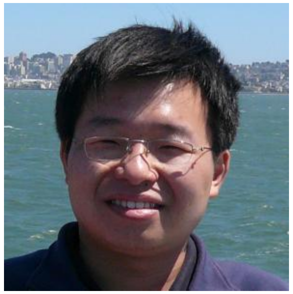
The 25th International Conference on Machine Learning (ICML) 2008, Helsinki, Finland

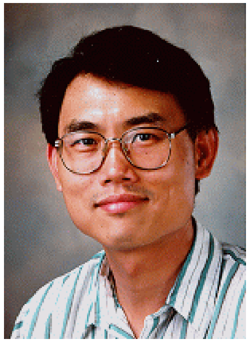
Adaptive *p*-Posterior Mixture Model Kernels for Multiple Instance Learning



Hua-Yan Wang
Peking University

Qiang Yang
Hong Kong University of Science and Technology

Hongbin Zha
Peking University



storyline

- **multiple instance learning (MIL)**
  – the concept
- ***different*** **applications of MIL**
  – how they *differ*
- **generalized MIL**
  – accommodate the *differences*
- **the *ppmm* kernels**
  – a simple yet effective solution to generalized MIL

storyline

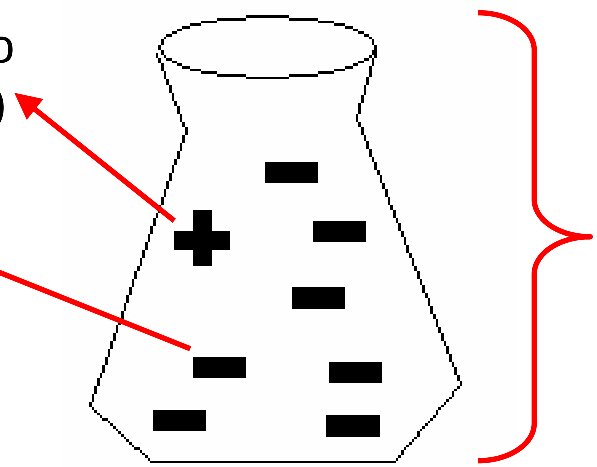
- **multiple instance learning (MIL)**
  - the concept
- ***different* applications of MIL**
  - how they *differ*
- **generalized MIL**
  - accommodate the *differences*
- **the *ppmm* kernels**
  - a simple yet effective solution to generalized MIL

Multiple Instance Learning

- original motivation of MIL (drug activity prediction)

a shape that binds well to
the target protein (active)

a shape that doesn't



a certain kind of drug molecule, which can adopt a number of shapes

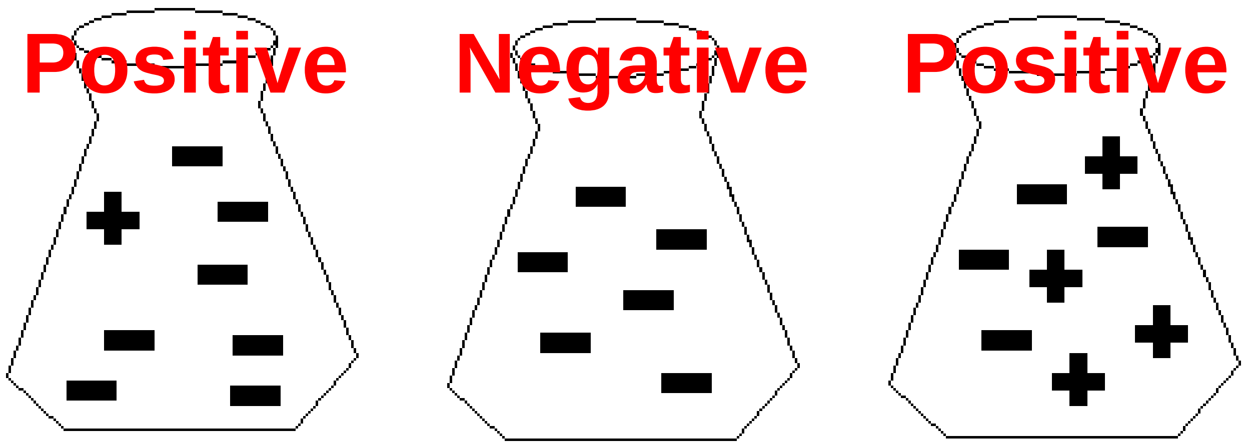
In wet-lab experiments, the molecule would be observed active.

Wet-lab experiments cost a lot time and $!!

With MIL, we predict (with mild confidence) activity of molecules in a dry-lab.

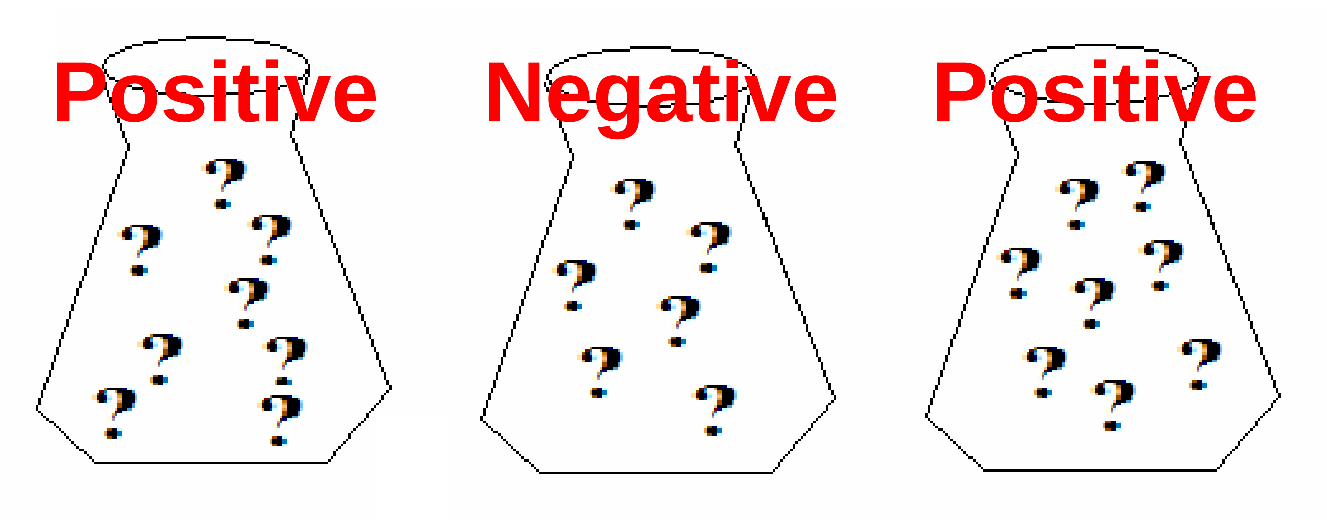
MIL saves $ for biologist and makes $ for computer scientists ☺

Multiple Instance Learning



- bags of instances
- instances labeled as *positive* or *negative*
- positive bag $\Leftrightarrow$ at least one positive instance

Multiple Instance Learning



- The learner *cannot* see the labels of the instances
- The learner is required to predict labels for previous unseen bags.

storyline

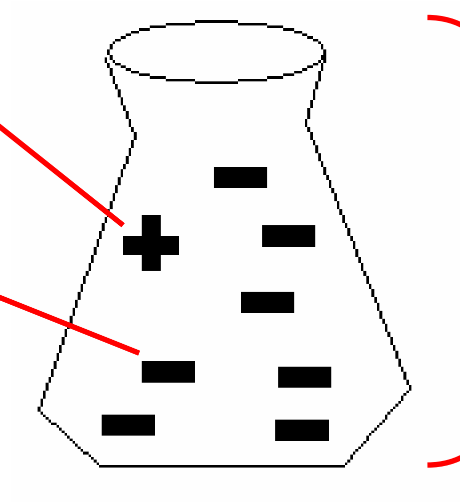
- **multiple instance learning (MIL)**
  – the concept
- ***different* applications of MIL**
  – how they *differ*
- **generalized MIL**
  – accommodate the *differences*
- **the *ppmm* kernels**
  – a simple yet effective solution to generalized MIL

different applications of MIL

- drug activity prediction

a shape that binds well to
the target protein (active)

a shape that doesn't



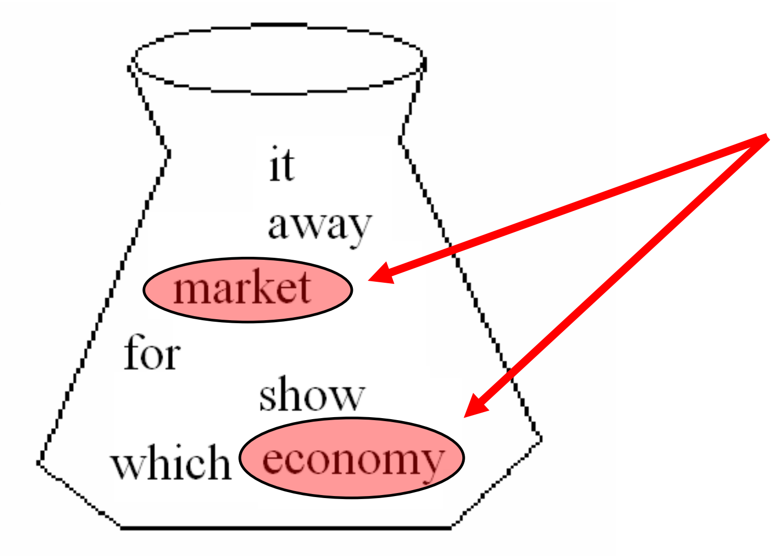
a certain kind of drug molecule, which can adopt a number of shapes

A positive instance is a **definite** evidence for a positive bag.

We need just **one** positive instance for labeling a bag as positive.

different applications of MIL

- document classification



document as a bag of words / phrases / sentences / paragraphs

For labeling the document as "economics", we have two positive instances here.

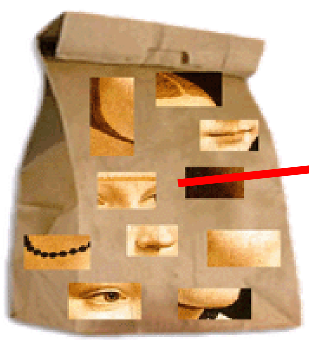
Positive instance are **strong** evidences for a positive bag.

We may need **several** positive instance for labeling a positive bag.

different applications of MIL

- image classification

image as a bag of
local observations



local image features

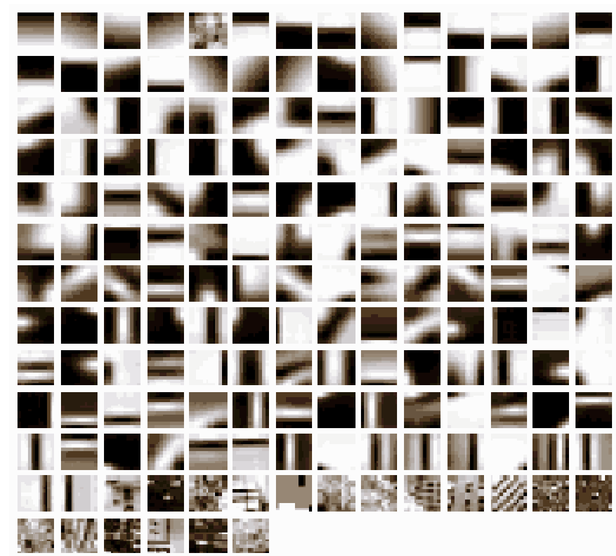


image courtesy of
Li Fei-Fei et al

Image features are low-level representations,

They serve as **weak** evidences for labeling a bag.

We need **many** positive instances for labeling a positive bag.

Introduction

- Multiple Instance Learning in **different application domains**
  - drug activity prediction: **bags**: molecules, **instances**: shapes
  - image classification: **bags**: images, **instances**: local features
  - document classification: **bags**: documents, **instances**: terms, sentences

drug activity prediction

document classification

image classification

'+' instance are ***strong*** evidences for '+' bags

Few '+' instances can determine a '+' bag.

'+' instance are ***weak*** evidences for '+' bags

Many '+' instances can determine a '+' bag.

storyline

- **multiple instance learning (MIL)**
  - the concept
- ***different* applications of MIL**
  - how they *differ*
- **generalized MIL**
  - accommodate the *differences*
- **the *ppmm* kernels**
  - a simple yet effective solution to generalized MIL

generalized MIL

drug activity prediction

document classification

image classification

‘+’ instance are ***strong*** evidences for ‘+’ bags

Few ‘+’ instances can determine a ‘+’ bag.

‘+’ instance are ***weak*** evidences for ‘+’ bags

Many ‘+’ instances can determine a ‘+’ bag.

<u>traditional</u> multiple instance learning

<u>generalized</u> multiple instance learning

generalized MIL

- MIL
  - The learner is presented with bags of instances, with labels (+/-) on bags, and required to predict labels on new bags.
  - A bag is ‘+’ *iff* at least one of its instance is ‘+’.
- Generalized MIL
  - …
  - A bag is ‘+’ *iff* more than **s%** of its instance is ‘+’.
  - **s%** is different and unknown for different applications.

generalized MIL

drug activity prediction

document classification

image classification

‘+’ instance are ***strong*** evidences for ‘+’ bags

Few ‘+’ instances can determine a ‘+’ bag.

‘+’ instance are ***weak*** evidences for ‘+’ bags

Many ‘+’ instances can determine a ‘+’ bag.

approximates this degree of freedom

- Generalized MIL
  - …
  - A bag is ‘+’ *iff* more than **s%** of its instance is ‘+’.
  - **s%** is different and unknown for different applications.

generalized MIL

- major challenges arisen from such a setting:
  - The underlying parameter **s%** varies across different datasets (application domains).
  - **s%** is unknown to the learner.

  - How can the learner, who is presented only with labeled bags, discover the underlying difference in **s%** across different datasets (application domains).
  - How can the learner automatically adapt itself to these different **s%** ?

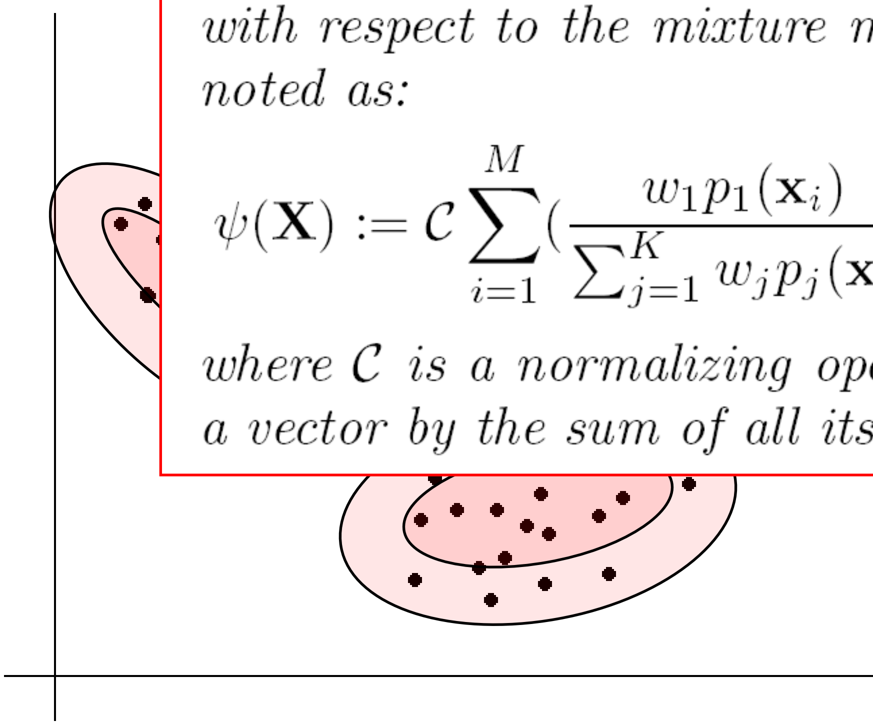
storyline

- **multiple instance learning (MIL)**
  - the concept
- ***different* applications of MIL**
  - how they *differ*
- **generalized MIL**
  - accommodates the *differences*
- **the *ppmm* kernels**
  - a simple yet effective solution to generalized MIL

the *ppmm* kernels

fit a mixture model

represent each instance



Definition 1 (Aggregate Posteriors) *The* ***aggregate posteriors*** *of a bag of instances* $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{M}$ *with respect to the mixture model* $\{(\boldsymbol{\Lambda}_i, w_i)\}_{i=1}^{K}$ *is denoted as:*

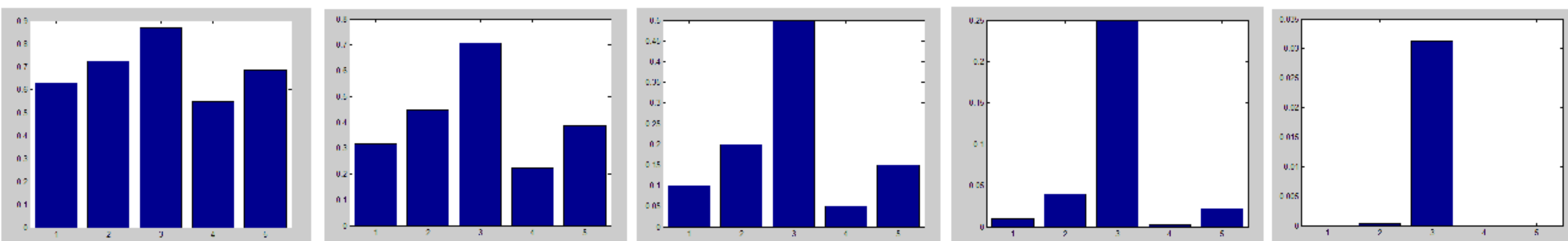
$$\psi(\mathbf{X}) := \mathcal{C} \sum_{i=1}^{M} \left(\frac{w_1 p_1(\mathbf{x}_i)}{\sum_{j=1}^{K} w_j p_j(\mathbf{x}_i)}, \cdots, \frac{w_K p_K(\mathbf{x}_i)}{\sum_{j=1}^{K} w_j p_j(\mathbf{x}_i)} \right)$$

where $\mathcal{C}$ *is a normalizing operator indicating dividing a vector by the sum of all its elements.*

the *ppmm* kernels

- Consider the mapping $\psi(\mathbf{X}) \Longrightarrow \psi(\mathbf{X})^p$

p = 0.2 | *p* = 0.5 | $\psi(\mathbf{X})$ | *p* = 2 | *p* = 5



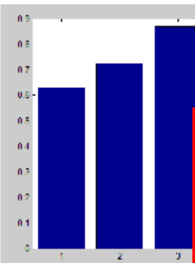
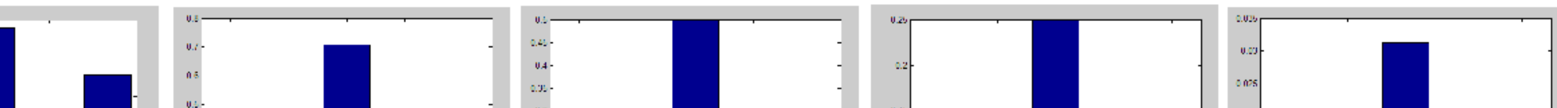
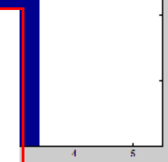
$0 < p < 1$ | $p = 1$ | $p > 1$

enhance minor patterns

attenuate minor patterns

the *ppmm* kernels

p = 0.2 *p* = 0.5 $\psi(\mathbf{X})$ *p* = 2 *p* = 5

dru

'+'
ev

Definition 2 (*p*-Posterior-Mixture-Model Kernel) *The p-posterior-mixture-model (ppmm) kernel function on a pair of bags* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *is defined as*

$$\kappa_p(\mathbf{X}_1, \mathbf{X}_2) := < \psi(\mathbf{X}_1)^p, \psi(\mathbf{X}_2)^p >$$

where $p \in (0, \infty)$, *and* $< \bullet, \bullet >$ *denotes the standard inner-product in* $\mathbb{R}^K$.

Few '+' instances can determine a '+' bag.

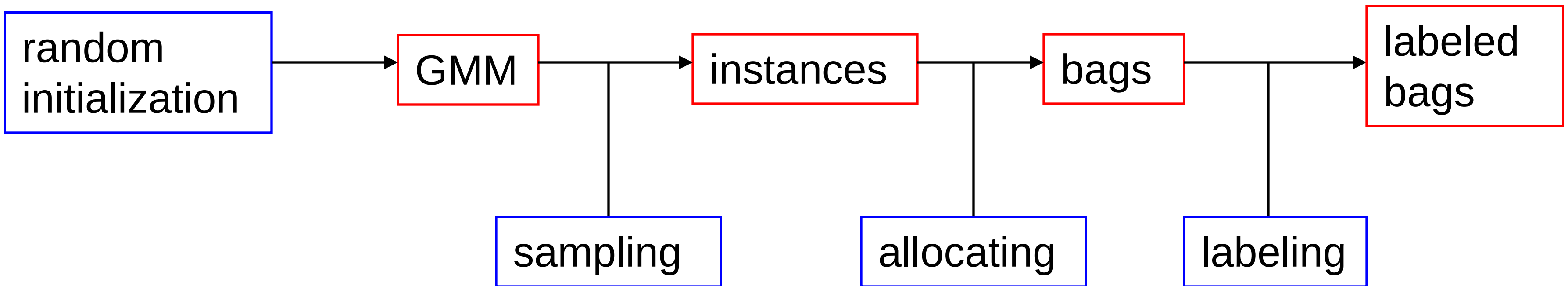
Minor patterns should be *enhanced* when comparing bags.

Many '+' instances can determine a '+' bag.

Minor patterns should be *attenuated* when comparing bags.

the *ppmm* kernels

- To validate our approach, we generate *3 synthetic datasets* as follows:



Instances from certain mixture components are labeled as ‘+’.

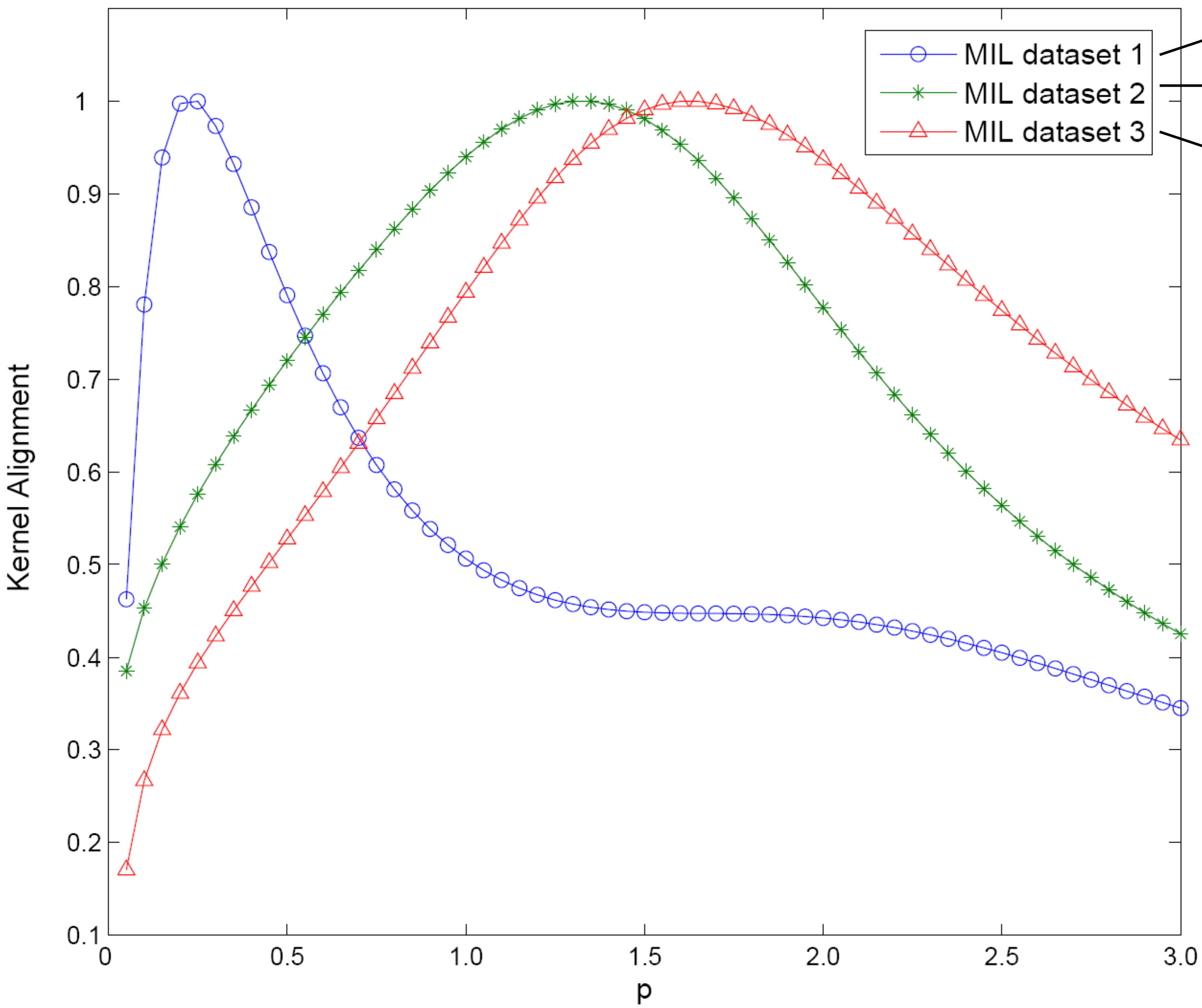
Bags with more than s% ‘+’ instances are labeled as ‘+’.

Setting s% = “at least one”, 20%, 50% yields MIL dataset 1, 2, 3.

The learner *only* sees bag labels and instances (without label).

We keep 50%-50% +/- bags for all datasets, which makes them appear *undistinguishable*.

the *ppmm* kernels



s% = "at least one"

s% = 20%

s% = 50%

The kernel alignment with ideal kernel assumes maximum at different *p* for different datasets.

The learner *revealed* the underlying difference among these 3 datasets, which "appear" to be undistinguishable.

the *ppmm* kernels

- comparison with state-of-the-art MIL techniques

Table 1. Empirical results of multiple instance learning methods, the last row shows the optimal p value learned in each task. MUSK1 and MUSK2 are drug activity prediction datasets. ELEPHANT, TIGER, FOX are image classification datasets. TREC1 and TREC2 are text classification datasets. Best performance in each task is in bold. The average performance over all tasks is shown in the last column.

DATESETS:	MUSK1	MUSK2	ELEPHANT	TIGER	FOX	TREC1	TREC2	**Average**
APR (DIETTERICH, 1997)	92.4%	89.2%	N/A	N/A	N/A	N/A	N/A	N/A
DD (MARON, 1998)	88.0%	84.0%	N/A	N/A	N/A	N/A	N/A	N/A
EM-DD (ZHANG, 2001)	84.8%	84.9%	78.3%	72.1%	56.1%	85.8%	84.0%	78.0%
CITATION k-NN (WANG, J., 2000)	91.3%	86.0%	80.5%	78.0%	60.0%	87.0%	81.0%	80.5%
mi-SVM (ANDREWS, 2003)	87.4%	83.6%	82.0%	78.9%	58.2%	93.6%	78.2%	80.3%
MI-SVM (ANDREWS, 2003)	77.9%	84.3%	81.4%	**84.0%**	59.4%	**93.9%**	**84.5%**	80.8%
MISS-SVM (ZHOU, 2007)	87.6%	80.0%	N/A	N/A	N/A	N/A	N/A	N/A
MG-ACC KERNEL (KWOK, 2007)	90.1%	**90.4%**	N/A	N/A	N/A	N/A	N/A	N/A
PPMM KERNEL (this paper)	**95.6%**	81.2%	**82.4%**	80.2%	**60.3%**	93.3%	79.5%	**81.8%**
OPTIMAL VALUE OF p	0.7	0.15	2.1	1.3	0.8	0.75	0.4	

Thanks!