_4\}$	2
$\{\lambda_1, \lambda_2, \lambda_3\}$	3

Labelset	Count
$\{\lambda_1\}$	13
$\{\lambda_2\}$	11
$\{\lambda_2, \lambda_3\}$	10
$\{\lambda_4\}$	9

- Start considering all labelsets → too many!! → Reduce labelsets
- Prune examples belonging to less frequent classes (e.g. $<$ 7)
- Distribute pruned examples → along more frequent subsets of their labelset
- Reduce the number of labelsets and focus on frequent ones
- Train a label Powerset multi-label classifier

PROBLEM TRANSFORMATION CLASSIFIER CHAINS - CC



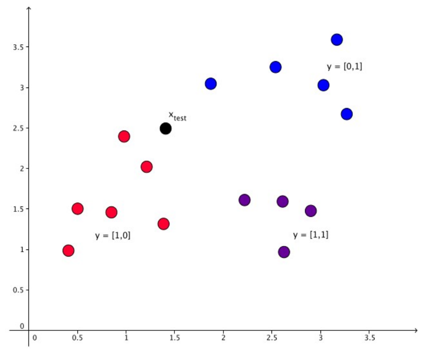
- Given a dataset with |L| labels: $y_1, y_2, \ldots, y_{|L|}$
- Dataset is transformed in |L| datasets where instances in the "j" dataset are of the form:

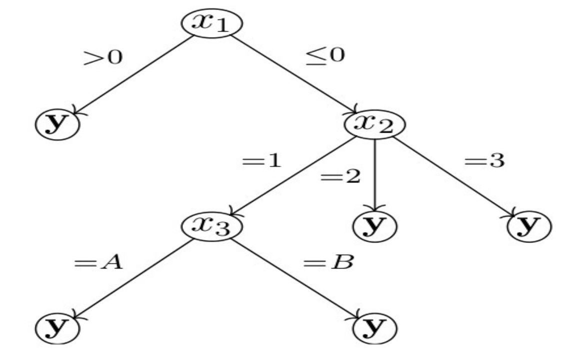
$$((x_i, y_1, y_2, \ldots, y_{j-1}), y_j)$$

- Classifiers build a CHAIN → each learns a binary classification of a single label
- Features in each classifier → EXTENDED with binary labels indicating the prediction of previous labels-classifiers in the chain
- Partial label dependence is maintained, but... order of the chain?

ALGORITHM ADAPTATION

- Extending supervised algorithms to deal with multi-label data
- Literature shows plenty of examples – Just a couple:
  - K-NN: assigns to **x** the most common labels of its K neighbours
  - Decision trees: extending the concept of multi-label entropy. Multiple labels at leaves





REMARKS

- Hot topic – specially in NLP
- Closely related with "tagging"-"annotation", news' categories, web 2.0, multiple outputs, learning from crowds, recommender systems…
- Many real world applications

- Software:
  - MULAN: WEKA-based library
  - MEKA: WEKA-based framework and GUI
  - Utiml, mldr, mldr.datasets R packages

- Datasets' repositories:
  - http://mulan.sourceforge.net/datasets.html
  - http://meka.sourceforge.net/#datasets
  - http://www.uco.es/kdis/mllresources/

EXTENDED INFO

- Talks-Tutorials:
  - C. Bielza, P. Larrañaga, UPM-Madrid [link]
  - J. Read, MEKA's programmer [link]
  - G. Tsoumakas et al., tutorial [link]

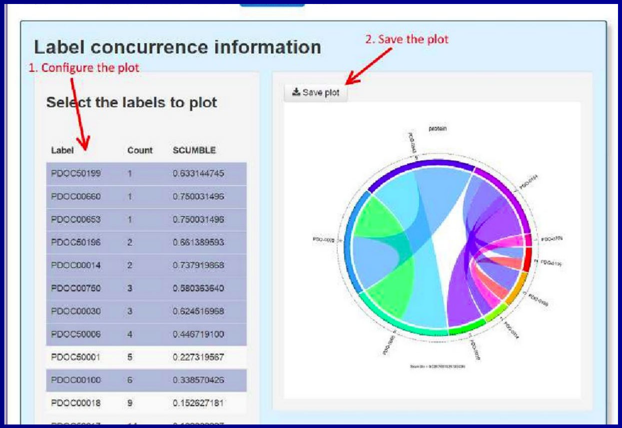
- Review:
  - M-L. Zhang, Z-H. Zhou (2013). "A review on multi-label learning algorithms". IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819-1837
  - F. Herrera, F. Charte, A.J. Rivera, M.J. Del Jesús (2016). Multi-label Classification. Springer

SOFTWARE

The utiml Package: Multi-label Classification in R

by Adriano Rivolli and Andre C. P. L. F. de Carvalho

Working with Multilabel Datasets in R: The **mldr** Package



Strategy function	Description	Approach[a]	Reference
`baseline`	Baseline	-	Metz et al. (2012)
`br`	Binary Relevance	BR	Tsoumakas et al. (2010)
`brplus`	BR+	BR, STA	Cherman et al. (2012)
`cc`	Classifier Chains	BR, CC	Read et al. (2009)
`clr`	Calibrated Label Ranking	PW	Brinker et al. (2006)
`ctrl`	ConTRolled Label correlation	BR, ENS	Li and Zhang (2014)
`dbr`	Dependent Binary Relevance	BR, STA	Montañes et al. (2014)
`ebr`	Ensemble of Binary Relevance	BR, ENS	Read et al. (2009)
`ecc`	Ensemble of Classifier Chains	BR, CC, ENS	Read et al. (2009)
`eps`	Ensemble of Pruned Set	ENS, PS	Read et al. (2008)
`homer`	Hierarchy Of Multi-label classifiER	HIE	Tsoumakas et al. (2008)
`lift`	Learning with Label specIfic FeaTures	BR, CLU	Zhang and Wu (2015)
`lp`	Label Powerset	PS	Tsoumakas and Katakis (2007)
`mbr`	Meta-BR, 2BR or stacking	BR, STA	Tsoumakas et al. (2009)
`mlknn`	Multi-label kNN	AD	Zhang and Zhou (2007)
`ns`	Nested Stacking	BR, CC	Senge et al. (2013)
`ppt`	Pruned Problem Transformation	PS	Read et al. (2008)
`prudent`	PRUned and confiDENT Stacking	BR, STA	Alali and Kubat (2015)
`ps`	Pruned Set	PS	Read (2008)
`rakel`	Random k-labelsets	ENS, PS	Tsoumakas and Vlahavas (2007)
`rdbr`	Recursive Dependent Binary Relevance	BR, ENS, STA	Rauber et al. (2014)
`rpc`	Ranking by Pairwise Comparison	PW	Hüllermeier et al. (2008)

[a] AD = Adaptation; BR = Binary transformation; CC = Chain of classifiers; CLU = Clustering based; ENS = Ensemble; HIE = Hierarchy; PS = Powerset transformation; PW = Pairwise transformation; STA = Stacking

`base.algorithm` value	Description	R function Called
`"C5.0"`	C5.0 Decision Trees	`C50::C5.0`
`"CART"`	Classification and regression trees	`rpart::rpart`
`"KNN"`	K Nearest Neighbor	`kknn::kknn`
`"NB"`	Naive Bayes	`e1071::naiveBayes`
`"RF"`	Random Forest	`randomForest::randomForest`
`"SMO"`	Sequential Minimal Optimization	`RWeka::SMO`
`"SVM"`	Support Vector Machine	`e1071::svm`
`"XGB"`	eXtreme Gradient Boosting	`xgboost::xgboost`
`"MAJORITY"`	Majority class prediction	-
`"RANDOM"`	Random prediction	-

SOFTWARE + EXERCISE

```r
# "utiml" and "mldr" packages for multi-label classification in R
# https://journal.r-project.org/archive/2018/RJ-2018-041/RJ-2018-041.pdf
# https://github.com/rivolli/utiml
# https://cran.r-project.org/web/packages/mldr/vignettes/mldr.pdf
# https://github.com/fcharte/mldr
install.packages("utiml")
library(utiml)
install.packages("mldr")
library(mldr)
# following package offers benchmarks for multi-label classification
install.packages("mldr.datasets")
library(mldr.datasets)
summary(ng20)
# ng20, a corpus with 19300 documents, 1006 words and 20 multi-labels
# Ken Lang, "Newsweeder: Learning to filter netnews", 12th ICML Conference
ng20$labels
# not practical, but part of its corpus can be viewed doing
ng20corpus = ng20$dataset
View(ng20corpus)
# ng20's bag of words
dim(ng20corpus)
colnames(ng20corpus)
# consult the help of the following function
ng20 <- remove_skewness_labels(ng20,10)
# label bat plot
plot(ng20, type="LB")
# visual relations among labels
plot(ng20, type="LC")
# an external GUI interface to explore the "ng20" dataset
mldrGUI() # press "escape" to exit GUI
# start modeling
set.seed(123)
# create a holdout partition: train and predict to evaluate
# I created a small test partition as the prediction step takes a long time
ds <- create_holdout_partition(ng20, c(train=0.90, test=0.10))
```

```r
# Binary relevance ML strategy with naive Bayes base classifier
model_BR_NB <- br(ds$train, "NB")
predictionsBR <- predict(model_BR_NB, ds$test)
head(predictionsBR)
resultsBRPerExamples <- multilabel_evaluate(ds$test, predictionsBR,
                                            c("example-based"))
resultsBRPerExamples
resultsBRPerLabel <- multilabel_evaluate(ds$test, predictionsBR,
                                         c("label-based"))
resultsBRPerLabel

# Classifier Chain ML strategy with naive Bayes base classifier
# Define the chain-order between labels: sample a random order
mychain <- sample(rownames(ng20$labels))
model_CC_NB <- cc(ds$train, "NB", mychain)
predictionsCC <- predict(model_CC_NB, ds$test)
resultsCCPerExamples <- multilabel_evaluate(ds$test, predictionsCC,
                                            c("example-based"))
resultsCCPerExamples
resultsCCPerLabel <- multilabel_evaluate(ds$test, predictionsCC,
                                         c("label-based"))
resultsCCPerLabel
```

```r
# PROPOSED EXERCISE
# Complete a similar work, for the popular Enron-Corpus of e-mails
# To create the corpus and the associated dataFrame
enron = enron()
enron$labels
# A description of its labels appears in the following lists:
# https://bailando.berkeley.edu/enron/enron_categories.txt
# https://data.world/brianray/enron-email-dataset
# where for example "label 2.13" in the lists  is "label B.B13" for enron-labels
```

EXERCISE

- "utiml" + "mldr" + "mldr.datasets" R packages
- Consult its R-vignette [R-Journal][CRAN]
- It is so linked with "mldr" package: loaded both together

- Choose a multilabel dataset (e.g. "Enron")→ understand the problem
- Understand its specific "multilabel preprocessing filters"
- Choose two multilabel strategies
- Choose a supervised base classifier type

- Create a train + test partition
- multilabel_evaluate( ) + predict( ) functions
- Understand its associated parameters

- Compare both multilabel strategies → types of offered metrics?