Localized Multiple Kernel Learning

Mehmet Gönen  Ethem Alpaydın

Department of Computer Engineering
Boğaziçi University, İstanbul

July 7, 2008

ICML 2008 25th 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

\[ f(x) = \langle w, \Phi(x) \rangle + b \]
\[ f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \Phi(x), \Phi(x_i) \rangle + b \]
\[ K(x, x_i) \]

- Multiple kernel learning

\[ f(x) = \sum_{m=1}^{p} \langle w_m, \Phi_m(x) \rangle + b \]
\[ f(x) = \sum_{m=1}^{p} \eta_m \sum_{i=1}^{n} \alpha_i y_i \langle \Phi_m(x), \Phi_m(x_i) \rangle + b \]
\[ K_m(x, x_i) \]
Motivation

\[ f(x) = \sum_{m=1}^{p} \eta_m \sum_{i=1}^{n} \alpha_i y_i K_m(x, x_i) + b \]

- Unweighted sum (Pavlidis et al., 2001; Moguerza et al., 2004)
  - \( \eta_m = 1 \ \forall m \)

- Weighted sum (Bach et al., 2004; Lanckriet et al., 2004b; Sonnenburg et al., 2006)
  - \( \sum_{m=1}^{p} \eta_m = 1 \) and \( \eta_m \geq 0 \ \forall m \)

- Generative model (Lewis et al., 2006)
- Compositional method (Lee et al., 2007)
- Localized multiple kernel learning
  - \( \eta_m(x|\Theta) \)
Motivation

- Linear and second degree polynomial kernels
Mathematical Model

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

\[ \min \frac{1}{2} \sum_{m=1}^{p} \| \mathbf{w}_m \|^2 + C \sum_{i=1}^{n} \xi_i \]

w.r.t. \( \mathbf{w}_m, b, \xi, \Theta \)

s.t. \( y_i \left( \sum_{m=1}^{p} \eta_m(\mathbf{x}_i|\Theta) \langle \mathbf{w}_m, \Phi_m(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i \quad \forall i \)

\( \xi_i \geq 0 \quad \forall i \)

- Not convex due to gating model
- Two-step alternate optimization algorithm, similar to Rakotomamonjy et al. (2007)
Kernel-Based Learning (Step 1)

\[ L_D = \frac{1}{2} \sum_{m=1}^{p} \|w_m\|^2 + \sum_{i=1}^{n} (C - \alpha_i - \beta_i) \xi_i + \sum_{i=1}^{n} \alpha_i \]

\[ - \sum_{i=1}^{n} \alpha_i y_i \left( \sum_{m=1}^{p} \eta_m(x_i|\Theta) \langle w_m, \Phi_m(x_i) \rangle + b \right) \]

\[ \frac{\partial L_D}{\partial w_m} \Rightarrow w_m = \sum_{i=1}^{n} \alpha_i y_i \eta_m(x_i|\Theta) \Phi_m(x_i) \quad \forall m \]

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

\[ \frac{\partial L_D}{\partial \xi_i} \Rightarrow C = \alpha_i + \beta_i \quad \forall i \]
Kernel-Based Learning (Step 1)

\[
\begin{align*}
\text{max} & \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_\eta(x_i, x_j) \\
\text{w.r.t.} & \quad \alpha \\
\text{s.t.} & \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \quad C \geq \alpha_i \geq 0 \quad \forall i
\end{align*}
\]

- **locally combined kernel matrix**

\[
K_\eta(x_i, x_j) = \sum_{m=1}^{p} \eta_m(x_i | \Theta) \left( \langle \Phi_m(x_i), \Phi_m(x_j) \rangle \right) \eta_m(x_j | \Theta) K_m(x_i, x_j)
\]
Gating Model Learning (Step 2)

Gating Model \( \eta_m(x|\Theta) \)

Update Step \( \Theta \leftarrow \Theta - \mu \frac{\partial J(\eta)}{\partial \Theta} \)

\[
J(\eta) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_\eta(x_i, x_j)
\]

- Linear gating model with soft-max

\[
\eta_m(x|\Theta) = \frac{\exp(\langle v_m, x \rangle + v_m 0)}{\sum_{k=1}^{p} \exp(\langle v_k, x \rangle + v_k 0)} \quad \text{where} \quad \Theta = \{v_1, v_{10}, \ldots, v_p, v_{p0}\}
\]
Complete Algorithm

**LMKL with linear gating model**

1: Initialize $v_m$ and $v_{m0}$ to small random numbers for $m = 1, \ldots, p$
2: repeat
3: Calculate $K_\eta(x_i, x_j)$ with gating model
4: Solve canonical SVM with $K_\eta(x_i, x_j)$
5: $v^{(t+1)}_m \leftarrow v^{(t)}_m - \mu(t) \frac{\partial J(\eta)}{\partial v_{m0}}$ for $m = 1, \ldots, p$
6: $v^{(t+1)}_m \leftarrow v^{(t)}_m - \mu(t) \frac{\partial J(\eta)}{\partial v_{m}}$ for $m = 1, \ldots, p$
7: until convergence

- After finding $\alpha$ and $\Theta$

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

\[
\begin{align*}
K_L(x_i, x_j) &= \langle x_i, x_j \rangle \\
K_P(x_i, x_j) &= (\langle x_i, x_j \rangle + 1)^2 \\
K_G(x_i, x_j) &= \exp \left( - \| x_i - x_j \|^2 / s^2 \right) \quad \text{where} \quad s = \frac{1}{n} \sum_{i=1}^{n} \| x_i - x_{nn(i)} \|
\end{align*}
\]
MKL ⇒ 0.3K_L + 0.7K_P
Combining Three Linear Kernels
Effect of Locality on Combined Kernel

\[
K_\eta(\mathbf{x}, \mathbf{x}) = \begin{pmatrix}
K_1(\mathbf{x}_1, \mathbf{x}_1) & 0 & \ldots & 0 \\
0 & K_2(\mathbf{x}_2, \mathbf{x}_2) & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & K_p(\mathbf{x}_p, \mathbf{x}_p)
\end{pmatrix}
\]

- \((K_L - K_P)\) combination
## Results on UCI Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SVM</th>
<th>MKL</th>
<th>LMKL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_P$</td>
<td>$K_G$</td>
<td>$(K_P-K_G)$</td>
</tr>
<tr>
<td></td>
<td>Acc.</td>
<td>SV</td>
<td>Acc.</td>
</tr>
<tr>
<td>BANANA</td>
<td>56.51</td>
<td>75.99</td>
<td>83.57</td>
</tr>
<tr>
<td>GERMANNUMERIC</td>
<td>71.80</td>
<td>54.17</td>
<td>68.65</td>
</tr>
<tr>
<td>HEART</td>
<td>72.78</td>
<td>73.89</td>
<td>77.67</td>
</tr>
<tr>
<td>IONOSPHERE</td>
<td>91.54</td>
<td>38.55</td>
<td>94.36</td>
</tr>
<tr>
<td>LIVERDISORDER</td>
<td>60.35</td>
<td>69.83</td>
<td>64.26</td>
</tr>
<tr>
<td>PIMA</td>
<td>66.95</td>
<td>24.26</td>
<td>71.91</td>
</tr>
<tr>
<td>RINGNORM</td>
<td>70.66</td>
<td>53.91</td>
<td>98.82</td>
</tr>
<tr>
<td>SONAR</td>
<td>65.29</td>
<td>67.54</td>
<td>72.71</td>
</tr>
<tr>
<td>SPAMBASE</td>
<td>84.18</td>
<td>47.92</td>
<td>79.80</td>
</tr>
<tr>
<td>WDBC</td>
<td>88.73</td>
<td>27.11</td>
<td>94.44</td>
</tr>
</tbody>
</table>

| 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 |

Gönen and Alpaydın (Boğaziçi University)  Localized Multiple Kernel Learning  ICML 2008, Helsinki, Finland
## Results on UCI Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SVM $K_L$</th>
<th>LMKL $(K_L-K_L-K_L)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>SV</td>
</tr>
<tr>
<td><strong>Banana</strong></td>
<td>59.18</td>
<td>93.99</td>
</tr>
<tr>
<td><strong>GermanNumeric</strong></td>
<td>74.58</td>
<td>97.09</td>
</tr>
<tr>
<td><strong>Heart</strong></td>
<td>78.33</td>
<td>67.00</td>
</tr>
<tr>
<td><strong>Ionosphere</strong></td>
<td>86.15</td>
<td>36.58</td>
</tr>
<tr>
<td><strong>LiverDisorder</strong></td>
<td>64.78</td>
<td>85.65</td>
</tr>
<tr>
<td><strong>Pima</strong></td>
<td>70.04</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>RingNorm</strong></td>
<td>76.91</td>
<td>78.68</td>
</tr>
<tr>
<td><strong>Sonar</strong></td>
<td>73.86</td>
<td>68.41</td>
</tr>
<tr>
<td><strong>Spambase</strong></td>
<td>85.98</td>
<td>77.43</td>
</tr>
<tr>
<td><strong>Wdbc</strong></td>
<td>95.08</td>
<td><strong>13.11</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5 × 2 cv Paired F Test</th>
<th>(W-T-L)</th>
<th>Direct Comparison</th>
<th>(W-T-L)</th>
<th>Wilcoxon’s Signed Rank Test</th>
<th>(W/T/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-7-0</td>
<td>7-1-2</td>
<td>8-0-2</td>
<td>W</td>
<td>T</td>
</tr>
</tbody>
</table>
Results on Bioinformatics Data Sets

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

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SVM $K_P$ Acc.</th>
<th>SVM $K_G$ Acc.</th>
<th>MKL $(K_P-K_G)$ Acc.</th>
<th>LMKL $(K_P-K_G)$ Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>74.30</td>
<td>77.41</td>
<td>80.10</td>
<td>80.82</td>
</tr>
<tr>
<td></td>
<td>68.08</td>
<td>42.36</td>
<td>89.96</td>
<td><strong>65.41</strong></td>
</tr>
<tr>
<td>Vertebrates</td>
<td>75.50</td>
<td>75.72</td>
<td>78.67</td>
<td>77.67</td>
</tr>
<tr>
<td></td>
<td>68.54</td>
<td>41.64</td>
<td>90.46</td>
<td>68.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th>SVM $K_L$ Acc.</th>
<th>LMKL $(K_L-K_L-K_L)$ Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>74.30</td>
<td><strong>81.29</strong></td>
</tr>
<tr>
<td></td>
<td>99.64</td>
<td><strong>68.66</strong></td>
</tr>
<tr>
<td>Vertebrates</td>
<td>75.50</td>
<td><strong>78.69</strong></td>
</tr>
<tr>
<td></td>
<td>99.02</td>
<td><strong>67.41</strong></td>
</tr>
</tbody>
</table>
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 (≈) support vectors (↓)
  - same kernels
    - accuracy (↑) support vectors (↓)
Conclusions

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

$$
\eta_m(x|\Theta) = \frac{\exp(\langle v_m, \Phi(x) \rangle + v_{m0})}{\sum_{k=1}^{p} \exp(\langle v_k, \Phi(x) \rangle + v_{k0})}
$$

- MATLAB implementation is available at http://www.cmpe.boun.edu.tr/~gonen/lmkl