电子科技大学
University of Electronic Science and Technology of China

Correlation-Based Methods in Multi-label Learning

Peiyan Li

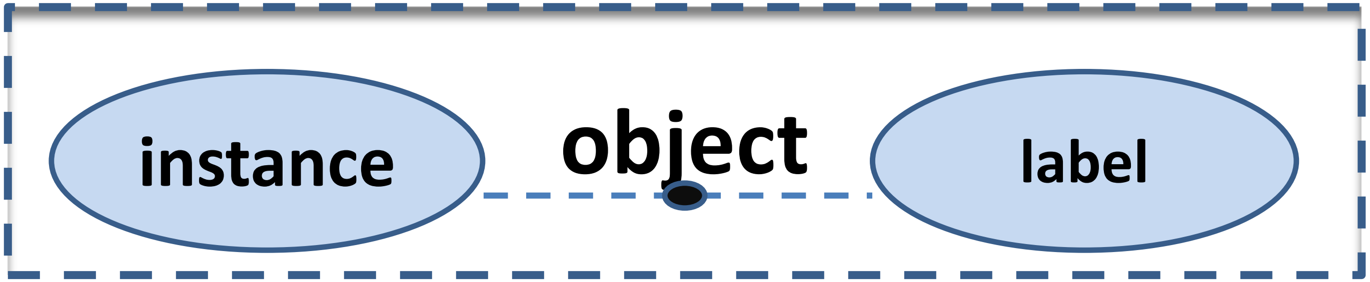
Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Outline

- What is MLL?
  - A brief
  - Formal Definition
- Challenge & Philosophy
- Order of Correlations
  - Three Levels
  - Calibrated Label Ranking
  - Random k-Labelsets
- Other Problem Transformation Style
  - Shared Subspace(Common Subspace)
  - Low-Rank Label Correlations
  - Local Label Correlations
- Summary

What is MLL???

Traditional Supervised Learning



- Input space: represented by a single instance (feature vector) characterizing its properties

- Output space: associated with a single label characterizing its semantics

Basic assumption

real-world objects are unambiguous

What is MLL???

数据挖掘实验室
Data Mining Lab

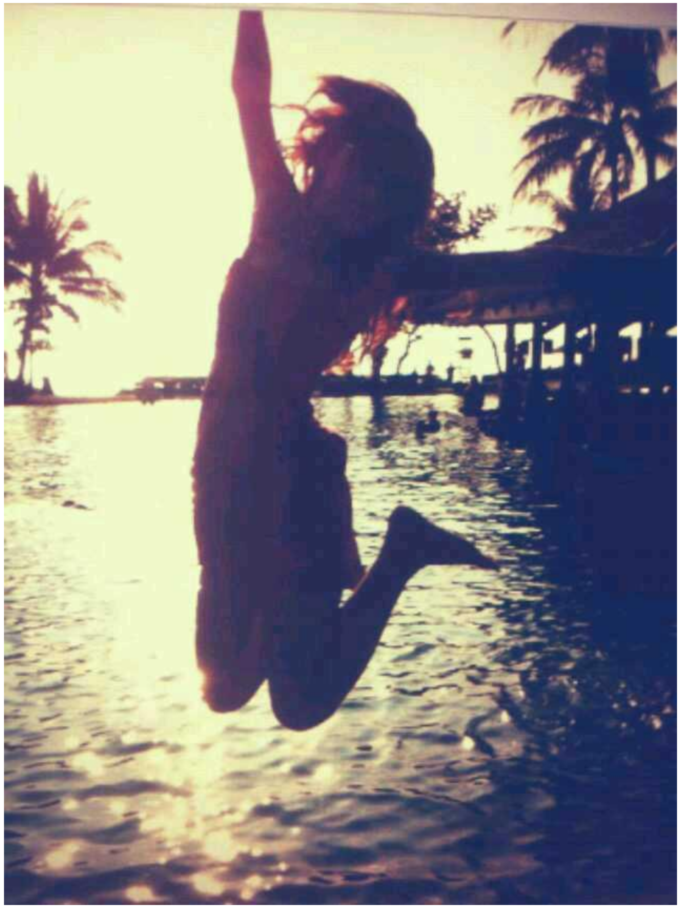
Multi-Label Objects

Clover
Adidas
Lucky
……

What is MLL???

数据挖掘实验室
Data Mining Lab

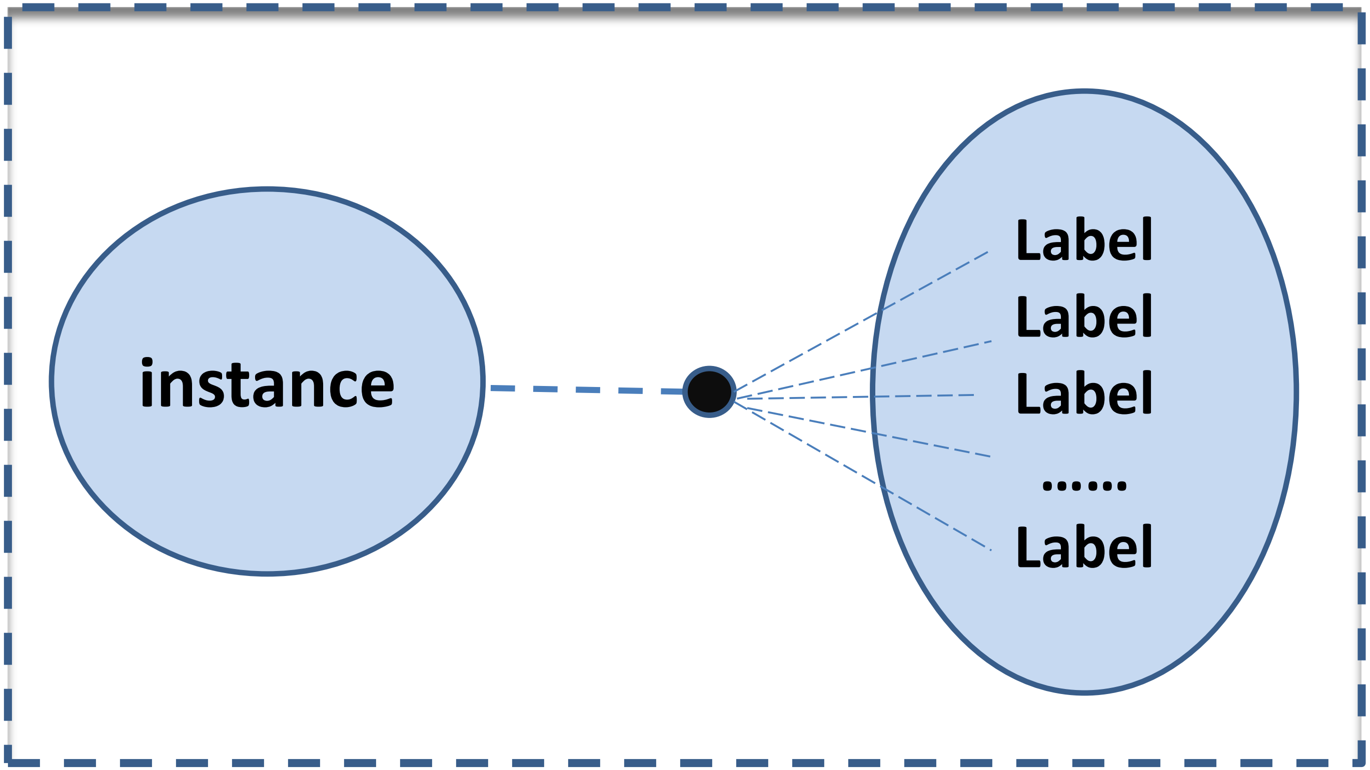
Multi-Label Objects - More



Jump
Water pool
Excited
……

Multi-label objects are ubiquitous !

What is MLL???



Multi-Label Learning (MLL)

What is MLL???

数据挖掘实验室
Data Mining Lab

Pick one

Label 1 ✓
Label 2

Binary

Pick one

Label 1
Label 2
Label 3
Label 4 ✓
…
…
Label L

Multi-class

Pick all applicable

Label 1
Label 2 ✓
Label 3
Label 4 ✓
…
…
Label L ✓

Multi-label

What is MLL???

数据挖掘实验室
Data Mining Lab

Snapshots on MLL - Applications

- Text Categorization
- Automatic annotation for multimedia contents
  - Image, Audio, Video
- Bioinformatics
- World Wide Web
- Information Retrieval
- Directed marketing
- ......

What is MLL: Formal Definition

数据挖掘实验室
Data Mining Lab

Settings

$\mathcal{X}$: d-dimensional feature space $\mathbb{R}^d$

$\mathcal{Y}$: label space with q labels $\{1,2,\cdots,q\}$

Inputs

$\mathcal{D}$: training set with m examples $\{(x_i, Y_i) | 1 \le i \le m\}$

$x_i \in \mathcal{X}$ is a d-dimensional feature vector $(x_{i1}, x_{i2}, \cdots, x_{id})^{\mathrm{T}}$

$Y_i \in \mathcal{Y}$ is the label set associated with x_i

Outputs

h : multi-label predictor $\mathcal{X} \to 2^{\mathcal{Y}}$

What is MLL: Formal Definition

数据挖掘实验室
Data Mining Lab

Alternative Outputs

$\boldsymbol{f}$: a ranking function $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

Here, $f(x, y)$ returns the "confidence" of labeling x with y

Given a threshold function $t: \mathcal{X} \to \mathbb{R}$, we have

$$h(x) = \{y | f(x, y) > t(x)\}$$

Here, $t(x)$ produces a bipartition of label space $\mathcal{Y}$ into relevant label set and irrelevant label set

Caveat here: MLL Label ≠ Ranking

The Major Challenge

input space

output space

features → label sets

The MLL Mapping

Exponential number of possible label sets !

q=5 → 32 label sets
q=10 → ~1k label sets
q=20 → ~1M label sets
……

How can we take on this challenge?

Exploiting Label Correlations

For instance:

An image labeled as lions and grassland would be likely annotated with label Africa.

A document labeled as politics would be unlikely labeled as entertainment.

A person labeled as ZhongZi would be an old driver.

Order of Correlations

First-Order Strategy

Tackle MLL problem in a label-by-label style, ignore the co-existence of other labels.

e.g.: decomposing MLL into q number of independent binary classification problems (BR, Binary Relevance)

Pros:

conceptually simple, efficient and easy to implement

Cons:

label correlations totally ignored, less effective

Order of Correlations

Second-Order Strategy

Tackle MLL problem by considering pairwise relations between labels.

e.g.: ranking between relevant and irrelevant labels, interaction between a pair of labels, etc. (Calibrated Label Ranking)

Pros:

correlations exploited, relatively effective

Cons:

correlations may go beyond second-order

Order of Correlations

数据挖掘实验室
Data Mining Lab

High-Order Strategy

Tackle MLL problem by considering high-order relations between labels.

e.g.: among all the possible labels, among a subset of labels, etc.

Pros:

more appropriate for realistic correlations

Cons:

high model complexity, less scalable

Calibrated Label Ranking

数据挖掘实验室
Data Mining Lab

Basic Idea

Transform MLL into a label ranking problem by pairwise comparison

Ranking by Pairwise Comparison

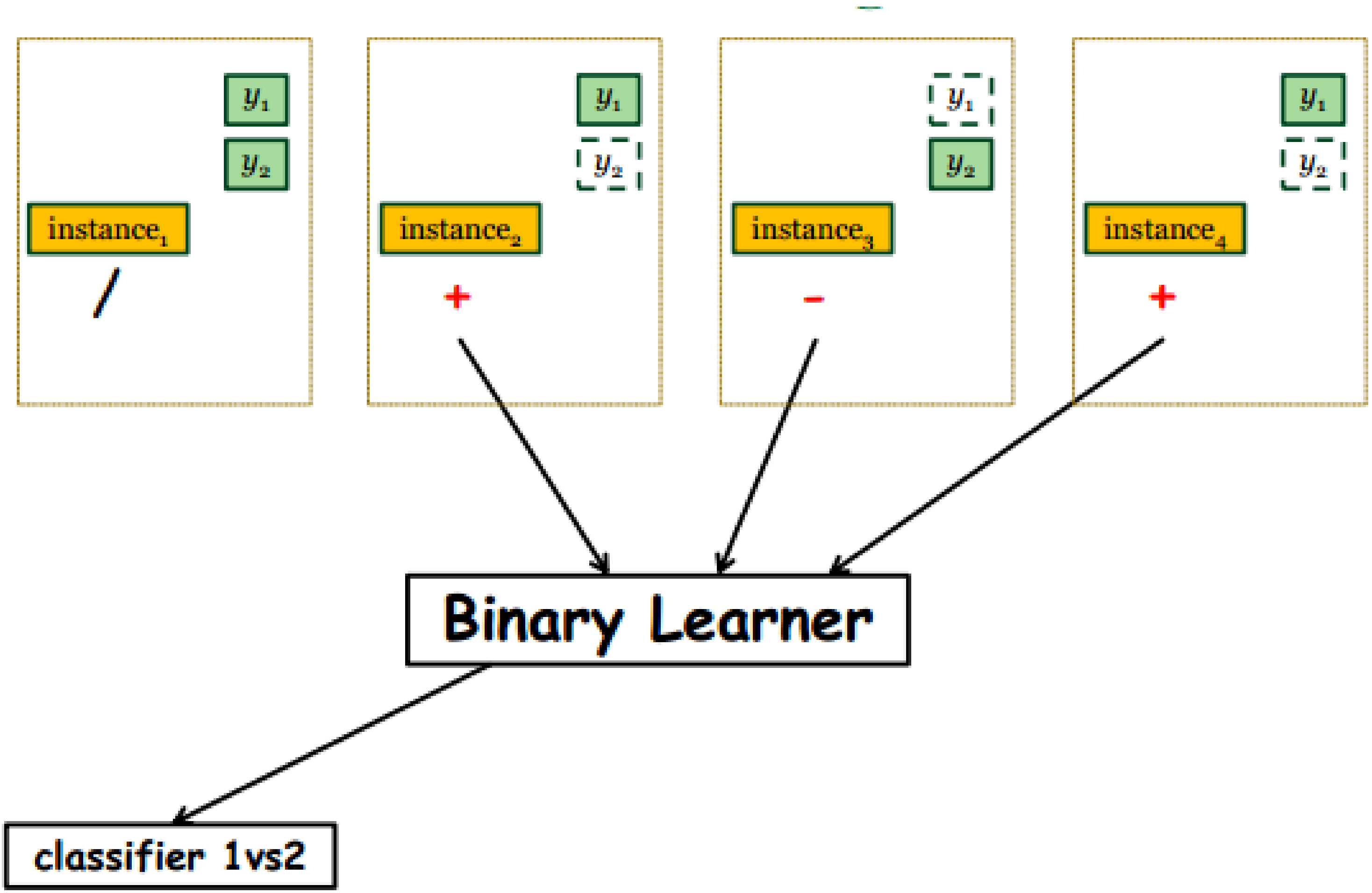
Learn $q(q-1)/2$ binary models, one for each label pair (y_j, y_k)， $1 \le j < k \le q$

Training set for binary model (y_j, y_k)

- x_i used as positive example if $y_j \in Y_i$ and $y_k \notin Y_i$
- x_i used as negative example if $y_j \notin Y_i$ and $y_k \in Y_i$
- Otherwise, x_i is ignored

[Fürnkranz et al. MLJ08]

Calibrated Label Ranking – Cont.



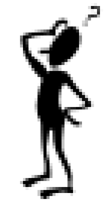
Made by MingLing Zhang

Calibrated Label Ranking – Cont.

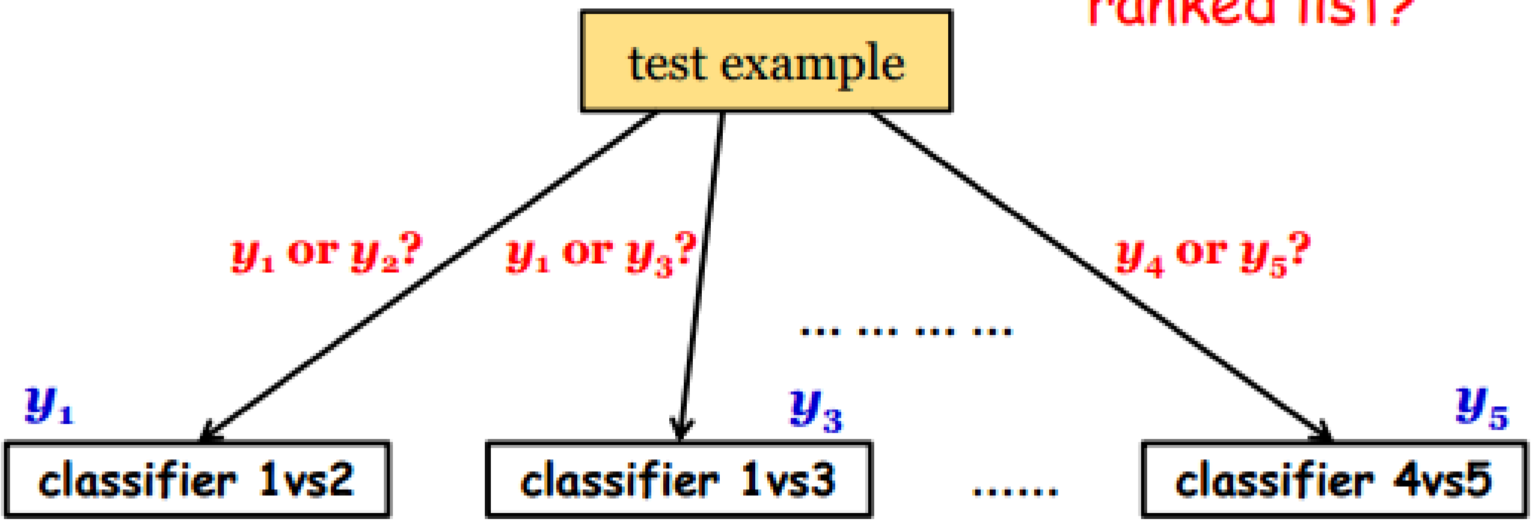
labels	votes
y_1	2
y_2	4
y_3	1
y_4	0
y_5	3

Ranking

$$y_2 \succ y_5 \succ y_1 \succ y_3 \succ y_4$$

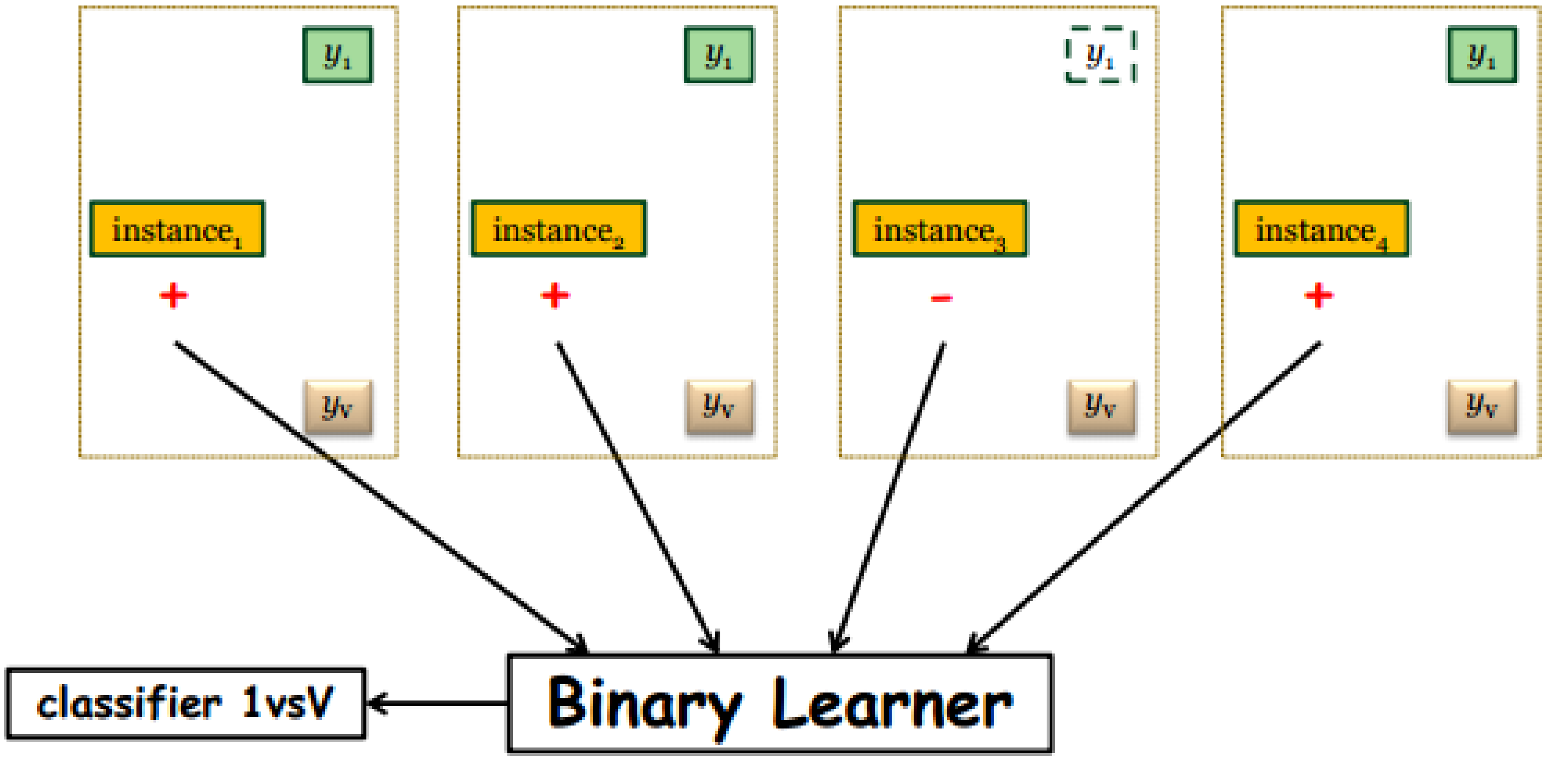


But, where should we bi-partition the ranked list?



Calibrated Label Ranking – Cont.

Add a virtual label y_v to each of the training examples, which serves as an artificial splitting point between relevant and irrelevant labels



Calibrated Label Ranking - Cont.

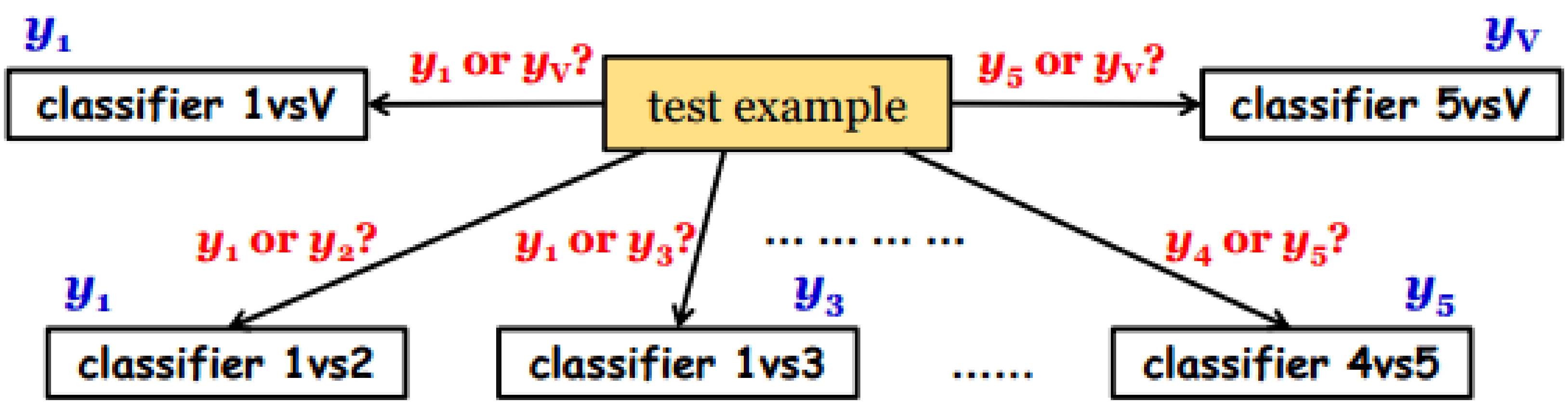
数据挖掘实验室
Data Mining Lab

labels	votes
y_1	2
y_2	5
y_3	1
y_4	0
y_5	4
y_V	3

Calibrated Ranking

bi-partition point

$$y_2 \succ y_5 \succ y_V \succ y_1 \succ y_3 \succ y_4$$



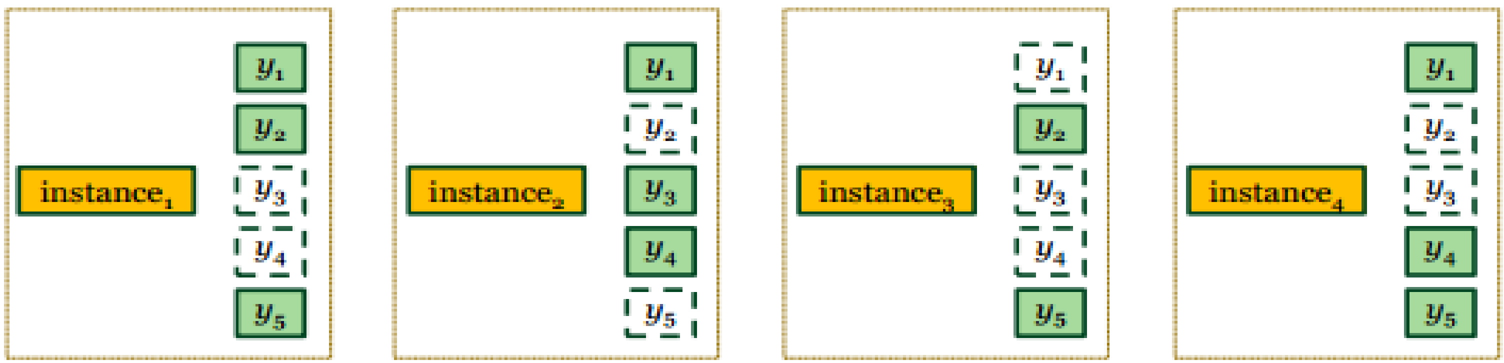
Random k-Labelsets

Basic Idea

transform MLL into an ensemble of single-label multi-class problems

Label Powerset(LP)

Treat each label set appearing in training set as a new class



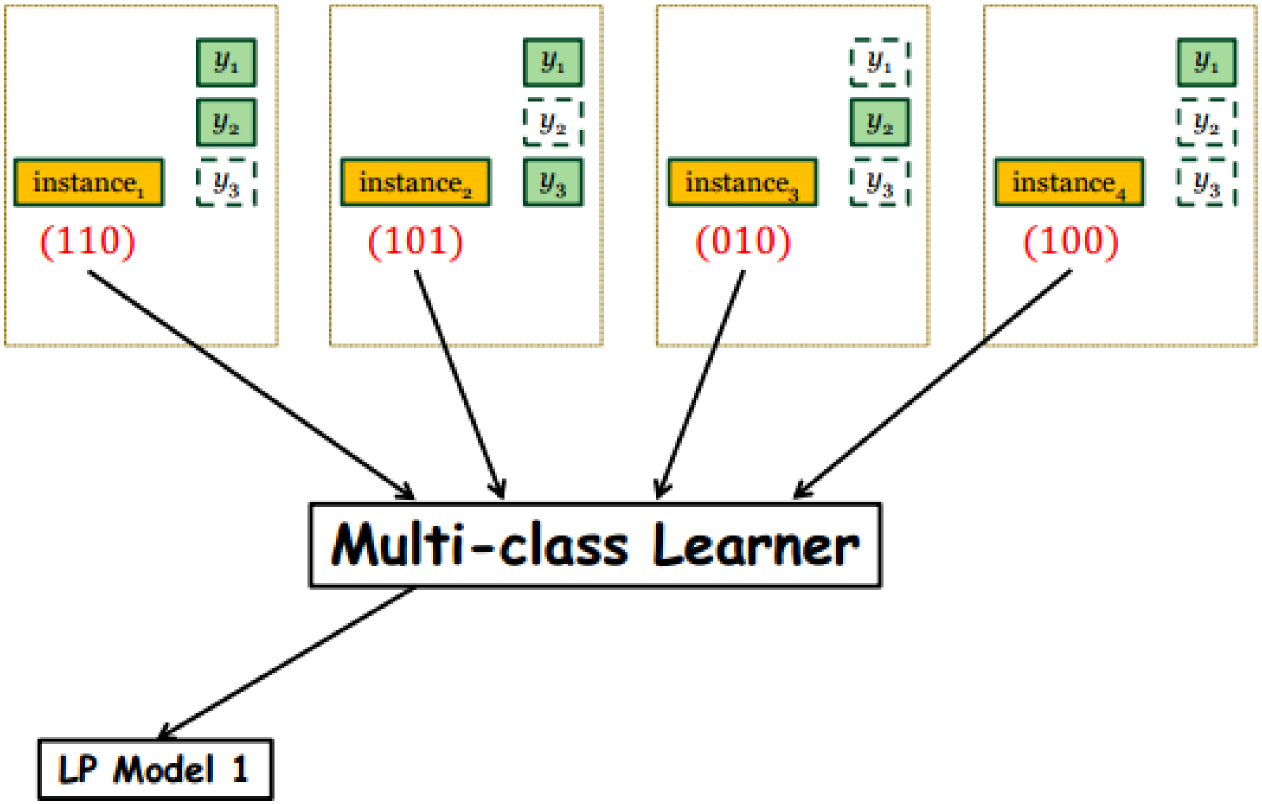
new class 1:	new class 2:	new class 3:	new class 4:
(11001)	(10110)	(01001)	(10011)

[Tsoumakas & Vlahavas, TKDE 11]

Random k-Labelsets - Cont.

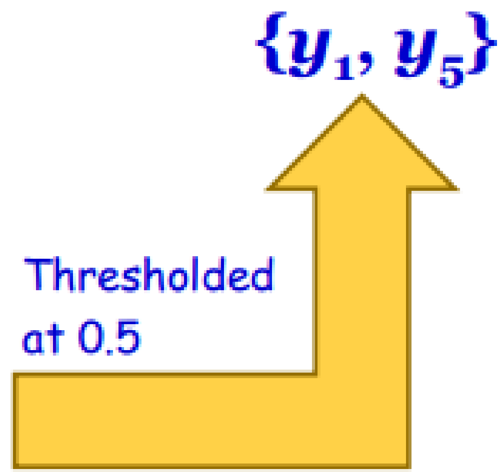
K-Labelsets

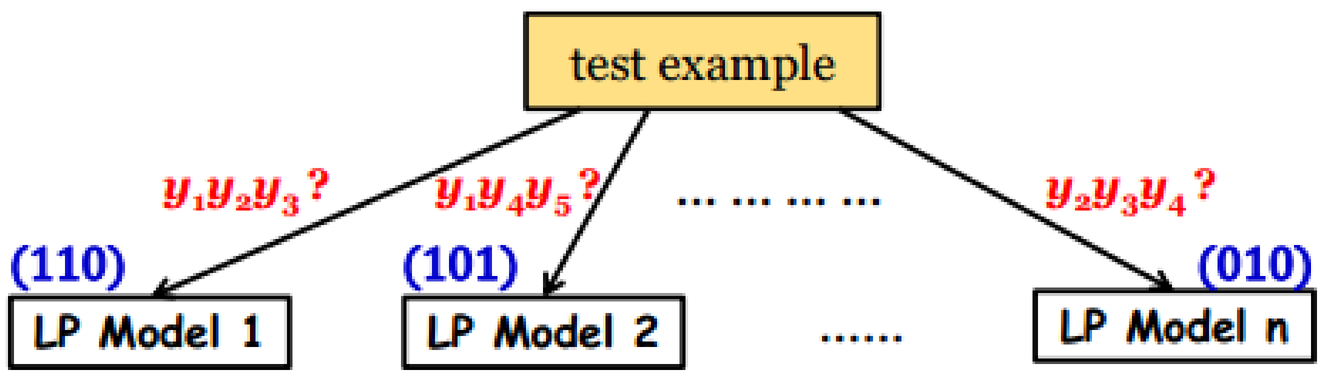
Randomly pick a subset of k labels (e.g. k=3), and invoke the LP method
Build an ensemble of LP models, and predict by voting and thresholding



Random k-Labelsets - Cont.

LP Model	k-labelsets	Prediction				
		y_1	y_2	y_3	y_4	y_5
h_1	$\{y_1, y_2, y_3\}$	1	1	0	-	-
h_2	$\{y_1, y_4, y_5\}$	1	-	-	0	1
h_3	$\{y_2, y_4, y_5\}$	-	0	-	1	1
h_4	$\{y_2, y_3, y_4\}$	-	0	0	0	-
averaged voting		2/2	1/3	0/2	1/3	2/2





Other Problem Transformation Style

数据挖掘实验室
Data Mining Lab

Traditional:

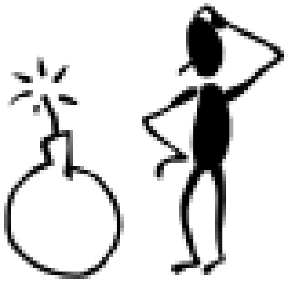
- First Order: Single label learning (Binary relevance)
- Second Order: Pairwise methods (Calibrated Label Ranking)
- Third Oder: Combine labels (Random K-Label sets)

Mention:

- Those methods are based on the number of labels we chosen (Output space).
- Algorithm adaption methods like Rank-SVM, Multi-label C4.5, BP-MLL and ML-KNN will not be mentioned here.

Other Problem Transformation Style

Multi-label Learning with feature selection or dimension reduction ?



Multi-label Learning via subspace methods !

Shared Subspace(Common Subspace)

- Extracting Shared Subspace for Multi-label Classification [Ji S, Tang L, Yu S, KDD08]

Basic Idea

A common subspace is assumed to be shared among multiple labels.

Shared Subspace(Common Subspace)

数据挖掘实验室
Data Mining Lab

The predictive function: $f_l(x) = w_l^T x + v_l^T \Theta x$

- one part is contributed from the original space
- the other part is derived from the shared subspace
- $\Theta\Theta^T = I$

Shared Subspace(Common Subspace)

数据挖掘实验室
Data Mining Lab

$$\sum_{l=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} L\left(\left(w_l + \Theta^T v_l\right)^T x_i, y_i^l\right) + \alpha\|w_l\|^2 + \beta\left\|w_l + \Theta^T v_l\right\|^2\right)$$

$$\min_{U,V,\Theta}\frac{1}{n}\|XU - Y\|_F^2 + \alpha\left\|U - \Theta^T V\right\|_F^2 + \beta\|U\|_F^2$$

$$s.t. \quad \Theta\Theta^T = I$$

Low-Rank Label Correlations

数据挖掘实验室 Data Mining Lab

- Learning Low-Rank Label Correlations for Multi-label Classification with Missing Labels [Xu L, Wang Z, Shen Z. ICDM 14]

Basic Idea

The multiple labels are usually correlated in some semantic space while sharing the same input space.

Low-Rank Label Correlations

	fish	ocean	sky	grass
Image1	1	0	0	0
Image2	0	0	0	1

×

	fish	ocean	sky	grass
fish	1	0.8	0.2	0.3
ocean	0.6	1	0.5	0.3
sky	0.3	0.3	1	0.3
grass	0.2	0.2	0.7	1

=

	fish	ocean	sky	grass
Image1	1	0.8	0.2	0.3
Image2	0.2	0.2	0.7	1

$$Y \times S = \hat{Y}$$

$$\min_{w,s,e} \|XW - YS\|_F^2 + \lambda_1 \|W\|_F^2 + \lambda_2 \|S\|_* + \lambda_3 \|E\|_{2,1}$$
$$s.t. \quad Y = YS + E$$

Local Label Correlations

数据挖掘实验室
Data Mining Lab

- Multi-Label Learning by Exploiting Label Correlations Locally [Huang S J, Zhou Z H, Zhou Z H. AAAI 12]

Basic Idea

Instances can be separated into different groups and each group share a subset of label correlations.

Instances with similar label vectors usually share the same correlations.

Local Label Correlations

数据挖掘实验室
Data Mining Lab

$$\min_{W,C,P} \sum_{i=1}^{n}\sum_{l=1}^{L}\xi_{il} + \lambda_1 \sum_{l=1}^{L}||\mathbf{w}_l||^2 + \lambda_2 \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}||\mathbf{y}_i - \mathbf{p}_j||^2 \tag{6}$$

$$\begin{aligned} s.t. \quad & y_{il}\langle \mathbf{w}_l, [\phi(\mathbf{x}_i), \mathbf{c}_i]\rangle \geq 1 - \xi_{il} \\ & \xi_{il} \geq 0 \quad \forall i \in \{1, \cdots, n\},\ l \in \{1, \cdots, L\} \\ & \sum\nolimits_{j=1}^{m} c_{ij} = 1 \quad \forall i \in \{1, \cdots, n\} \\ & 0 \leq c_{ij} \leq 1 \quad \forall i \in \{1, \cdots, n\},\ j \in \{1, \cdots, m\} \end{aligned}$$

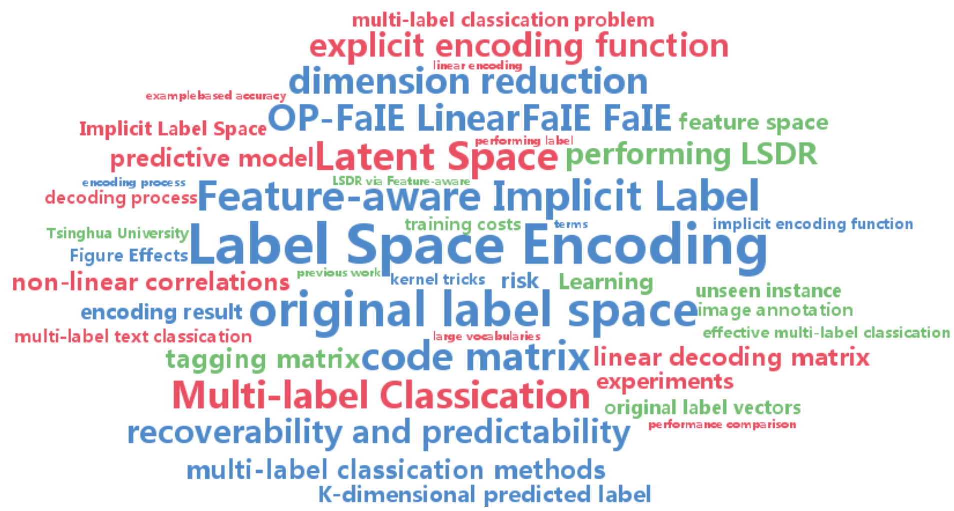
c_{ij} measures the probability that S_j is helpful to x_i, thus, it is constrained to be in the interval $[0,\ 1]$, and the sum of each $\mathbf{c}_i$ is constrained to be 1.

Summary

- A lot of work has been done on the label space and the transformation space. Why ??

- What can we do with the input space ?

Thanks



Peiyan Li

11/1/2017