MULTI-LABEL CLASSIFICATION

Pedro Larrañaga, Concha Bielza

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Machine Learning

Master Data Science + Master HMDA

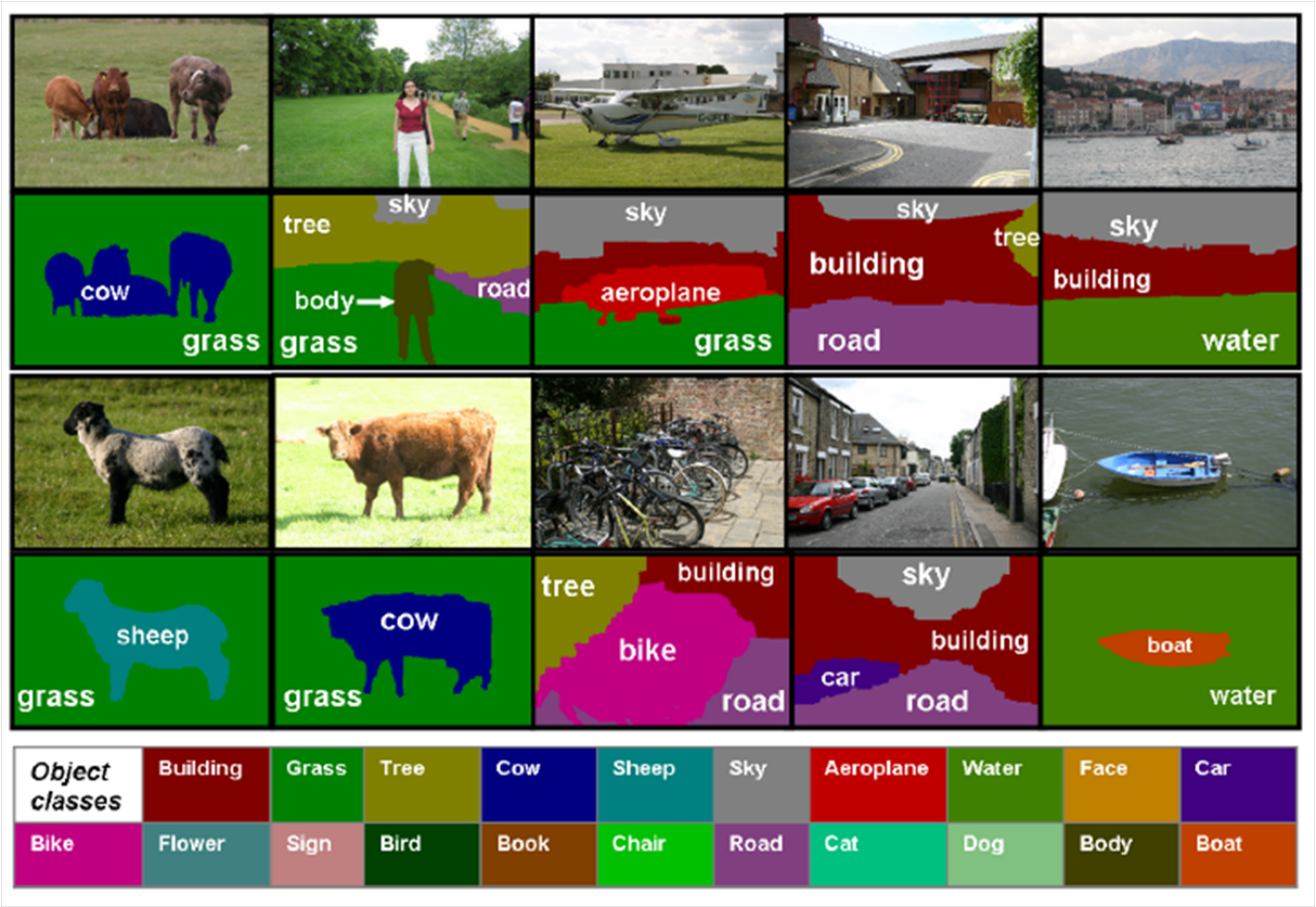
Outline

Introduction. Single Label 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

Introduction. Image. Simultaneous Object Class Recognition



- Image understanding by multi-label object recognition and segmentation (Shotton et al. (2007). International Journal on Computer Vision)

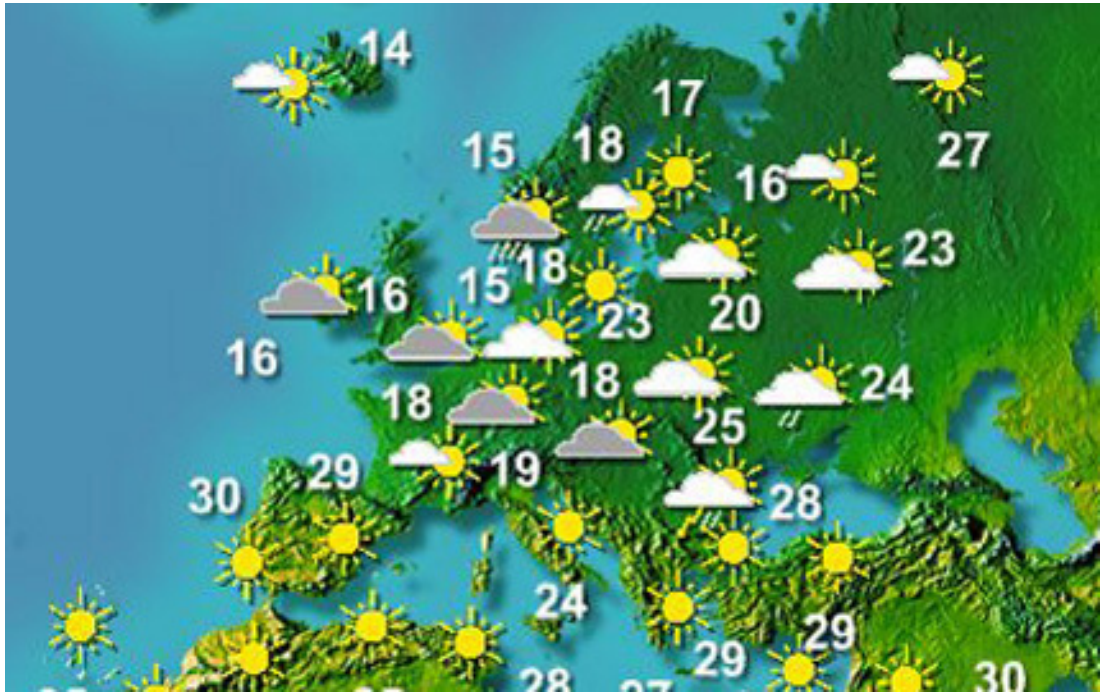
Introduction. Call-type Categorization

- Automatic call-type identification (Schapire and Singer (2000))
- 8000 examples in the training set, and 1000 in the test set
- 14 call types + an "other" label
- Some calls can be of more than one type (e.g. `collect` and `person-to-person`)

Introduction. Medical Diagnosis

- Features: medical history, symptoms
- Classes: diseases

Introduction. Weather Forecast

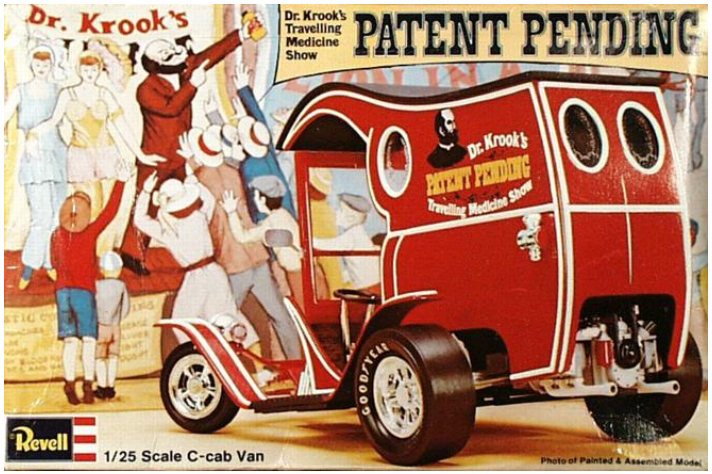


- Features: from meteorological stations: thermometer, barometer, hygrometer, anenometer, wind vane, rain gauge, disdrometer, transmissometer, ceilling projector, ...
- Classes: temperature, humidity, rain, wind, cloud, fog,...
- Temporal and spatial characteristics

Introduction. Emotional Categorization of Music

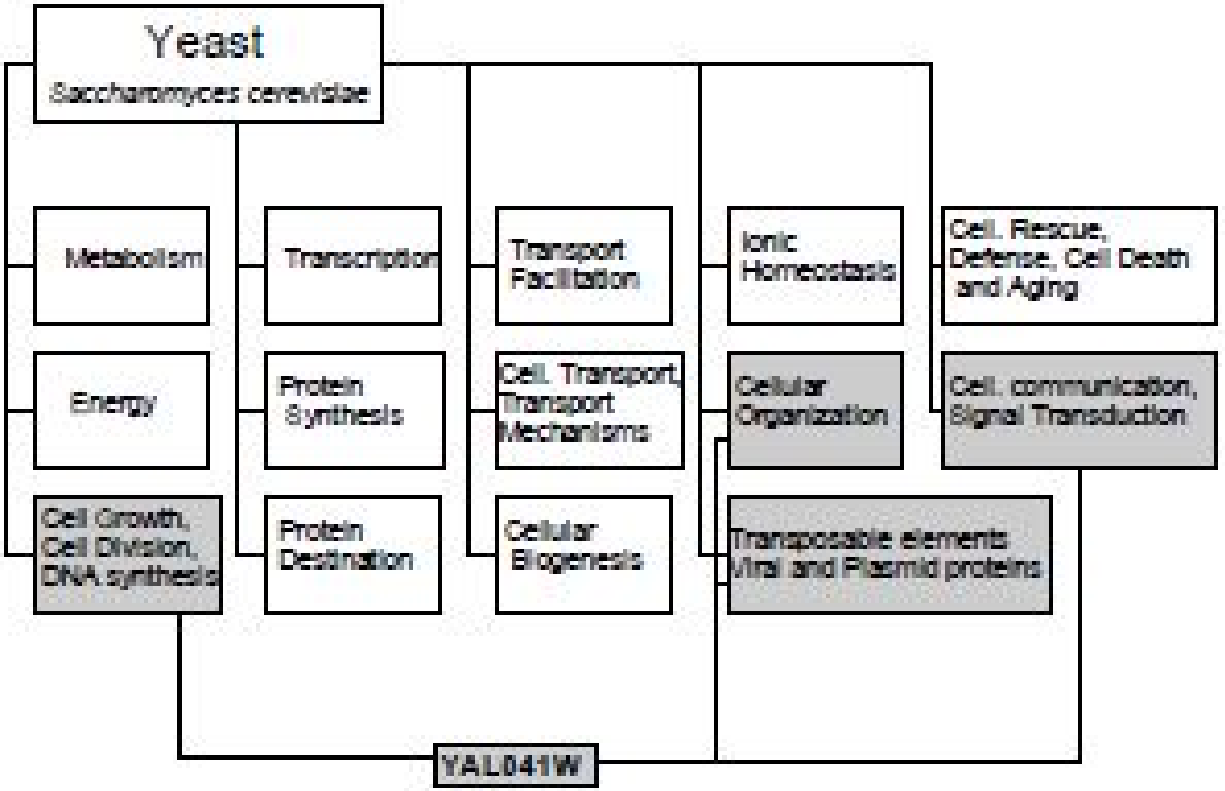
- Data set `Emotions`: 593 tracks (100 songs) from each of the following 7 different genres: classical, reggae, rock, pop, hip-Hop, techno and jazz (Trohidis et al. (2008) ISMIR)
- Features: 72 (8 rhythmic + 64 timbre)
- Labels: 6 following the Tellegen-Watson-Clark model of mood (happy, calm, sad, angry, quiet, amazed)

Introduction. International Patent Classification



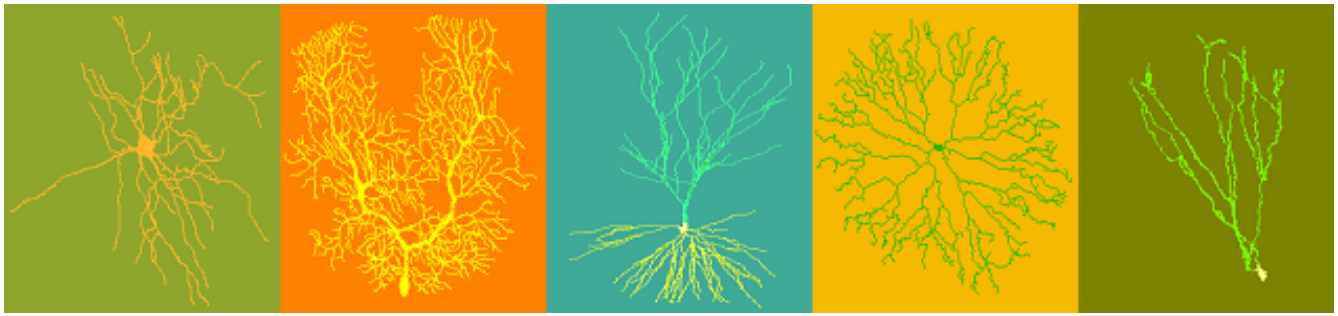
- Automating the attribution of international patent classification codes to patent applications (Fall et al. 2003)
- Patent applications: title, list of inventors, list of applicants, abstract, claim section, long description converted to electronic form by OCR
- International Patent Classification (IPC): standard, complex hierarchical taxonomy covers all areas of technology currently used by more 90 countries
- IPC taxonomy: 8 sections, 120 classes, 630 subclasses, 69000 groups, more than 40 million of documents
- Document collection: 46324 for training and 28926 for testing at `http://www.wipo.int/ibis/datasets`

Introduction. Bioinformatics



- Gene functional classification problem
- 2417 examples (1500 for training and 917 for testing)
- Features: 103 characteristics of the gene
- Labels: 14 (in the first functional catalogue (FunCAt) level)
- One gene, for instance the gene YAL041W, can belong to different groups (shaded in grey)

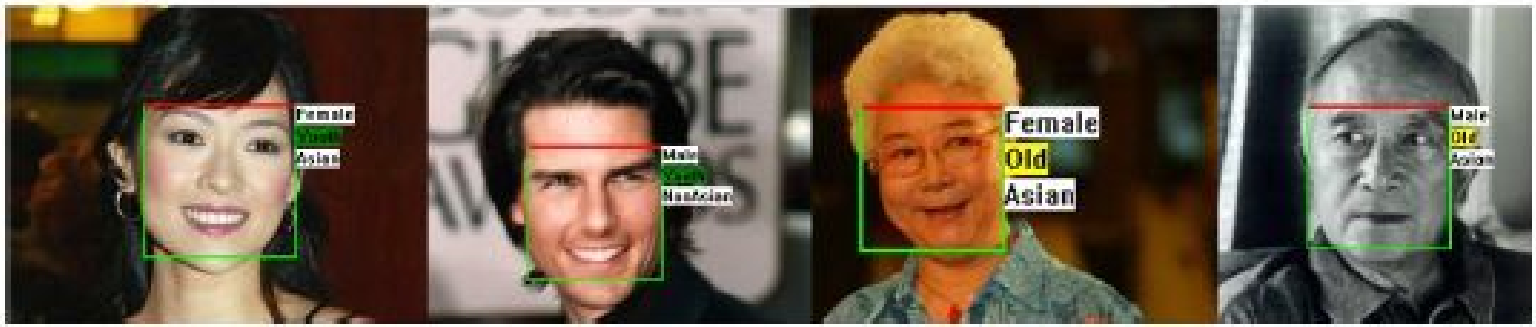
Introduction. Neurology



- Features: 65 morphological variables
- Labels:
  - Animal specie (rat, mouse, monkey, cat, human, ...)
  - Cell (neuron) type: pyramidal, interneuron, Purkinje, Martinotti, ...)
  - Brain region (amygdala, cerebral cortex, hippocampus, ...)
  - Age of the animal
- Collaboration with Howard Hughes Medical Institute (Columbia University)

Introduction. Demographic Classification from Images

- Demographic classification (sex, age, ethnicity, ...) from digital facial images



Yang and Ai (ICB 2007)

- Developing our own dataset in collaboration with Universidad Católica del Norte (Chile)

Introduction. Multi-output Response

- Problems with multiple target variables
- What can the type of targets be?
  - Categorical targets
    - Binary targets: Multi-label classification
    - Multi-class targets: Multi-dimensional classification (nominal or ordinal)
  - Numerical: Multi-output regression
  - Combination of categorical and numerical

Introduction. Standard Notation

- An *m*-dimensional input space: $\Omega_{\boldsymbol{X}}$ for $\boldsymbol{X} = (X_1, ..., X_m)$ with $\Omega_{\boldsymbol{X}} = \prod_{i=1}^{m} \Omega_{X_i}$ where $\Omega_{X_i} \subseteq \mathbb{N}$ (for nominal features) or $\Omega_{X_i} \subseteq \mathbb{R}$ (for numeric features)
- A set of *d* possible output labels: $\mathcal{Y} = \{\lambda_1, ..., \lambda_d\}$
- A multi-label data set with *N* training examples: $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \mathrm{Y}^{(1)}), ..., (\boldsymbol{x}^{(N)}, \mathrm{Y}^{(N)})\}$ where $\boldsymbol{x}^{(i)} \in \Omega_{\boldsymbol{X}}$ and $Y^{(i)} \subseteq \mathcal{Y}$ for all $i \in \{1, ..., N\}$
- Example of $\mathcal{D}$ with $m = 5$, $d = 4$ and $N = 5$

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	$\{\lambda_1, \lambda_2, \lambda_3\}$

Introduction. Standard Notation. Classification

- Classification produces a bipartition of the set of labels into a relevant (positive) and an irrelevant (negative) set
  - Example. Given $\mathcal{Y} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ an instance $\boldsymbol{x} = (x_1, ..., x_m)$ produce a bipartition $Pos_{\boldsymbol{x}} = \{\lambda_2, \lambda_5\}$ and $Neg_{\boldsymbol{x}} = \{\lambda_1, \lambda_3, \lambda_4\}$
- The learning task consists of obtaining a function h

$$h : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \to Y \subseteq \mathcal{Y}$$

$$(x_1, ..., x_m) \mapsto \{\lambda_r, ..., \lambda_u\} \subseteq \mathcal{Y}$$

that generalizes well, in the sense of minimizing the expected prediction loss with respect to a specific loss function

Introduction. Standard Notation. Ranking

- Ranking produces a total strict order of all labels according to relevance to the given instance
  - Example. Given $\mathcal{Y} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ an instance $\boldsymbol{x} = (x_1, ..., x_m)$ produce a ranking: $r_{\boldsymbol{x}}(\lambda_2) < r_{\boldsymbol{x}}(\lambda_5) < r_{\boldsymbol{x}}(\lambda_3) < r_{\boldsymbol{x}}(\lambda_1) < r_{\boldsymbol{x}}(\lambda_4)$ where $r_{\boldsymbol{x}}(\lambda_i)$ denotes the position of label λ_i in the ranking associated with $\boldsymbol{x}$
- The learning task consists of obtaining a function f

$$f : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \to \Pi(\mathcal{Y})$$

$$(x_1, ..., x_m) \mapsto \Pi(\lambda_1, ..., \lambda_d)$$

  whose associated classification generalizes well
- The classification associated with the ranking should be consistent: $\lambda_i \in Pos_{\boldsymbol{x}}, \lambda_j \in Neg_{\boldsymbol{x}} \Rightarrow r_{\boldsymbol{x}}(\lambda_i) < r_{\boldsymbol{x}}(\lambda_j)$
  - The ranking: $r_{\boldsymbol{x}}(\lambda_2) < r_{\boldsymbol{x}}(\lambda_5) < r_{\boldsymbol{x}}(\lambda_3) < r_{\boldsymbol{x}}(\lambda_1) < r_{\boldsymbol{x}}(\lambda_4)$ can produce the bipartition: $Pos_{\boldsymbol{x}} = \{\lambda_2, \lambda_5\}$ and $Neg_{\boldsymbol{x}} = \{\lambda_3, \lambda_1, \lambda_4\}$

Introduction. Notation with Labels as Variables

- A class variable for label obtaining a d-dimensional variable: $\boldsymbol{C} = (C_1, ..., C_d)$ where C_i is the binary variable associated with label λ_i with $i \in \{1, ..., d\}$
- The learning task consists of obtaining a function h

$$h : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \rightarrow \Omega_{C_1} \times \cdots \times \Omega_{C_d}$$

$$(x_1, ..., x_m) \mapsto (c_1, ..., c_d)$$

  that generalizes well, in the sense of minimizing the expected prediction loss with respect to a specific loss function
- Allows to works with multi-dimensional classification problems, where $\exists j$ such that $|\Omega_{C_j}| \neq 2$

Introduction. Equivalent Notations

X_1	X_2	X_3	X_4	X_5	C_1	C_2	C_3	C_4	$Y \subseteq \mathcal{Y}$
3.2	1.4	4.7	7.5	3.7	1	0	0	1	$\{\lambda_1, \lambda_4\}$
2.8	6.3	1.6	4.7	2.7	0	0	1	1	$\{\lambda_3, \lambda_4\}$
7.7	6.2	4.1	3.3	7.7	1	0	0	1	$\{\lambda_1, \lambda_4\}$
9.2	0.4	2.8	0.5	3.9	0	1	0	0	$\{\lambda_2\}$
5.5	5.3	4.9	0.6	6.6	1	1	1	0	$\{\lambda_1, \lambda_2, \lambda_3\}$

Outline

Multi-label Data Sets Statistics

- An m-dimensional input space: $\Omega_{\boldsymbol{X}}$ for $\boldsymbol{X} = (X_1, ..., X_m)$
- A set of d possible output labels: $Y \subseteq \mathcal{Y} = \{\lambda_1, ..., \lambda_d\}$
- A class variable for label obtaining a d-dimensional variable: $\boldsymbol{C} = (C_1, ..., C_d)$ where C_i is the binary variable associated with label λ_i
- A multi-label data set:
  - $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \mathrm{Y}^{(1)}), ..., (\boldsymbol{x}^{(N)}, \mathrm{Y}^{(N)})\}$
  - $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{c}^{(1)}), ..., (\boldsymbol{x}^{(N)}, \boldsymbol{c}^{(N)})\}$

STATISTICS

- Label cardinality: $l_c = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{d} c_i^{(j)}$
  Average number of labels for example
- Label density $l_d = \frac{l_c}{d}$
  Label cardinality divided by the total number of labels
- Distinct labelsets: Number of different label combinations
- Diversity: $div = l_c \cdot N$
  Total number of labels

Evaluation Metrics: A Taxonomy

TSOUMAKAS AND VLAHAVAS (2007). ECML/PKDD

- Based on calculation
  - Example-based are calculated separately for each test example and averaged across the test set
  - Label-based are calculated separately for each label and then averaged across all labels
- Based on the output of the learner
  - Binary prediction for each label
  - Ranking of the labels (example-based)
  - Probability or score for each label

Example-based and Binary. Subset Accuracy

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

$$\text{SUBSET ACCURACY} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{Y}^{(i)} = Y^{(i)})$$

where $\mathbb{I}(\text{true}) = 1$ and $\mathbb{I}(\text{false}) = 0$

- SUBSET ACCURACY = $\frac{1}{5}(0 + 1 + 1 + 0 + 0)$
- Also called classification accuracy
- Very strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels

ZHU ET AL. (2005). ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL

Example-based and Binary. Hamming Loss

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

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

where Δ stands for the symmetric difference of two sets (XOR operation)

- HAMMING LOSS = $\frac{1}{4} \cdot \frac{1}{5}(2 + 0 + 0 + 1 + 1)$
- Average binary classification error

SCHAPIRE AND SINGER (2000). MACHINE LEARNING

Example-based and Ranking. One-error

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

$$\text{ONE-ERROR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\arg\min_{\lambda \in \mathcal{Y}} r_{\boldsymbol{x}^{(i)}}(\lambda) \notin Y^{(i)})$$

where $\mathbb{I}(\text{true}) = 1$ and $\mathbb{I}(\text{false}) = 0$

- $\text{ONE-ERROR} = \frac{1}{5}(0 + 0 + 0 + 0 + 1)$
- Evaluates how many times the top-ranked label is not in the set of proper labels of the example

SCHAPIRE AND SINGER (2000). MACHINE LEARNING

Outline

Overview of Techniques

- Problem transformation methods
  - They transform the learning task into one or more single-label classification tasks
  - They are algorithm independent
  - Some could be used for feature selection as well
- Algorithm adaptation methods
  - They extend specific learning algorithms in order to handle multi-label data directly
  - Boosting, generative (mixtures), SVM, decision tree, neural network, k-NN, probabilistic chain classifier, Bayesian networks, ...

Problem Transformation

- Binary relevance
- Ranking via single-label learning
- Pairwise methods
  - Ranking by pairwise comparison
  - Calibrated label ranking
- Methods that combine labels
  - Label powerset
  - Pruned sets
- Ensemble methods
  - RAkEL
  - Ensemble of pruned sets
- Identifying label dependencies
  - Correlation-based pruning of stacked BR models
  - Chi-Dep
  - Chi-Dep ensemble
  - Hierarchy of multi-label classifiers

Problem Transformation. Binary Relevance

GODBOLE AND SARAWAGI (2004). PAKDD

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

$\boldsymbol{x}$	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
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(1)}$	true

$\boldsymbol{x}$	λ_3
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	true

$\boldsymbol{x}$	λ_4
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(1)}$	true
$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(1)}$	false

- Learns one binary classifier for each label independently of the rest of labels
- Outputs the concatenation of their predictions
- Does not consider label relationships

Problem Transformation. Pairwise Methods. Ranking by Pairwise Comparison

HÜLLERMEIER ET AL. (2008). ARTIFICIAL INTELLIGENCE

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

λ_1-λ_2		λ_1-λ_3		λ_1-λ_4		λ_2-λ_3		λ_2-λ_4		λ_3-λ_4	
$\boldsymbol{x}^{(1)}$	true	$\boldsymbol{x}^{(1)}$	true	$\boldsymbol{x}^{(2)}$	false	$\boldsymbol{x}^{(2)}$	false	$\boldsymbol{x}^{(1)}$	false	$\boldsymbol{x}^{(1)}$	false
$\boldsymbol{x}^{(3)}$	true	$\boldsymbol{x}^{(2)}$	false	$\boldsymbol{x}^{(5)}$	true	$\boldsymbol{x}^{(4)}$	true	$\boldsymbol{x}^{(2)}$	false	$\boldsymbol{x}^{(3)}$	false
$\boldsymbol{x}^{(4)}$	false	$\boldsymbol{x}^{(3)}$	true					$\boldsymbol{x}^{(3)}$	false	$\boldsymbol{x}^{(5)}$	true
								$\boldsymbol{x}^{(4)}$	true		
								$\boldsymbol{x}^{(5)}$	true		

- It learns $d(d-1)/2$ binary models, one for each pair of labels
- Each model is trained based on examples that are annotated by at least one of the labels, but not both
- Given a new instance, all models are invoked and a ranking is obtained by counting the votes receives by each label

$\boldsymbol{x}$	λ_1-λ_2	λ_1-λ_3	λ_1-λ_4	λ_2-λ_3	λ_2-λ_4	λ_3-λ_4
	λ_1	λ_3	λ_1	λ_3	λ_4	λ_3

The ranking: $r_{\boldsymbol{x}}(\lambda_3) < r_{\boldsymbol{x}}(\lambda_1) < r_{\boldsymbol{x}}(\lambda_4) < r_{\boldsymbol{x}}(\lambda_2)$

Problem Transformation. Pairwise Methods. Calibrated Label Ranking

HÜLLERMEIER ET AL. (2008). ARTIFICIAL INTELLIGENCE

λ_1-λ_0	
$x^{(1)}$	true
$x^{(2)}$	false
$x^{(3)}$	true
$x^{(4)}$	false
$x^{(5)}$	true

λ_2-λ_0	
$x^{(1)}$	false
$x^{(2)}$	false
$x^{(3)}$	false
$x^{(4)}$	true
$x^{(5)}$	true

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

λ_3-λ_0	
$x^{(1)}$	false
$x^{(2)}$	true
$x^{(3)}$	false
$x^{(4)}$	false
$x^{(5)}$	true

λ_4-λ_0	
$x^{(1)}$	true
$x^{(2)}$	true
$x^{(3)}$	true
$x^{(4)}$	false
$x^{(5)}$	false

λ_1-λ_2	
$x^{(1)}$	true
$x^{(3)}$	true
$x^{(4)}$	false

λ_1-λ_3	
$x^{(1)}$	true
$x^{(2)}$	false
$x^{(3)}$	true

λ_1-λ_4	
$x^{(2)}$	false
$x^{(5)}$	true

λ_2-λ_3	
$x^{(2)}$	false
$x^{(4)}$	true

λ_2-λ_4	
$x^{(1)}$	false
$x^{(2)}$	false
$x^{(3)}$	false
$x^{(4)}$	true
$x^{(5)}$	true

λ_3-λ_4	
$x^{(1)}$	false
$x^{(3)}$	false
$x^{(5)}$	true

- Extends ranking by pairwise comparison by introducing an additional virtual label (λ_0)
- Pairwise models that include the virtual label correspond to models of binary relevance (all examples are used)
- The final ranking includes the virtual label, used as splitting point between positive and negative labels

x	λ_1-λ_2	λ_1-λ_3	λ_1-λ_4	λ_2-λ_3	λ_2-λ_4	λ_3-λ_4	λ_1-λ_0	λ_2-λ_0	λ_3-λ_0	λ_4-λ_0
	λ_1	λ_3	λ_1	λ_3	λ_4	λ_3	λ_1	λ_0	λ_3	λ_0

The ranking: $r_{\boldsymbol{x}}(\lambda_3) < r_{\boldsymbol{x}}(\lambda_1) < r_{\boldsymbol{x}}(\lambda_0) < r_{\boldsymbol{x}}(\lambda_4) < r_{\boldsymbol{x}}(\lambda_2) \Rightarrow \{\lambda_1, \lambda_3\}$

Problem Transformation. Combining Labels. Label Powerset

BOUTELL ET AL. (2004). PATTERN RECOGNITION

$\boldsymbol{x}$	$Y \subseteq \mathcal{Y}$	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 different set of labels becomes a different class in a new single-label classification task
- Most implementations of label powerset classifiers essentially ignore label combination that are not presented in the training set (cannot predict unseen labelsets)
- Limited training examples for many classes

Problem Transformation. Ensemble Methods. RAndom k-labELsets (RAkEL)

TSOUMAKAS AND VLAHAVAS (2007). ECML/PKDD

- Randomly (without replacement) break a large set of labels into a number (n) of subsets of small size (k), called k-labelsets
- For each of them train a multi-label classifier using the label powerset method
- Given a new instance, query models and average their decisions per label (thresholding to assign the labelset)

		predictions							
model	3-labelsets	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
h_1	$\{\lambda_1, \lambda_2, \lambda_8\}$	1	0	-	-	-	-	-	1
h_2	$\{\lambda_3, \lambda_4, \lambda_7\}$	-	-	0	1	-	-	1	-
h_3	$\{\lambda_2, \lambda_5, \lambda_6\}$	-	1	-	-	1	0	-	-
h_4	$\{\lambda_1, \lambda_7, \lambda_8\}$	1	-	-	-	-	-	1	0
h_5	$\{\lambda_3, \lambda_4, \lambda_6\}$	-	-	1	1	-	0	-	-
h_6	$\{\lambda_2, \lambda_6, \lambda_8\}$	-	0	-	-	-	0	-	1
average votes		2/2	1/3	1/2	2/2	1/1	0/3	2/2	2/3
prediction (threshold= 0.5)		1	0	1	1	1	0	1	1

P.T. Identifying Label Dependencies. Chi-Dep

TENENBOIM ET AL. (2010). ICML WORKSHOP ML

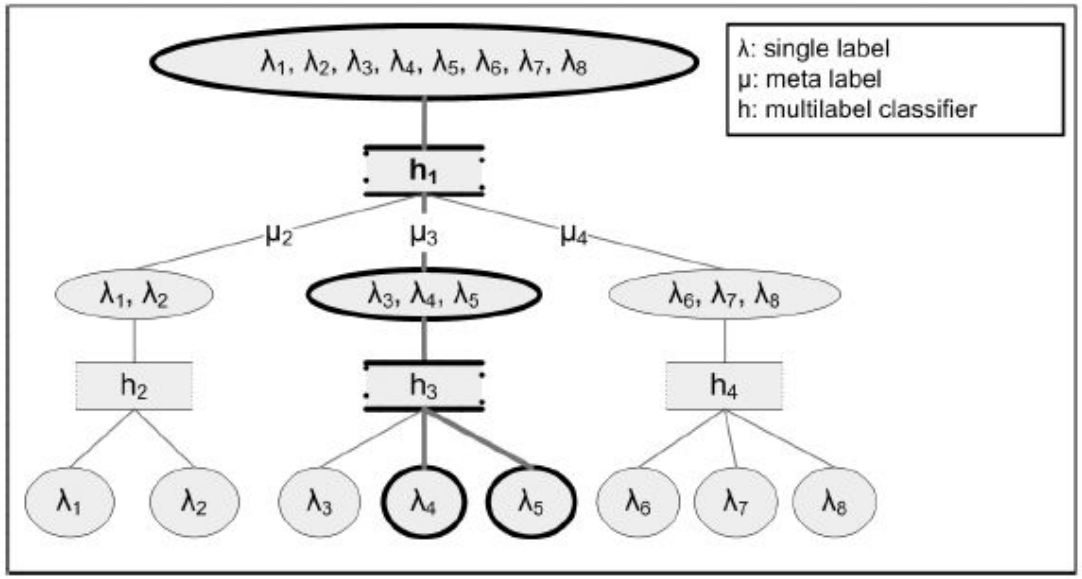
LPBR

- A model allowing to balance between BR and LP by modeling existing dependencies between labels
- LPBR algorithm
  - Step 1. Start from total independence (BR model)
  - Step 2. Cluster the pair of most dependent labels (chi-square test)
  - Step 3. Build and evaluate the new model
  - Step 4. Compare accuracy to the previous model. If improved, accept the new model
  - Repeat Steps 2-4 until stopping condition is meet
  - Return the latest accepted model
- Stopping condition:
  - Number of allowed ”non-improvement”label pair is exceeded
  - Chi-square value is bellow threshold
  - No more pairs to consider
  - All labels are in the same cluster

P.T. Identifying Label Dependencies. Hierarchy of Multi-label Classifiers

TSOUMAKAS ET AL. (2008). ECML/PKDD2008 WORKSHOP ON MMD'08

Hierarchy Of Multi-label classifiERs (HOMER)



- HOMER constructs a hierarchy of multi-label classifiers each one dealing with a much smaller set of labels and a more balanced example distribution
- Balanced *k*-means: approximately equal number of labels in each cluster

Adaptation Methods

- Boosting
- Rank-SVM
- Decision Tree
- Neural Networks
- K-Nearest Neighbor
- Instance Based Using Mallows Model
- Instance Based + Logistic Regression
- Probabilistic Chain Classifiers
- Bayesian Networks

Adaptation Methods. Probabilistic Chain Classifiers

DEMBCZYŃSKI ET AL. (2010). ICML

- Overcoming the label independence assumption of the binary relevance method
- Learn d functions h_i on augmented input spaces $\Omega_{\boldsymbol{X}} \times \{0,1\}^{i-1}$, respectively, taking $c_1, ..., c_{i-1}$ as additional features:

$$h_i : \Omega_{\boldsymbol{X}} \times \{0,1\}^{i-1} \rightarrow [0,1]$$

$$(\boldsymbol{x}, c_1, ..., c_{i-1}) \mapsto p(c_i | \boldsymbol{x}, \hat{c}_1, ..., \hat{c}_{i-1})$$

- Classifier chain was first introduced in multi-label classification by READ ET AL. (2009) in ECML/PKDD albeit without a probabilistic interpretation

Adaptation Methods. Bayesian Networks

MULTIPLE DIAGNOSIS PROBLEM

	X_1	$\ldots$	X_m	C_1	$\ldots$	C_d
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_m^{(1)}$	$c_1^{(1)}$	$\ldots$	$c_d^{(1)}$
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_m^{(2)}$	$c_1^{(2)}$	$\ldots$	$c_d^{(2)}$
$\ldots$		$\ldots$			$\ldots$	
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_m^{(N)}$	$c_1^{(N)}$	$\ldots$	$c_d^{(N)}$

Optimal diagnosis as abductive inference (MPE)

$$(c_1^*, \ldots, c_d^*) = arg \quad max_{(c_1,\ldots,c_d)} p(C_1 = c_1, \ldots, C_d = c_d | X_1 = x_1, \ldots, X_m = x_m)$$

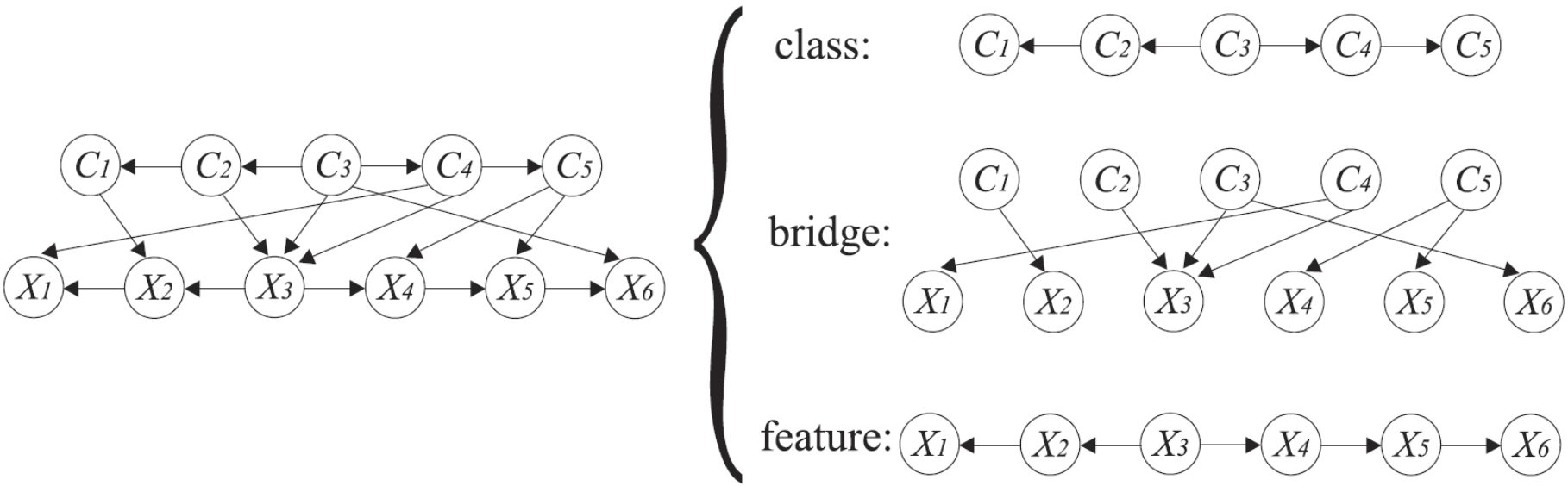
$$= arg \quad max_{(c_1,\ldots,c_d)} p(C_1 = c_1, \ldots, C_d = c_d) p(X_1 = x_1, \ldots, X_m = x_m | C_1 = c_1, \ldots, C_d = c_d)$$

Number of parameters to be estimated: $2^d - 1 + 2^d(2^m - 1)$

d	m		number parameters
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

Adaptation Methods. Bayesian Networks

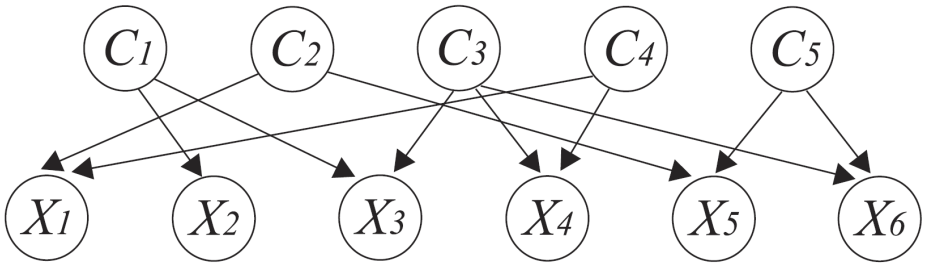
Multi-dimensional Bayesian Network Classifier (MBC)



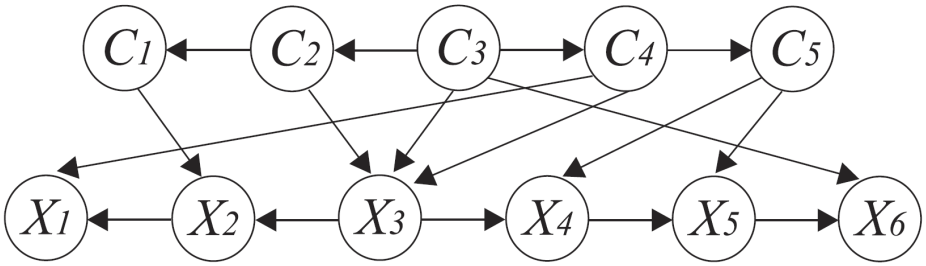
- The set of vertices $\mathcal{V}$ is partitioned into:
  - $\mathcal{V}_C = \{C_1, ..., C_d\}$ of class variables and
  - $\mathcal{V}_X = \{X_1, ..., X_m\}$ of feature variables, with $(d + m = n)$

Adaptation Methods. Bayesian Networks

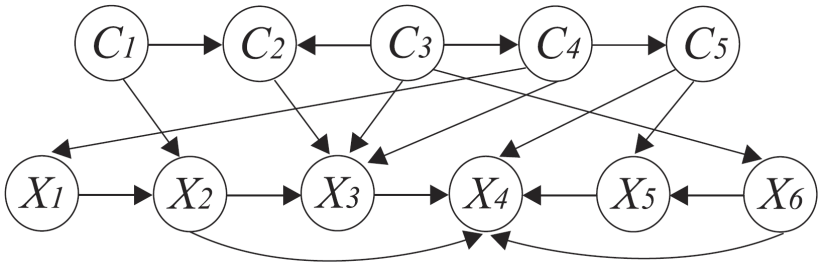
Multi-dimensional Bayesian Network Classifier (MBC)



(a) `Empty-empty` MBC



(b) `Tree-tree` MBC



(c) `Polytree-DAG` MBC

Outline

Software

- Mulan a open-source software in Weka at http://mulan.sourceforge.net contains:
  - Methods: Binary relevance, calibrated label ranking, label powerset, ppt, multi-lable stacking, RAkEL, HOMER, HMC, ML-kNN, BRkNN, IBLR, BPMLL
  - Metrics: Subset accuracy, Accuracy, Hamming loss, precision, recall, F-measure, macro, micro
- Matlab code at http://lamda.nju.edu.cn/datacode/MLkNN.htm contains: MLKNN, BPMLL
- Meka library at http://www.cs.waikato.ac/nz/∽jmr30/#software contains: Pruned sets, Classifier chains
- Multi-label alternating decision tree at http://www.grappa.univ-lille3.fr/grappa/index.php3?info=logiciels

Outline

MULTI-LABEL CLASSIFICATION

- Hot topic in machine learning
- Possibility of generalize standard methods and evaluation procedures
- Challenging real word problems

MULTI-LABEL CLASSIFICATION

Pedro Larrañaga, Concha Bielza

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Machine Learning
Master Data Science + Master HMDA