Learning Kernels for variants of Normalized Cuts: Convex Relaxations and applications

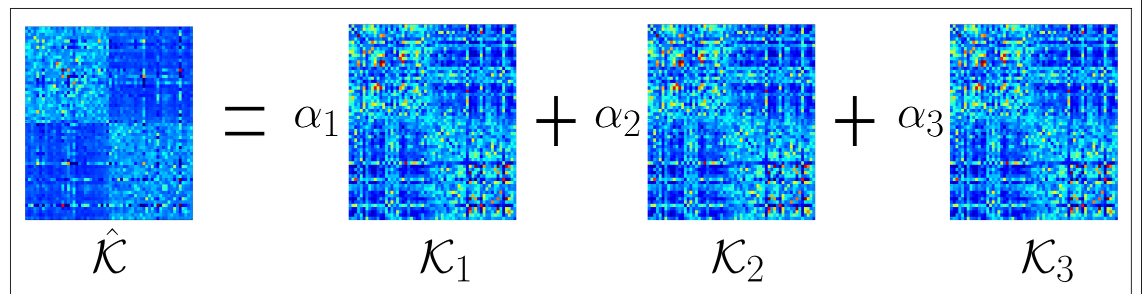
Lopamudra Mukherjee[1], Vikas Singh[2], Jiming Peng[3], Chris Hinrichs[2]

[1]Mathematics & Computer Science
University of Wisconsin Whitewater
mukherjl@uww.edu

[2]Biostatistics & Med. Info., Computer Sc.
University of Wisconsin – Madison
{vsingh,hinrichs}@cs.wisc.edu

[3]Industrial and Enterprise Systems Eng.
University of Illinois Urbana Champaign
pengj@uiuc.edu

Problem: Learn a weighted combination of basis kernels from training data to produce *N-Cuts favorable partitions*



Motivation

1. Choice of similarity measure or kernel for clustering and classification is typically user dependent
2. Kernel engineering may be non-obvious or difficult
3. **Rather,** learn weighted combination of diverse kernels

Advantages

1. Seamless incorporation of heterogenous kernels (e.g., diverse features)
2. Explicitly solve for *the* mixture of kernels that leads to the best separation of classes
3. Kernel selection, if many features are uninformative

Problem Statement

Given: d basis kernels: $\mathcal{K}^1, \cdots, \mathcal{K}^d$ each of size $n \times n$
Given: A training set $\mathcal{X}$ with known partitions
Find: Sub-kernel weights $\alpha = (\alpha_1, \alpha_2 \cdots \alpha_d)^T$ where the combined kernel $\mathcal{K}^\alpha = \sum_{l=1}^d \alpha_l \mathcal{K}^l$ maximizes

$$f(\alpha) = \sum_{t=1}^{\hat{k}} \frac{\sum_{p,q \in \mathcal{V}_t} K^\alpha_{pq}}{\sum_{p \in \mathcal{V}_t, q \notin \mathcal{V}_t} K^\alpha_{pq}}$$

where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \cdots \cup \mathcal{V}_{\hat{k}}$ are the different classes in training data, $\mathcal{X}$.

Related Work

- Target Similarity (*Meila 2001*)
- Target Eigenvector (*Cour 2005*)
- Spectral Semi-supervised Learning (*Shortreed 2006*)
- Spectral learning via orthogonal projection (*Bach 2006*)
- Much recent work on Multi-kernel learning (*Bach 2004, Rakotomamonjy 2008, Sonnenburg 2006,2010, Varma 2008, Gehler 2008,2009*)

Our Approach: Key Points

Departs from other methods: avoids spectral relaxation, avoids descent on functions of eigen-vectors Rather,

- Proceed with original discrete specification of N-cuts type functions
- Rearrange and express as a non-convex QP; then relax into a (small sized) SDP
- Final running time depends only on the **number of kernels**, and *not* the size of training set.

Model

(measure of inter-class similarity)
$\mathbf{U} = [u(t,l)] \in R^{\hat{k} \times d}$ where $u(t,l) = \sum_{p \in V_t, q \notin V_t} \mathcal{K}^l_{pq}$ and
(measure of intra-class similarity)
$\mathbf{V} = [v(t,l)] \in R^{\hat{k} \times d}$ where $v(t,l) = \sum_{p,q \in V_t} \mathcal{K}^l_{pq}$
Expressing our objective in terms of **U** and **V**:

$$\max_\alpha \quad f(\alpha) = \sum_{t=1}^{\hat{k}} \frac{\sum_{l=1}^d v(t,l)\alpha_l}{\sum_{l=1}^d u(t,l)\alpha_l}$$
$$\text{subject to} \quad \sum_{l=1}^d \alpha_l = 1, \alpha_l \geq 0$$

For two classes denote:

$$\hat{v}(l) = \sum_{t=1}^2 v(t,l) \;\text{ and }\; \hat{u}(l) = u(1,l) = u(2,l),$$

Substituting,

$$\max_\alpha f(\alpha) = \max \frac{\hat{v}^T\alpha}{\hat{u}^T\alpha} = \min \frac{\hat{u}^T\alpha}{\hat{v}^T\alpha} \quad \text{s.t.} \sum_{l=1}^d \alpha_l = 1, \alpha_l \geq 0.$$

$\mathcal{X} = \{X^{(1)}, \cdots, \mathcal{X}^{(N)}\}$ comes with "correct" partition(s)
Create $\hat{u}_{(j)}$ and $\hat{v}_{(j)}$ for each training example $x_j \in \mathcal{X}$.
Then we obtain a function of multiple ratios,

$$\max_\alpha f(\alpha) = \sum_{j \in X} \frac{\hat{u}_j^T\alpha}{\hat{v}_j^T\alpha}$$
$$\text{subject to} \sum_{l=1}^d \alpha_l = 1, \quad \alpha_l \geq 0.$$

Towards Convex Relaxations

Create a set of training example "pairs" for $\mathcal{X}$ as

$$\Phi = \{(g,h) | \mathcal{X}^{(g)}, \mathcal{X}^{(h)} \in \mathcal{X}, g \neq h\}$$

continued...

Then the objective is equivalent to,

$$\min \sum_{j \in X} \frac{\hat{u}_j^T\alpha}{\hat{v}_j^T\alpha}(|X|-1) = \min \sum_{(g,h)\in\Phi} \frac{\hat{u}_g^T\alpha}{\hat{v}_g^T\alpha} + \frac{\hat{u}_h^T\alpha}{\hat{v}_h^T\alpha}$$
$$= \min \sum_{(g,h)\in\Phi} \frac{\alpha^T(\hat{u}_g\hat{v}_h^T + \hat{u}_h\hat{v}_g^T)\alpha}{\alpha^T\hat{v}_g\hat{v}_h^T\alpha}$$
$$= \min \sum_{(g,h)\in\Phi} \frac{\alpha^T A_{gh}\alpha}{\alpha^T B_{gh}\alpha}$$

Problem: Multiple ratio optimization is difficult to solve
Strategy: Minimizes the gap between the numerator and the denominator (from single ratio optimization)

$$\textbf{(Minimize Gap)} \quad \min \sum_{g \neq h} \delta_{gh}$$
$$\text{subject to} \quad \alpha^T(A_{gh} - B_{gh})\alpha \leq \delta_{gh}$$
$$\sum_{l=1}^d \alpha_l = 1$$

Using $\mathcal{J}_{gh} = (A_{gh} - B_{gh})$, we obtain a Standard Quadratic Program (StQP)

$$\textbf{(StQP)} \quad \min \sum_{g \neq h} \alpha^T \mathcal{J}_{gh}\alpha$$
$$\text{subject to} \quad \sum_{l=1}^d \alpha_l = 1, \alpha \geq 0$$

SDP relaxations

Let $\mathcal{J} = \sum_{g \neq h} \mathcal{J}_{gh}$ and $Q = (\mathcal{J} + \mathcal{J}^T)/2$,

$$(\text{SDP}_1) \min \quad \text{tr}(QZ)$$
$$\text{subject to} \quad \sum_{l=1}^d \sum_{l'=1}^d Z_{ll'} = 1,$$
$$Z \succeq 0, \quad Z \geq 0.$$

Rounding algorithm

Step 1. For the optimal solution $\mathbf{Z}^*$ to problem (SDP_1), construct a vector by the following procedure

$$\alpha_i^* = \sqrt{\mathbf{Z}^*_{ii}}, \quad i = 1, \cdots, n.$$

Step 2. Rescale α^* to make it feasible for (StQP).

Theorem 1. *Suppose that the optimal solution of problem (SDP_1), i.e., $\mathbf{Z}^*$ has only positive elements. Then, the proposed algorithm will provide an optimal solution to the StQP.*

Theorem 2. *If $d \leq 3$, the optimal solution of problem (SDP_1) can be achieved at a rank one matrix* **Z**.

Caltech 101 - Object Categorization

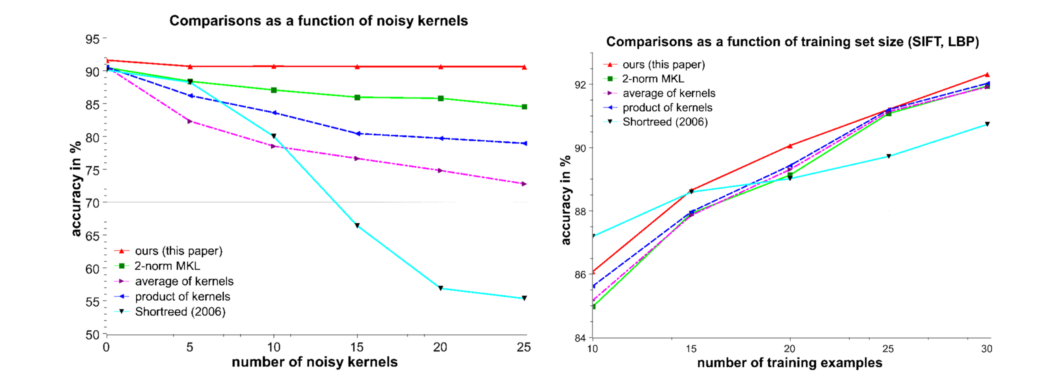


Fig. Performance on Caltech101 for 2-class for performance w.r.t. introduction of noisy kernels (left) and different feature types (right).

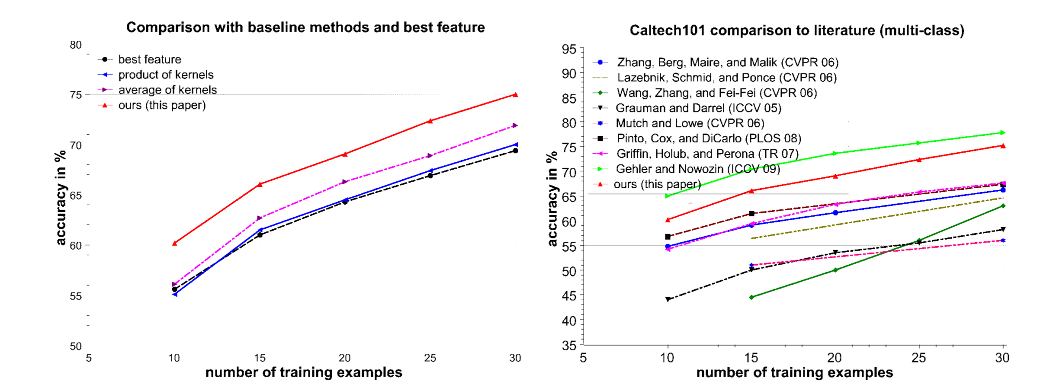


Fig. Performance on Caltech101 for the 102-class setting for performance w.r.t. baseline methods and best feature (left), other algorithms from the literature (right).

ADNI Brain Imaging dataset

Classification of Alzheimer's disease subjects and controls using kernels from MR image volumes, FDG-PET images, and cognitive/clinical biomarkers.

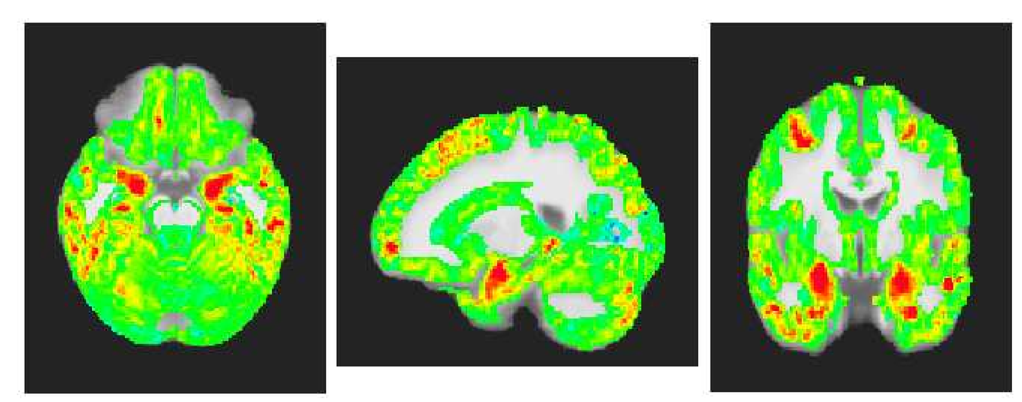


Fig. Brain regions selected when using gray matter probabilities derived from MR images.

Method	accuracy	AUC
ours (this paper)	84.61%	0.9149
2-norm MKL	83.61%	0.9130
average of kernels	83.44%	0.9114

Fig. Summary of accuracy of our method, 2-norm MKL, and average of kernels on ADNI data.