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)

In wet-lab experiments, the molecule would be observed active.
Wet-lab experiments cost a lot of 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 \textit{positive} or \textit{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

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.

document as a bag of words / phrases / sentences / paragraphs
different applications of MIL

• image classification

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.
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

‘+’ 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

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

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

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

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

**traditional** multiple instance learning

**generalized** multiple instance learning

*drug activity prediction*  *document classification*  *image classification*
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
  – …
  – A bag is ‘+’ *iff* more than $s\%$ of its instance is ‘+’.
  – $s\%$ is different and unknown for different applications.

‘+’ 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

• 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 \textit{ppmm} kernels

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

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

where $C$ is a normalizing operator indicating dividing a vector by the sum of all its elements.
the **ppmm** kernels

- Consider the mapping \( \psi(X) \rightarrow \psi(X)^p \)

\[
egin{align*}
\text{\( p = 0.2 \)} & \quad \text{\( p = 0.5 \)} & \quad \text{\( \psi(X) \)} & \quad \text{\( p = 2 \)} & \quad \text{\( p = 5 \)} \\
& \quad & \quad & \quad & \\
0 < p < 1 & \quad & \quad & p = 1 & \quad & p > 1 \\
\text{enhance minor patterns} & \quad & \quad & \text{attenuate minor patterns}
\end{align*}
\]
the $ppmm$ kernels

$$p = 0.2 \quad p = 0.5 \quad \psi(X) \quad p = 2 \quad p = 5$$

**Definition 2 (p-Posterior-Mixture-Model Kernel)**

The $p$-posterior-mixture-model ($ppmm$) kernel function on a pair of bags $X_1$ and $X_2$ is defined as

$$\kappa_p(X_1, X_2) := \langle \psi(X_1)^p, \psi(X_2)^p \rangle$$

where $p \in (0, \infty)$, and $\langle \bullet, \bullet \rangle$ 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.
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

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.

<table>
<thead>
<tr>
<th>DATESETS:</th>
<th>MUSK1</th>
<th>MUSK2</th>
<th>ELEPHANT</th>
<th>TIGER</th>
<th>FOX</th>
<th>TREC1</th>
<th>TREC2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>apr (Dietterich, 1997)</td>
<td>92.4%</td>
<td>89.2%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>dd (Maron, 1998)</td>
<td>88.0%</td>
<td>84.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>em-dd (Zhang, 2001)</td>
<td>84.8%</td>
<td>84.9%</td>
<td>78.3%</td>
<td>72.1%</td>
<td>56.1%</td>
<td>85.8%</td>
<td>84.0%</td>
<td>78.0%</td>
</tr>
<tr>
<td>citation k-NN (Wang, J., 2000)</td>
<td>91.3%</td>
<td>86.0%</td>
<td>80.5%</td>
<td>78.0%</td>
<td>60.0%</td>
<td>87.0%</td>
<td>81.0%</td>
<td>80.5%</td>
</tr>
<tr>
<td>mi-svm (Andrews, 2003)</td>
<td>87.4%</td>
<td>83.6%</td>
<td>82.0%</td>
<td>78.9%</td>
<td>58.2%</td>
<td>93.6%</td>
<td>78.2%</td>
<td>80.3%</td>
</tr>
<tr>
<td>MI-svm (Andrews, 2003)</td>
<td>77.9%</td>
<td>84.3%</td>
<td>81.4%</td>
<td>84.0%</td>
<td>59.4%</td>
<td>93.9%</td>
<td>84.5%</td>
<td>80.8%</td>
</tr>
<tr>
<td>Miss-svm (Zhou, 2007)</td>
<td>87.6%</td>
<td>80.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>mg-acc kernel (Kwok, 2007)</td>
<td>90.1%</td>
<td>90.4%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>PPMM KERNEL (this paper)</td>
<td>95.6%</td>
<td>81.2%</td>
<td>82.4%</td>
<td>80.2%</td>
<td>60.3%</td>
<td>93.3%</td>
<td>79.5%</td>
<td>81.8%</td>
</tr>
</tbody>
</table>

**Optimal value of $p$**

| 0.7 | 0.15 | 2.1 | 1.3 | 0.8 | 0.75 | 0.4 |

---
Thanks!