Localized Multiple Kernel Learning

Mehmet Gönen   Ethem Alpaydın

Department of Computer Engineering
Boğaziçi University, İstanbul

July 7, 2008

ICML 2008 25^{th} International Conference on Machine Learning

Outline

1. Introduction and Motivation
2. Localized Multiple Kernel Learning
3. Discussions
4. Experiments
5. Conclusions

Introduction

- Single kernel learning

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b$$

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i \underbrace{\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}_i) \rangle}_{K(\boldsymbol{x}, \boldsymbol{x}_i)} + b$$

- Multiple kernel learning

$$f(\boldsymbol{x}) = \sum_{m=1}^{p} \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}) \rangle + b$$

$$f(\boldsymbol{x}) = \sum_{m=1}^{p} \eta_m \sum_{i=1}^{n} \alpha_i y_i \underbrace{\langle \Phi_m(\boldsymbol{x}), \Phi_m(\boldsymbol{x}_i) \rangle}_{K_m(\boldsymbol{x}, \boldsymbol{x}_i)} + b$$

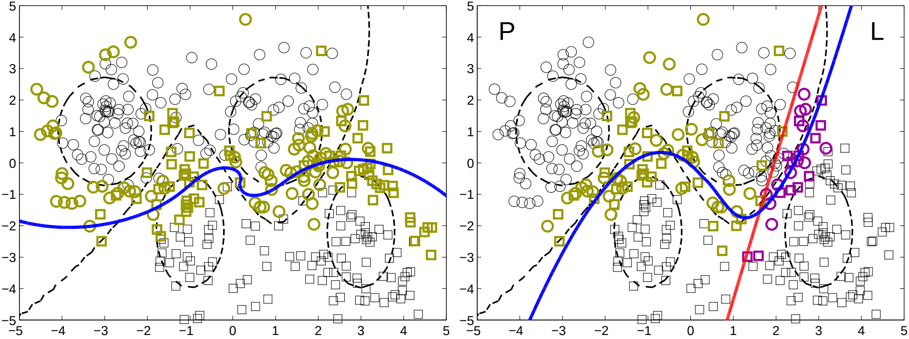
Motivation

$$f(\boldsymbol{x}) = \sum_{m=1}^{p} \eta_m \sum_{i=1}^{n} \alpha_i y_i K_m(\boldsymbol{x}, \boldsymbol{x}_i) + b$$

- Unweighted sum (Pavlidis et al., 2001; Moguerza et al., 2004)
  - $\eta_m = 1 \quad \forall m$
- Weighted sum (Bach et al., 2004; Lanckriet et al., 2004b; Sonnenburg et al., 2006)
  - $\sum_{m=1}^{p} \eta_m = 1$ and $\eta_m \geq 0 \quad \forall m$
- Generative model (Lewis et al., 2006)
- Compositional method (Lee et al., 2007)
- Localized multiple kernel learning
  - $\eta_m(\boldsymbol{x}|\boldsymbol{\Theta})$

Motivation

- Linear and second degree polynomial kernels



Mathematical Model

$$f(\boldsymbol{x}) = \sum_{m=1}^{p} \eta_m(\boldsymbol{x}|\boldsymbol{\Theta}) \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}) \rangle + b$$

$$\begin{aligned}
\text{min} \quad & \frac{1}{2} \sum_{m=1}^{p} \|\boldsymbol{w}_m\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{w.r.t.} \quad & \boldsymbol{w}_m, b, \boldsymbol{\xi}, \boldsymbol{\Theta} \\
\text{s.t.} \quad & y_i \left(\sum_{m=1}^{p} \eta_m(\boldsymbol{x}_i|\boldsymbol{\Theta}) \langle \boldsymbol{w}_m, \Phi_m(\boldsymbol{x}_i) \rangle + b \right) \geq 1 - \xi_i \quad \forall i \\
& \xi_i \geq 0 \quad \forall i
\end{aligned}$$

- Not convex due to gating model
- Two-step alternate optimization algorithm, similar to Rakotomamonjy et al. (2007)

Kernel-Based Learning (Step 1)

$$L_D = \frac{1}{2}\sum_{m=1}^{p} \|\boldsymbol{w}_m\|^2 + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\xi_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i \left(\sum_{m=1}^{p} \eta_m(\boldsymbol{x}_i|\boldsymbol{\Theta})\langle\boldsymbol{w}_m, \Phi_m(\boldsymbol{x}_i)\rangle + b\right)$$

$$\frac{\partial L_D}{\partial \boldsymbol{w}_m} \Rightarrow \boldsymbol{w}_m = \sum_{i=1}^{n}\alpha_i y_i \eta_m(\boldsymbol{x}_i|\boldsymbol{\Theta})\Phi_m(\boldsymbol{x}_i) \quad \forall m$$

$$\frac{\partial L_D}{\partial b} \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial L_D}{\partial \xi_i} \Rightarrow C = \alpha_i + \beta_i \quad \forall i$$

Kernel-Based Learning (Step 1)

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{w.r.t. } \boldsymbol{\alpha}$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0 \quad \forall i$$

- *locally combined kernel matrix*

$$K_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{p} \eta_m(\boldsymbol{x}_i|\boldsymbol{\Theta}) \underbrace{\langle \Phi_m(\boldsymbol{x}_i), \Phi_m(\boldsymbol{x}_j) \rangle}_{K_m(\boldsymbol{x}_i, \boldsymbol{x}_j)} \eta_m(\boldsymbol{x}_j|\boldsymbol{\Theta})$$

Gating Model Learning (Step 2)

Gating Model $\eta_m(\boldsymbol{x}|\boldsymbol{\Theta})$

Update Step $\boldsymbol{\Theta} \Leftarrow \boldsymbol{\Theta} - \mu \dfrac{\partial J(\eta)}{\partial \boldsymbol{\Theta}}$

$$J(\eta) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Linear gating model with soft-max

$$\eta_m(\boldsymbol{x}|\boldsymbol{\Theta}) = \frac{\exp(\langle \boldsymbol{v}_m, \boldsymbol{x} \rangle + v_{m0})}{\sum\limits_{k=1}^{p} \exp(\langle \boldsymbol{v}_k, \boldsymbol{x} \rangle + v_{k0})} \quad \text{where} \quad \boldsymbol{\Theta} = \{\boldsymbol{v}_1, v_{10}, \ldots, \boldsymbol{v}_p, v_{p0}\}$$

Complete Algorithm

LMKL with linear gating model

```
Initialize $\boldsymbol{v}_m$ and $v_{m0}$ to small random numbers for $m = 1, \ldots, p$
**repeat**
    Calculate $K_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with gating model
    Solve canonical SVM with $K_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j)$
    $v_{m0}^{(t+1)} \Leftarrow v_{m0}^{(t)} - \mu^{(t)} \frac{\partial J(\eta)}{\partial v_{m0}}$ for $m = 1, \ldots, p$
    $\boldsymbol{v}_m^{(t+1)} \Leftarrow \boldsymbol{v}_m^{(t)} - \mu^{(t)} \frac{\partial J(\eta)}{\partial \boldsymbol{v}_m}$ for $m = 1, \ldots, p$
**until** convergence
```

- After finding $\boldsymbol{\alpha}$ and $\boldsymbol{\Theta}$

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{m=1}^{p} \alpha_i y_i \eta_m(\boldsymbol{x}|\boldsymbol{\Theta}) K_m(\boldsymbol{x}, \boldsymbol{x}_i) \eta_m(\boldsymbol{x}_i|\boldsymbol{\Theta}) + b$$

Discussions

- Mixture of Experts (MoE) (Jacobs et al., 1991)
    - LMKL with multiple linear kernels is similar to MoE.

- Mixture of SVMs (Collobert et al., 2001)
    - LMKL couples SVM training and clustering.

- LMKL can be generalized for regression and one-class classification.

Discussions

- Computational complexity
    - training complexity
        - complexity of canonical SVM solver
        - number of iterations
    - testing complexity
        - number of support vectors
        - gating model outputs

- Knowledge extraction
    - MKL extracts global importances of kernels.
    - LMKL extracts local importances of kernels.

Experiments

- Algorithm is implemented with C++ and MOSEK.

- 2/3 for training, 1/3 for test
- 5×2 cross-validation with stratification
- C values from $\{0.01,\ 0.1,\ 1,\ 10,\ 100\}$

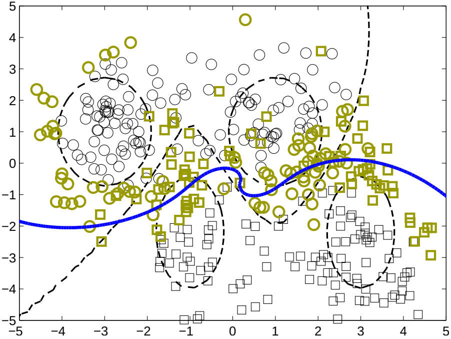
- We use three kernels:

$$K_L(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

$$K_P(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + 1)^2$$

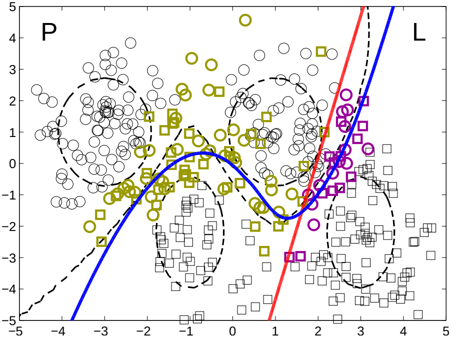
$$K_G(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / s^2\right) \quad \text{where} \quad s = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \boldsymbol{x}_{nn(i)}\right\|$$

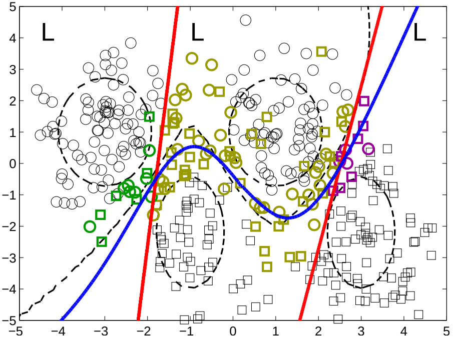
Combining Linear and Polynomial Kernels



- MKL $\Rightarrow 0.3K_L + 0.7K_P$

Combining Three Linear Kernels

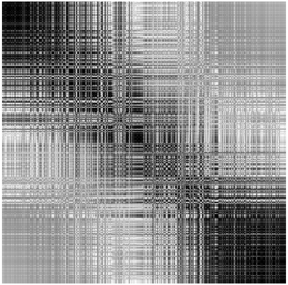


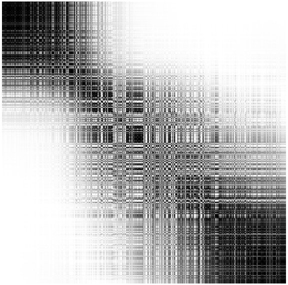


Effect of Locality on Combined Kernel

$$K_\eta(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}) = \begin{pmatrix} K_1(\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_1) & 0 & \dots & 0 \\ 0 & K_2(\boldsymbol{\mathcal{X}}_2, \boldsymbol{\mathcal{X}}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & K_p(\boldsymbol{\mathcal{X}}_p, \boldsymbol{\mathcal{X}}_p) \end{pmatrix}$$

- (K_L-K_P) combination





Results on UCI Data Sets

	SVM				MKL		LMKL	
	K_P		K_G		$(K_P\text{-}K_G)$		$(K_P\text{-}K_G)$	
Data Set	Acc.	SV	Acc.	SV	Acc.	SV	Acc.	SV
BANANA	56.51	75.99	83.57	92.67	81.99	93.39	83.84	83.97
GERMANNUMERIC	71.80	54.17	68.65	58.44	73.32	84.89	73.92	80.90
HEART	72.78	73.89	77.67	79.11	75.78	87.89	79.44	**81.44**
IONOSPHERE	91.54	38.55	94.36	61.71	93.68	64.10	93.33	53.33
LIVERDISORDER	60.35	69.83	64.26	74.43	63.39	93.57	64.87	92.52
PIMA	66.95	24.26	71.91	74.26	72.62	80.39	72.89	**73.63**
RINGNORM	70.66	53.91	98.82	40.68	98.86	57.68	98.69	56.69
SONAR	65.29	67.54	72.71	73.48	80.29	89.57	79.57	90.00
SPAMBASE	84.18	47.92	79.80	49.50	90.46	57.47	91.41	58.24
WDBC	88.73	27.11	94.44	54.74	95.50	58.11	95.98	**42.95**
5×2 cv Paired F Test (W-T-L)							0-10-0	3-7-0
Direct Comparison (W-T-L)							7-0-3	8-0-2
Wilcoxon's Signed Rank Test (W/T/L)							T	W

Results on UCI Data Sets

	SVM		LMKL	
	K_L		$(K_L$-K_L-$K_L)$	
Data Set	Acc.	SV	Acc.	SV
Banana	59.18	93.99	**81.39**	**54.03**
Germannumeric	74.58	97.09	75.09	**57.21**
Heart	78.33	67.00	77.00	**58.44**
Ionosphere	86.15	36.58	87.86	49.06
Liverdisorder	64.78	85.65	64.78	78.35
Pima	70.04	100.00	**73.98**	**53.09**
Ringnorm	76.91	78.68	78.92	**52.53**
Sonar	73.86	68.41	77.14	60.43
Spambase	85.98	77.43	**91.18**	**34.93**
Wdbc	95.08	**13.11**	94.34	21.89
5×2 cv Paired F Test		(W-T-L)	3-7-0	6-3-1
Direct Comparison		(W-T-L)	7-1-2	8-0-2
Wilcoxon's Signed Rank Test		(W/T/L)	W	T

Results on Bioinformatics Data Sets

- Two translation initiation site data sets (Pedersen & Nielsen, 1997)

	SVM				MKL		LMKL	
	K_P		K_G		$(K_P$-$K_G)$		$(K_P$-$K_G)$	
Data Set	Acc.	SV	Acc.	SV	Acc.	SV	Acc.	SV
ARABIDOPSIS	74.30	68.08	77.41	42.36	80.10	89.96	80.82	**65.41**
VERTEBRATES	75.50	68.54	75.72	41.64	78.67	90.46	77.67	68.14

	SVM		LMKL	
	K_L		$(K_L$-K_L-$K_L)$	
Data Set	Acc.	SV	Acc.	SV
ARABIDOPSIS	74.30	99.64	**81.29**	**68.66**
VERTEBRATES	75.50	99.02	**78.69**	**67.41**

Conclusions

- Introduces a localized multiple kernel learning framework
  - a parametric gating model
  - a kernel-based learning algorithm
- Coupled optimization with a two-step alternate optimization procedure
- Allows using multiple copies of the same kernel

- On experiments
  - different kernels
    - accuracy ($\approx$) support vectors ($\Downarrow$)
  - same kernels
    - accuracy ($\Uparrow$) support vectors ($\Downarrow$)

Conclusions

- Kernel-based gating model
  - Use one or a combination of $\Phi_m(\boldsymbol{x})$
  - Nonvectorial data

$$\eta_m(\boldsymbol{x}|\boldsymbol{\Theta}) = \frac{\exp(\langle \boldsymbol{v}_m, \Phi(\boldsymbol{x})\rangle + v_{m0})}{\sum\limits_{k=1}^{p} \exp(\langle \boldsymbol{v}_k, \Phi(\boldsymbol{x})\rangle + v_{k0})}$$

- MATLAB implementation is available at

  `http://www.cmpe.boun.edu.tr/~gonen/lmkl`