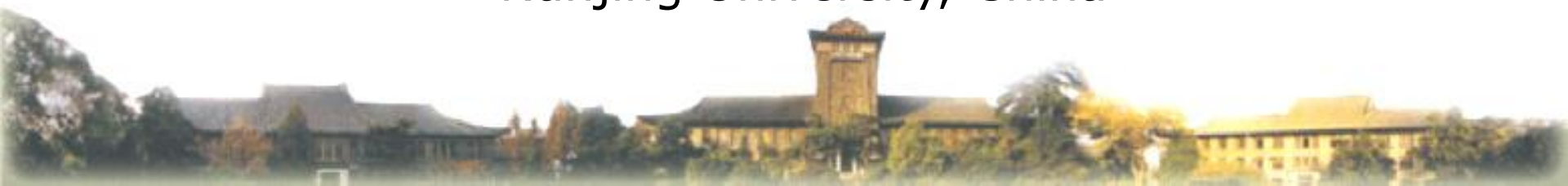
A Learning Framework for Data Objects with Complex Semantics

Zhi-Hua Zhou

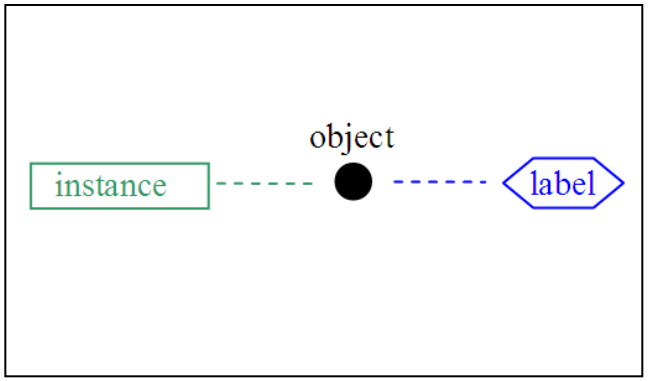
http://cs.nju.edu.cn/zhouzh/
Email: zhouzh@nju.edu.cn

LAMDA Group
National Key Laboratory for Novel Software Technology
Nanjing University, China



Traditional learning setting

- A real-world object is represented by an instance (feature vector)
- The instance is associated with a label which indicates the concerned characteristics (such as categorization) of the object



$\mathcal{X}$ - the instance space
$\mathcal{Y}$ - the set of class labels

The task:

To learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a given data set $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$ where $\boldsymbol{x}_i \in \mathcal{X}$ is an instance and $y_i \in \mathcal{Y}$ is the known label of $\boldsymbol{x}_i$

Outline

- MIML: A New Learning Framework
  - The framework
  - Why MIML?
- Learning Algorithms
- A Real Application
  - The problem
  - Solution and results

A question

What is the most difficult task currently ?

Semantics-related tasks

"Semantic gap"

The gap between low-level features and high-level semantics

This is the "source of difficulty" of many difficult tasks

Why there is the "semantic gap"?

— data objects with complex semantics

Data objects with complex semantics

Elephant ? *Lion ?* *Grassland?*
Tropic ? *Africa ?* *… …*

Data objects with complex semantics

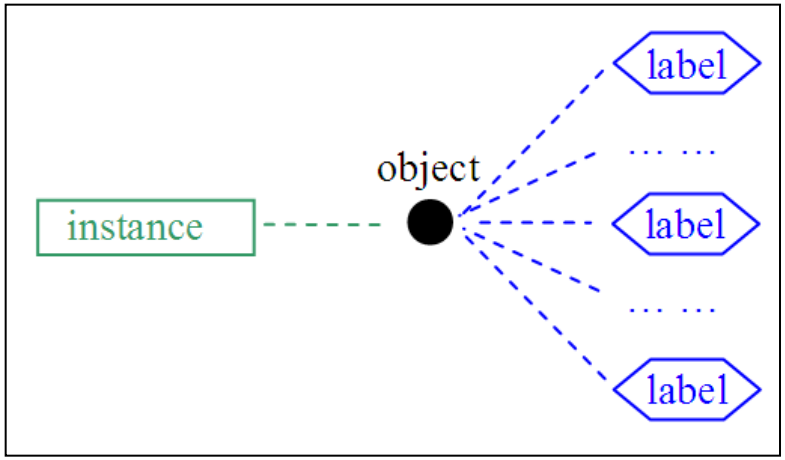
Scientific novel

Jules Verne's writing

Book on traveling

… …

Multi-label learning



MLL task:

To learn a function $f_{MLL} : \mathcal{X} \to 2^{\mathcal{Y}}$ from a given data set $\{(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \cdots, (\boldsymbol{x}_m, Y_m)\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is an instance and $Y_i \subseteq \mathcal{Y}$ is a set of labels $\{y_1^{(i)}, y_2^{(i)}, \cdots, y_{l_i}^{(i)}\}$, $y_k^{(i)} \in \mathcal{Y}$ $(k = 1, 2, \cdots, l_i)$.

$\mathcal{X}$ - the instance space

$\mathcal{Y}$ - the set of class labels

l_i - the number of labels in Y_i

Multi-label learning algorithms

- Decomposing the task into multiple binary classification problems each for a class
  - ✓ MLSVM [Boutell et al., PR 2004]
  - ✓ . . . . . .
- Considering the ranking among labels
  - ✓ BoosTexter [Schapire & Singer, MLJ 2000]
  - ✓ BP-MLL [Zhang & Zhou, TKDE 2006]
  - ✓ RankSVM [Elisseeff & Weston, NIPS'01]
  - ✓ . . . . . .
- Exploring the class correlation
  - ✓ Probabilistic generative models [McCallum, AAAI'99w; Ueda & Saito, NIPS'02]
  - ✓ Maximum entropy methods [Ghamrawi & McCallum, CIKM'05; Zhu et al., SIGIR'05]
  - ✓ . . . . . .

The problem

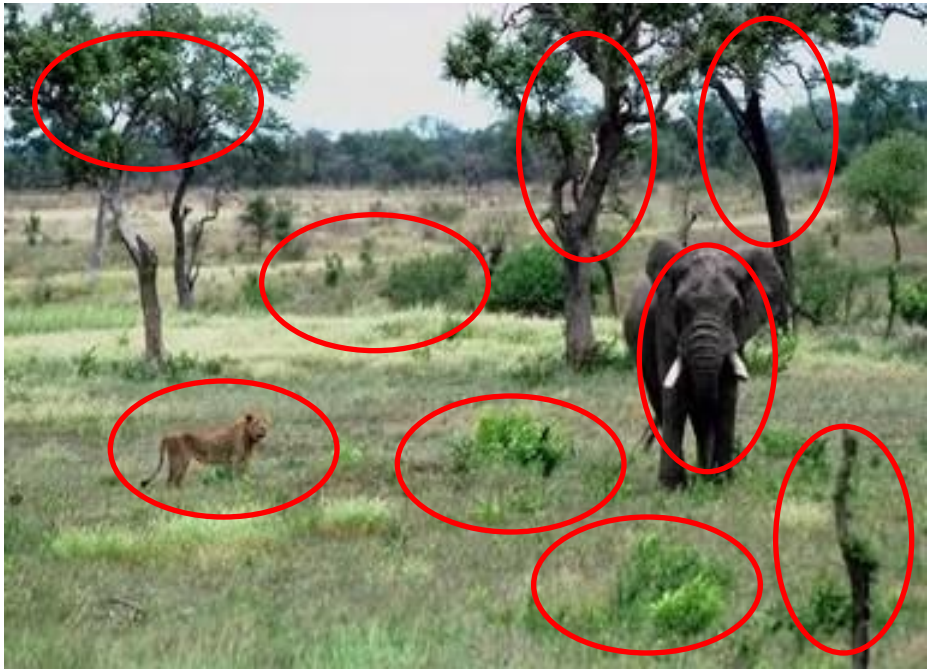
$[x_1, x_2, \ldots, x_d]^{\mathrm{T}}$

one-to-many mapping

- *Elephant*
- *Lion*
- *Grassland*
- *Tropic*
- *Africa*

Consider …

An image usually contains multiple regions each can be represented by an instance



The image can simultaneously belong to multiple classes

Elephant

Lion

Grassland

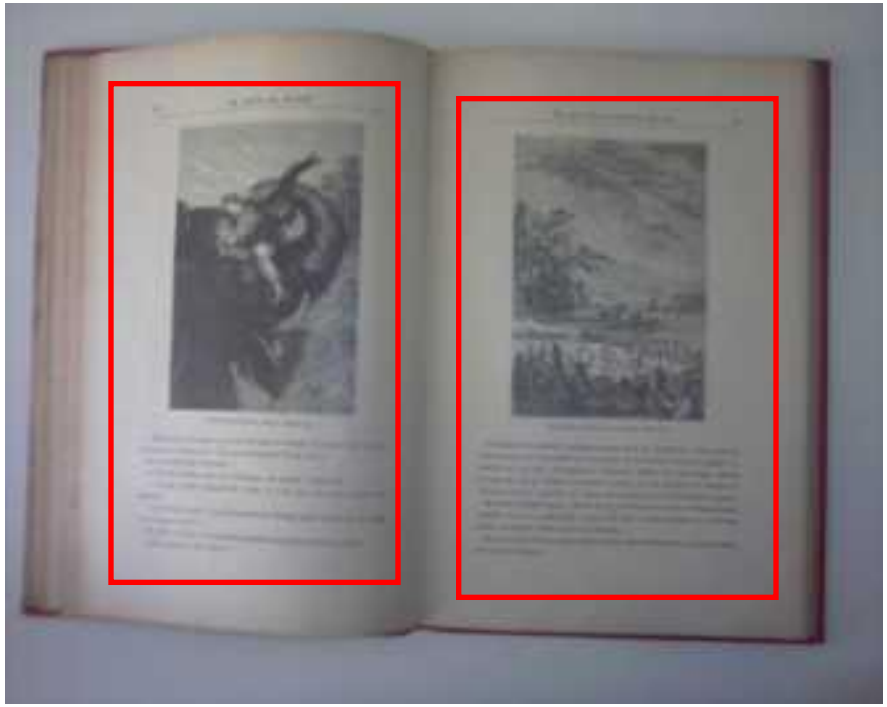
Tropic

Africa

… …

Consider …

A document usually contains multiple sections each can be represented by an instance



The document can simultaneously belong to multiple categories

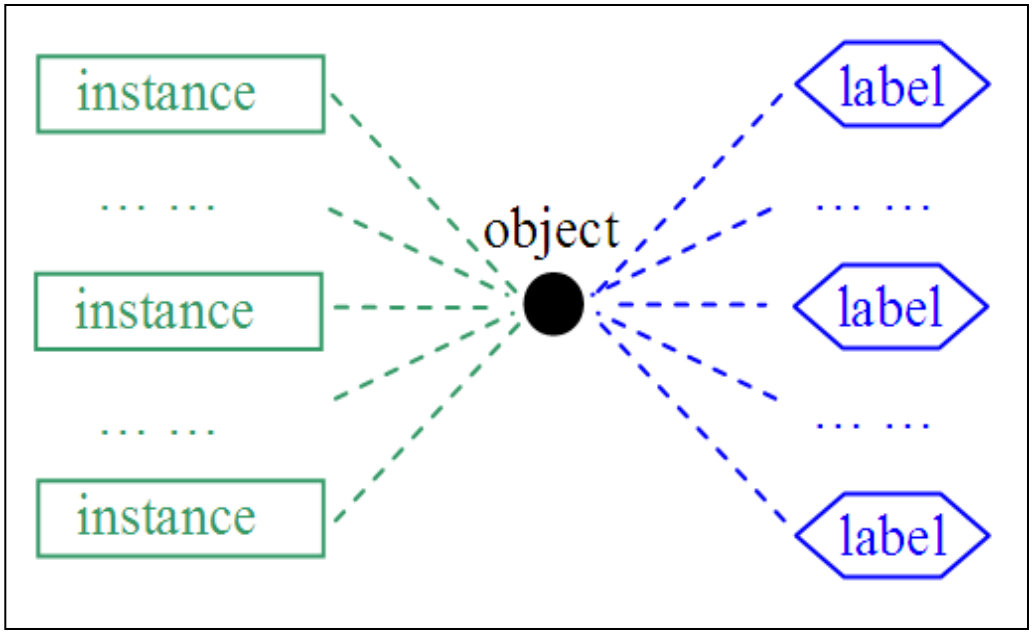
Scientific novel

Jules Verne's writing

Book on traveling

… …

MIML



Multi-Instance Multi-Label (MIML) Learning

[Zhou & Zhang, NIPS'06]

Outline

- MIML: A New Learning Framework
  - The framework
  - **Why MIML?**
- Learning Algorithms
- A Real Application
  - The problem
  - Solution and results

Why MIML?

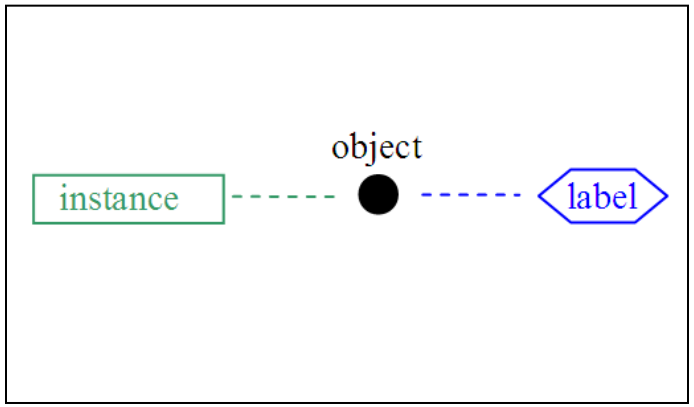
Adequate representation is important

Having an adequate representation is as important as having a strong learning algorithm

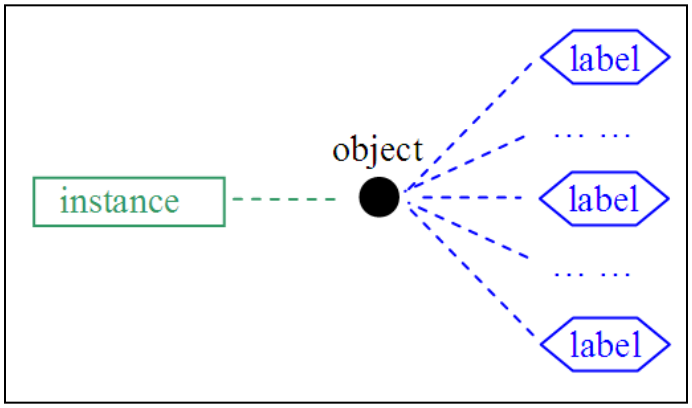
MIML captures more information of ambiguous data

Traditional supervised learning, multi-instance learning and multi-label learning are degenerated versions of MIML

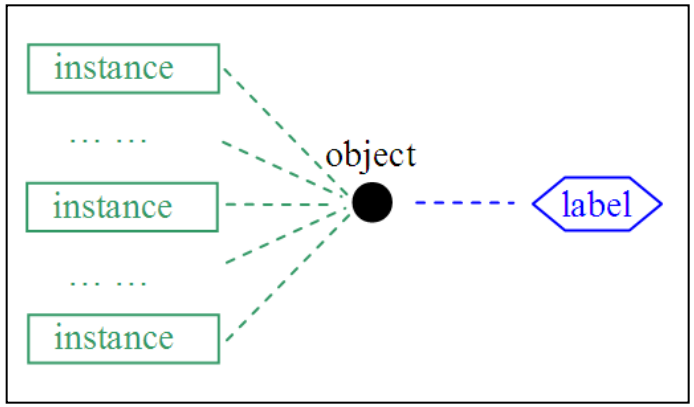
Why MIML? (cont')



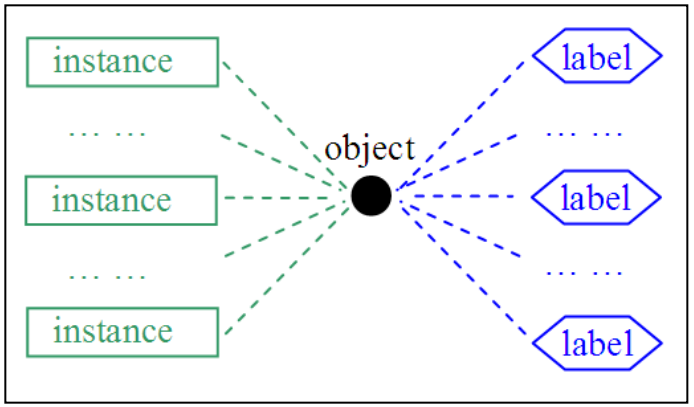
Traditional supervised learning



Multi-label learning



Multi-instance learning



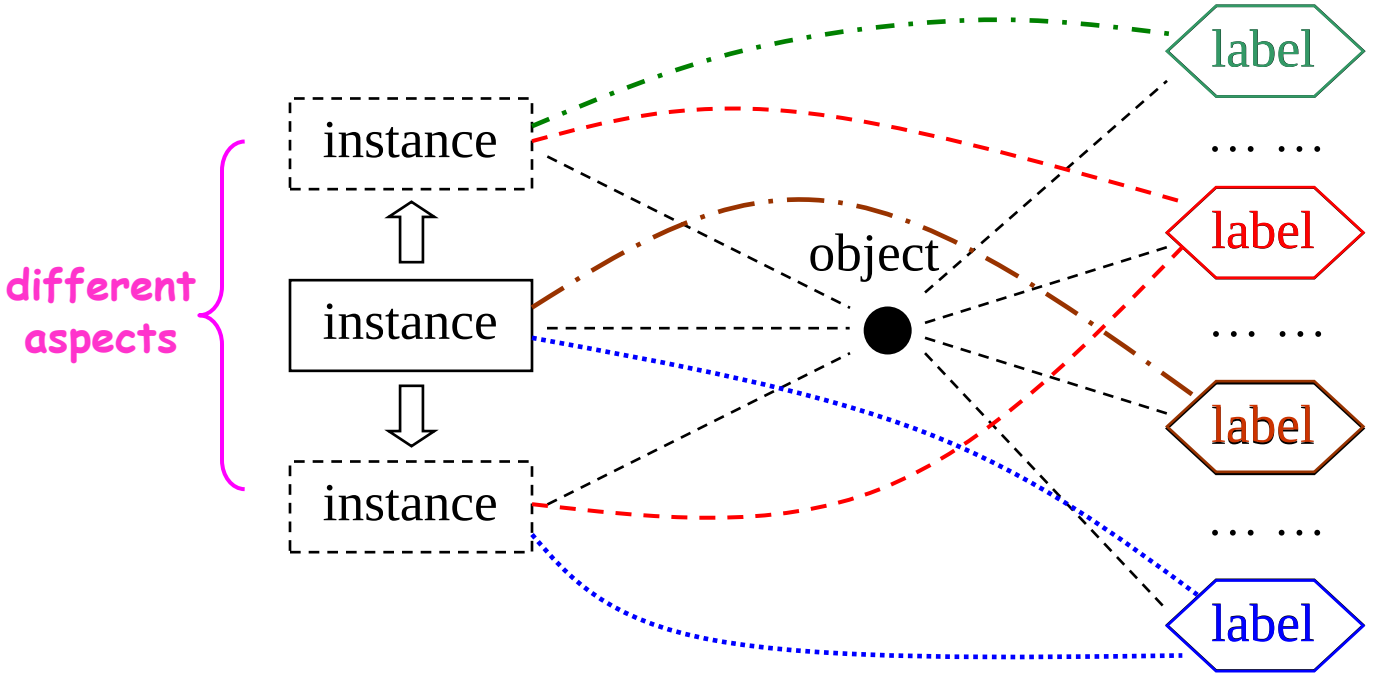
MIML

[Zhou & Zhang, NIPS'06]

Why MIML? (cont')

To learn an **one-to-many** mapping is an ill-posed problem

many-to-many mapping is better; moreover, **MIML offers the possibility for understanding the relationship between input feature patterns and output semantics**



[Zhou et al., AIJ 2012]

Why MIML? (cont')

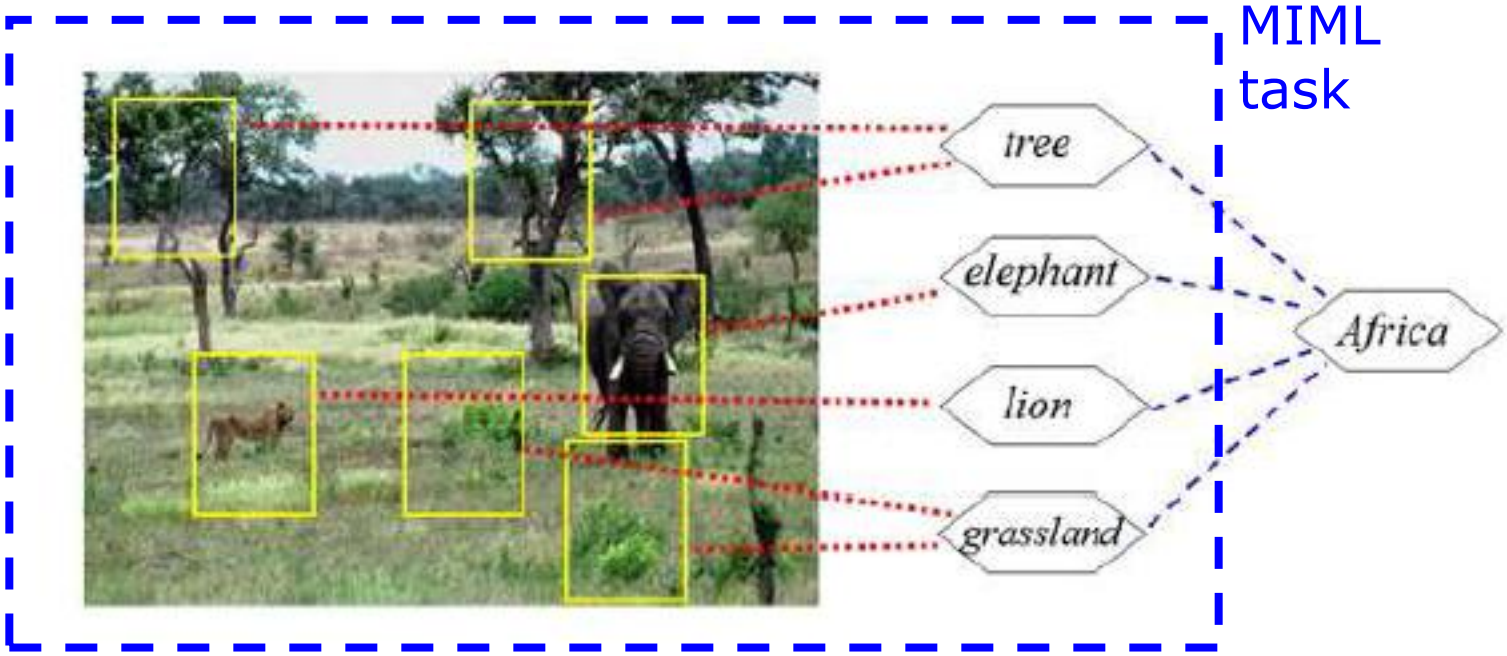
MIML can also be helpful for learning single-label examples involving complicated high-level concepts

(a) *Africa* is a complicated high-level concept

[Zhou et al., AIJ 2012]

Why MIML? (cont')

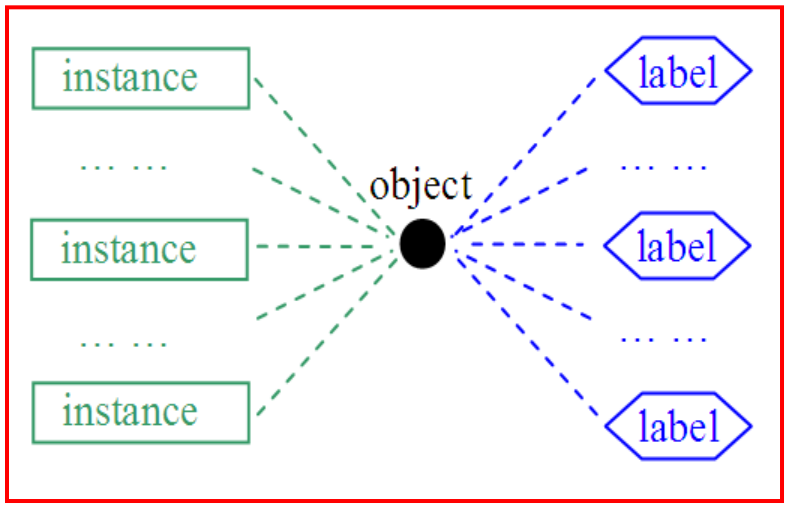
MIML can also be helpful for learning single-label examples involving complicated high-level concepts



(b) The concept *Africa* may become easier to learn through exploiting some sub-concepts

[Zhou et al., AIJ 2012]

Multi-Instance Multi-Label learning



MIML task:

To learn a function $f_{MIML} : 2^{\mathcal{X}} \rightarrow 2^{\mathcal{Y}}$ from a given data set $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_m, Y_m)\}$, where $X_i \subseteq \mathcal{X}$ is a set of instances $\{\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \cdots, \boldsymbol{x}_{n_i}^{(i)}\}$, $\boldsymbol{x}_j^{(i)} \in \mathcal{X}$ $(j = 1, 2, \cdots, n_i)$, and $Y_i \subseteq \mathcal{Y}$ is a set of labels $\{y_1^{(i)}, y_2^{(i)}, \cdots, y_{l_i}^{(i)}\}$, $y_k^{(i)} \in \mathcal{Y}$ $(k = 1, 2, \cdots, l_i)$.

MIML:

多示例多标记学习

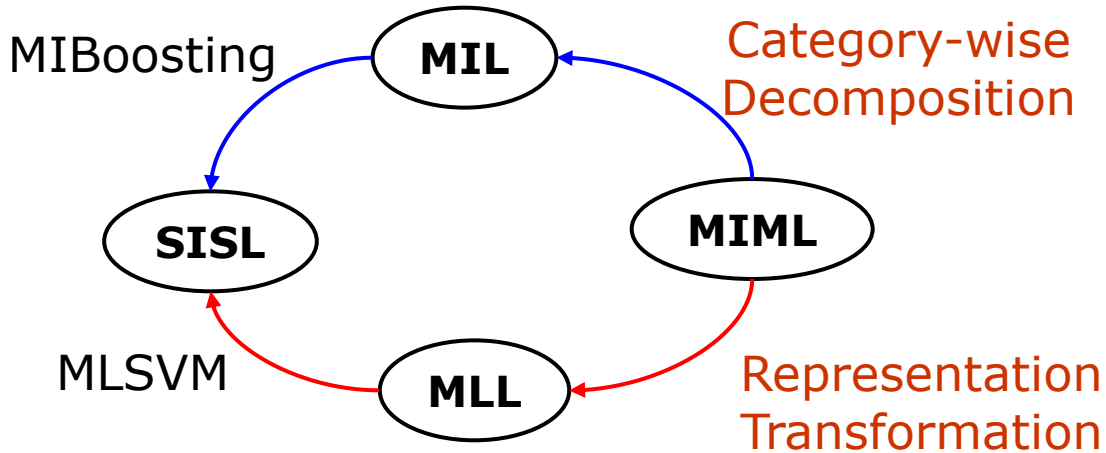
$\mathcal{X}$ - the instance space
$\mathcal{Y}$ - the set of class labels
n_i - the number of instances in X_i
l_i - the number of labels in Y_i

Outline

- MIML: A New Learning Framework
  - The framework
  - Why MIML?
- **Learning Algorithms**
- A Real Application
  - The problem
  - Solution and results

Solving MIML by degeneration

MIMLBoost (an illustration of Solution 1)



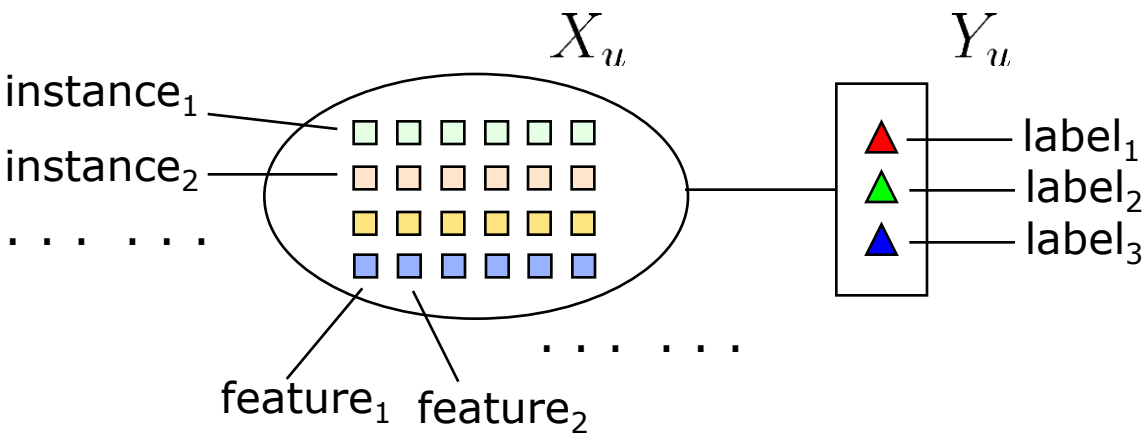
MIMLSVM (an illustration of Solution 2)

unambiguous ambiguous

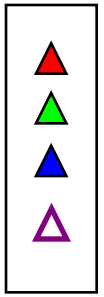
MIMLBoost

Illustration of the category-wise decomposition:

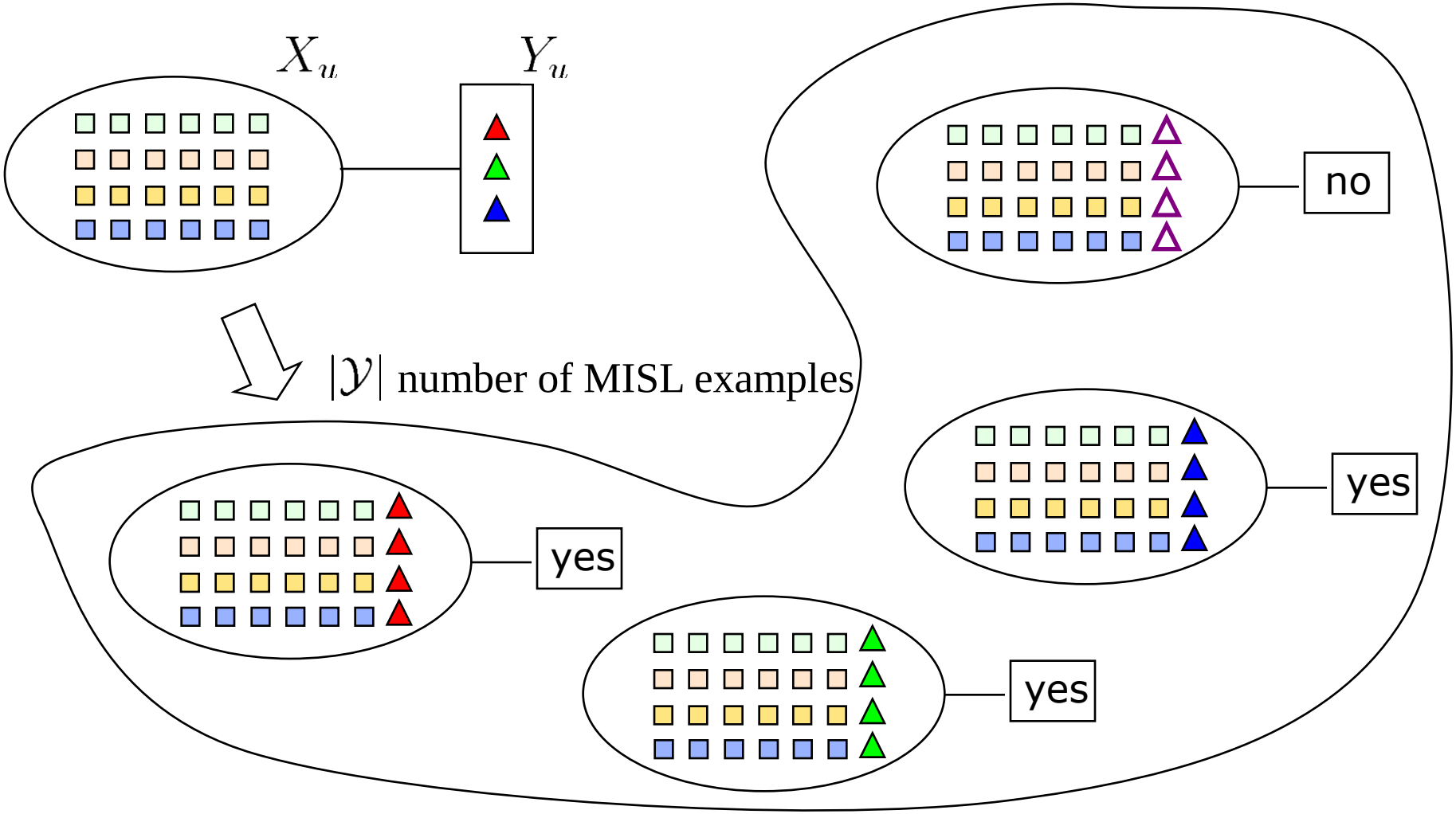
An MIML example (X_u, Y_u)



Label set $\mathcal{Y}$



MIMLBoost (con't)

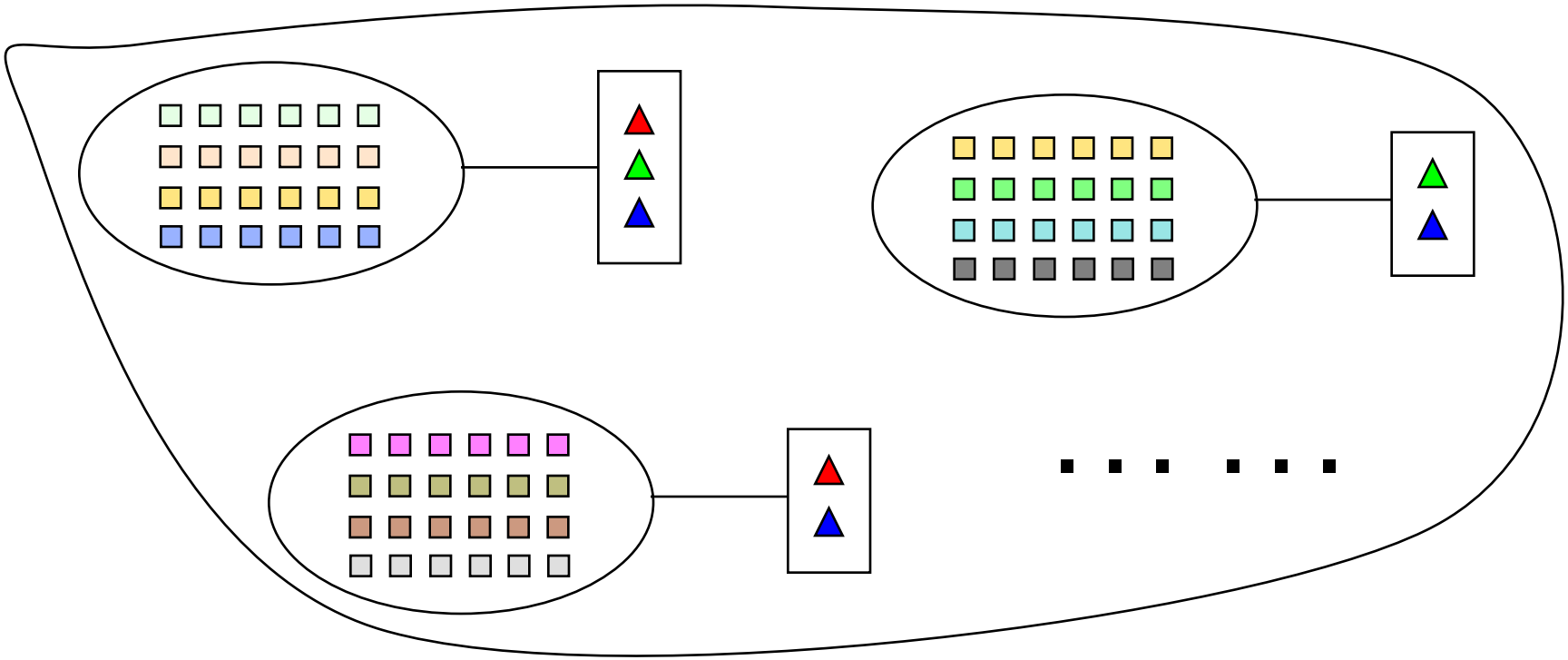


[Zhou & Zhang, NIPS'06]

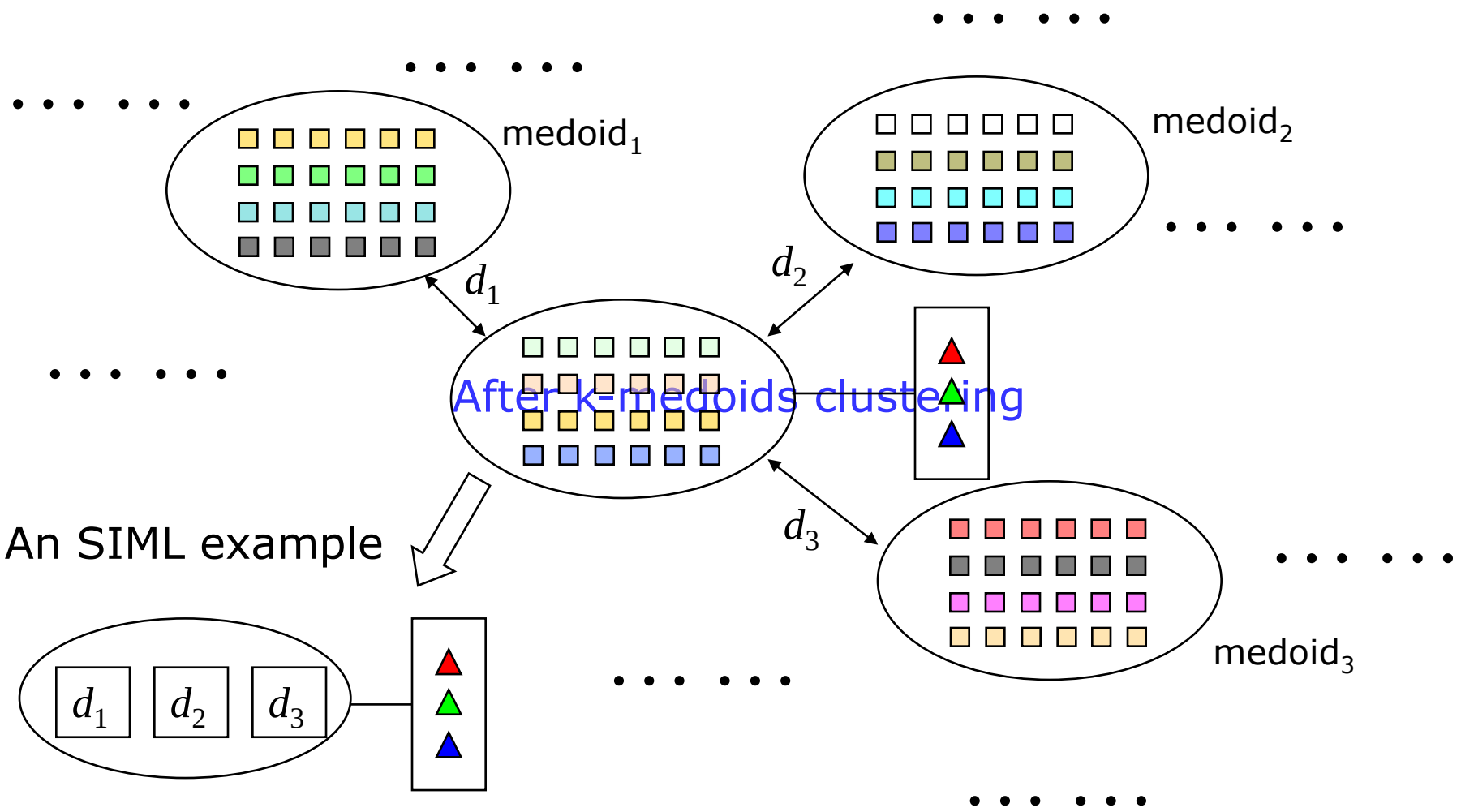
MIMLSVM

Illustration of the representation transformation:

A set of MIML examples

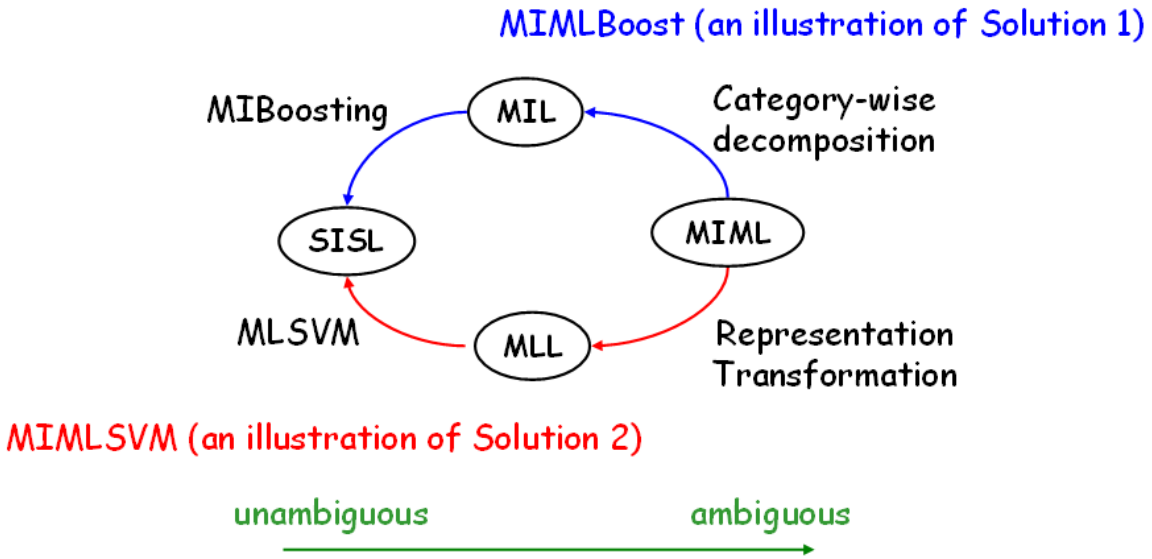


MIMLSVM (con't)



[Zhou & Zhang, NIPS'06]

Again, Why MIML?



- The MIML framework incorporates more information (+)
- These solutions degenerate MIML to solve, while the degeneration loses information (-)

If (+) > (-), then it is worth doing

Scene classification: Result

Table 3
Results (mean±std.) on scene classification ('↓' indicates 'the smaller the better'; '↑' indicates 'the larger the better')

Compared Algorithms	Evaluation Criteria				
	hloss↓	*one-error*↓	*coverage*↓	*rloss*↓	*aveprec*↑
MIMLBOOST	.192±.004	**.349±.016**	**.986±.041**	**.179±.008**	**.778±.009**
MIMLSVM	**.190±.009**	.350±.020	1.083±.050	.201±.001	.766±.013
ADTBOOST.MH	.210±.006	.436±.019	1.223±.049	N/A	.718±.012
RANKSVM	.219±.020	.400±.062	1.177±.160	.225±.041	.739±.040
ML-*k*NN	.191±.006	.370±.017	1.085±.047	.203±.010	.759±.010

The MIML algorithms are apparently superior to non-MIML algorithms

[Zhou et al, CORR abs/0808.3231]

Text categorization: Result

Table 4
Results (mean±std.) on text categorization ('↓' indicates 'the smaller the better'; '↑' indicates 'the larger the better')

Compared Algorithms	Evaluation Criteria				
	hloss↓	*one-error*↓	*coverage*↓	*rloss*↓	*aveprec*↑
MIMLBOOST	.054±.004	.092±.013	.401±.035	.037±.004	.937±.007
MIMLSVM	**.034±.003**	**.071±.009**	**.315±.029**	**.024±.003**	**.955±.006**
ADTBOOST.MH	.055±.004	.120±.016	.409±.046	N/A	.925±.010
RANKSVM	.093±.007	.205±.055	.639±.161	.078±.027	.867±.037
ML-*k*NN	.067±.005	.191±.017	.683±.052	.085±.008	.871±.010

The MIML algorithms are apparently superior to non-MIML algorithms

[Zhou et al, CORR abs/0808.3231]

Some theoretical results

Theorem 3 *Suppose the l labels are independent to each other and the conditions for the multi-instance learning algorithm in Theorem 2 are met, let* $d_{B_{\max}} = \max\{d_{B_1}, \cdots, d_{B_l}\}$, *the following bound holds for all* $\theta > 0$

$$error(f_{MM}) \leq \left(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\right)^{\frac{T}{2}} + O\left(\frac{1}{\sqrt{m}}\left(\frac{d_{B_{\max}} \log^2(m/d_{B_{\max}})}{\theta^2} + \log(1/\delta)\right)^{\frac{1}{2}}\right).$$

Roughly speaking, if the labels are independent, the difficulty of MIML is no larger than traditional multi-instance learning, and we can bound the error according to the above

Some theoretical results (con't)

Theorem 4 *Suppose the l labels satisfy above correlation assumption and the conditions for the multi-instance learning algorithm in Theorem 2 are met, let $d_{B_{\max}} = \max\{d_{B_1}, \cdots, d_{B_l}\}$, the following bound holds for all $\theta > 0$*

$$error(f_{MM}) \leq \left(\frac{4n^{1+\theta}(n-1)^{1-\theta}}{(2n-1)^2}\right)^{\frac{T}{2}} + \frac{1}{l}\sum_{r=1}^{\varrho}\left(\sum_{q\neq r}|C_q|e_{r,q}^{+} - |C_r|e_r^{-}\right)$$
$$+O\left(\frac{1}{\sqrt{m}}\left(\frac{d_{B_{\max}}\log^2(m/d_{B_{\max}})}{\theta^2} + \log(1/\delta)\right)^{\frac{1}{2}}\right).$$

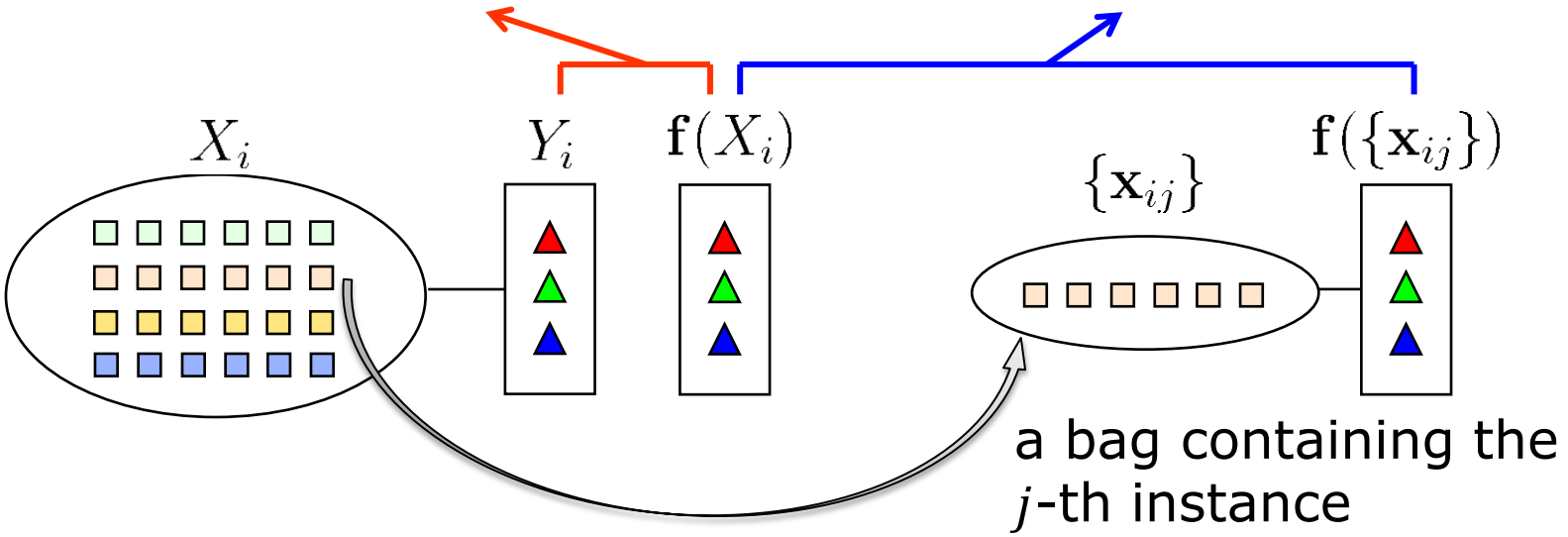
If the labels are correlated, the situation is much difficult; however, it can still be learnable. If the labels can be divided into different groups, where labels in the same group are positively correlated while labels in different groups are negatively correlated, the error can be bounded as above

D-MIMLSVM

Solving MIML directly in regularization framework

The loss function

$$V(\{X_i\}_{i=1}^m, \{Y_i\}_{i=1}^m, \mathbf{f}) = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}(1 - y_{it}f_t(X_i))_+ + \frac{\lambda}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T} l(f_t(X_i), \max_{j=1,\cdots,n_i} f_t(\mathbf{x}_{ij}))$$



[Zhou et al., AIJ 2012]

D-MIMLSVM (con't)

The labels associated with the same example should have some relatedness

Assume the $\boldsymbol{w}_t$'s come from a particular Gaussian distribution, with the mean $\boldsymbol{w}_0$:

$$f_t(\boldsymbol{x}) = \langle \boldsymbol{w}_t, \phi(\boldsymbol{x}) \rangle \qquad \boldsymbol{w}_0 = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{w}_t$$

We want to minimize $\sum_{t=1}^{T} \|\boldsymbol{w}_t\|^2$ and $\|\boldsymbol{w}_0\|^2$ simultaneously, and thus we have:

$$\min_{f \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} \|f_t\|_{\mathcal{H}}^2 + \mu \| \frac{\sum_{t=1}^{T} f_t}{T} \|_{\mathcal{H}}^2 + \gamma \cdot V\left(\{X_i\}_{i=1}^m, \{Y_i\}_{i=1}^m, \boldsymbol{f}\right)$$

[Zhou et al., AIJ 2012]

D-MIMLSVM (con't)

Assume the bags and instances are ordered as:

$$(X_1, \cdots, X_m, \boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1,n_1}, \cdots, \boldsymbol{x}_{m,1}, \cdots, \boldsymbol{x}_{m,n_m})$$

Thus each object (bags or instances) can be indexed by:

$$\begin{cases} \mathcal{I}(X_i) = i \\ \mathcal{I}(\boldsymbol{x}_{ij}) = m + \sum_{l=1}^{i-1} n_l + j \end{cases}$$

We can obtain the $(m+n) \times (m+n)$ kernel matrix $\boldsymbol{K}$ with the i-th column denoted by $\boldsymbol{k}_i$. We have:

$$f_t(X_i) = \boldsymbol{k}'_{\mathcal{I}(X_i)}\boldsymbol{\alpha}_t + b_t \qquad f_t(\boldsymbol{x}_{ij}) = \boldsymbol{k}'_{\mathcal{I}(\boldsymbol{x}_{ij})}\boldsymbol{\alpha}_t + b_t$$

[Zhou et al., AIJ 2012]

D-MIMLSVM (con't)

$$\min_{\boldsymbol{A},\boldsymbol{\xi},\boldsymbol{\delta},\boldsymbol{b}} \frac{1}{2T}\sum_{t=1}^{T}\boldsymbol{\alpha}_t'\boldsymbol{K}\boldsymbol{\alpha}_t + \frac{\mu}{T^2}\mathbf{1}'\boldsymbol{A}'\boldsymbol{K}\boldsymbol{A}\mathbf{1} + \frac{\gamma}{mT}\boldsymbol{\xi}'\mathbf{1} + \frac{\gamma\lambda}{mT}\boldsymbol{\delta}'\mathbf{1}$$

$$\text{s.t. } y_{it}(\boldsymbol{k}'_{\mathcal{I}(X_i)}\boldsymbol{\alpha}_t + b_t) \geq 1 - \xi_{it},$$

$$\boldsymbol{\xi} \geq \mathbf{0},$$

$$\boldsymbol{k}'_{\mathcal{I}(\boldsymbol{x}_{ij})}\boldsymbol{\alpha}_t - \delta_{it} \leq \boldsymbol{k}'_{\mathcal{I}(X_i)}\boldsymbol{\alpha}_t,$$

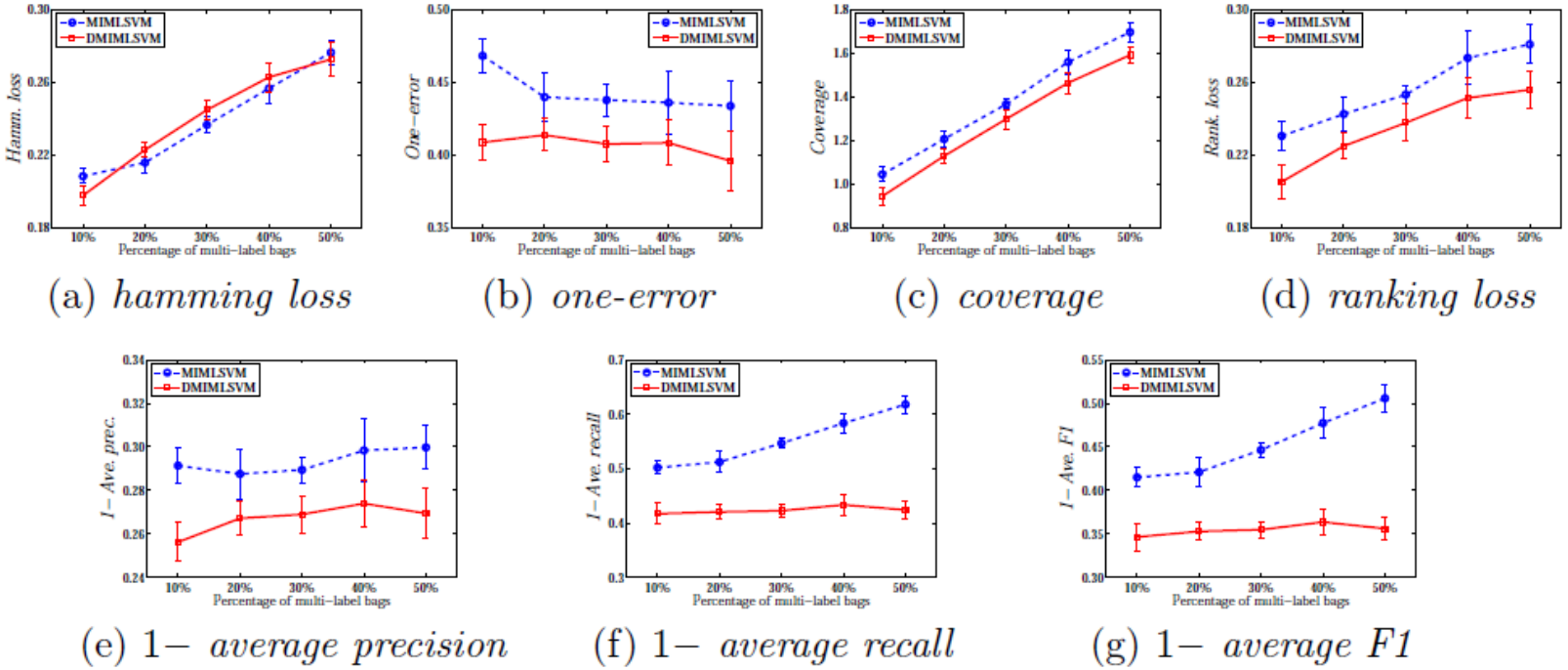
$$\boldsymbol{k}'_{\mathcal{I}(X_i)}\boldsymbol{\alpha}_t - \max_{j=1,\cdots,n_i} \boldsymbol{k}'_{\mathcal{I}(\boldsymbol{x}_{ij})}\boldsymbol{\alpha}_t \leq \delta_{it},$$

where $\boldsymbol{\xi} = [\xi_{11}, \xi_{12}, \cdots, \xi_{it}, \cdots, \xi_{mT}]'$ are slack variables for the errors on the training bags for each label, $\boldsymbol{\delta} = [\delta_{11}, \delta_{12}, \cdots, \delta_{it}, \cdots, \delta_{mT}]'$, and $\mathbf{0}$ and $\mathbf{1}$ are all-zero and all-one vector, respectively. $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_T]$ and $\boldsymbol{b} = [b_1, b_2, \cdots, b_T]'$.

This can be solved by CCCP (concave-convex procedure), and the efficiency can be further improved by adopting a cutting-plane algorithm

[Zhou et al., AIJ 2012]

D-MIMLSVM (con't)



(a) *hamming loss* (b) *one-error* (c) *coverage* (d) *ranking loss*

(e) 1− *average precision* (f) 1− *average recall* (g) 1− *average F1*

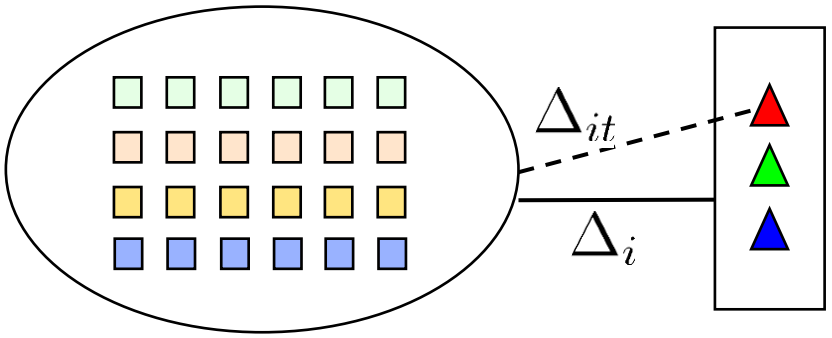
Fig. 7. Results on scene classification with different percentage of multi-label data. The lower the curve, the better the performance.

D-MIMLSVM is apparently superior to MIMLSVM

[Zhou et al, CORR abs/0808.3231]

An maximum margin MIML method

The key: How to define the margin of MIML examples



$$\Delta_{it} = \frac{y_{it} \cdot \max_{\mathbf{x} \in X_i}(\langle \mathbf{w}_t, \mathbf{x} \rangle + b_t)}{||\mathbf{w}_t||}$$

$$\Delta_i = \min_{1 \le t \le T} \Delta_{it}$$

Margin on the whole data:

$$\Delta = \min_{1 \le i \le m} \Delta_i = \min_{1 \le t \le T} \frac{1}{||\mathbf{w}_t||}$$

M3MIML (con't)

$$\max_{1\le t\le T} ||\mathbf{w}_t||^2 \le \sum_{t=1}^{T} ||\mathbf{w}_t||^2 \qquad \max_{\mathbf{x}\in X_i}(\langle \mathbf{w}_t, \mathbf{x}\rangle + b_t) \ge \frac{\sum_{j=1}^{n_i}(\langle \mathbf{w}_t, \mathbf{x}_{ij}\rangle + b_t)}{n_i}$$

Thus, the maximum margin problem can be approximated by:

$$\min_{W,\mathbf{b},\Xi,\Theta} \frac{1}{2}\sum_{t=1}^{T} ||\mathbf{w}_t||^2 + C\sum_{t=1}^{T}\left(\sum_{y_{it}=1}\xi_{it} + \sum_{y_{it}\neq 1}\sum_{j=1}^{n_i}\theta_{itj}\right)$$

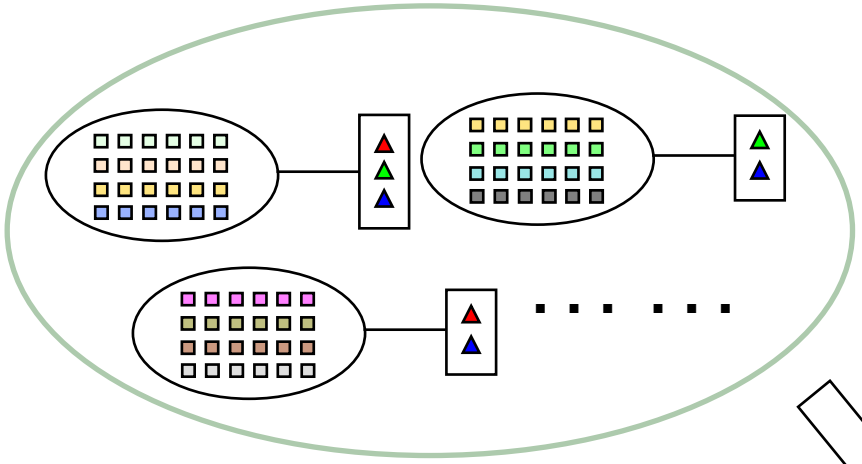
$$\text{subject to: } \forall i \in \{1,\cdots,m\}, t \in \{1,\cdots,T\}, \text{such that}$$

$$\begin{cases} \frac{\sum_{j=1}^{n_i}(\langle \mathbf{w}_t,\mathbf{x}_{ij}\rangle + b_t)}{n_i} \ge 1-\xi_{it} & \text{if } y_{it}=1 \\ -\langle \mathbf{w}_t, \mathbf{x}_{ij}\rangle - b_t \ge 1-\theta_{itj}(1\le j\le n_i) & \text{if} y_{it}\neq 1 \end{cases}$$

$$\xi_{it}\ge 0, \theta_{itj}\ge 0(1\le j\le n_i)$$

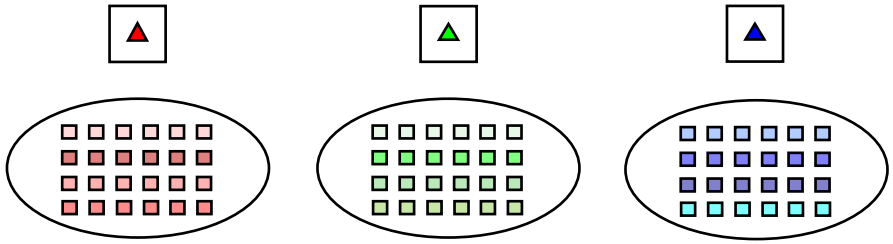
[Zhang & Zhou, ICDM'08]

MIML distance metric learning



A set of MIML examples

Each class is assumed to have K centers (sub-classes), and the corresponding instances comprise a bag

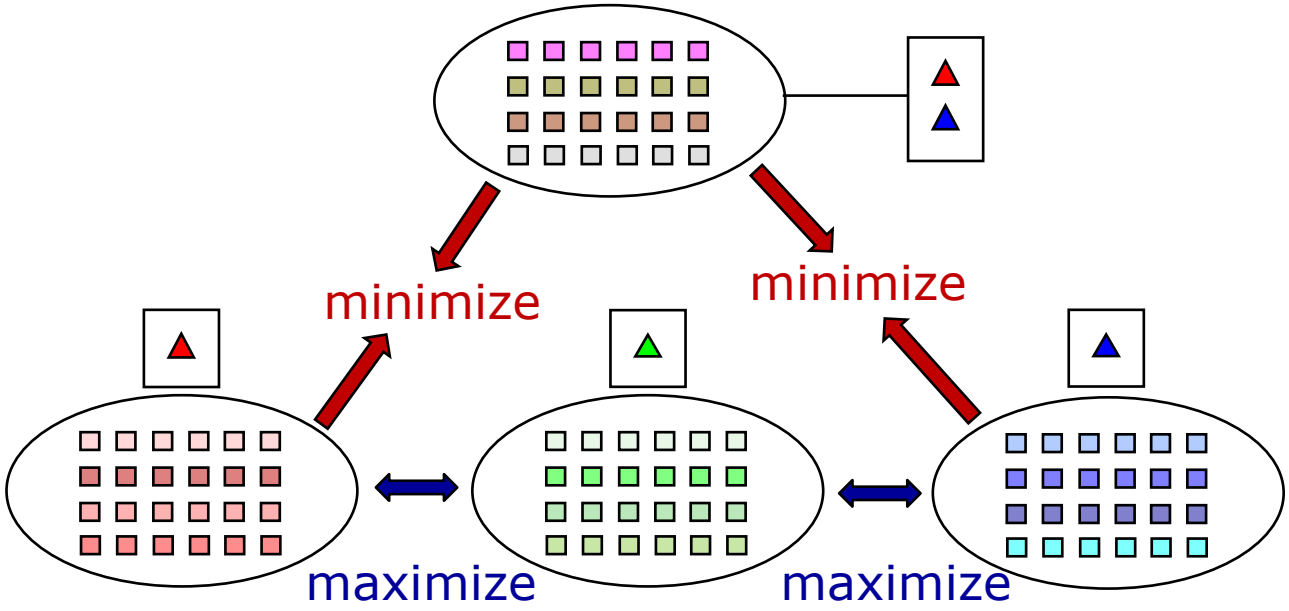


class centers

[Jin et al., CVPR'09]

MIML distance metric learning (con't)

- ✓ To **minimize** the distances between each bag and its classes
- ✓ To **maximize** the distances between classes



[Jin et al., CVPR'09]

MIML distance metric learning (con't)

Distance between two bags X_i and X_j :

$$D(X_i, X_j) = \min_{1 \le k \le n_i, 1 \le l \le n_j} |x_i^k - x_j^l|_A^2$$

Distance between a bag X_i and a class c_j :

$$d(X_i, c_j) = D(X_i, Z_j) = \min_{1 \le k \le n_i, 1 \le l \le K} |x_i^k - z_j^l|_A^2$$

Distance between two classes:

$$D(Z_i, Z_j) = \min_{1 \le k, l \le K} |z_i^k - z_j^l|_A^2$$

A is the metric to be learned:

To **minimize** the distance between each bag and its classes, and **maximize** the distances between classes

[Jin et al., CVPR'09]

MIML distance metric learning (con't)

Objective function:

$$\min_{\mathrm{tr}(A)=r, A\succeq 0, Z} \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_i^j D(X_i, Z_j)}{\sum_{i,j=1}^{m} D(Z_i, Z_j)(1-\delta(i,j))}$$

$$A = \sum_{i=1}^{r} w_i w_i^\top \qquad w_i^\top w_j = \delta(i,j)$$

which can be re-written as:

$$\min_{A\in\Lambda_r, Q, P, Z} \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_i^j \sum_{k=1}^{n_i}\sum_{l=1}^{K} Q_{k,l}^{(i,j)} |x_i^k - z_j^l|_A^2}{\sum_{i,j=1}^{m}(1-\delta(i,j))\sum_{k,l=1}^{K} P_{k,l}^{(i,j)} |z_i^k - z_j^l|_A^2}$$

3 groups of variables

$$\text{s. t.} \quad Q^{(i,j)} \in \mathbb{R}_+^{n_i\times n_j}, Q^{(i,j)}\mathbf{1} = \mathbf{1}, i,j = 1,\ldots,n$$

$$P^{(i,j)} \in \mathbb{R}_+^{K\times K}, P^{(i,j)}\mathbf{1} = \mathbf{1}, i,j = 1,\ldots,K$$

$P^{(i,j)}$: Distance indicator between two bags X_i and c_j

$Q^{(i,j)}$: Distance indicator between two bags c_i and c_j

A solution: Iteratively, solving one by fixing two

[Jin et al., CVPR'09]

More …

MIML based on Hidden Conditional Random Fields (HCRFs)

- MLMIL [Zha et al., CVPR'08]

MIML based on Dirichlet-Bernoulli Alignment

- DBA [Yang et al., NIPS'09]

MIML based on assuming each instance with one label:

- SISL-MIML [Nguyen, ICDM'10]

When there is no access to raw objects:

- INSDIF [Zhang & Zhou, AAAI'07]

To help the learning of complicated high-level concepts:

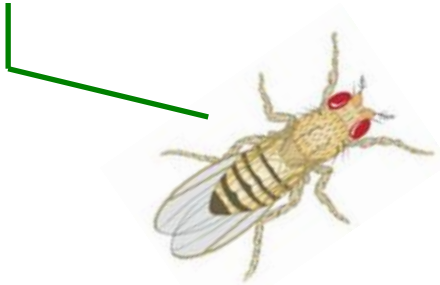
- SUBCOD [Zhou et al., AIJ 2012]

Outline

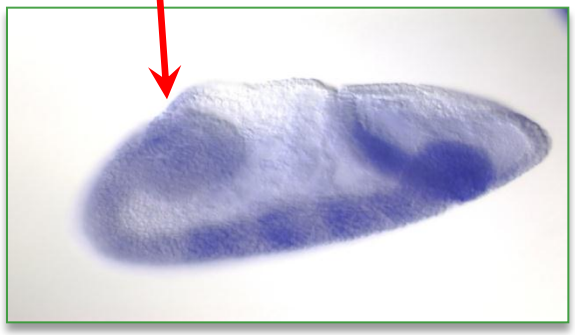
- MIML: A New Learning Framework
  - The framework
  - Why MIML?
- Learning Algorithms
- **A Real Application**
  - The problem
  - Solution and results

Drosophila gene expression pattern

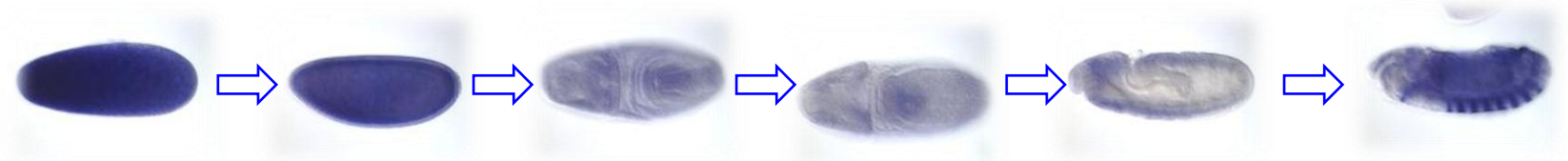
Drosophila, or *fruit fly*, is a model organism widely studied in developmental biology



Gene RhoGAP71E expressed stage: 7-8



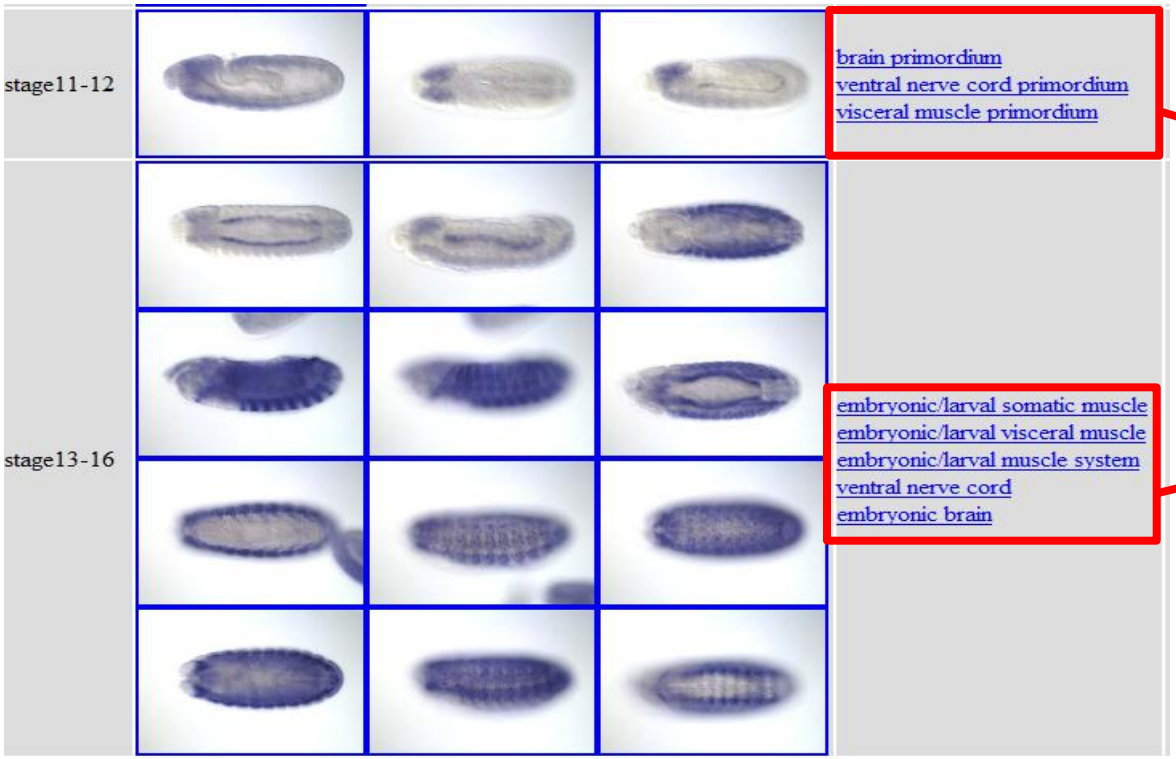
Gene expression pattern by RNA *in situ* hybridization during Drosophila embryogenesis



stage (1-3) stage (4-6) stage (7-8) stage (9-10) stage (11-12) stage (13-16)

The BDGP project

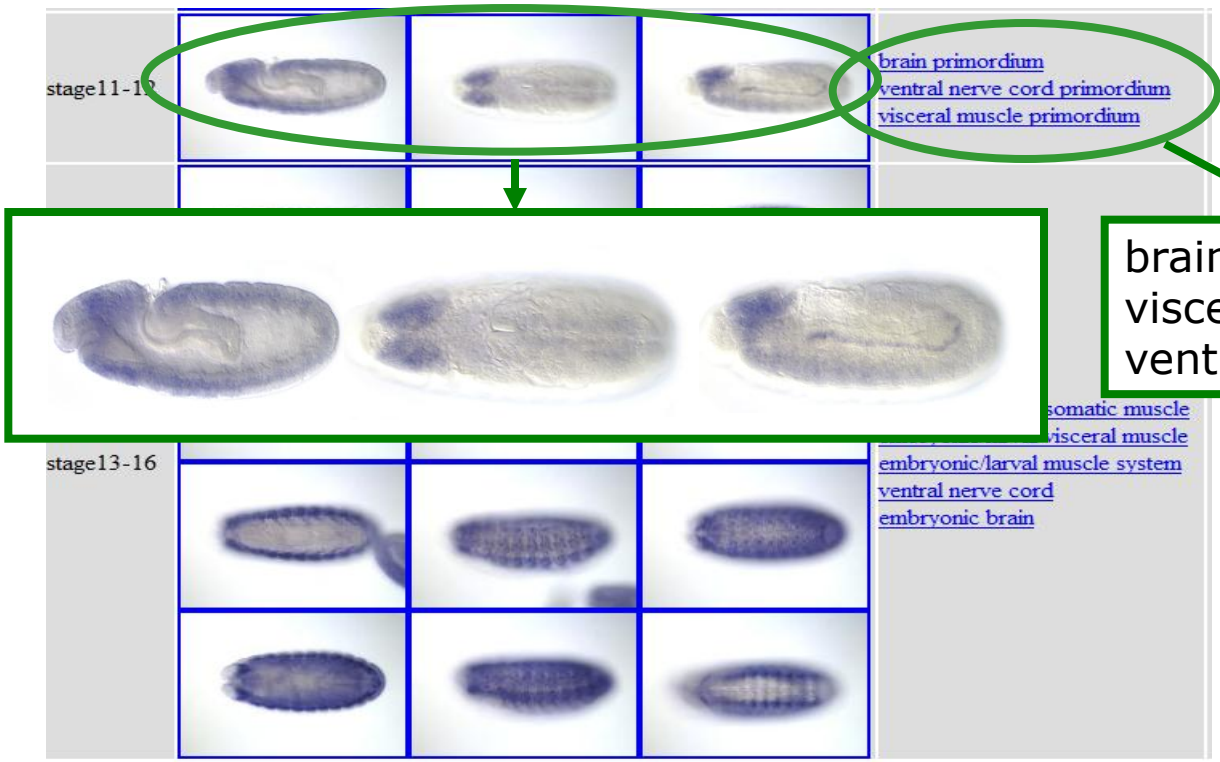
The Berkeley *Drosophila* Genome Project (BDGP) produced a large amount of spatial-temporal gene expression images



anatomical and developmental ontology terms manually labeled by human curators

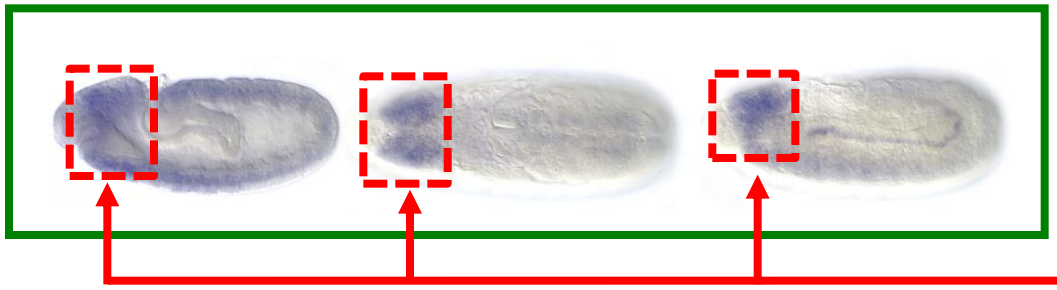
Gene: Actn

Difficulty for automatic annotation



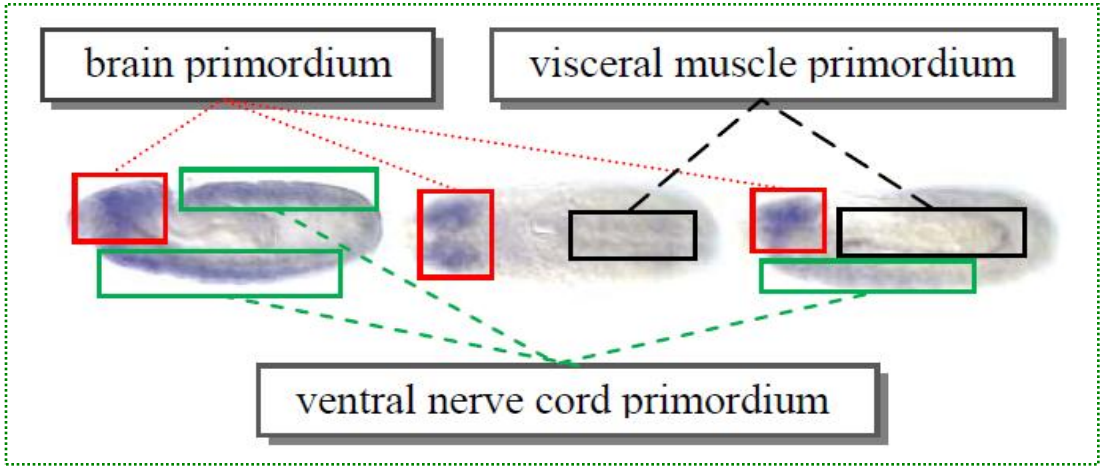
brain primordium
visceral muscle primordium
ventral nerve cord primordium

Difficulty for automatic annotation



brain primordium
visceral muscle primordium
ventral nerve cord primordium

The terms are body-part related



We do not know which term is associated with which region in the images !!

Generality of the problem

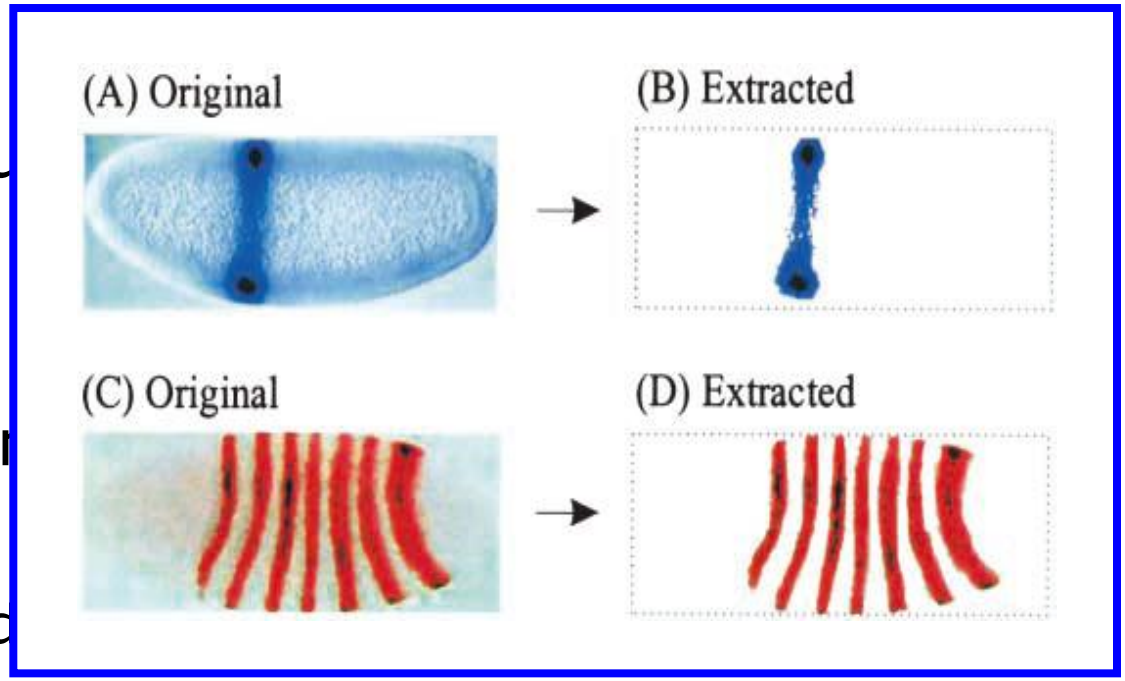
A good solution to the *Drosophila* gene expression pattern annotation task will also benefit other bio-problems

e.g., Protein functional prediction

- ✓ many conformations, varying functions
- ✓ lack knowledge of which conformation is responsible for a specific function

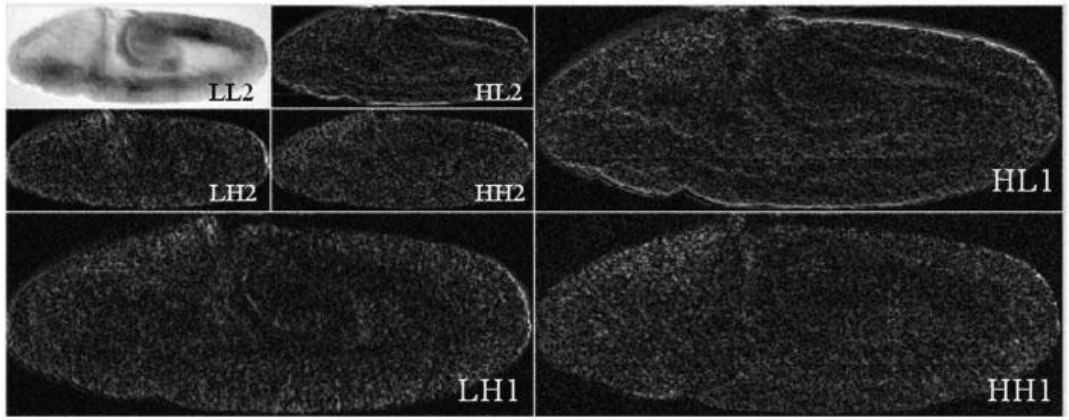
Previous solutions

- ✓ BESTi Algorithm [Kumar et al., Genetics02]
  - use images from literatures
  - use binary feature vector
- ✓ 2D Wavelet featu
  - use BDGP images
- ✓ Multi-kernel learn
  - use BDGP images
  - use multi-pyramid



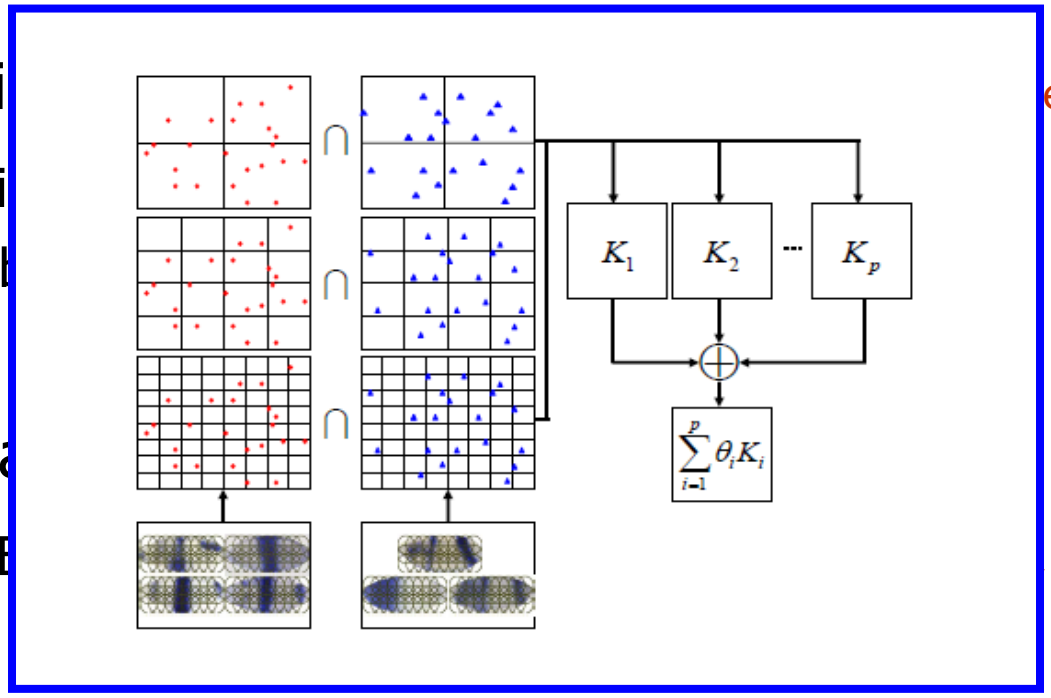
Previous solutions

- ✓ BESTi Algorithm [Kumar et al., Genetics02]
  - use images from literatures
  - use binary feature vector
- ✓ 2D Wavelet features, LDA classifier
  - use BDGP image
- ✓ Multi-kernel lea
  - use BDGP image
  - use multi-pyram



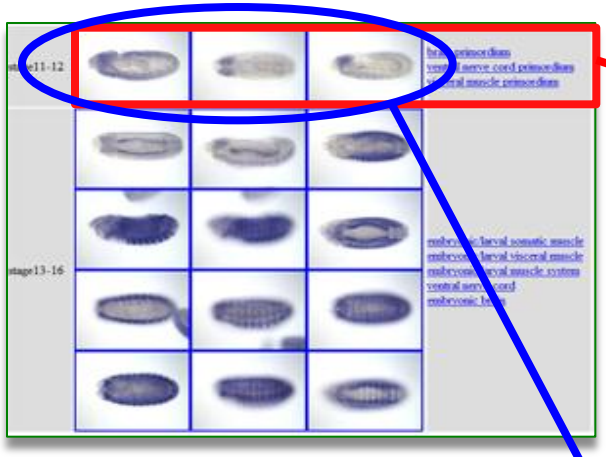
Previous solutions

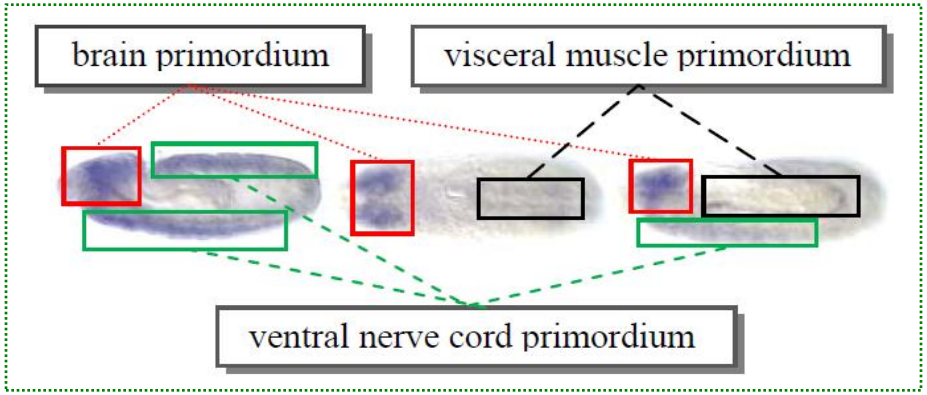
- ✓ BESTi [et al., Genetics02]
  - use i
  - use b
- ✓ 2D Wa
  - use B [Bioinformatics07]

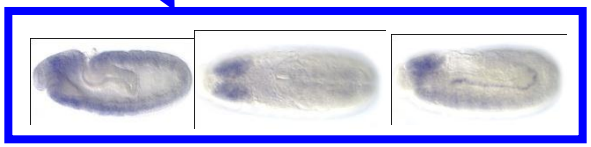


- ✓ Multi-kernel learning with hypergraph
  - use BDGP images [Ji et al., Bioinformatics08]
  - use multi-pyramid match kernel and hypergraph learning

Formulated as an MIML problem



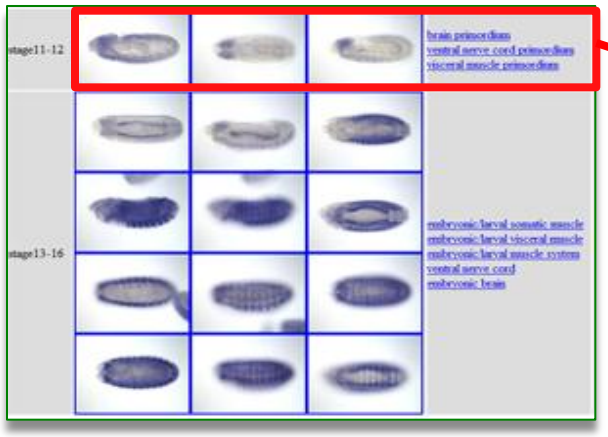


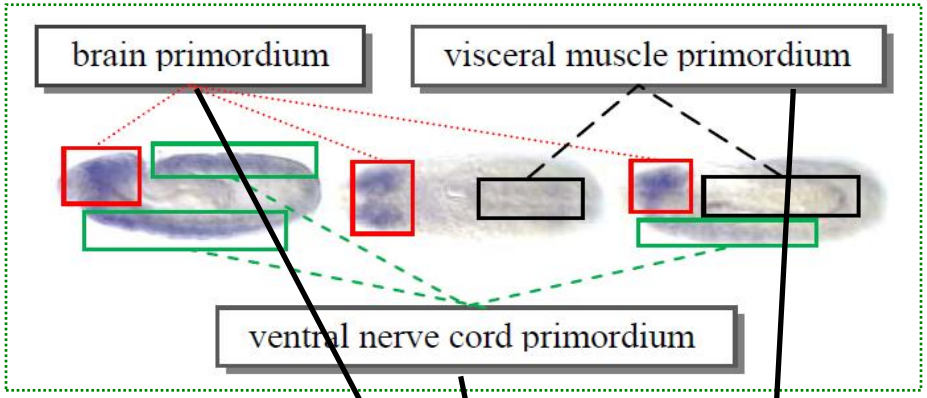


object

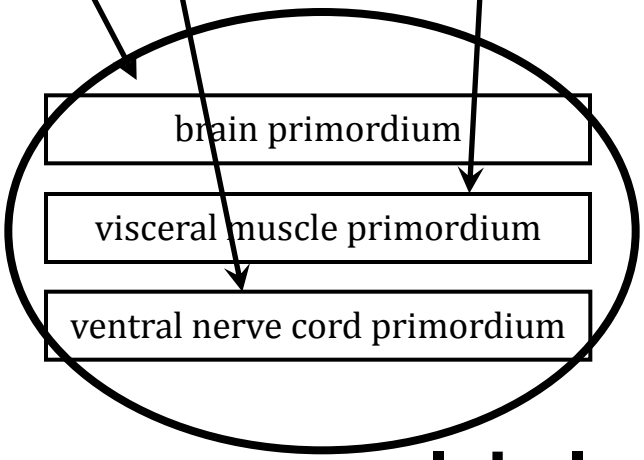
[Li et al., IJCAI'09]

Formulated as an MIML problem



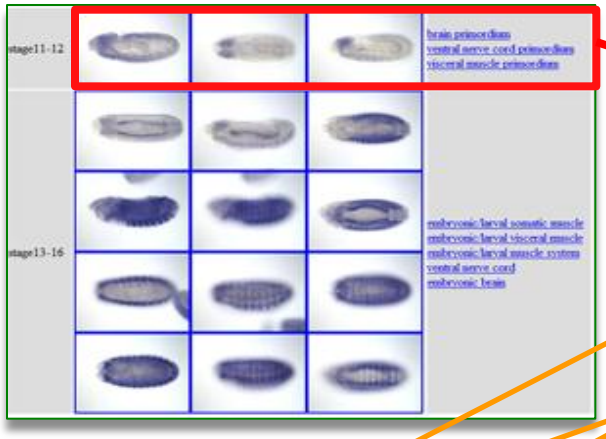


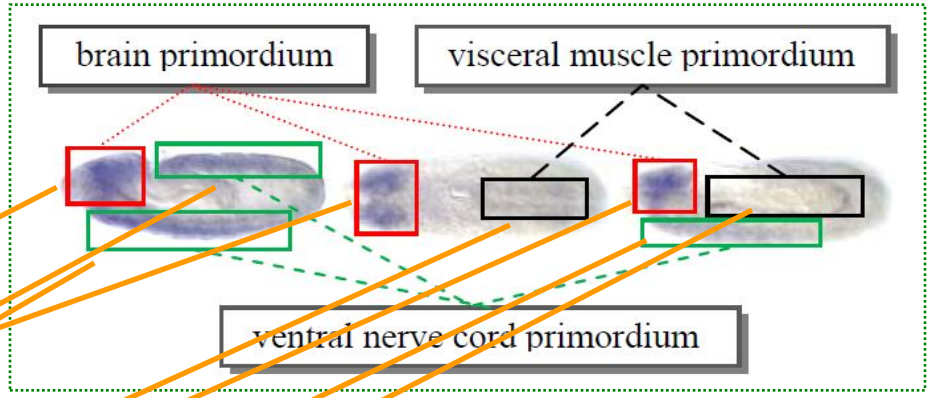
object

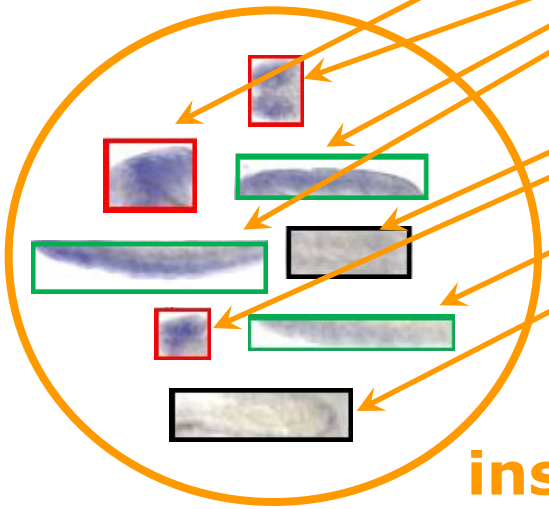


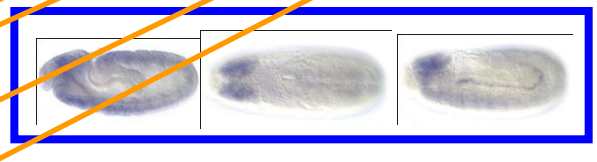
labels

Formulated as an MIML problem

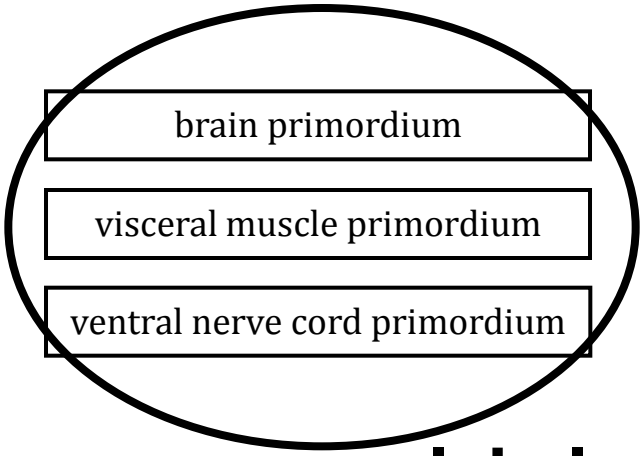








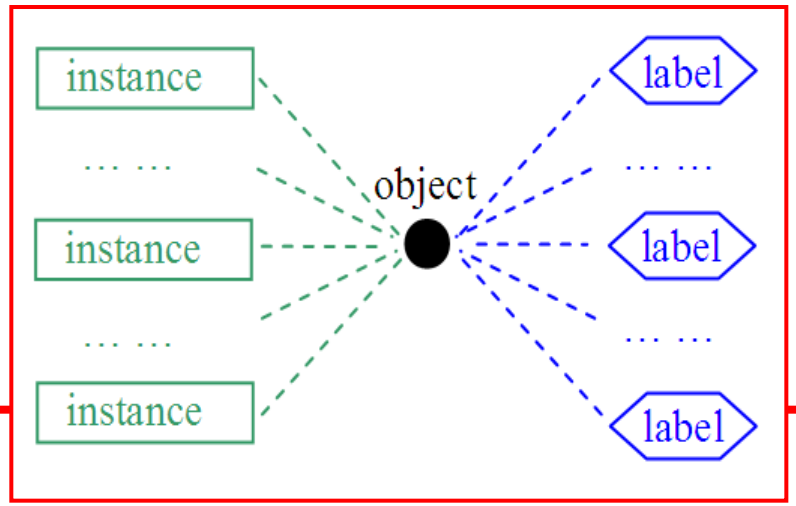
object

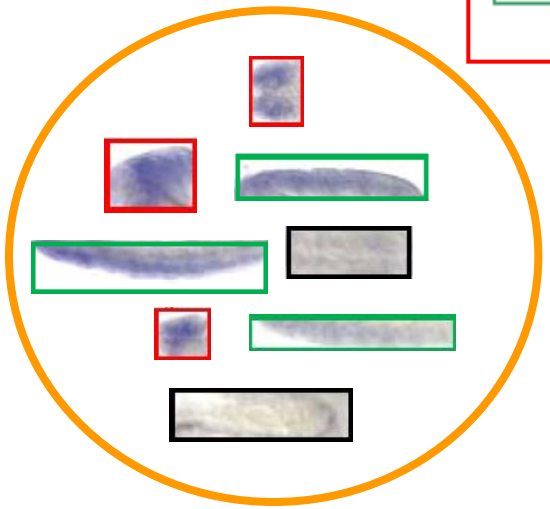


instances

labels

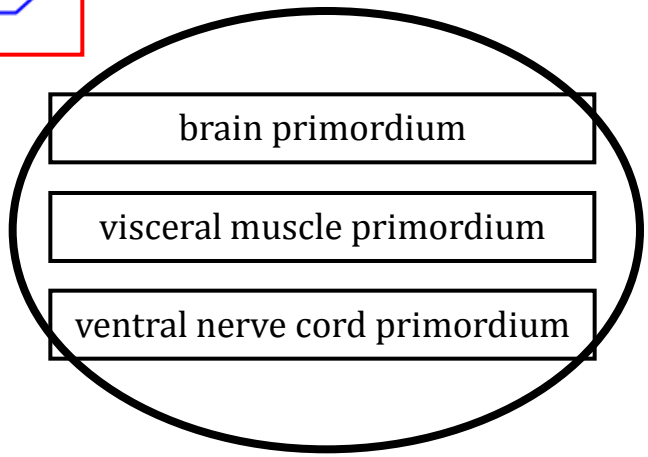
Formulated as an MIML problem (con't)





instances

object



labels

Outline

- MIML: A New Learning Framework
  - Why MIML?
  - Advances of MIML
- A Real Application
  - The problem
  - **Solution and results**

The MIMLSVM+ algorithm

For each label $y \in \mathcal{Y}$, let $\varphi(X_i, y) = +1$ if $y \in Y_i$ and -1 otherwise

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C^+ \sum_{\varphi(X_i,y)=1} \xi_i + C^- \sum_{\varphi(X_i,y)=-1} \xi_i$$

$$\text{subject to: } \varphi(X_i, y)(w' \phi(X_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \qquad (i = 1, 2, \cdots, n)$$

We set $C^+ > C^-$ to make the classifier biased toward positive class

[Li et al., IJCAI'09]

The MIMLSVM+ algorithm

For each label $y \in \mathcal{Y}$, let $\varphi(X_i, y) = +1$ if $y \in Y_i$ and -1 otherwise

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C^+ \sum_{\varphi(X_i,y)=1} \xi_i + C^- \sum_{\varphi(X_i,y)=-1} \xi_i$$

$$\text{subject to: } \varphi(X_i, y)(w' \phi(X_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \qquad (i = 1, 2, \cdots, n)$$

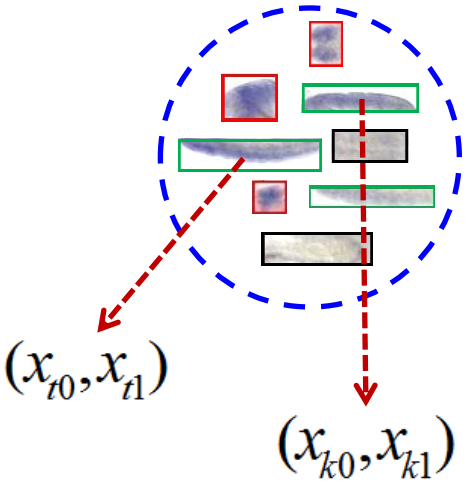
This involves a kernel function mapping a bag of instances into kernel space. We simply use the *set kernel*:

$$K_{SET}(X, X') = \sum_{i=1}^{n} \sum_{j=1}^{m} K(x_i, x'_j)$$

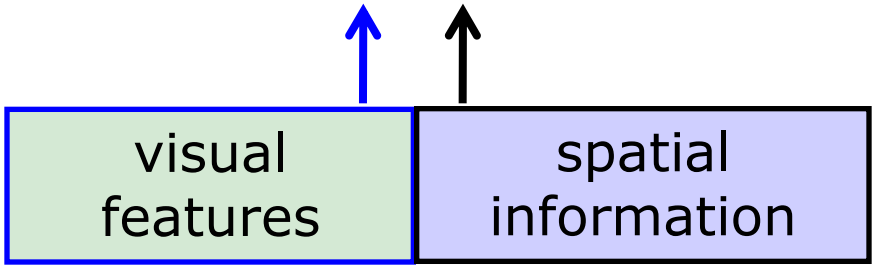
[Li et al., IJCAI'09]

Features used to describe instances

- ✓ visual features of gene expression of patches
- ✓ spatial information of patches



$$X_i = \{x_t\} = \{x_{t0}, x_{t1}\}$$



$$K(x_t, x_k) = e^{-\gamma_1 \|x_{t0} - x_{k0}\|^2 - \gamma_2 \|x_{t1} - x_{k1}\|^2}$$

Experimental configuration

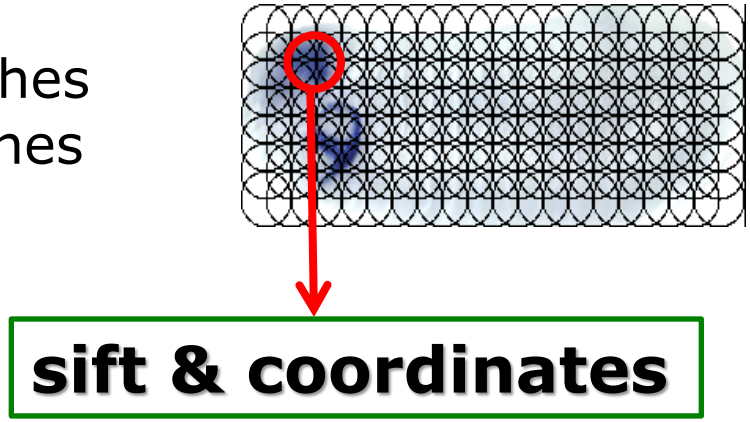
Dataset

2,816 bags, **2,052,722 instances** (15,434 x 133), **119 labels**

(2,816 image groups, 15,434 images, 133 instances per image, 119 terms)

Feature

SIFT on dense regular patches
Center coordinates of patches



Evaluation measures

Extended from traditional measures

- ✓ Macro-F1 the larger, the better
- ✓ Micro-F1 the larger, the better
- ✓ AUC (Area under ROC curve) the larger, the better

Multi-Label measures

- ✓ Average precision the larger, the better
- ✓ One-error the smaller, the better
- ✓ Coverage the smaller, the better
- ✓ Ranking loss the smaller, the better
- ✓ Hamming loss the smaller, the better

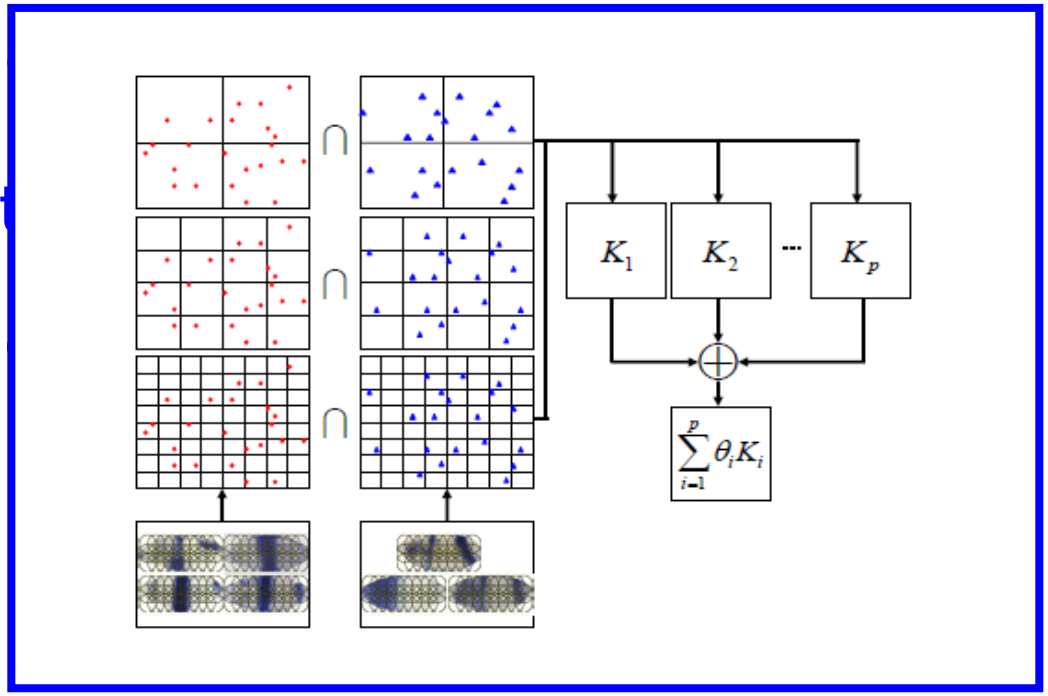
Compared methods

Existing methods

- ✓ MKL-PMK [Ji et al., Bioinformatics08]
- ✓ MIML-SVM [Zhou

Degenerated variant

- ✓ MIML-SVM $^{+}_{sv}$: Con
- ✓ MIML-SVM $^{+}_{v}$: Use



Experimental results

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

50% train 50% test, 30 runs with random partitions

# terms	# groups	Algorithms	macro-F1 ↑	micro-F1 ↑	AUC ↑	Ave. Precision ↑	one-error ↓	coverage ↓	Rankloss ↓	Hammloss ↓
10	222	MIMLSVM$^+$	**0.643±0.011**	**0.689±0.007**	**0.883±0.004**	**0.779±0.005**	**0.272±0.008**	**2.994±0.056**	**0.150±0.006**	**0.150±0.004**
		MIMLSVM$^+_{SV}$	0.627±0.010	0.676±0.006	0.869±0.004	0.773±0.005	0.277±0.011	3.073±0.048	0.157±0.004	0.156±0.003
		MIMLSVM$^+_{V}$	0.619±0.011	0.667±0.007	0.863±0.004	0.764±0.005	0.291±0.009	3.139±0.044	0.164±0.004	0.160±0.003
		MKL-PMK	0.584±0.009	0.621±0.009	0.825±0.006	0.722±0.007	0.343±0.011	3.483±0.072	0.198±0.006	0.196±0.006
# terms	# groups	Algorithms	macro-F1 ↑	micro-F1 ↑	AUC ↑	Ave. Precision ↑	one-error ↓	coverage ↓	Rankloss ↓	Hammloss ↓
20	247	MIMLSVM$^+$	**0.468±0.015**	**0.587±0.007**	**0.862±0.003**	**0.673±0.008**	**0.357±0.011**	**6.189±0.117**	**0.152±0.005**	**0.114±0.002**
		MIMLSVM$^+_{SV}$	0.454±0.012	0.574±0.008	0.845±0.003	0.660±0.009	0.364±0.013	6.481±0.119	0.163±0.005	0.118±0.003
		MIMLSVM$^+_{V}$	0.445±0.012	0.566±0.006	0.840±0.004	0.651±0.008	0.377±0.011	6.609±0.114	0.169±0.004	0.119±0.002
		MKL-PMK	0.410±0.007	0.506±0.006	0.771±0.006	0.580±0.007	0.445±0.009	8.082±0.122	0.230±0.005	0.144±0.003
# terms	# groups	Algorithms	macro-F1 ↑	micro-F1 ↑	AUC ↑	Ave. Precision ↑	one-error ↓	coverage ↓	Rankloss ↓	Hammloss ↓
30	2646	MIMLSVM$^+$	**0.368±0.012**	**0.541±0.007**	**0.850±0.003**	**0.623±0.007**	**0.377±0.010**	**9.406±0.173**	**0.153±0.003**	**0.087±0.002**
		MIMLSVM$^+_{SV}$	0.354±0.001	0.527±0.006	0.829±0.004	0.605±0.007	0.388±0.010	9.964±0.195	0.166±0.004	0.090±0.002
		MIMLSVM$^+_{V}$	0.340±0.012	0.517±0.007	0.822±0.004	0.596±0.007	0.399±0.010	10.183±0.189	0.171±0.004	0.091±0.002
		MKL-PMK	0.310±0.008	0.455±0.008	0.741±0.007	0.511±0.008	0.488±0.011	13.010±0.2413	0.243±0.006	0.142±0.003

MIMLSVM+ achieves the best performance on ALL cases and ALL evaluation measures

[Li et al., IJCAI'09]

Experimental results (con't)

Since MIMLSVM could not work on the previous large data sets, we extract a smaller data set via random sampling

167 bags, **57,323 instances** (431 x 133), **10 labels**

(167 image groups, 431 images, 133 inst per image, 10 terms)

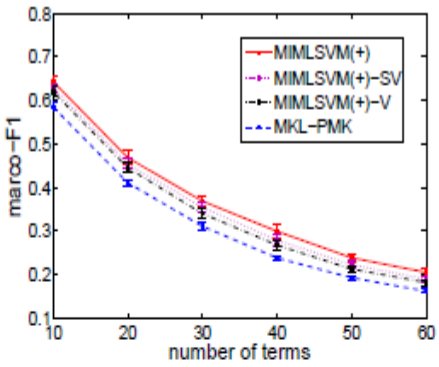
20 runs with random splits of training/test sets

# terms	# groups	Algorithms	macro-F1 ↑	micro-F1 ↑	AUC ↑	Ave. Precision ↑	one-error ↓	coverage ↓	Rankloss ↓	Hammloss ↓
10	167	MIMLSVM^{+}	0.460±0.041	0.606±0.026	0.807±0.191	0.733±0.019	0.311±0.034	3.508±0.262	0.186±0.015	0.171±0.019
		MIMLSVM$^{+}_{SV}$	0.424±0.049	0.569±0.033	0.774±0.017	0.710±0.027	0.354±0.047	3.667±0.199	0.204±0.016	0.191±0.015
		MIMLSVM	0.176±0.047	0.367±0.054	0.629±0.041	0.592±0.028	0.468±0.060	4.792±0.300	0.318±0.029	0.241±0.097

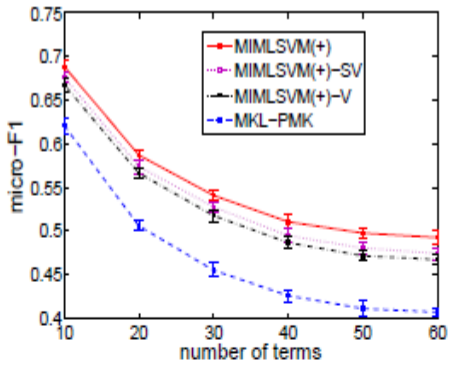
MIMLSVM+ achieves the best performance on ALL evaluation measures

[Li et al., IJCAI'09]

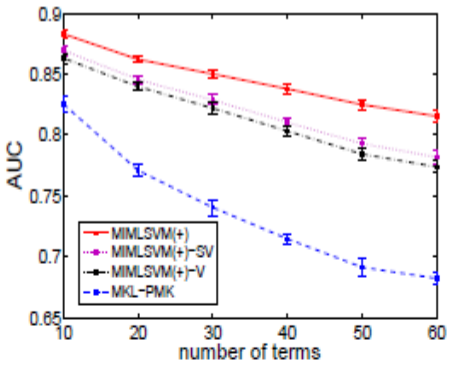
Experimental results (con't)



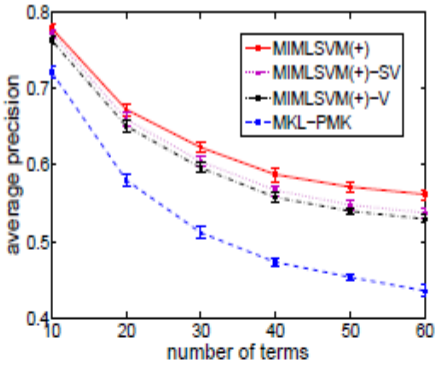
(a) macro-F1 ↑



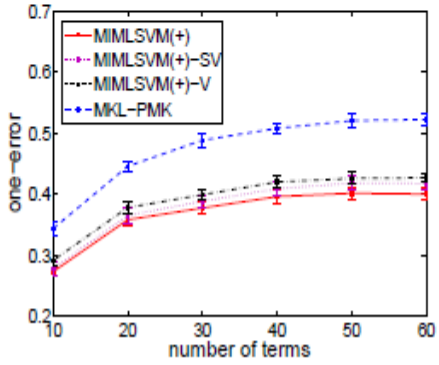
(b) micro-F1 ↑



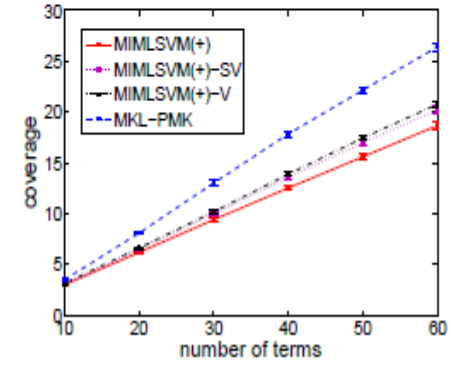
(c) AUC ↑



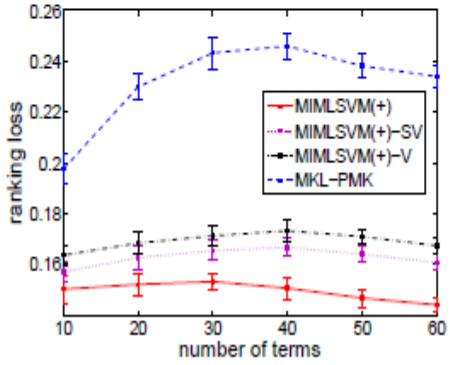
(d) average precision ↑



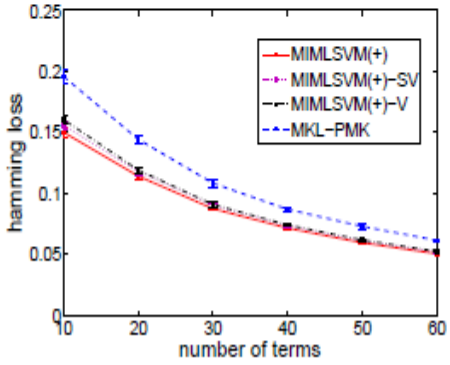
(e) one-error ↓



(f) coverage ↓



(g) ranking loss ↓



(h) hamming loss ↓

The comparison under different number of labels (annotation terms)

[Li et al., IJCAI'09]

Further examination

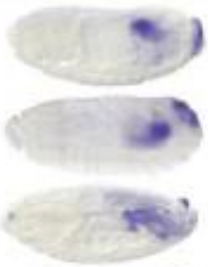
Among 1,438 test image groups, 427 cases are found whose top-predicted terms are not exactly the BDGP annotation terms

this may owe to mistakes of our approach, or human annotators

So, we pick a small number of such cases to invite human domain expert to carefully re-examine the annotations

[Li et al., ACM/IEEE TCBB 2012]

Further examination (con't)

Genes	Images	BDGP terms	Predicted terms
r-1		embryonic midgut embryonic Malpighian tubule embryonic hindgut embryonic anal pad embryonic/larval muscle system dorsal prothoracic pharyngeal muscle	embryonic midgut embryonic Malpighian tubule embryonic hindgut embryonic anal pad embryo
cad		embryonic Malpighian tubule embryonic hindgut	embryonic midgut embryonic anal pad embryonic Malpighian tubule embryonic hindgut

Missed by our approach

Missed by human experts

A very practical value of our technique is to help human experts to "double-check" during the expensive manual annotation process

[Li et al., ACM/IEEE TCBB 2012]

Acknowledgement

The talk involves some joint work with

My students :

- Sheng-Jun Huang
- Yin-Xing Li
- Yu-Feng Li
- Min-Ling Zhang
- . . .

My collaborators:

- Shuiwang Ji
- Rong Jin
- Sudir Kumar
- Shijun Wang
- Jieping Ye
- … …

Some MIML papers

MIMLBoost & MIMLSVM

- ✓ Z.-H. Zhou, M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In: Advances in Neural Information Processing Systems 19 (NIPS'06), Cambridge, MA: MIT Press, 2007, pp.1609-1616.

INSDIF

- ✓ M.-L. Zhang, Z.-H. Zhou. Multi-label learning by instance differentiation. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI'07), Vancouver, Canada, 2007, pp.669-674.

M3MIML

- ✓ M.-L. Zhang, Z.-H. Zhou. M3MIML: A maximum margin method for multi-instance multi-label learning. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy, 2008, pp.688-697.

  Code: http://lamda.nju.edu.cn/code_M3MIML.ashx

Some MIML papers

MIML Distance Metric Learning

- ✓ S. Wang, R. Jin, Z.-H. Zhou. Learn a distance metric from multi-instance multi-label data. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, 2009, pp.896-902.

MIML Ensemble for Video Annotation

- ✓ X.-S. Xu, X. Xue, Z.-H. Zhou. Ensemble multi-instance multi-label learning approach for video annotation task. In: Proceedings of the 19th ACM International Conference on Multimedia (MM'11), Scottsdale, AZ, 2011.

MIML for Drosophila

- ✓ Y.-X. Li, S. Ji, J. Ye, S. Kumar, Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. IEEE/ACM Trans. Computational Biology and Bioinformatics, 2012, 9(1): 98-112. (early version at IJCAI'09)

  Code: http://lamda.nju.edu.cn/code_MIMLdros.ashx

Some MIML papers

Full Description of MIML:

- ✓ Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. Artificial Intelligence, 2012, 176(1): 2291-2320.

Code: http://lamda.nju.edu.cn/code_MIML.ashx

Data-1: http://lamda.nju.edu.cn/data_MIMLimage.ashx

Data-2: http://lamda.nju.edu.cn/data_MIMLtext.ashx

Thanks !