Conformal Multi-Instance Kernels

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

December 8, 2006

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: $p = \{x_1, \dots, x_N\}$ and $p' = \{x'_1, \dots, x'_{N'}\}$

Related Work

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

$$k(p, p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \kappa(x_i, x'_j)^{\rho} \tag{1}$$

- Bhattacharyya Kernel (Kondor and Jebara, 2003)

$$k(p, p') = \int \sqrt{p(x)}\sqrt{p'(x)}dx \tag{2}$$

$p(x)$ in this case is a Gaussian distribution computed from kernel-PCA.

Related Work (continued)

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

$$k(p,p') = \frac{1}{2}\left(\hat{k}(p,p') + \hat{k}(p',p)\right) \tag{3}$$

$$\hat{k}(p,p') = \frac{1}{N}\sum_{i=1}^{N} \max_{j\in 1,\ldots,N'} \kappa(x_i, x'_j) \tag{4}$$

- Pyramid Match Kernel (Grauman and Darrell, 2005)

$$k(p,p') = \frac{\hat{k}(p,p')}{\sqrt{\hat{k}(p,p)\cdot\hat{k}(p',p')}} \tag{5}$$

$$\hat{k}(p,p') = \sum_{i=0}^{\lceil \log 2r \rceil} \alpha_i \left(|H_{p,i} \cap H_{p,i}| - |H_{p,i-1} \cap H_{p',i-1}|\right) \tag{6}$$

Kernels between distributions

- A general form

$$k(p,p') = \int p(x)p'(x)dx = E_p[p'(x)] = E_{p'}[p(x)] \tag{7}$$

- Gaussian distribution (spherical)

$$\int_{\mathbb{R}^D} p(x)p'(x)dx = \frac{1}{(4\pi\sigma^2)^{D/2}} e^{-\|\mu'-\mu\|^2/(4\sigma^2)} \tag{8}$$

  This is a Gaussian kernel to a constant factor.

- Gaussian distribution in general case ($\rho \neq 1$, arbitrary covariance matrix) is also known in closed form

Kernel Density Estimation Over Bags

$$k(p,p') = \int \left(\frac{1}{N}\sum_{i=1}^{N}\kappa(x_i,x)\right)\cdot\left(\frac{1}{N'}\sum_{j=1}^{N'}\kappa(x'_j,x)\right)dx \tag{9}$$

$$= \frac{1}{N\cdot N'}\frac{1}{(4\pi\sigma^2)^{D/2}}\sum_{i=1}^{N}\sum_{j=1}^{N'}e^{-\|x'_j-x_i\|^2/(4\sigma^2)} \tag{10}$$

$$\propto \frac{1}{N\cdot N'}\sum_{i=1}^{N}\sum_{j=1}^{N'}\kappa(x_i,x'_j) \tag{11}$$

Conformal Kernels (Amari and Wu, 1999)

Metric tensor induced by mapping, φ

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

The volume for in a Riemannian space is defined as

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

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

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

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

where $c_i(x) = \partial c(x)/\partial x_i$.

Conformal Multi-Instance Kernels

$$\tilde{\kappa}(x_i, x'_j) = c_\theta(x_i) c_\theta(x'_j) \kappa(x_i, x'_j) \tag{16}$$

General form:

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

where θ are parameters of the function f that can be optimized to maximize discriminability.

Implementation details

$$c_\theta(x) = \sum_{i=1}^{|\theta|} \theta_i e^{||x-\mu_i||^2/2\sigma^2} \tag{18}$$

$$\tilde{k}(p,p') = \frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \left(\sum_{k=1}^{|\theta|} \theta_k \tilde{\kappa}(x_i, \mu_k) \right) \left(\sum_{l=1}^{|\theta|} \theta_l \tilde{\kappa}(x'_j, \mu_l) \right) \kappa(x_i, x'_j) \tag{19}$$

μ_i are chosen using k-means with the buckshot heuristic. $|\theta|$ is chosen according to how much computation is available. σ is currently optimized using cross-validation.

Gradient Descent on the Radius-Margin Bound

Algorithm:

- 1. Initialize θ to some value
- 2. Solve for $\alpha^*(\theta)$ using standard SVM algorithm
- 3. Update the parameters θ 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

$$w_{C,\tau}(\alpha, \theta) = \max_{\alpha,\theta} 2\alpha^T e - \alpha^T \left(G(K_\theta) + \tau I\right) \alpha \tag{20}$$

$$C \geq \alpha \geq 0, \ \alpha^T y = 0 \tag{21}$$

When $K_\theta = \sum_{l=1}^{q} \theta_l K_l, \quad \theta > 0$, we can solve for θ using a QCQP or SILP

Diagonalization of conformal transformation

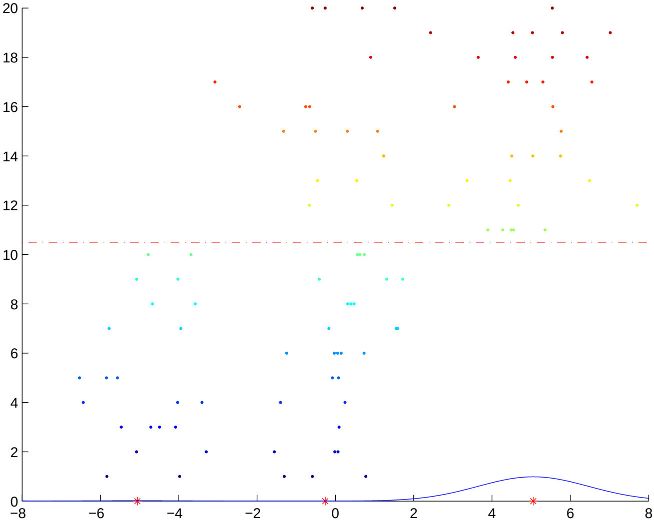
When $l \neq m$

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

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

$$= \sum_{l=1}^{q} \theta_l^2 \left(\frac{1}{N \cdot N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \tilde{\kappa}(x_i, \mu_l)\tilde{\kappa}(x'_j, \mu_l)\kappa(x_i, x'_j) \right) \tag{24}$$

A toy example



Experimental Results

	MUSK 1	MUSK 2
Conformal Kernels	90.22	86.96
Multi-instance SVM	92.4 (IAPR)	89.2 (IAPR)
EM Discriminative Density	84.8	84.9

	Elephant	Fox	Tiger
Conformal Kernels	83.5	61.5	84.5
Multi-instance SVM	82.2	59.4	84
EM Discriminative Density	78.3	56.1	72.1

	TREC 1	TREC 2	TREC 3
Conformal Kernels	94	76.25	86
Multi-instance SVM	93.9	84.5	87
EM Discriminative Density	85.8	84.0	69

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.