Conformal Multi-Instance Kernels

Matthew Blaschko and Thomas Hofmann
Max Planck Institute for Biological Cybernetics and Google Switzerland

December 8, 2006
1 Introduction

2 Multi-Instance Kernels

3 Conformal Kernels

4 Conformal Multi-Instance Kernels

5 Generalization Bound Minimization

6 Diagonalization Approximation of the Trace-Margin Bound

7 Results
Multiple Instance Learning

- We wish to learn a mapping from bags of patterns to output labels
- two kinds of ambiguity
  - intrinsic variability of feature vectors
  - identifying implicitly or explicitly characteristic vectors in a bag
- Witness assumption: if any single pattern in a bag is positive, the bag inherits a positive label
- notation: $p = \{x_1, \ldots, x_N\}$ and $p' = \{x'_1, \ldots, x'_{N'}\}$
Related Work

- Multi-instance kernels (Gärtner, et al., 2002)

\[ k(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \kappa(x_i, x'_j)^\rho \]  
\hspace{1cm} (1)

- Bhattacharyya Kernel (Kondor and Jebara, 2003)

\[ k(p, p') = \int \sqrt{p(x)} \sqrt{p'(x)} \, dx \]  
\hspace{1cm} (2)

\( p(x) \) in this case is a Gaussian distribution computed from kernel-PCA.
Related Work (continued)

- Matching Kernel (Wallraven, Caputo, and Graf, 2003)

\[
k(p, p') = \frac{1}{2} \left( \hat{k}(p, p') + \hat{k}(p', p) \right)
\]

(3)

\[
\hat{k}(p, p') = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in 1, \ldots, N'} \kappa(x_i, x_j')
\]

(4)

- Pyramid Match Kernel (Grauman and Darrell, 2005)

\[
k(p, p') = \frac{\hat{k}(p, p')}{\sqrt{\hat{k}(p, p) \cdot \hat{k}(p', p')}}
\]

(5)

\[
\hat{k}(p, p') = \sum_{i=0}^{\lceil \log 2r \rceil} \alpha_i \left( |H_{p,i} \cap H_{p,i'}| - |H_{p,i-1} \cap H_{p',i-1}| \right)
\]

(6)
Kernels between distributions

- A general form

\[ k(p, p') = \int p(x)p'(x)dx = E_p[p'(x)] = E_{p'}[p(x)] \]  

(7)

- Gaussian distribution (spherical)

\[ \int_{\mathbb{R}^d} p(x)p'(x)dx = \frac{1}{(4\pi\sigma^2)^{D/2}}e^{-\|\mu' - \mu\|^2/(4\sigma^2)} \]  

(8)

This is a Gaussian kernel to a constant factor.

- Gaussian distribution in general case \((\rho \neq 1, \text{ arbitrary covariance matrix})\) is also known in closed form
Kernel Density Estimation Over Bags

\[
k(p, p') = \int \left( \frac{1}{N} \sum_{i=1}^{N} \kappa(x_i, x) \right) \cdot \left( \frac{1}{N'} \sum_{j=1}^{N'} \kappa(x'_j, x) \right) \, dx
\]

\[
= \frac{1}{N \cdot N'} \frac{1}{(4\pi\sigma^2)^{D/2}} \sum_{i=1}^{N} \sum_{j=1}^{N'} e^{-\|x'_j - x_i\|^2 / (4\sigma^2)}
\]

\[
\propto \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \kappa(x_i, x'_j)
\]
Conformal Kernels (Amari and Wu, 1999)

Metric tensor induced by mapping, $\varphi$

$$g_{ij}(x) = \left( \frac{\partial}{\partial x_i} \varphi(x) \right) \cdot \left( \frac{\partial}{\partial x_j} \varphi(x) \right)$$ (12)

The volume for in a Riemannian space is defined as

$$dV = \sqrt{g(x)} dx_1 \ldots dx_n$$ (13)

where $g(x) = \det |g_{ij}(x)|$.

$$\tilde{k}(x, x') = c(x)c(x')k(x, x')$$ (14)

$$\tilde{g}_{ij}(x) = c_i(x)c_j(x) + c(x)^2 g_{ij}(x)$$ (15)

where $c_i(x) = \partial c(x)/\partial x_i$. 
Conformal Multi-Instance Kernels

\[ \tilde{\kappa}(x_i, x'_j) = c_\theta(x_i)c_\theta(x'_j)\kappa(x_i, x'_j) \]  \hspace{1cm} (16)

General form:

\[ \tilde{k}(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} c_\theta(x_i)c_\theta(x'_j)\kappa(x_i, x'_j) \]  \hspace{1cm} (17)

where \( \theta \) are parameters of the function \( f \) that can be optimized to maximize discriminability.
Implementation details

\[ c_\theta(x) = \sum_{i=1}^{\theta} \theta_i e^{-\|x - \mu_i\|^2 / 2\sigma^2} \]  

(18)

\[ \tilde{k}(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \left( \sum_{k=1}^{\theta} \theta_k \tilde{k}(x_i, \mu_k) \right) \left( \sum_{l=1}^{\theta} \theta_l \tilde{k}(x'_j, \mu_l) \right) \kappa(x_i, x'_j) \]  

(19)

The \( \mu_i \) are chosen using k-means with the buckshot heuristic. \( |\theta| \) is chosen according to how much computation is available. \( \sigma \) is currently optimized using cross-validation.
Gradient Descent on the Radius-Margin Bound

Algorithm:

1. Initialize $\theta$ to some value
2. Solve for $\alpha^*(\theta)$ using standard SVM algorithm
3. Update the parameters $\theta$ using a gradient step $(\partial R^2 \|w\|^2 / \partial \theta)$
4. Go to step 2 or stop when minimum is reached

Advantage: only requirement is that the kernel be differentiable
Problem: slow as molasses
Optimizing the Trace-Margin Bound

\[ w_{C,\tau}(\alpha, \theta) = \max_{\alpha, \theta} 2\alpha^T e - \alpha^T (G(K_\theta) + \tau I) \alpha \]  

(20)

\[ C \geq \alpha \geq 0, \quad \alpha^T y = 0 \]  

(21)

When \( K_\theta = \sum_{l=1}^{q} \theta_l K_l \), \( \theta > 0 \), we can solve for \( \theta \) using a QCQP or SILP
Diagonalization of conformal transformation

When \( l \neq m \)

\[
\tilde{\kappa}(x_i, \mu_l) \tilde{\kappa}(x'_j, \mu_m) \kappa(x_i, x'_j) \approx 0
\] (22)

\[
k(p, p') \approx \frac{1}{NN'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \left( \sum_{l=1}^{q} \theta_i^2 \tilde{\kappa}(x_i, \mu_l) \tilde{\kappa}(x'_j, \mu_l) \right) \kappa(x_i, x'_j)
\] (23)

\[
= \sum_{l=1}^{q} \theta_i^2 \left( \frac{1}{NN'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \tilde{\kappa}(x_i, \mu_l) \tilde{\kappa}(x'_j, \mu_l) \kappa(x_i, x'_j) \right)
\] (24)
A toy example
## Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>MUSK 1</th>
<th>MUSK 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformal Kernels</td>
<td>90.22</td>
<td>86.96</td>
</tr>
<tr>
<td>Multi-instance SVM</td>
<td>92.4 (IAPR)</td>
<td>89.2 (IAPR)</td>
</tr>
<tr>
<td>EM Discriminative Density</td>
<td>84.8</td>
<td>84.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Elephant</th>
<th>Fox</th>
<th>Tiger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformal Kernels</td>
<td>83.5</td>
<td>61.5</td>
<td>84.5</td>
</tr>
<tr>
<td>Multi-instance SVM</td>
<td>82.2</td>
<td>59.4</td>
<td>84</td>
</tr>
<tr>
<td>EM Discriminative Density</td>
<td>78.3</td>
<td>56.1</td>
<td>72.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TREC 1</th>
<th>TREC 2</th>
<th>TREC 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conformal Kernels</td>
<td>94</td>
<td>76.25</td>
<td>86</td>
</tr>
<tr>
<td>Multi-instance SVM</td>
<td>93.9</td>
<td>84.5</td>
<td>87</td>
</tr>
<tr>
<td>EM Discriminative Density</td>
<td>85.8</td>
<td>84.0</td>
<td>69</td>
</tr>
</tbody>
</table>
Future Work

- Better selection of RBF centers
- Alternate basis for conformal function - e.g. spectral decompositions
- Scaling up to thousands of bags with hundreds of patterns per bag
- Application to Computer Vision applications
- More public datasets
Thank you

- I’ll be glad to answer any questions.
- This work is funded in part by the EC projects CLASS and PerAct.