Tutorial on Multilabel Classification

Eneldo Loza Mencía
Johannes Fürnkranz
Knowledge Engineering Group, TU Darmstadt
{eneldo,juffi}@ke.tu-darmstadt.de

Outline

- Introduction
  - Multilabel Setting
  - Applications & Datasets
- Theoretical Foundations
  - Probabilities in Multilabel
    - joint vs. marginal
  - Losses
    - Ranking
- Programming in MULAN
  - data loading
  - training and evaluation
  - implementation of new approach
- Algorithms
  - Transformation vs. Holistic
  - Transformational Approaches
    - BR, LP, Pairwise
  - Label Dependencies
    - Classifier Chains
  - Holistic Approaches
    - Overview
  - Large Number of Labels
    - Adaptations
    - HOMER
    - Label Space Transformation

Multilabel setting

- assignment of an object x to a **subset** of a set of label $\mathcal{Y}$
- in contrast to
  - (single-label) multiclass classification: mapping to exactly one class
  - two-class/binary classification: mapping to one of only two classes

Typical application areas

- text: tagging/indexing of news, web pages, blogs, … with keywords, topics, genres, authors, languages, writing styles, …
- multimedia: detection of scenes/object (images), instruments, emotions, music styles (audio)
- biology: classification of functions of genomes and protein

Image annotation

{Fall foliage, Field}

{Beach, Urban}

scene dataset consists of 2407 images assigned to 6 labels

Matthew R. BOUTELL, Jiebo LUO, Xipeng SHEN, C. M. Christopher M. BROWN: Learning Multi-Label Scene Classification. In: Pattern Recognition, vol. 37 (9): pp. 1757–1771, 2004.

Movies

Mapping of movies (e.g. plot summaries) to genres (labels)

e.g. available at http://meka.sourceforge.net/#datasets

Formal definition

TECHNISCHE UNIVERSITÄT DARMSTADT

Given input:

- a set of training objects $x_1, \ldots, x_m, x_i$ vectors in R^a
- a set of label mappings $y_1, \ldots, y_m$, each a subset of $Y=\{\lambda_1, \ldots, \lambda_n\}$

i	x_1	x_2	x_3	...	x_a	**y**
1	A	1	0	...	0.1	$\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\{\lambda_2\}$
3	C	3	0	...	0.5	$\{\}$
4	D	4	1	...	0.6	$\{\lambda_1\}$
...						

Objective:

- find a function $h: R^a \rightarrow Y$ which maps x_i to y_i
- as accurately as possible, as efficiently as possible

Formal definition

TECHNISCHE UNIVERSITÄT DARMSTADT

Alternative view: Multitarget Prediction

- a set of training objects $x_1, \ldots, x_m$, x_i vectors in R^a
- a number of n binary Target variables $y_i = \{0,1\}$

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	$\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\{\lambda_2\}$
3	C	3	0	...	0.5	$\{\}$
4	D	4	1	...	0.6	$\{\lambda_1\}$
...						

i	x_1	x_2	x_3	...	x_a	y_1	y_2	...	y_n
1	A	1	0	...	0.1	**1**	**0**	**...**	**1**
2	B	2	1	...	0.3	**0**	**1**	**...**	**0**
3	C	3	0	...	0.5	**0**	**0**	**...**	**0**
4	D	4	1	...	0.6	**1**	**0**	**...**	**0**
...									

Objective:

- find a function $h: R^a \rightarrow Y = \{0,1\}^n$ which maps x_i to a binary vector
- as accurately as possible, as efficiently as possible

Challenges in multilabel learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Dimensionality of input:
  - the number of features
- Quantity of data:
  - the number of examples
- Availability of data
  - real-time processing

not specific to multilabel classification, but common challenges in multilabel learning

- Structure of the output space
  - flat and hierarchical structures
- Dimensionality of output
  - the number of labels
- Dependencies between the Labels
  - correlations, implications, exclusions

specific to multilabel learning (and multitarget prediction), subject of research

News categorization

```xml
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="477551" id="root" date="1997-03-31" xml:lang="en">
<title>SPAIN: Spain's Banesto issue $150 mln in subordinated loaN.</title>
<headline>Spain's Banesto issue $150 mln in subordinated loaN.</headline>
<dateline>MADRID 1997-03-31</dateline>
<text>
<p>Banco Espanol de Credito Banesto said on Monday it issued $150 million in subordinated 10-year 7.5 percent debt. Lead manager is Lehman Brothers.</p>
<p>The statement added that this is the first international issue Banesto has launched since 1993.</p>
<p>Banco Santander has a 50 percent stake in Banesto.</p>
<p>– Madrid Newsroom, + 341 585 8340</p>
</text>
<code code="C17">
```

Funding/Capital

```xml
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
<code code="C172">
```

Bonds/Debt issues

```xml
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
<code code="CCAT">
```

Corporate/Industrial

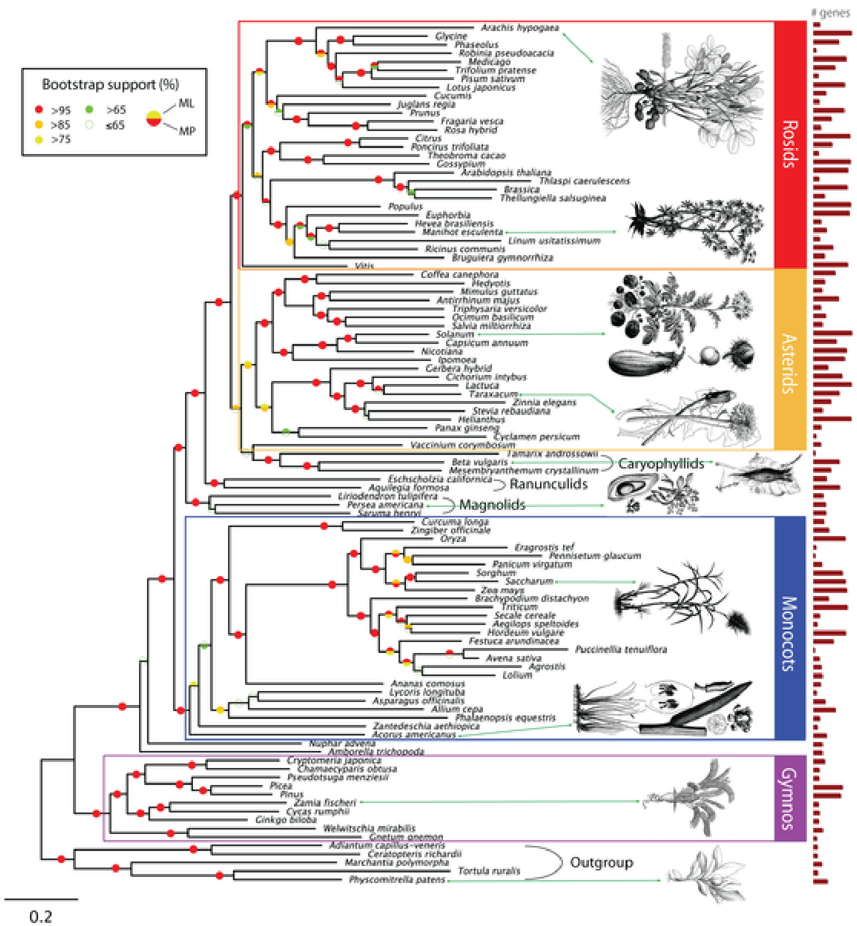
```xml
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
```

The Reuters RCV1 dataset has in total 103 assignable news categories for 804.414 news articles

David Dolan LEWIS, Yiming YANG, Tony G. ROSE, Fan LI: RCV1: A New Benchmark Collection for Text Categorization Research. In: Journal of Machine Learning Research, 2004.

Main challenges: number of **instances & features**, hierarchy

Biology



Mapping of proteins to their functions, e.g. according to FunCAT hierarchy

- *yeast* dataset contains 2417 instances assigned to 14 different labels

Challenges:
input data, hierarchy, dependencies

André ELISSEEFF, Jason WESTON: A Kernel Method for Multi-Labelled Classification. In: Advances in Neural Information Processing Systems, vol. 14, 2001

EUR-Lex repository

- 19328 (freely accessible) documents of the *Directory of Community legislation in force* of the European Union
  - documents available in several European languages
- multiple classifications of the same documents

available at http://www.ke.tu-darmstadt.de/resources/eurlex/

EUR-Lex repository

TECHNISCHE UNIVERSITÄT DARMSTADT

Title and reference

Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs

Classications

EUROVOC descriptor

- **data-processing law**
- **computer piracy**
- **copyright**
- **software**
- **approximation of laws**

Directory Code:

- **Law relating to undertakings/IPR Law**

Subject matter:

- **Internal market**
- **Industrial and commercial property**

Text

COUNCIL DIRECTIVE of 14 May 1991 on the legal protection of computer programs (91/250/EEC)
THE COUNCIL OF THE EU,
Having Regard to the Treaty establishing the European Economic Community and in particular Article 100a thereof,
Having regard to the proposal of the Commission (1), ...

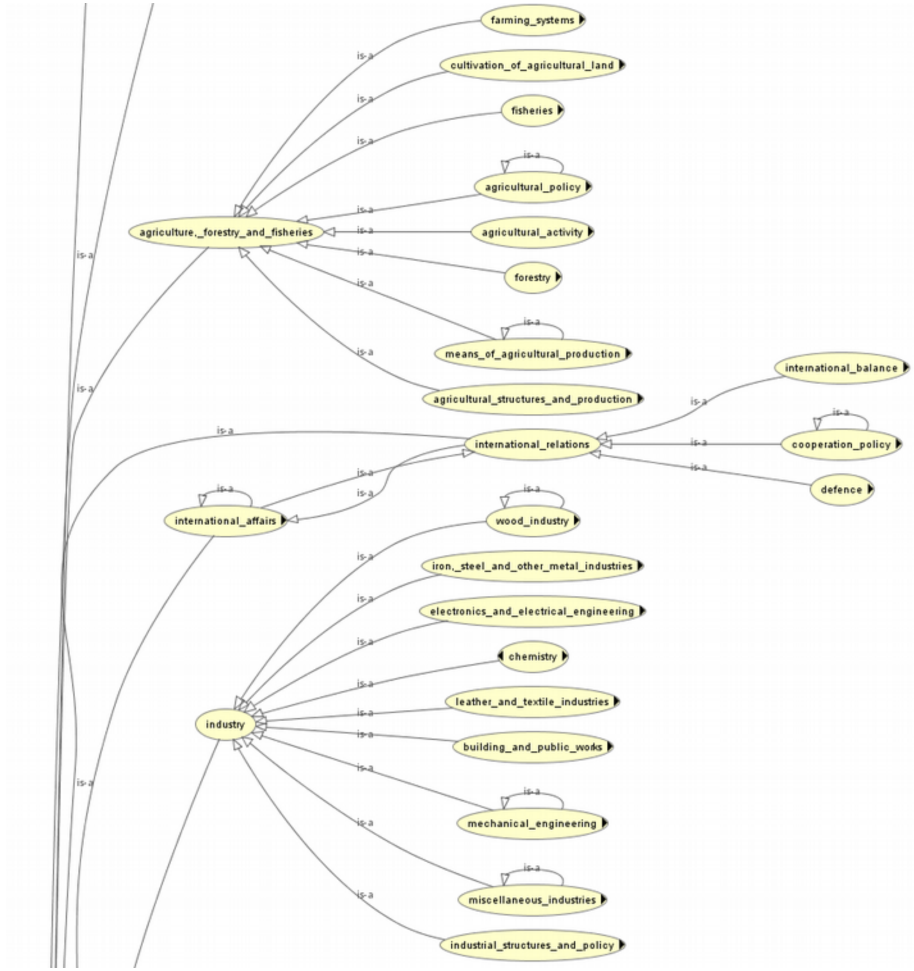
EUR-Lex repository

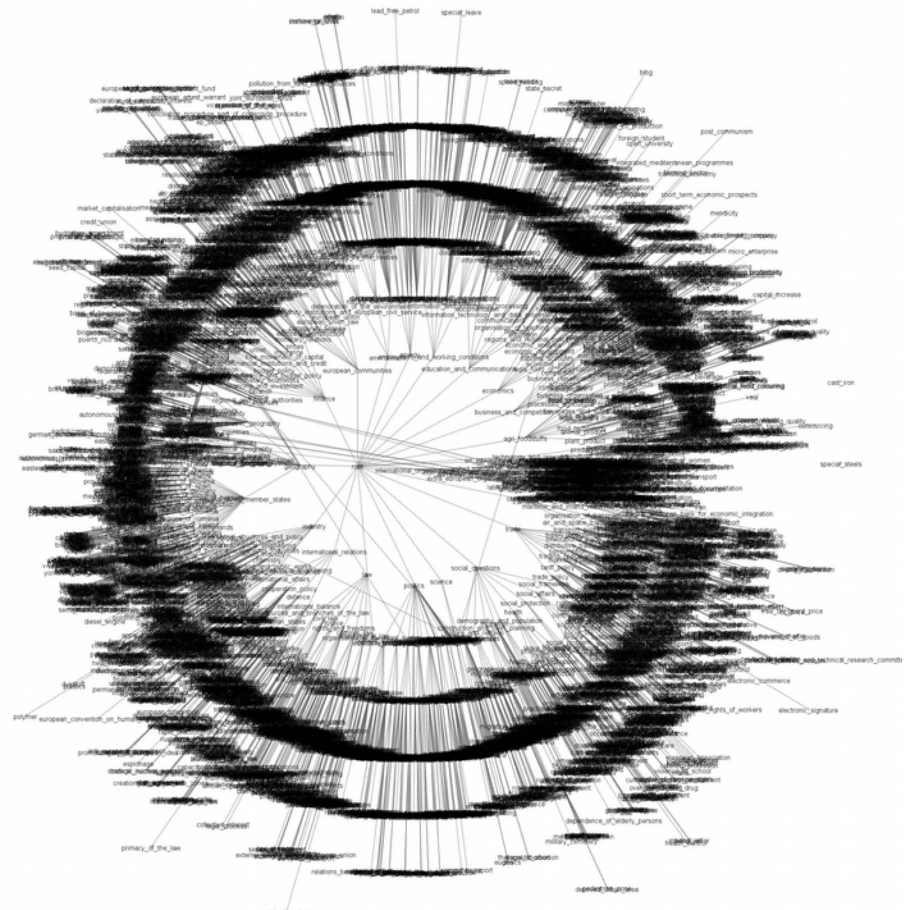
TECHNISCHE UNIVERSITÄT DARMSTADT

- 19328 (freely accessible) documents of the *Directory of Community legislation in force* of the European Union
  - documents available in several European languages
- multiple classifications of the same documents
- most challenging one: **EUROVOC** descriptors associated to each document
  - **3965** descriptors, on average 5.37 labels per document
  - descriptors are organized in a hierarchy with up to 7 levels

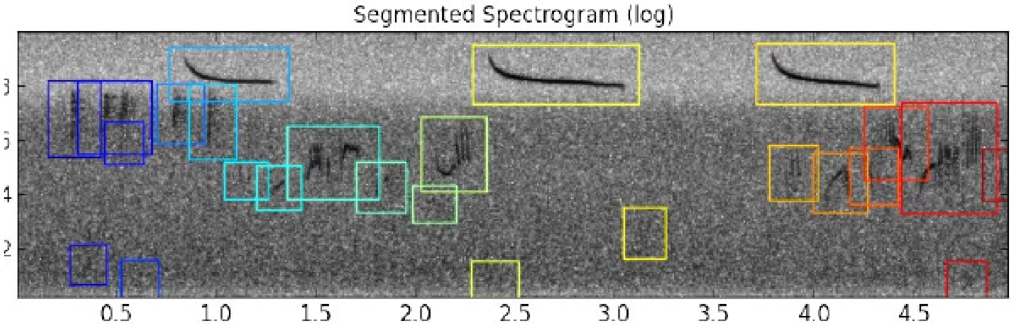
Challenges:
number of labels,
hierarchy

EUR-Lex repository





Audio



NIPS4B competition:687 audio samples recording sounds of 87 different bird species

emotions dataset: 30 secs samples from songs with spectral and rhythmic features extracted, each labeled with induced emotions:

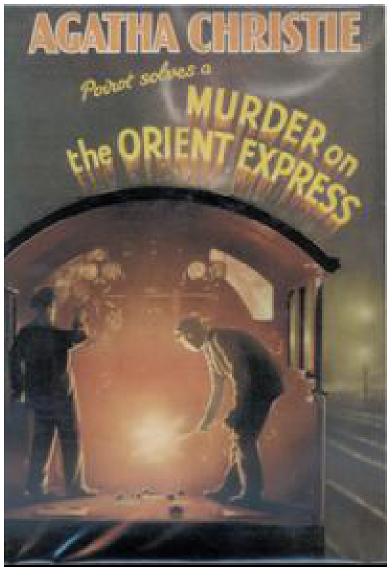
{amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely, angry-aggressive}

Challenges:
input data,
dependencies

http://sabiod.univ-tln.fr/nips4b/challenge1.html
TROHIDIS, TSOUMAKAS, KALLIRIS, Ioannis P. VLAHAVAS: Multilabel Classification of Music into Emotions. In: ISMIR 2008

Book Scenario

TECHNISCHE UNIVERSITÄT DARMSTADT



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Text: It was five o'clock on a winter's morning in Syria. … "Then," said Poirot, "having placed my solution before you, I have the honour to retire from the case."

Author:
Agatha Christie

Genres:
Crime, Mystery, Thriller

Subjects (LOC):
Private Investigators, Orient Express, ...

Keywords:
mystery, fiction, crime, murder, british, poirot, ...

Rate:
4 of 5 stars

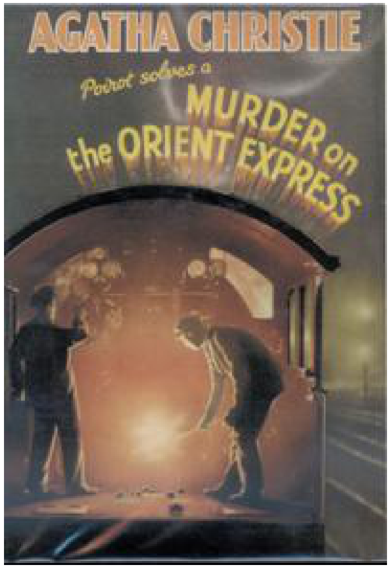
Epoch:
1930ies

Country:
UK

...

Challenges:
dependencies

Book Scenario



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Text: It was five o'clock on a winter's morning in Syria. … "Then," said Poirot, "having placed my solution before you, I have the honour to retire from the case."

Author:
Agatha Christie

Genres:
Crime, Mystery, Thriller

Subjects (LOC):
Private Investigators, Orient Express, ...

Keywords:
mystery, fiction, crime, murder, british, poirot, ...

Rate:
4 of 5 stars

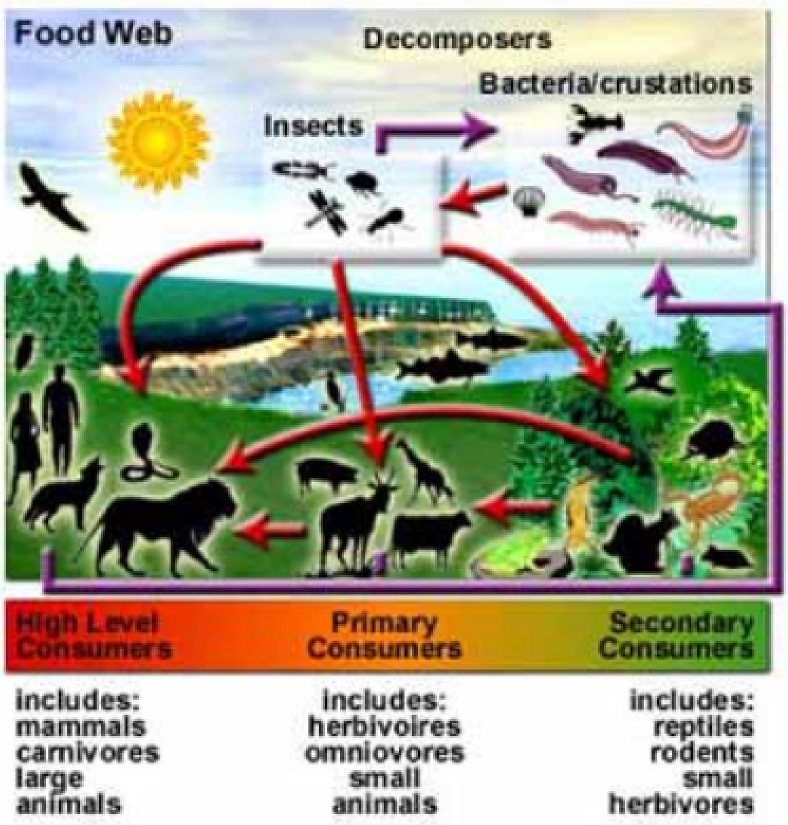
Epoch:
1930ies

Country:
UK

...

Challenges:
dependencies

Dependencies



prediction of presence or absence of species

→ there are obvious dependencies

Available benchmark datasets

dataset name	domain	#instances m	#attributes a	#labels n	labelset size d	density $\frac{d}{n}$	distinct $\lvert\{P_{\mathbf{x}}\}\rvert$
scene	image	2407	294	6	1.074	17.9 %	15
emotions	music	593	72	6	1.869	31.1 %	27
yeast	biology	2417	103	14	4.237	30.3 %	198
tmc2007	text	28596	49060	22	2.158	9.8 %	1341
genbase	biology	662	1186	27	1.252	4.6 %	32
medical	text	978	1449	45	1.245	2.8 %	94
enron	text	1702	1001	53	3.378	6.4 %	753
mediamill	video	43907	120	101	4.376	4.3 %	6555
rcv1	text	804414	231188	101	3.241	3.1 %	13922
r21578	text	11367	21474	120	1.258	1.0 %	533
jmlr2003	image	65362	46	153	3.071	2.0 %	3115
bibtex	text	7395	1836	159	2.402	1.5 %	2856
eccv2002	image	47065	36	374	3.525	0.9 %	3175
hifind	music	32971	98	623	37.304	6.0 %	32734
delicious	text	16105	500	983	19.020	1.9 %	15806
EUR-Lex	text	19348	166448				
subject matter	"	"	"	201	2.213	1.1 %	2504
directory code	"	"	"	410	1.292	0.3 %	1615
EUROVOC	"	"	"	3956	5.317	0.1 %	16467

Sources: http://mulan.sourceforge.net/datasets.html
http://www.csie.ntu.edu.tw/%7Ecjlin/libsvmtools/datasets/multilabel.html
http://meka.sourceforge.net/#datasets
http://www.ke.tu-darmstadt.de/resources/eurlex/

Available benchmark datasets

TECHNISCHE UNIVERSITÄT DARMSTADT

General characteristics

- low label cardinality (in general <= 5)
- hence, low label density (the more labels, the less dense)
- low number of distinct label combinations in relation to potential 2^n
  - the lower the diversity, the more dependencies between labels
- number of possible labels < 1000
  - exception: EUROVOC
  - in real applications more labels are, in principle, available

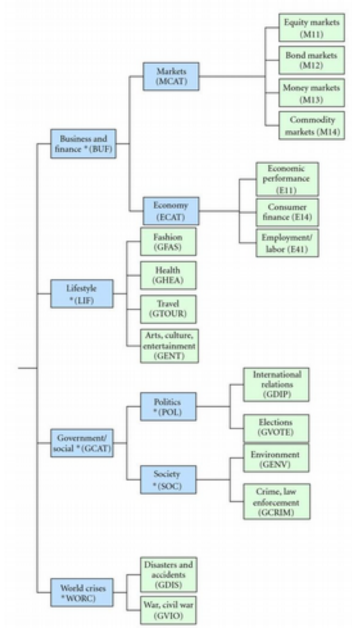
- oldest dataset is from 1991 (Reuters 21578)
- recent development: datasets with large number of labels (e.g. extracted from keyword tagging / Web 2.0)

Philip J. HAYES, Steven P. WEINSTEIN: CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. In: Proceedings of the 2nd Conference on Innovative Applications of Artificial Intelligence (IAAI-90)

Related tasks

Hierarchical Multilabel Classification

- usually solved via "flattening" problem
  - structure is considered via label dependencies
- but: often different losses used



Label Ranking

- learn from and predict rankings on labels
- Multilabel Ranking:
  - get labelset for each example (=bipartite ranking!),
  - predict a label ranking (see later)

$$\{\lambda_1\} \succ_{\mathcal{T}} \{\lambda_2\} \succ_{\mathcal{T}} \{\lambda_3\} \succ_{\mathcal{T}} \{\lambda_4\}$$

(a) total label ranking

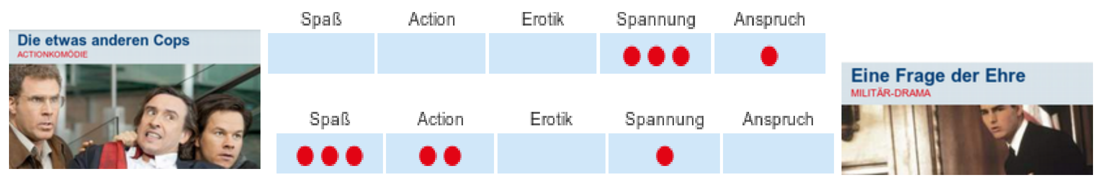
$$\{\lambda_1\} \succ_{\mathcal{T}} \{\lambda_2, \lambda_3, \lambda_4\}$$

(b) bipartite

Related tasks

Graded multilabel classification

- labels can have (ordered) degrees



Collaborative Filtering

- only some output variables are missing, usually no input data

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C	?	X	X	?	?	?
Customer D		X				X
Customer E	X				X	

Multivariate regression

- likewise several outputs, but real valued instead of binary

X_1	X_2	X_3	X_4
0.34	0	10	174
1.45	0	32	277
1.22	1	46	421
0.74	1	25	165
0.95	1	72	273
1.04	0	33	158
0.92	1	81	382

Y_1	Y_2	Y_3	Y_4
14	0.3	10	10
15	1.4	30	50
23	0.7	20	17
19	1.2	40	60
12	0.6	60	48
17	0.9	61	29
16	1.1	71	54

Multi-task learning

- general concept of learning multiple tasks in parallel

Multi-target prediction

Outline

- Introduction
  - Multilabel Setting
  - Applications & Datasets
- Theoretical Foundations
  - Probabilities in Multilabel
    - joint vs. marginal
  - Losses
    - Ranking
- Programming in MULAN
  - data loading
  - training and evaluation
  - implementation of new approach
- Algorithms
  - Transformation vs. Holistic
  - Transformational Approaches
    - BR, LP, Pairwise
  - Label Dependencies
    - Classifier Chains
  - Holistic Approaches
    - Overview
  - Large Number of Labels
    - Adaptations
    - HOMER
    - Label Space Transformation

Probabilistic Model

TECHNISCHE UNIVERSITÄT DARMSTADT

Joint probability distribution

- joint probability of event $\mathbf{y}$: $P(\mathbf{y}|\mathbf{x})$
- $\mathbf{y}$ is the joint event of seeing the label combination $y_1, y_2, y_3, \ldots y_n$ together
- Can it be reduced to modeling probability $P(y_i \mid \mathbf{x})$ of individual labels?

Marginal probability distribution

- marginal probability of event $y_i = b \in \{0,1\}$:

$$P(y_i = b|\mathbf{x}) = \sum_{y \in \mathcal{Y}, y_i = b} P(\mathbf{y}|\mathbf{x})$$

- note that it does not hold $\sum_i P(y_i = 1) = 1$ but $\sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{y}) = 1$

Probabilistic Model

Distinction between joint and marginal probability is very important in multilabel classification, since predicting according to one or the other may give quite different results:

y_1	y_2	y_3	$P(\mathbf{y} \mid \mathbf{x})$
0	0	0	0
0	0	1	0
0	1	0	0.4
0	1	1	0.3
1	0	0	0
1	0	1	0.3
1	1	0	0
1	1	1	0

Probabilistic Model

Distinction between joint and marginal probability is very important in multilabel classification, since predicting according to one or the other may give quite different results:

- mode of joint distribution = (0,1,0)

y_1	y_2	y_3	$P(\mathbf{y} \mid \mathbf{x})$
0	0	0	0
0	0	1	0
0	**1**	**0**	**0.4**
0	1	1	0.3
1	0	0	0
1	0	1	0.3
1	1	0	0
1	1	1	0

Probabilistic Model

Distinction between joint and marginal probability is very important in multilabel classification, since predicting according to one or the other may give quite different results:

- mode of joint distribution $= (0,1,0)$
- mode of marginal distribution $= (0,1,1)$
- question to answer:
  - do I want to predict the correct label combination
  - or do I want to predict each label itself correctly

→ different loss functions

y_1	y_2	y_3	$P(\mathbf{y} \mid \mathbf{x})$
0	0	0	0
0	0	1	0
0	**1**	**0**	**0.4**
0	1	1	0.3
1	0	0	0
1	0	1	0.3
1	1	0	0
1	1	1	0
0.7	0.3	0.4	$P(y_i=0 \mid \mathbf{x})$
0.3	**0.7**	**0.6**	$P(y_i=1 \mid \mathbf{x})$

example adapted from: Tutorial on Multi-target prediction at ICML 2013, http://www.ngdata.com/knowledge-base/icml-2013-tutorial-multi-target-prediction/

Subset Accuracy vs. Hamming Loss

TECHNISCHE UNIVERSITÄT DARMSTADT

Subset Accuracy

- ratio of correctly predicted label combinations. Compute

$$\textsc{Acc}(\mathbf{y}, \hat{\mathbf{y}}) = [[\mathbf{y} = \hat{\mathbf{y}}]], \ [[x]] = 1 \text{ if } x \text{ is correct, } 0 \text{ otherwise}$$

  for each test instance and average over the whole test set
- the whole predicted label vector $\hat{\mathbf{y}}$ has to be equal!
- the *risk minimizer* is the **joint mode**

Hamming Loss

- percentage of labels that are misclassified

$$\textsc{HamLoss}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|\mathcal{Y}|}|\mathbf{y} \triangle \hat{\mathbf{y}}|, \quad \triangle \text{ is the symmetric difference}$$

- can also be seen as macro-averaged classification error:

$$\textsc{HamLoss}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{fp + fn}{fp + fn + tp + tn}$$

(tp,tn,fp,fn computed for each text example)

- the *risk minimizer* is the **marginal mode**

Krzysztof DEMBCZYNSKI, Willem WAEGEMAN, Weiwei CHENG, Eyke HÜLLERMEIER: Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'10)

Subset Accuracy vs. Hamming Loss

For non-deterministic data (noise, typically all data available) it is usually not possible to optimize both measures simultaneously

- otherwise probabilities $P(y_i \mid \mathbf{x}), i=1..n, P(\mathbf{y} \mid \mathbf{x})$ would be 1 for the correct $\mathbf{y}$

  → joint and marginal modes would coincide

Subset Accuracy vs. Hamming Loss of different multilabel classifiers on the yeast dataset:

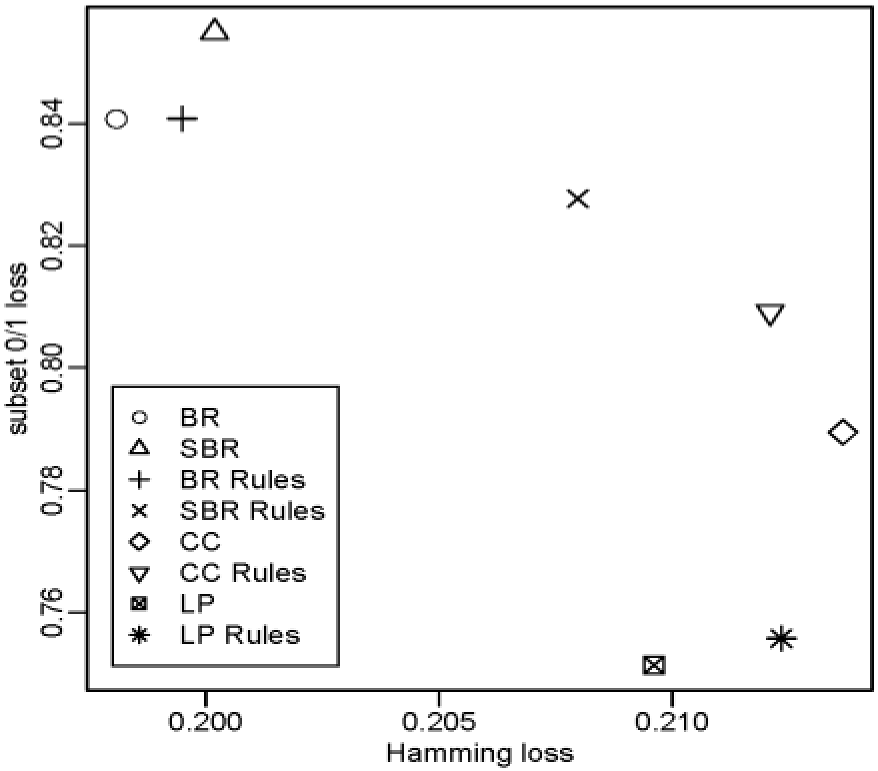


image taken from: Tutorial on Multi-target prediction at ICML 2013, http://www.ngdata.com/knowledge-base/icml-2013-tutorial-multi-target-prediction/

Multilabel Loss Functions

- the risk minimizers for subset accuracy and hamming loss are the same, (i.e. optimizing one measure also optimizes the other), only if
  - labels are (conditionally) independent, or
  - the probability of the joint mode is greater than 0.5

- there is a large variety of metrics in multilabel classification
  - even more when counting hierarchical ML losses
- therefore, in multilabel classification, it is important to know the objective (the loss to optimize) and the appropriate approach for it
  - in general, there is no such as one approach best for all measures
  - although this is often suggested in experimental results (“our approach is best on almost all losses”)

Multilabel Loss Functions

We can discriminate between two groups of loss functions:

- Bipartition Measures
  - measure how good the separation into relevant and irrelevant labels is
  - essentially adaptations of measures for classification error to the label space
- Ranking Measures
  - some algorithms sort the labels before they partition them
  - ranking measures estimate how well the labels are sorted
  - ideally all relevant labels should be sorted before all irrelevant labels

Bipartation Losses

computed is based on a confusion matrix in label space

C	predicted	not predicted
relevant	tp	fn
irrelevant	fp	tn

- Recall
  - fraction of retrieved relevant labels

$$\textsc{Rec}(C) := \frac{tp}{tp + fn}$$

- Precision
  - fraction of retrieved labels that are relevant

$$\textsc{Prec}(C) := \frac{tp}{tp + fp}$$

- F1
  - harmonic average of recall and precision

$$\text{F1}(C) := \frac{2}{\frac{1}{\textsc{Rec}(C)} + \frac{1}{\textsc{Prec}(C)}}$$

- Error
  - fraction of incorrectly classified labels

$$\textsc{HamLoss}(C) = \frac{fp + fn}{fp + fn + tp + tn}$$

Bipartition Losses - Averaging

The confusion matrix can be computed in different ways:

- Micro-averaging: (most common) i : labels, j : test instances
  - compute confusion matrix for each example and each label
  - add them up
  - compute the measures from the result

$$\delta\left(\sum_{j=1}^{m}\sum_{i=1}^{n} C_i^j\right)$$

- Example-based:
  - sum up for each label
  - compute measure for each example
  - and average them

$$\frac{1}{m}\sum_{j=1}^{m}\delta\left(\sum_{i=1}^{n} C_i^j\right)$$

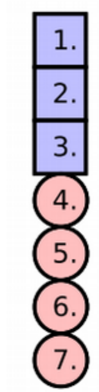
- Macro-averaging:
  - sum up for each example
  - compute measure for each label and then average
  - gives all labels, regardless of size, equal weight

$$\frac{1}{n}\sum_{i=1}^{n}\delta\left(\sum_{j=1}^{m} C_i^j\right)$$

Ranking Losses

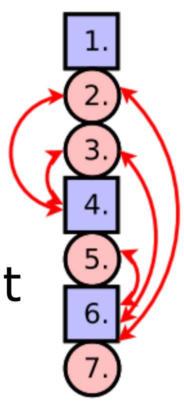
IsError-Loss:

- 0 if all positive labels are on top, otherwise 1
- 1-IsError upper bounds subset accuracy



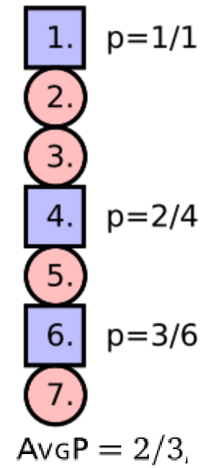
Ranking-Loss

- fraction of pairs of positive and negative label which are incorrectly ordered
- corresponds to Kendall's tau coefficient or 1-AUC



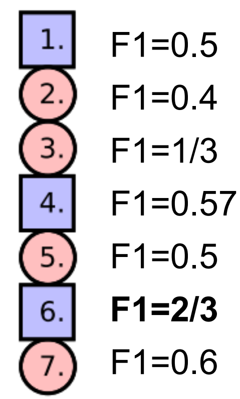
Average Precision

- the average of the precision values at positions of positive labels
- rough interpretation: positive label density at the top of the ranking
- focuses on good results on higher ranks (ranking loss treats all ranks the same)



MaxF1

- the maximum F1-score at all positions in the ranking
- upper bounds F1



Outline

- Introduction
  - Multilabel Setting
  - Applications & Datasets
- Theoretical Foundations
  - Probabilities in Multilabel
    - joint vs. marginal
  - Losses
    - Ranking
- Programming in MULAN
  - data loading
  - training and evaluation
  - implementation of new approach
- Algorithms
  - Transformation vs. Holistic
  - Transformational Approaches
    - BR, LP, Pairwise
  - Label Dependencies
    - Classifier Chains
  - Holistic Approaches
    - Overview
  - Large Number of Labels
    - Adaptations
    - HOMER
    - Label Space Transformation

Approaches for learning multilabel data

TECHNISCHE UNIVERSITÄT DARMSTADT

Main solutions in order to solve multilabel problems:

Holistic approaches

- solve problem globally and jointly, e.g. solving one single optimization problem
- also called *single-machine* (Rifkin), *all-at-once* (Rueda) or *algorithm adaptation* approaches (Tsoumakas)
- not trivial and often not possible

Transformation of multilabel problems into single-label problems

- well known problem setting, clear semantics
- many state-of-the-art binary learners usable: SVMs, rule learners, decision trees
- usually out-of-the-box usage: no additional parameter settings

RIFKIN, KLAUTAU: In Defense of One-Vs-All Classification. In: JMLR, vol. 5, 2004.
RUEDA, OOMMEN, HENRÍQUEZ: Multi-class pairwise linear dimensionality reduction using heteroscedastic schemes. In: Pattern Recognition, vol. 43 (7), 2010.
G. Tsoumakas, I. Katakis, "Multi-Label Classification: An Overview", International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.

Transformational approaches

Three main competing transformational approaches:

- **binary relevance** decomposition: learn one classifier for ***each label***
  - aka one-against-all
    → solve a linear number of binary problems

- **pairwise decomposition:** learn one classifier for ***each pair of labels***
  - aka one-against-one, round robin, all-pairs
    → solve a quadratic number of binary problems

- **label powerset** transformation: learn one classifier for ***each label combination***
    → solve one single-label multiclass problem

Binary Relevance Decomposition

- learn one classifier per label
  - positive examples are the ones for which the label is positive
  - negatives are all the remaining ones

i	x_1	x_2	x_3	...	x_a	y_1	y_2	...	y_n
1	A	1	0	...	0.1	**1**	**0**	**...**	**1**
2	B	2	1	...	0.3	**0**	**1**	**...**	**0**
3	C	3	0	...	0.5	**0**	**0**	**...**	**0**
4	D	4	1	...	0.6	**1**	**0**	**...**	**0**
...									

for label λ_1

i	x_1	x_2	x_3	...	x_a	y_1
1	A	1	0	...	0.1	**1**
2	B	2	1	...	0.3	**0**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**1**

for label λ_2

i	x_1	x_2	x_3	...	x_a	y_2
1	A	1	0	...	0.1	**0**
2	B	2	1	...	0.3	**1**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**0**

...

for label λ_n

i	x_1	x_2	x_3	...	x_a	y_n
1	A	1	0	...	0.1	**1**
2	B	2	1	...	0.3	**0**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**0**

Binary Relevance Decomposition

predict the union of the base classifiers' predictions

- can also produce rankings if classifiers output scores

i	x_1	x_2	x_3	...	x_a	y_1	y_2	...	y_n
1	A	1	0	...	0.1	**1**	**0**	**...**	**1**
2	B	2	1	...	0.3	**0**	**1**	**...**	**0**
3	C	3	0	...	0.5	**0**	**0**	**...**	**0**
4	D	4	1	...	0.6	**1**	**0**	**...**	**0**
...									

for label λ_1

for label λ_2

for label λ_n

i	x_1	x_2	x_3	...	x_a	y_1
1	A	1	0	...	0.1	**1**
2	B	2	1	...	0.3	**0**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**1**

i	x_1	x_2	x_3	...	x_a	y_2
1	A	1	0	...	0.1	**0**
2	B	2	1	...	0.3	**1**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**0**

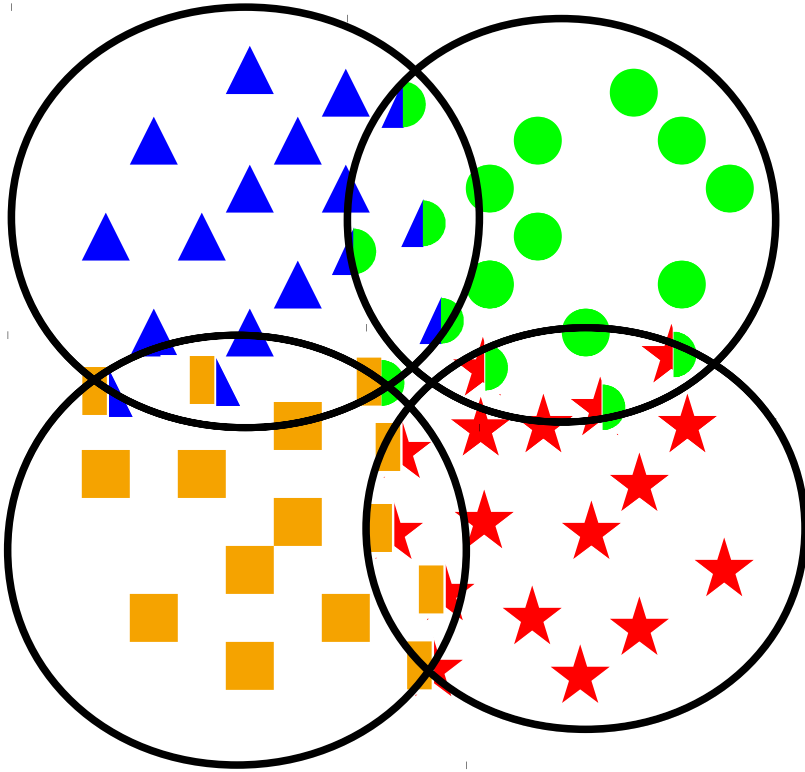
...

i	x_1	x_2	x_3	...	x_a	y_n
1	A	1	0	...	0.1	**1**
2	B	2	1	...	0.3	**0**
3	C	3	0	...	0.5	**0**
4	D	4	1	...	0.6	**0**

Binary Relevance Decomposition

Simple and straight-forward approach

- corresponds to concept learning
  - learn each label as separate concept learning problem
- most popular approach, often used as baseline



Complexity

- training: n subproblems with each m training examples
- testing: evaluation of n classifiers

  → efficient and scalable

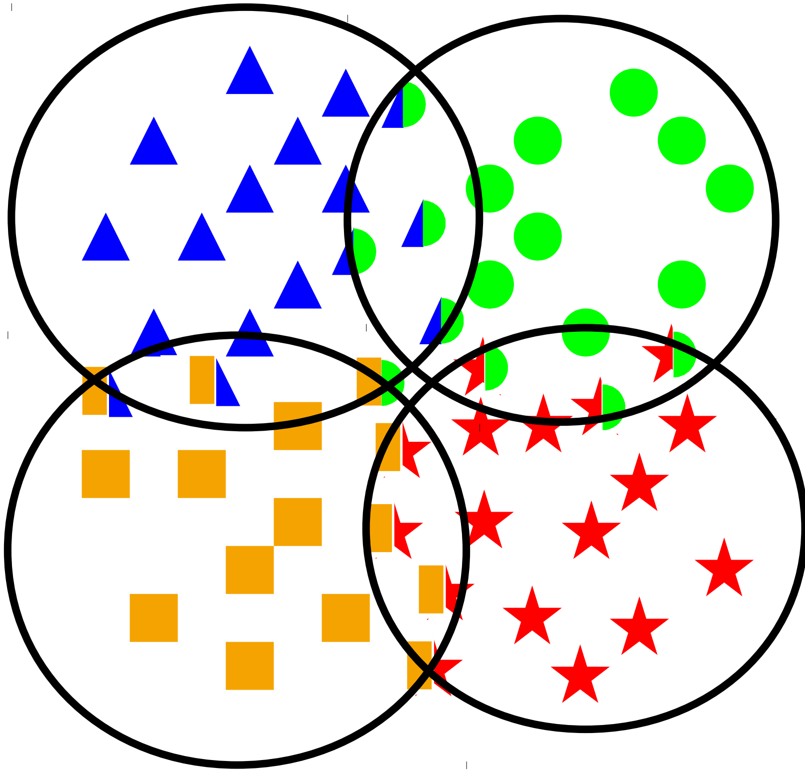
First employment of BR decomposition known: JOACHIMS: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: ECML-98
First appearance of term BR: Klaus BRINKER, Johannes FÜRNKRANZ, Eyke HÜLLERMEIER: A Unified Model for Multilabel Classification and Ranking. In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI-06),

Binary Relevance Decomposition

TECHNISCHE UNIVERSITÄT DARMSTADT

Limitations

- not considering label dependencies
  - each target label is learned separately
- but consistent with Hamming Loss
  - training each base classifier corresponds to learning marginal class probabilities $P(y_i | x)$
  - moreover: ranking labels with respect to probability estimates $P(y_i | x)$ is sufficient to minimize the Ranking Loss[1]
    - but good estimations are difficult to get!



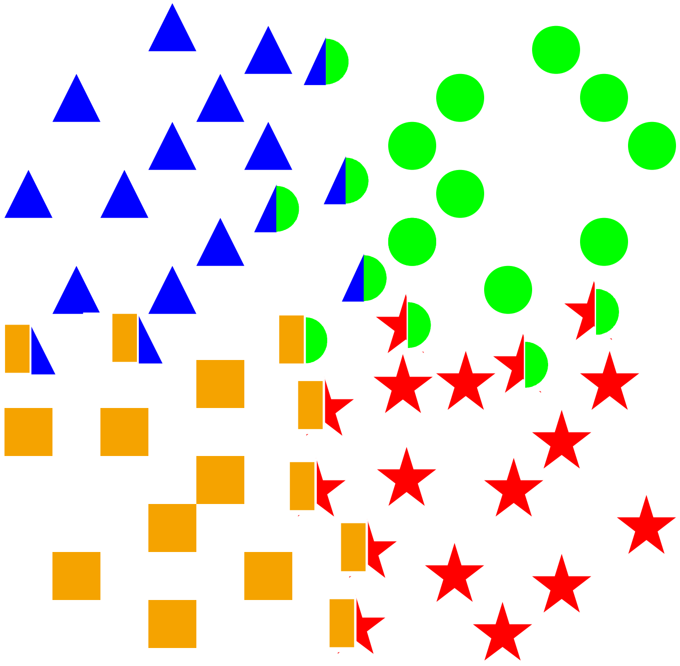
[1] W. Kotlowski, K. Dembczynski, and E. Hüllermeier: Bipartite Ranking through Minimization of Univariate Loss. In: ICML-11
K. Dembczynski, W. Kotlowski, and E. Hüllermeier: Consistent multilabel ranking through univariate losses. In ICML, 2012

Pairwise Decomposition
Pairwise Multilabel Ranking

TECHNISCHE UNIVERSITÄT DARMSTADT

Pairwise decomposition learns a binary classifier for each pair of labels { λ_p, λ_q }

- base classifiers learn to discriminate between two labels



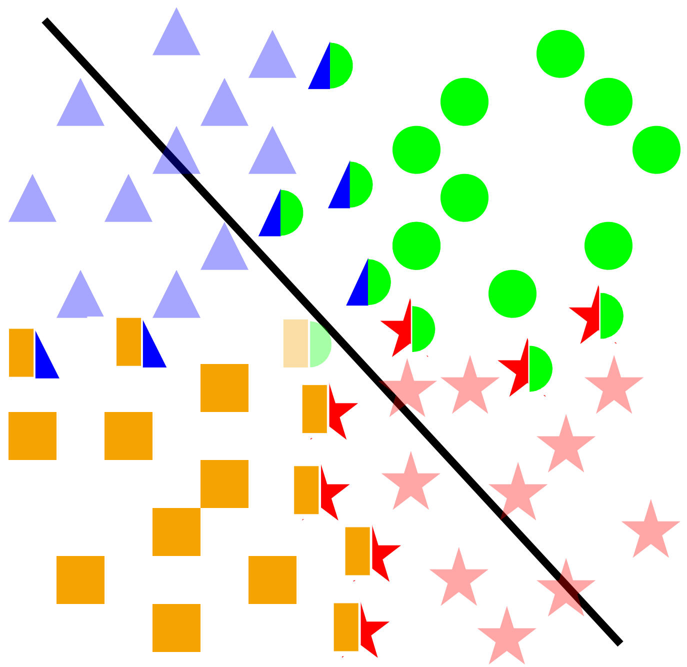
Johannes FÜRNKRANZ: Round Robin Classification. In: JMLR 2002.
Johannes FÜRNKRANZ, Eyke HÜLLERMEIER: Pairwise Preference Learning and Ranking. In: Proceedings of the 14th European Conference on Machine Learning (ECML-03)
Eneldo LOZA MENCÍA, Johannes FÜRNKRANZ: Pairwise Learning of Multilabel Classifications with Perceptrons. In: IEEE IJCNN-08

Pairwise Decomposition
Pairwise Multilabel Ranking

Pairwise decomposition learns a binary classifier for each pair of labels { λ_p, λ_q }

- base classifiers learn to discriminate between two labels
- during prediction, each base classifier gives a vote for one of the two labels

→ label relevance ranking according to obtained votes for each label



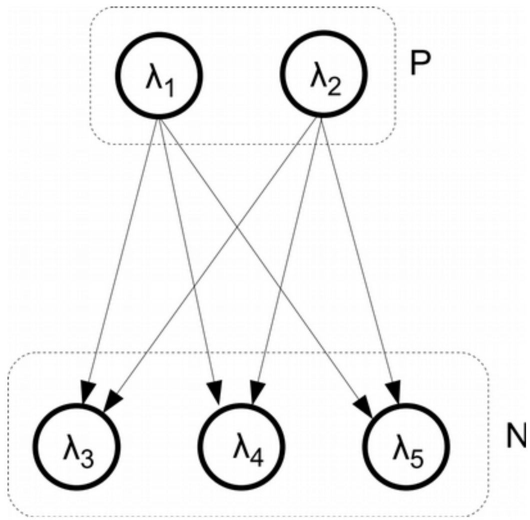
Relation to Preference Learning:

- each base learner learns and predicts whether $\lambda_p > \lambda_q$ or $\lambda_p < \lambda_q$

Johannes FÜRNKRANZ: Round Robin Classification. In: JMLR 2002.
Johannes FÜRNKRANZ, Eyke HÜLLERMEIER: Pairwise Preference Learning and Ranking. In: Proceedings of the 14th European Conference on Machine Learning (ECML-03)
Eneldo LOZA MENCÍA, Johannes FÜRNKRANZ: Pairwise Learning of Multilabel Classifications with Perceptrons. In: IEEE IJCNN-08

Pairwise Decomposition
Pairwise Multilabel Ranking

Training:



relevant labels

|P|·|N| preferences

irrelevant labels

Prediction:

$h_{1,2}=1$	$h_{2,1}=0$	$h_{3,1}=0$	$h_{4,1}=0$	$h_{5,1}=0$
$h_{1,3}=1$	$h_{2,3}=1$	$h_{3,2}=0$	$h_{4,2}=0$	$h_{5,2}=0$
$h_{1,4}=1$	$h_{2,4}=1$	$h_{3,4}=1$	$h_{4,3}=0$	$h_{5,3}=0$
$h_{1,5}=1$	$h_{2,5}=1$	$h_{3,5}=1$	$h_{4,5}=1$	$h_{5,4}=0$
$v_1=4$	$v_2=3$	$v_3=2$	$v_4=1$	$v_5=0$

→ Ranking: $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$

during prediction many "incompetent" classifiers vote, but there are guarantees that irrelevant labels cannot obtain more votes than relevant ones
(given good base predictions)

Learning by pairwise comparison

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	$\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\{\lambda_2\}$
3	C	3	0	...	0.5	$\{\}$
4	D	4	1	...	0.6	$\{\lambda_1\}$
...						

in each subproblem, only instances are used with either

$$\lambda_p > \lambda_q \text{ or } \lambda_p < \lambda_q$$

(λ_p=1, λ_q=0 or λ_p=0, λ_q=1)

the remaining ones are ignored

λ_1 vs. λ_2

i	x_1	x_2	x_3	...	x_a	
1	A	1	0	...	0.1	**1**
2	B	2	1	...	0.3	**0**
~~3~~	~~C~~	~~3~~	~~0~~	~~...~~	~~0.5~~	~~?~~
4	D	4	1	...	0.6	**1**
...						

λ_1 vs. λ_3

i	x_1	x_2	x_3	...	x_a	
1	A	1	0	...	0.1	**1**
~~2~~	~~B~~	~~2~~	~~1~~	~~...~~	~~0.3~~	~~?~~
~~3~~	~~C~~	~~3~~	~~0~~	~~...~~	~~0.5~~	~~?~~
4	D	4	1	...	0.6	**1**
...						

...

λ_{n-1} vs. λ_n

i	x_1	x_2	x_3	...	x_a	
1	A	1	0	...	0.1	**0**
~~2~~	~~B~~	~~2~~	~~1~~	~~...~~	~~0.3~~	~~?~~
~~3~~	~~C~~	~~3~~	~~0~~	~~...~~	~~0.5~~	~~?~~
~~4~~	~~D~~	~~4~~	~~1~~	~~...~~	~~0.6~~	~~?~~
...						

Pairwise Decomposition
Pairwise Multilabel Ranking

Advantages

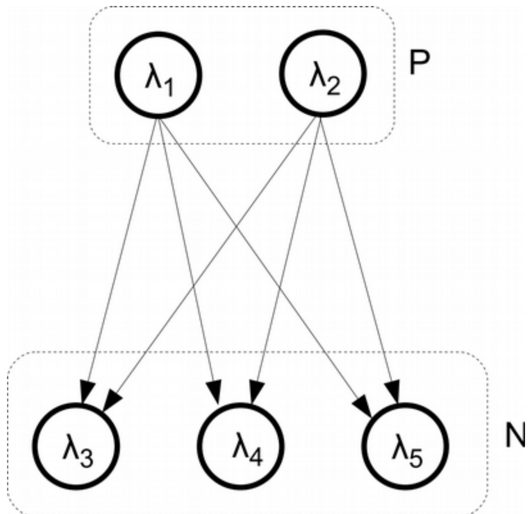
- much smaller sub-problems
  → easier to learn, faster to train
- consideration of pairwise label relationships
  - but loss of information in the label intersections
- high degree of parallelization

Disadvantages

- only ranking, but we may want labelsets
- quadratic number of sub-problems
  - high memory costs
  - high prediction costs

Pairwise Decomposition Calibrated Label Ranking[1]

Training:



relevant labels

irrelevant labels

Idea: introduce a virtual label which indicates the boundary between relevant and irrelevant labels

Prediction:

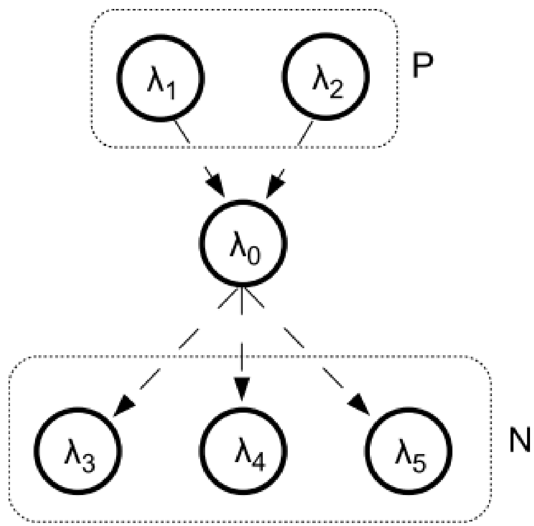
$h_{1,2} = 1$	$h_{2,1} = 0$	$h_{3,1} = 0$	$h_{4,1} = 0$	$h_{5,1} = 0$
$h_{1,3} = 1$	$h_{2,3} = 1$	$h_{3,2} = 0$	$h_{4,2} = 0$	$h_{5,2} = 0$
$h_{1,4} = 1$	$h_{2,4} = 1$	$h_{3,4} = 1$	$h_{4,3} = 0$	$h_{5,3} = 0$
$h_{1,5} = 1$	$h_{2,5} = 1$	$h_{3,5} = 1$	$h_{4,5} = 1$	$h_{5,4} = 0$
$v_1 = 4$	$v_2 = 3$	$v_3 = 2$	$v_4 = 1$	$v_5 = 0$

→ Ranking: $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$

[1] J. Fürnkranz, E. Hüllermeier, E. Loza, K. Brinker: *Multilabel Classification via Calibrated Label Ranking*. Machine Learning, vol. 73 (2): pp. 133–153

Pairwise Decomposition
Calibrated Label Ranking

Training:



relevant labels

virtual label

irrelevant labels

Idea:
introduce a virtual label which indicates the boundary between relevant and irrelevant labels

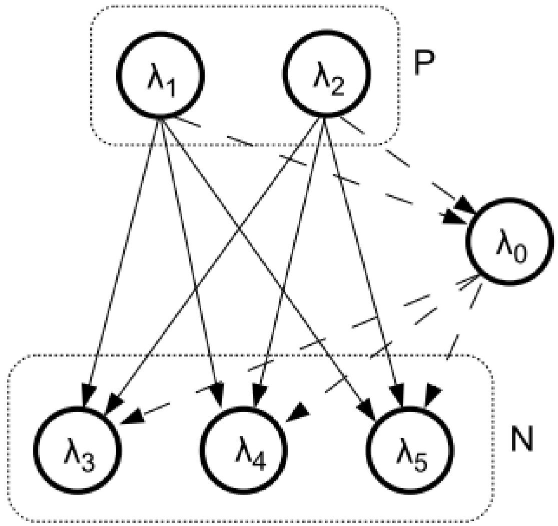
Prediction:

$h_{1,2}=1$	$h_{2,1}=0$	$h_{3,1}=0$	$h_{4,1}=0$	$h_{5,1}=0$
$h_{1,3}=1$	$h_{2,3}=1$	$h_{3,2}=0$	$h_{4,2}=0$	$h_{5,2}=0$
$h_{1,4}=1$	$h_{2,4}=1$	$h_{3,4}=1$	$h_{4,3}=0$	$h_{5,3}=0$
$h_{1,5}=1$	$h_{2,5}=1$	$h_{3,5}=1$	$h_{4,5}=1$	$h_{5,4}=0$
$v_1=4$	$v_2=3$	$v_3=2$	$v_4=1$	$v_5=0$

→ Ranking: $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$

Pairwise Decomposition
Calibrated Label Ranking

Training:



relevant labels

virtual label

irrelevant labels

Idea:
introduce a virtual label which indicates the boundary between relevant and irrelevant labels

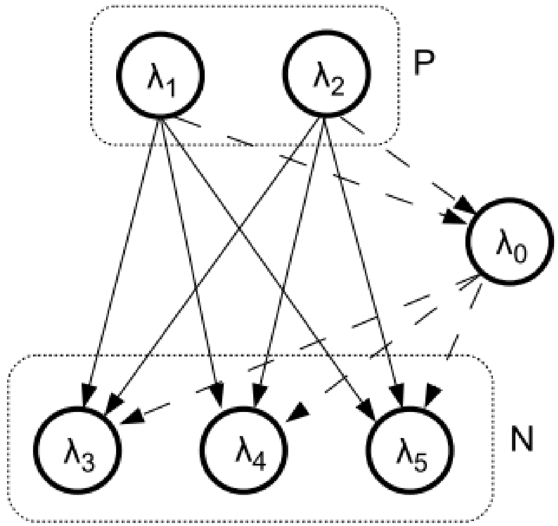
Prediction:

$h_{1,2}=1$	$h_{2,1}=0$	$h_{3,1}=0$	$h_{4,1}=0$	$h_{5,1}=0$
$h_{1,3}=1$	$h_{2,3}=1$	$h_{3,2}=0$	$h_{4,2}=0$	$h_{5,2}=0$
$h_{1,4}=1$	$h_{2,4}=1$	$h_{3,4}=1$	$h_{4,3}=0$	$h_{5,3}=0$
$h_{1,5}=1$	$h_{2,5}=1$	$h_{3,5}=1$	$h_{4,5}=1$	$h_{5,4}=0$
$v_1=4$	$v_2=3$	$v_3=2$	$v_4=1$	$v_5=0$

→ Ranking: $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$

Pairwise Decomposition Calibrated Label Ranking

Training:



relevant labels

virtual label

irrelevant labels

Idea:
introduce a virtual label which indicates the boundary between relevant and irrelevant labels

Prediction:

$h_{0,1}=0$	$h_{1,0}=1$	$h_{2,0}=1$	$h_{3,0}=0$	$h_{4,0}=0$	$h_{5,0}=0$
$h_{0,2}=0$	$h_{1,2}=1$	$h_{2,1}=0$	$h_{3,1}=0$	$h_{4,1}=0$	$h_{5,1}=0$
$h_{0,3}=1$	$h_{1,3}=1$	$h_{2,3}=1$	$h_{3,2}=0$	$h_{4,2}=0$	$h_{5,2}=0$
$h_{0,4}=1$	$h_{1,4}=1$	$h_{2,4}=1$	$h_{3,4}=1$	$h_{4,3}=0$	$h_{5,3}=0$
$h_{0,5}=1$	$h_{1,5}=1$	$h_{2,5}=1$	$h_{3,5}=1$	$h_{4,5}=1$	$h_{5,4}=0$
$v_0=3$	$v_1=5$	$v_2=4$	$v_3=2$	$v_4=1$	$v_5=0$

→ Ranking: $\lambda_1 > \lambda_2 > \lambda_0 > \lambda_3 > \lambda_4 > \lambda_5$

Pairwise Decomposition Complexity

Training:

- only [avg. labelset size] times more training examples needed than BR
  - usually <5
- due to smaller subproblems: can be even faster than BR for base learners which need more than linear O(m) time in the number of training examples
- but: calibration basically learns an additional BR ensemble

Prediction:

- quadratic number of base predictions (n(n-1)/2 votes)
- but: *Quick Voting* reduces costs to log-linear evaluations[1]

Memory:

- quadratic number of base classifiers
- but: reformulation allows applying it on up to 4000 labels[2]
  - despite 8 million base classifiers (see later)

[1] E. Loza, S.-H. Park, J. Fürnkranz: Efficient Voting Prediction for Pairwise Multilabel Classification. In: Neurocomputing, vol. 73 (7-9): pp. 1164 –1176, 2010.
[2] E. Loza, S.-H. Park, J. Fürnkranz: Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain. In: ECML 2008

Pairwise Decomposition Predictive Quality

- pairwise approach (presumably) consistent with Ranking Loss
  - but advantage over BR makes it consistently better than BR also on the other measures

rcv1

	BR	CMLPP
IsErr	35.87	**27.36**
ErrSetSize	7.614	**1.904**
RankLoss	2.529	**0.472**
Margin	5.833	**1.438**
OneErr	4.022	**2.902**
AvgP	90.00	**93.81**
$F1_{\lvert P \rvert}$	81.40	**87.99**
$Prec_d$	78.86	**82.74**
Rec_d	73.24	**76.85**
$F1_d$	75.95	**79.68**
$1-$HamLoss	98.74	**98.97**
Prec	80.15	**86.77**
Rec	**79.70**	79.33
F1	79.93	**82.88**

r21578

	BR	CMLPP
RankLoss	2.977	**0.239**
AvgP	91.59	**95.89**
Prec	78.38	**87.98**
Rec	**85.59**	83.79

scene

	BR	CMLPP
RankLoss	8.165	**7.285**
AvgP	85.64	**86.79**
Prec	71.80	**71.83**
Rec	71.21	**74.20**

yeast

	BR	CMLPP
RankLoss	22.73	17.54
AvgP	70.41	74.98
Prec	60.47	**62.37**
Rec	59.07	**63.31**

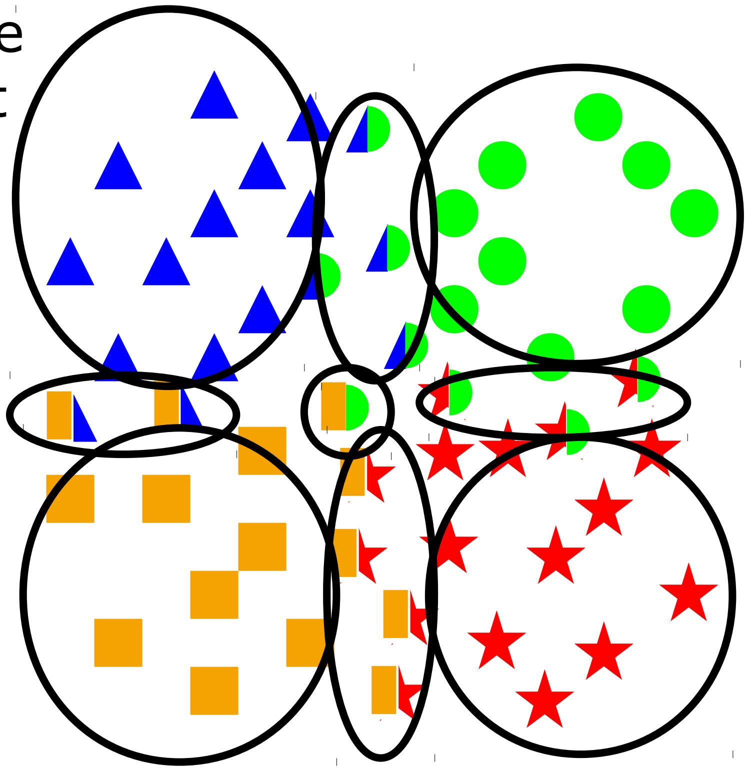
		HamLoss		Prec		Rec		F1	
dataset	n	BR	CMLPP	BR	CMLPP	BR	CMLPP	BR	CMLPP
scene	6	10.42	**10.00**	71.80	**71.83**	71.21	**74.20**	71.19	**72.76**
emotions	6	35.64	**34.08**	46.78	**48.62**	60.15	**61.90**	52.63	**54.47**
yeast	14	24.09	**22.67**	60.47	**62.37**	59.07	**63.31**	59.76	**62.83**
tmc2007	22	7.37	**6.78**	62.57	**64.16**	66.47	**73.61**	64.46	**68.56**
genbase	27	**0.26**	0.48	99.22	**99.59**	**95.49**	90.60	**97.32**	94.88
medical	45	**1.51**	1.51	71.72	**76.02**	**75.84**	66.75	**73.72**	71.08
enron	53	7.56	**6.01**	41.56	**52.82**	47.05	**49.51**	44.13	**51.11**
mediamill	101	4.52	**4.16**	42.28	**56.66**	10.05	**19.70**	16.24	**29.23**
rcv1	103	1.26	**1.03**	80.15	**84.89**	79.70	**81.61**	79.93	**83.22**
r21578	120	0.78	**0.55**	59.98	**72.89**	**78.36**	76.68	67.92	**74.63**
bibtex	159	1.57	**1.35**	46.53	**57.97**	**36.30**	34.84	40.78	**43.53**
eurlex_sm	201	0.76	**0.54**	63.39	**77.88**	**74.11**	71.57	68.32	**74.59**
eurlex_dc	410	0.26	**0.17**	56.26	**79.21**	**70.54**	61.98	62.58	**69.54**
delicious	983	5.58	**3.48**	11.88	**19.77**	**29.59**	26.51	16.95	**22.65**

taken from: Eneldo Loza Mencía: "Efficient Pairwise Multilabel Classification", 2012, http://www.ke.tu-darmstadt.de/bibtex/publications/show/2337

Label Powerset Transformation

Straight-forward approach: create one meta-class for each occurring labelset

- train a multiclass learner, i.e. learn each label***set*** independently
  - e.g. using Decision Tree learner, but also one-against-all or pairwise



first appearance: Matthew R. BOUTELL, Jiebo LUO, Xipeng SHEN, C. M. Christopher M. BROWN: Learning Multi-Label Scene Classification. In: Pattern Recognition, vol. 37 (9): pp. 1757–1771,2004.

Label Powerset Transformation

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	$\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\{\lambda_2\}$
3	C	3	0	...	0.5	$\{\}$
4	D	4	1	...	0.6	$\{\lambda_1\}$
...						

multiclass problem

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	$\mu_1=\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\mu_2=\{\lambda_2\}$
3	C	3	0	...	0.5	$\mu_3=\{\}$
4	D	4	1	...	0.6	$\mu_4=\{\lambda_1\}$
...						

binary, multi-target representation:

i	x_1	x_2	x_3	...	x_a	μ_1	μ_2	μ_3	μ_4
1	A	1	0	...	0.1	1	0	0	0
2	B	2	1	...	0.3	0	1	0	0
3	C	3	0	...	0.5	0	0	1	0
4	D	4	1	...	0.6	0	0	0	1
...									

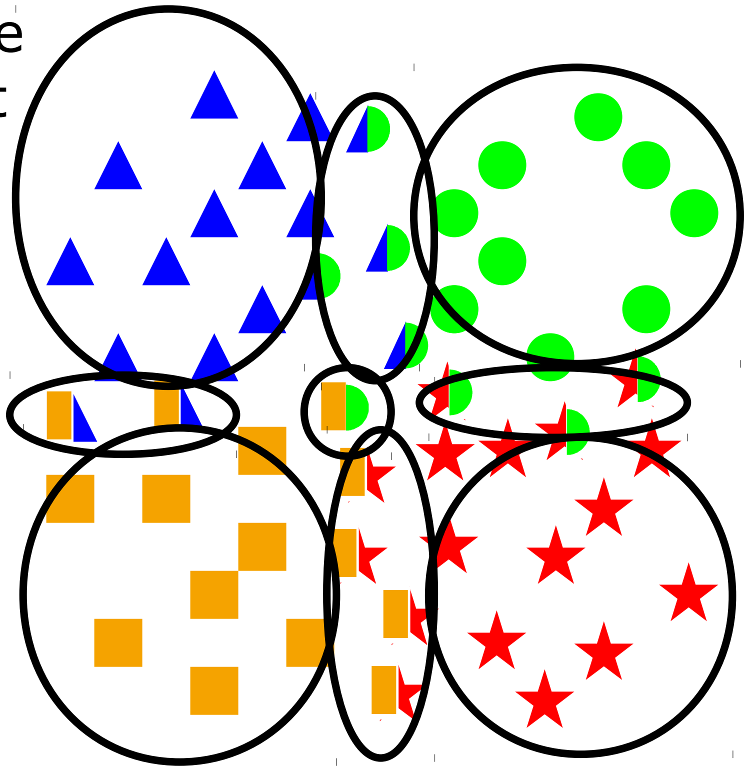
Label Powerset Transformation

Straight-forward approach: create one meta-class for each occurring labelset

- train a multiclass learner, i.e. learn each label***set*** independently
  - e.g. using Decision Tree learner, but also one-against-all or pairwise
- corresponds to learning the joint class probabilities $P(y_1,\ldots,y_n \mid x)$
  - predicts the most likely joint event y

  → consistent with Subset Accuracy
  - moreover: if we have probability estimates, we can obtain marginals $P(y_1,\ldots,y_n \mid x)$

  → also consistent with Hamming Loss and Ranking Loss



Label Powerset Transformation

Complexity

- high number of meta-classes
  - upper bounded by $\min(m,2^n)$
  - problematic for many base learners

Dataset	#training ex. m	#labels n	Distinct Labelsets $\min(m,2^n)$	Actual	Diversity
emotions	593	6	64	27	0.42
enron	1702	53	1702	753	0.44
hifind	32971	632	32971	32734	**0.99**
mediamill	43907	101	43907	6555	0.15
medical	978	45	978	94	**0.1**
scene	2407	6	64	15	0.23
tmc2007	28596	22	28596	1341	**0.05**
yeast	2417	14	2417	198	**0.08**

taken from: Multilabel Learning Tutorial by Greg Tsoumakas at ECML 2009,
http://www.ecmlpkdd2009.net/program/tutorials/learning-from-multi-label-data/

Label Powerset Transformation Limitations

TECHNISCHE UNIVERSITÄT DARMSTADT

- computationally expensive: possible labelsets may grow exponentially
  - solutions exist: Pruned Sets[1], RakEL[2]
  - but: ensemble approaches (costly, more parameters) and no clear objective anymore
- limited training examples for many labelsets

  → often reduced prediction quality
- prediction of unseen label combinations in training data impossible
- learn co-occurrences, but no explicit interdependencies ("implications")
  - though we can compute any $P(y_{i1},y_{i2},..| y_{j1},y_{j2},..,x)$ we want for each test instance separately
    - but no global model, not represented in model

[1] Read, Jesse ; Pfahringer, Bernhard ; Holmes, Geoffrey: Multi-label classification using ensembles of pruned sets. In ICDM 2008
[2] Grigorios Tsoumakas, Ioannis Katakis, Ioannis Vlahavas: Random k-Labelsets for Multi-Label Classification. IEEE Transactions on Knowledge and Data Engineering. 2011

Label (In-)Dependence

Differentiation between two types of dependencies[1]:

Unconditional dependency: $$P(\mathbf{y}) \neq \prod_{i=1}^{n} P(y_i)$$

- unconditional on the instance at hand
  → "global" dependency
- e.g. hierarchical constraints: *P*(*parent category* | *child category*)=1

sidenote: independence would exist if *P*(*parent* , *child*) = *P*(*parent*) P(*child*), i.e. *P*(*parent* | *child*) = *P*(*parent*)

Conditional dependency: $$P(\mathbf{y} \mid \mathbf{x}) \neq \prod_{i=1}^{n} P(y_i \mid \mathbf{x})$$

- conditional on the instance at hand
  → "local" dependency
- e.g. *P*(*foreign affairs* | *politics*, "text about Euro crisis") > *P*(*foreign affairs* | *politics*)

[1] Krzysztof DEMBCZY NSKI, Willem WAEGEMAN, Weiwei CHENG, Eyke HÜLLERMEIER: On label dependence in multi-label classification. In: Proceedings of the ICML-10 Workshop on Learning from Multi-Label Data

Label (In-)Dependence

- there does not have an implication between conditional (in)dependence and unconditional (in)dependence
  - but unconditional is the "average" conditional dependence: $P(\mathbf{y}) = \int_{\mathcal{X}} P(\mathbf{y} \mid \mathbf{x}) d\mathbf{x}$

Exploitation of label dependencies

- typically: exploit unconditional dependencies, e.g. via regularization, for predicting conditional distributions
- but: the effect of exploiting label dependence is often difficult to isolate, and difficult to distinguish from other reasons of improvement
  - often improvement is due to using a more complex model than in the baseline

Classifier Chains[1]

Idea: instead of learning models $h_i(x)$ for predicting label y_i (like BR), why not learning $h_i(x,y_j)$

- would capture conditional dependence $P(\ y_i \mid y_j, x)$

$h(x_1,x_2,x_3,x_4)=y_1$

$h(x_1,x_2,x_3,x_4,y_1)=y_2$

$h(x_1,x_2,x_3,x_4,y_1,y_2)=y_3$

$h(x_1,x_2,x_3,x_4,y_1,y_2,y_3)=y_4$

→ CC stacks predictions of previous binary single-label classifiers (BR classifiers)

- explicitly models label dependencies
- but: fixed ordering, **learns dependencies only in one direction**
- corresponds to learning conditional label probabilities $P(y_i \mid y_1 \ldots y_{i-1}, x)$
  - but only dependencies in direction $y_1 \ldots y_{i-1} \rightarrow y_i$

[1] Jesse READ, Bernhard PFAHRINGER, Geoff HOLMES, Eibe FRANK: Classifier chains for multi-label classification. In: Machine Learning, vol. 85 (3): pp. 333–359, 2011.

Classifier Chains

TECHNISCHE UNIVERSITÄT DARMSTADT

CC explicitly models label dependencies

- *modelling* in the sense of explicitly capturing the interdependencies in the model
  - with chain rule of probability, it is possible to compute P(y | x)[1], and hence any $P(y_{i1},y_{i2},..| y_{j1},y_{j2},..,x)$ (like for LP)
- but: fixed ordering, **learns dependencies only in one direction**
  - only in predetermined direction $y_1 \ldots y_{i-1} \rightarrow y_i$

→ Ensemble CC merges prediction of m independent CC with different ordering of labels in the chain (often m=50)

- increases complexity

[1] Krzysztof DEMBCZYNSKI, Weiwei CHENG, Eyke HÜLLERMEIER: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML 2010

Classifier Chains Limitations

- CC is only approximation of finding the most likely combination y
  - compute full $P(y \mid x)$[1] (2^n combinations!) or use Monte Carlo search approaches[2]
- for n>50, CC does not improve over BR (chains too long)
- it is not clear whether improvement of CC due to exploiting dependencies or increase of expressivity of the model in stacking
- general critics on stacking label information:
  - CC learns a function $h_1(x,y_2)$ for predicting y_1
  - but y_2 is not known, so a second function $h_2(x)$ is learned, in order to predict y_2, which is then put into h_1: $h_1(x,h_2(x))$
  - but then, why not directly learning a function $h_1'(x)$ instead of $h_1(x,h_2(x))$ since $h_2(x)$ and $h_1(x,h_2(x))$ all only depend on input x?

[1] Krzysztof DEMBCZYNSKI, Weiwei CHENG, Eyke HÜLLERMEIER: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML 2010
[2] Jesse Read, Luca Martino, David Luengo: Efficient Monte Carlo Methods for Multi-Dimensional Learning with Classifier Chains. Submitted to Pattern Recognition

Comparisons

Own experiments on three datasets emotions, scene, yeast mainly confirm our analyses:

- LP best in Subset Accuracy, followed by CC
- pairwise approach (CLR) best for ranking measures (Ranking Loss and Average Precision, statistically significant)
- but BR only good w.r.t. Precision, also worst for Hamming Loss!
  - predicts too conservative? Why ...?
- CC not better than LP at Subset Accuracy, and very bad at ranking
  - it is not clear how to correctly do ranking for CC at all

average rankings (following Friedman test):

Measure	CLR		LP		CC		BR	CD
ACC	3.400	<	**1.489**	=	1.722	>	3.389	0.700
HAMLOSS	2.967	<	**1.787**	=	2.160	>	3.087	0.700
PREC	1.989	>	3.467	=	3.111	<	**1.433**	"
REC	2.156	=	**1.956**	=	2.422	>	3.467	"
AVGP	**1.000**	>	2.778	=	3.111	=	3.111	"
RANKLOSS	**1.000**	>	2.622	=	3.133	=	3.244	"

Wouter DUIVESTEIJN, Eneldo LOZA MENCÍA, Johannes FÜRNKRANZ, Arno J. KNOBBE: Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns. In: IDA-2011

Holistic Approaches “Classical” ones

Rank-SVM (!= SVMrank)

- incorporates pairwise label constraints directly in the optimization problem
- classical approach, but slow and not scalable

Multilabel C4.5 decision tree learner

- defines new splitting criterion based on multi-label entropy

BP-MLL

- extension of BP neural network, which uses error function based on pairwise Ranking Loss
- but new findings suggest that error function is not consistent!
- our own extension with Hinge-loss based error function works is consistent and works better (contact Jinseok Nam!)

ML-kNN

- combines label distribution of k neighbors and a priori distribution

ELISSEEFF, WESTON: A Kernel Method for Multi-Labelled Classification. In: NIPS 2001
ZHANG, ZHOU: Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. In: IEEE Transactions on Knowledge and Data Engineering, 2006
Clare, King: Knowledge Discovery in Multi-label Phenotype Data. In: PKDD 2001
Zhang,Zhou: ML-KNN: A lazy learning approach to multi-label learning.Pattern Recognition 2007

Holistic Approaches
Newer ones

TECHNISCHE UNIVERSITÄT DARMSTADT

Ensembles of Random Decision Trees (RDT)

- generate k random RDT with random tests at inner nodes
- leaf nodes contain observed label distribution of arrived training examples
- very fast to train and to apply, very memory efficient (for k=O(1))

Parametric mixture models

- probabilistic generative models for each label in form of prototypes (basically word distributions)
- labelsets are modeled on top with respect to label prototypes

Topic Models

- assume that a label corresponds to a topic, but additional LDA process on top samples topics and hence models dependencies

Xiatian ZHANG, Quan YUAN, Shiwan ZHAO, Wei FAN, Wentao ZHENG, Zhong WANG: Multi-label Classification without the Multi-label cost. In: SIAM ICDM 2010
Naonori UEDA, Kazumi SAITO: Parametric Mixture Models for Multi-Labeled Text. In: NIPS 2002
RUBIN, CHAMBERS, SMYTH, STEYVERS: Statistical Topic Models for Multi-Label Document Classification. To be published in: The Machine Learning Journal

Large Number of Labels

Keyword tagging: common setting for multilabel problems

- from Web 2.0, wikis, archives, …
- dataset examples:
  - delicious[1]: 16105 web sites tagged in the social bookmarking platform
    - 983 keywords, on average 19 labels per document
  - EUR-Lex: 19328 legal documents tagged with EUROVOC descriptors
    - 3965 descriptors, on average 5.37 labels per document
  - ECML 2012 Discovery Challenge[2]: 2.4 mio. documents from Wikipedia!
    - 325000 possible categories!
    - reset 2014 as 4 LSHTC Challenge
  - and … Twitter data annotated with mio. of hashtags

[1] Grigorios TSOUMAKAS, I. KATAKIS, Ioannis P. VLAHAVAS: Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In: Proceedings ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08), 2008.
[2] http://www.ecmlpkdd2012.net/info/discovery-challenge/ , http://lshtc.iit.demokritos.gr/

Large Number of Labels Solutions

Adaptation

- e.g.: dual reformulation of pairwise ensemble of linear classifiers

  → rough idea: save each of the quadratic number of linear classifiers as linear combination of its support vectors

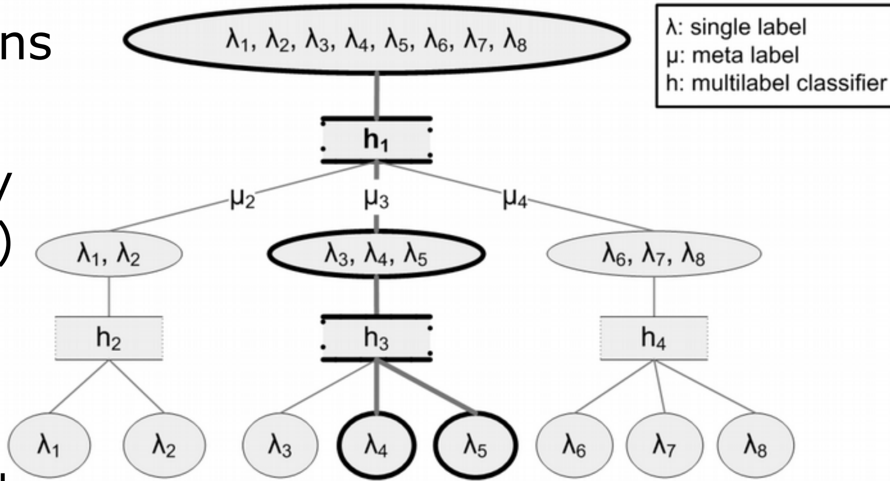
  - memory costs now limited by size of the training set
  - DMLPP was able to solve EUR-Lex problem with 4000 labels (→ usually 8 mio. pairwise classifiers needed!)
    - training is also done in the dual → online training possible
    - predictive quality was much better than BR approaches
  - Multilabel LibSVM
    - simple modifications of LibSVM for pairwise multilabel classification
    - but more than 100 times less time and memory!
    - www.ke.tu-darmstadt.de/resources/multilabellibsvm or contact Eneldo
  - but of course limited scalability!

E. Loza, S.-H. Park, J. Fürnkranz: Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain. In: ECML 2008
Eneldo Loza Mencía and Johannes Fürnkranz: Efficient multilabel classification algorithms for large-scale problems in the legal domain. In Semantic Processing of Legal Texts, pages 192-215

Large Number of Labels Solutions

Structured Decompositions

- e.g. **HOMER: Hierarchy of Multilabel Classifiers**
  - breaks up the problem into subproblems organized in a hierarchy
  - k labels are joined to one multilabel, which in turn is one possible label in the parent multilabel problem
  - labels are joined by balanced k-means
- Own results:
  - HOMER and pairwise harmonize very well: accurate and fast(-er than BR!)
  - HOMER enables to apply pairwise to potentially arbitrarily large datasets
    - margin to BR reduced to a user-defined constant factor k
    - though, problem transformation is not equivalent anymore



TSOUMAKAS, KATAKIS, P. VLAHAVAS: Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In: Proceedings ECML/PKDD MMD'08, 2008.
G. Tsoumakas, E. Loza, I. Katakis, S.-H. Park, J. Fürnkranz: *On the Combination of Two Decompositive Multi-Label Classification Methods*. In: Proceedings of the ECML PKDD 2009 Workshop on Preference Learning

Large Number of Labels
Solutions

Label Output Space Transformations

- Starting Point: sparsity of label space
  - only little labels relevant even for large number of labels
- Idea: compress label vector y to less dimensional vector y' and solve new problem $x \rightarrow y'$: $y' = A\,y$
  - different techniques for building projection Matrix A:
    - randomly (compressed sensing[1])
    - singular value decomposition[2]
    - Kernel Principal Component Analysis[3]
  - predicting y' usually solved by using multivariate regression
    - nature of problem is completely changed
  - predicting y: inverse projection of $y'' = A^{-1}\,y'$, then find closest y using e.g. error correcting output codes (y'' is still numeric)

[1] HSU, KAKADE, LANGFORD, ZHANG: Multi-Label Prediction via Compressed Sensing.NIPS 2009
[2] Farbound TAI, Hsuan-tien LIN: Multi-label Classification with Principle Label Space Transformation. to appear in Neural Computation
[3] Wei BI, James Tin-Yau KWOK: Multi-Label Classification on Tree- and DAG-Structured Hierarchies. In: ICML 2011

Continued in MULAN slides

- Introduction
  - Multilabel Setting
  - Applications & Datasets
- Theoretical Foundations
  - Probabilities in Multilabel
    - joint vs. marginal
  - Losses
    - Ranking
- Programming in MULAN
  - data loading
  - training and evaluation
  - implementation of new approach
- Algorithms
  - Transformation vs. Holistic
  - Transformational Approaches
    - BR, LP, Pairwise
  - Label Dependencies
    - Classifier Chains
  - Holistic Approaches
    - Overview
  - Large Number of Labels
    - Adaptations
    - HOMER
    - Label Space Transformation

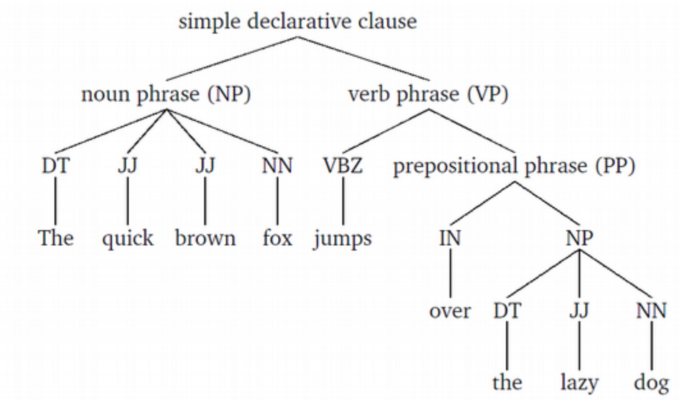
Current and Future Work

Pairwise decomposition

- build in pairwise formulation directly in Neural Networks
  - save computational costs, improve accuracy
- take label intersections into consideration
  - better exploit label dependencies
  - adapt pairwise voting to other losses rather than ranking specific

Syntactic Parsing

- exploit annotation dependencies
- consider all annotations at once instead of separately
  - use e.g. Dependent BR
- collaboration is welcome!



token	The	quick	brown	fox	jumps	over	the	lazy	dog
token features	the=1 +1.quick=1	quick=1 +1.brown=1 -1.the=1	brown=1 +1.fox=1 -1.quick=1	fox=1 +1.jumps=1 -1.brown=1	jumps=1 +1.over=1 -1.fox=1	over=1 +1.the=1 -1.jumps=1	the=1 +1.lazy=1 -1.over=1	lazy=1 +1.dog=1 -1.the=1	dog=1 -1.lazy=1
POS	[DT, DT]	[JJ, JJ]	[JJ, JJ]	[NN, NN]	[VBZ, VBZ]	[IN, IN]	[DT, DT]	[JJ, JJ]	[NN, NN]
syntactic	[NP			NP]	[VP	[PP	[NP		NP], PP], VP]

Thank you for your attention

Questions?

References and Further Reading

- Tutorial given at MLKDD 2013 by Jesse Read
  - http://www.tsc.uc3m.es/~jesse/
- Tutorial on Multi-target prediction at ICML 2013
  - http://www.ngdata.com/knowledge-base/icml-2013-tutorial-multi-target-prediction/
- Tutorial by Greg Tsoumakas at ECML 2009
  - http://www.ecmlpkdd2009.net/program/tutorials/learning-from-multi-label-data/
- Survey papers
  - G. Tsoumakas, I. Katakis, "Multi-Label Classification: An Overview", International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.
  - G. Tsoumakas, I. Katakis, I. Vlahavas, "Mining Multi-label Data", Data Mining and Knowledge Discovery Handbook, O. Maimon, L. Rokach (Ed.), Springer, 2nd edition, 2010.
- Dissertation of Eneldo :)
  - "Efficient Pairwise Multilabel Classification", 2012, http://www.ke.tu-darmstadt.de/bibtex/publications/show/2337