Supplement to: An Automated Combination of Kernels for Predicting Protein Subcellular Localization

Cheng Soon Ong[1,2] and Alexander Zien[1,3]

[1] Friedrich Miescher Laboratory, Tübingen, Germany
[2] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3] Fraunhofer Institute FIRST, Berlin, Germany

1 Multiclass Multiple Kernel Learning

This section is provides the detailed derivation for the specific case of soft margin. Refer to [1] for the general formulation.

Corollary 1. *When choosing the hinge loss,* $\ell(t) := C \max(0, 1-t)$*, the optimum* $\mathbf{w}$ *can be computed as*

$$\forall k : \mathbf{w}_k = \sum_i \sum_{u \in \mathcal{Y}} \alpha_{iu} \Phi_k(\mathbf{x}_i, u) \quad ,$$

where $\boldsymbol{\alpha} \in \mathbf{R}^{n \times \mathcal{Y}}$ *is the solution of the quadratically constrained linear program defined by*

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \ & \gamma - \sum_i \alpha_{iy_i} \\
\text{s.t.} \ & \forall i : 0 \leqslant \alpha_{iy_i} \leqslant C \\
& \forall i : \forall u \neq y_i : \alpha_{iu} \leqslant 0 \\
& \forall i : \sum_{u \in \mathcal{Y}} \alpha_{iu} = 0 \qquad \textit{and} \qquad \forall u : \sum_i \alpha_{iu} = 0 \\
& \forall k : \gamma \geqslant \frac{1}{2} \sum_{i,j,u,v} \alpha_{iu} \alpha_{jv} \left\langle \Phi_k(\mathbf{x}_i, u), \Phi_k(\mathbf{x}_j, v) \right\rangle
\end{aligned} \tag{1}$$

Proof. We begin at the primal:

$$\begin{aligned}
\min_{\boldsymbol{\beta}, \mathbf{w}, \mathbf{b}} \ & \frac{1}{2} \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|^2 \\
& + \sum_{i=1}^{n} \max_{u \neq y_i} \ell\left(f_{\mathbf{w},\mathbf{b},\boldsymbol{\beta}}(\mathbf{x}_i, y_i) - f_{\mathbf{w},\mathbf{b},\boldsymbol{\beta}}(\mathbf{x}_i, u)\right) \\
\text{s.t.} \ & \sum_{k=1}^{p} \beta_k = 1, \qquad \text{and} \qquad \forall k : 0 \leqslant \beta_k.
\end{aligned}$$

Consider the hinge loss,

$$\ell(t) := C \max(0, 1-t)$$

where we obtain:

$$\begin{aligned}\min_{\boldsymbol{\beta},\mathbf{w},\mathbf{b}} \; & \frac{1}{2}\sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|^2 + \sum_{i=1}^{n}\xi_i \\ \text{s.t. } & \xi_i = \max_{u\neq y_i} s_{iu}, \qquad s_{iu} \geqslant 0 \\ & s_{iu} \geqslant 1 - f_{\mathbf{w},\mathbf{b},\boldsymbol{\beta}}(\mathbf{x}_i, y_i) + f_{\mathbf{w},\mathbf{b},\boldsymbol{\beta}}(\mathbf{x}_i, u) \\ & \sum_{k=1}^{p}\beta_k = 1, \qquad \text{and} \qquad \forall k: 0 \leqslant \beta_k\end{aligned}$$

Substituting,

$$f_{\mathbf{w},\mathbf{b},\boldsymbol{\beta}}(\mathbf{x}, y) = \sum_{k=1}^{p}\beta_k \left\langle \mathbf{w}_k, \varPhi_k(x,y)\right\rangle + b_y.$$

we get where we obtain:

$$\begin{aligned}\min_{\boldsymbol{\beta},\mathbf{w},\mathbf{b}} \; & \frac{1}{2}\sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|^2 + \sum_{i=1}^{n}\xi_i \\ \text{s.t. } & \xi_i = \max_{u\neq y_i} s_{iu}, \qquad s_{iu} \geqslant 0 \\ & \sum_{k=1}^{p}\beta_k \left\langle \mathbf{w}_k, (\varPsi_{kiu})\right\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu} \\ & \sum_{k=1}^{p}\beta_k = 1, \qquad \text{and} \qquad \forall k: 0 \leqslant \beta_k\end{aligned}$$

Note that the constraints in the second line are quadratic and are non-convex. We substitute $\mathbf{v}_k := \beta_k\mathbf{w}_k$, and since a quadratic over a linear function is convex (i.e. $\frac{x^2}{y}$ is convex), We obtain the following convex optimization problem.

$$\begin{aligned}\min_{\boldsymbol{\beta},\mathbf{v},\mathbf{b}} \; & \frac{1}{2}\sum_{k=1}^{p}\frac{1}{\beta_k}\|\mathbf{v}_k\|^2 + \sum_{i=1}^{n}\xi_i \\ \text{s.t. } & \xi_i \geqslant s_{iu}, \qquad s_{iu} \geqslant 0 \\ & \sum_{k=1}^{p}\left\langle \mathbf{v}_k, \varPsi_{kiu}\right\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu} \\ & \sum_{k=1}^{p}\beta_k = 1, \qquad \text{and} \qquad \forall k: 0 \leqslant \beta_k.\end{aligned}$$

Observe that the constraints are now linear.

The Lagrangian of this is given by

$$\begin{aligned}\mathcal{L} = & \frac{1}{2}\sum_k \frac{1}{\beta_k}\|\mathbf{v}_k\|^2 + C\sum_i \xi_i + \sum_i \sum_{u\neq y_i} \eta_{iu}(s_{iu} - \xi_i) \\ & -\mu_{iu}s_{iu} + \gamma\left(\sum_k \beta_k - 1\right) - \sum_k \epsilon_k\beta_k \\ & + \sum_i \sum_{u\neq y_i} \tilde{\alpha}_{iu}\left(1 - s_{iu} - \sum_k \left\langle \mathbf{v}_k, \varPsi_{kiu}\right\rangle - b_{y_i} + b_u\right),\end{aligned}$$

with Lagrange variables $\tilde{\boldsymbol{\alpha}} \in \mathbf{R}^{m \times \mathcal{Y}}$, $\mathbf{0} \leqslant \boldsymbol{\epsilon}, \mu \in \mathbf{R}^p$, and $\mathbf{0} \leqslant \boldsymbol{\eta} \in \mathbf{R}^{m \times \mathcal{Y}}$.

We find the stationary points by setting the partial derivatives with respect to the primal variables

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = -\frac{1}{2}\frac{1}{\beta_k^2}\|\mathbf{v}_k\|^2 + \gamma - \epsilon_k \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_k} = \frac{1}{\beta_k}\mathbf{v}_k - \sum_i \sum_{u \neq y_i} \tilde{\alpha}_{iu} \left(\Psi_{kiu}\right), \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial b_v} = -\sum_i \delta_{y_i v} \sum_{u \neq y_i} \tilde{\alpha}_{iu} + \sum_i \sum_{u \neq y_i} \delta_{uv} \tilde{\alpha}_{iu} \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \sum_{u \neq y_i} \eta_{iu}, \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial s_{iu}} = \eta_{iu} - \mu_{iu} - \tilde{\alpha}_{iu} \tag{6}$$

to zero. Substituting 6 into the Lagrangian, the linear terms in s_{iu} cancel out,

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\sum_k \frac{1}{\beta_k}\|\mathbf{v}_k\|^2 + C\sum_i \xi_i - \sum_i \sum_{u \neq y_i} \eta_{iu}\xi_i \\ +\gamma\left(\sum_k \beta_k - 1\right) - \sum_k \epsilon_k \beta_k \\ +\sum_i \sum_{u \neq y_i} \tilde{\alpha}_{iu}\left(1 - \sum_k \langle \mathbf{v}_k, \Psi_{kiu}\rangle - b_{y_i} + b_u\right).\end{aligned}$$

Substituting 5, the linear terms in ξ_i cancel out,

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\sum_k \frac{1}{\beta_k}\|\mathbf{v}_k\|^2 \\ +\gamma\left(\sum_k \beta_k - 1\right) - \sum_k \epsilon_k \beta_k \\ +\sum_i \sum_{u \neq y_i} \tilde{\alpha}_{iu}\left(1 - \sum_k \langle \mathbf{v}_k, \Psi_{kiu}\rangle - b_{y_i} + b_u\right).\end{aligned}$$

Move all the terms involving $\mathbf{v}_k$ to the first row, and substituting 3,

$$\begin{aligned}\mathcal{L} = -\frac{1}{2}\sum_k \beta_k \sum_{i,j}\sum_{u,v} \tilde{\alpha}_{iu}\tilde{\alpha}_{jv}\langle \Psi_{kiu}, \Psi_{kjv}\rangle \\ +\gamma\left(\sum_k \beta_k - 1\right) - \sum_k \epsilon_k \beta_k \\ +\sum_i \sum_{u \neq y_i} \tilde{\alpha}_{iu}\left(1 - b_{y_i} + b_u\right).\end{aligned}$$

Substituting 3 into 2,

$$-\frac{1}{2}\sum_{i,j}\sum_{u,v}\tilde{\alpha}_{iu}\tilde{\alpha}_{jv}\left\langle \Psi_{kiu},\Psi_{kjv}\right\rangle+\gamma-\epsilon_k=0, \tag{7}$$

and hence we obtain

$$\mathcal{L}=-\gamma+\sum_{i}\sum_{u\neq y_i}\tilde{\alpha}_{iu}\left(1-b_{y_i}+b_u\right).$$

Using 4 we obtain the dual Lagrangian,

$$\mathcal{L}=\sum_{i}\sum_{u\neq y_i}\tilde{\alpha}_{iu}-\gamma.$$

From 6, and the fact that $\mu_{iu}\geqslant 0$, we have $\eta_{iu}\geqslant\tilde{\alpha}_{iu}$, and substituting this into 5, we get the constraint

$$\sum_{u\neq y_i}\tilde{\alpha}_{iu}\leqslant C.$$

From 7, and the fact that $\epsilon_k\geqslant 0$, we obtain

$$\frac{1}{2}\sum_{u,v}\tilde{\alpha}_{iu}\tilde{\alpha}_{jv}\left\langle \Psi_{kiu},\Psi_{kjv}\right\rangle\leqslant\gamma.$$

2 Experiments

To be clear about the definitions of the various performance measures used to report multiclass results, we collect all the definitions in this section. Most measures are defined with respect to a particular class, say A, and can be calculated from the corresponding confusion matrix:

		Predicted Label	
		A	$\neg A$
Actual	A	True Positive (TP)	False Negative (FN)
Label	$\neg A$	False Positive (FP)	True Negative (TN)

From the confusion matrix above, we can define the various performance measures, including the Matthews Correlation Coefficient (MCC).

Measure	Formula
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall / Sensitivity	$\frac{TP}{(TP+FN)}$
Specificity	$\frac{TN}{(TN+TP)}$
MCC	$\frac{TP\,TN-FP\,FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$

2.1 Comparison on TargetP dataset

Data	Class	Our Method					TargetP			TargetLoc		
		Accuracy	Precision	Recall	F1-Score	MCC	Recall	SP	MCC	Recall	SP	MCC
plant	ch	96.7 ± 0.4	95.4	84.4	89.5 ± 1.4	87.8 ± 1.5	85	69	72	88	76	78
	mi	95.3 ± 0.4	92.0	97.3	94.6 ± 0.4	90.5 ± 0.8	82	90	77	87	94	84
	SP	97.4 ± 0.3	96.0	94.5	95.2 ± 0.7	93.5 ± 0.9	91	95	90	93	97	93
	OT	95.6 ± 0.3	87.3	86.7	86.9 ± 1.4	84.3 ± 1.6	85	78	77	92	84	86
nonplant	mi	96.9 ± 0.2	87.8	90.1	88.9 ± 0.9	87.1 ± 1.0	89	67	73	91	77	81
	SP	96.8 ± 0.3	94.4	93.6	94.0 ± 0.6	91.8 ± 0.8	96	92	92	95	92	91
	OT	94.9 ± 0.3	95.9	95.7	95.8 ± 0.3	89.3 ± 0.7	88	97	82	91	97	86

Table 1. Comparing with TargetP and TargetLoc on sensitivity (SE), specificity (SP) and Matthew's Correlation Coefficient (MCC) on the plant and non-plant dataset. The classes are chloroplast (ch), mitochondria (mi), secretory pathway (SP) and other (OT).

times selected	**mean β_k**	**kernel**
10	26.53%	RBF on log BLAST E-value, $\sigma = 10^5$
10	19.77%	RBF on BLAST E-value, $\sigma = 10^3$
10	16.53%	RBF on inv phyl. profs, $\sigma = 300$
10	11.12%	RBF on lin phyl. profs, $\sigma = 1$
10	5.50%	motif (●,○,○,○,○) on $[1, 15]$
10	4.68%	motif (●,○,○,○,●) on $[1, 15]$
10	3.48%	motif (●,○,○,○,○) on $[1, 60]$
8	3.17%	motif (●,●,○,○,●) on $[1, 60]$
9	2.56%	motif (●,○,○,○,○) on $[1, Inf]$
5	1.44%	motif (●,○,●,○,●) on $[1, 60]$
7	1.05%	motif (●,○,○,●,○) on $[1, 15]$
7	0.95%	motif (●,●,○,○,○) on $[1, Inf]$
3	0.65%	motif (●,●,●,○,●) on $[1, 60]$
5	0.64%	motif (●,○,○,○,●) on $[1, Inf]$
2	0.40%	motif (●,○,○,●,●) on $[1, 60]$
6	0.38%	motif (●,○,●,○,○) on $[-15, Inf]$
7	0.29%	motif (●,○,○,○,○) on $[-15, Inf]$
3	0.26%	motif (●,○,●,○,●) on $[1, 15]$
2	0.18%	motif (●,○,○,●,○) on $[1, 60]$
3	0.12%	linear kernel on BLAST E-value
2	0.12%	motif (●,○,○,●,●) on $[1, 15]$
2	0.08%	motif (●,○,●,○,●) on $[-15, Inf]$
1	0.07%	motif (●,●,●,○,●) on $[-15, Inf]$
1	0.03%	motif (●,●,○,○,○) on $[1, 60]$
1	0.02%	motif (●,●,○,○,●) on $[1, 15]$

Table 2. Kernels selected in the ten repetitions of experiments on the plant dataset, sorted by importance as indicated by the averaged coefficient β_k. Note that the selection is very consistent across the repetitions, and that only a small fraction of kernels obtained a positive weight in any repetition. The first column shows the considered region of the protein, starting with 1 at the N-terminus; ∞ means that the region extends to the C-terminus. The second column shows the pattern associated with the kernel.

times selected	**mean β_k**	**kernel**
9	30.69%	RBF on log BLAST E-value, $\sigma = 10^5$
9	29.46%	RBF on BLAST E-value, $\sigma = 10^3$
10	11.85%	RBF on inv phyl. profs, $\sigma = 300$
9	7.15%	RBF on lin phyl. profs, $\sigma = 1$
9	4.48%	motif (●,○,○,○,○) on $[1, 15]$
10	3.23%	motif (●,○,○,●,●) on $[1, 15]$
9	2.32%	motif (●,○,○,○,○) on $[1, Inf]$
9	2.17%	motif (●,○,○,○,○) on $[1, 60]$
8	1.92%	motif (●,○,○,●,○) on $[1, 60]$
9	1.53%	linear kernel on BLAST E-value
9	1.48%	motif (●,●,●,○,○) on $[1, Inf]$
8	0.94%	motif (●,○,●,○,○) on $[1, 15]$
4	0.54%	motif (●,●,○,○,●) on $[1, 60]$
3	0.38%	motif (●,○,●,●,●) on $[1, 60]$
3	0.35%	motif (●,●,●,○,●) on $[1, 60]$
3	0.32%	motif (●,○,○,●,○) on $[1, 15]$
5	0.29%	motif (●,○,○,○,●) on $[1, 60]$
2	0.23%	motif (●,○,○,○,●) on $[1, 15]$
3	0.21%	motif (●,○,○,●,●) on $[1, 60]$
5	0.11%	motif (●,○,○,○,●) on $[-15, Inf]$
1	0.08%	motif (●,○,○,○,●) on $[1, Inf]$
2	0.08%	motif (●,●,○,○,●) on $[-15, Inf]$
1	0.06%	motif (●,●,○,○,○) on $[1, Inf]$
2	0.06%	motif (●,●,●,●,●) on $[1, 60]$
1	0.04%	motif (●,●,●,●,●) on $[1, Inf]$
1	0.04%	motif (●,●,○,○,○) on $[1, 60]$

Table 3. Analoguous to Table 2, but for nonplant data.

2.2 Comparison on PSORTdb dataset

Data	Class	Our Method					PSORTb v2.0		
		Accuracy	Precision	Recall	F1-Score	MCC	Precision	Recall	F1-Score
PSORT+	C	97.7 ± 0.7	95.4	99.1	97.1 ± 0.8	95.3 ± 1.3	97.1	86.6	91.6
	CM	98.1 ± 0.5	98.9	90.8	94.3 ± 1.8	93.5 ± 1.8	96.9	91.3	94.0
	CW	97.9 ± 0.4	93.6	86.9	89.7 ± 1.7	88.8 ± 1.8	94.7	88.5	91.5
	EC	96.2 ± 0.5	94.2	94.9	94.4 ± 0.7	91.6 ± 1.0	93.9	67.8	78.7
PSORT-	C	97.9 ± 0.3	90.6	99.6	94.8 ± 0.7	93.7 ± 0.8	92.9	70.1	79.9
	CM	98.9 ± 0.3	100.0	94.9	97.3 ± 0.6	96.7 ± 0.7	95.3	92.6	93.9
	P	98.0 ± 0.2	94.1	94.8	94.4 ± 0.6	93.2 ± 0.7	95.5	69.2	80.3
	OM	99.3 ± 0.1	99.9	97.7	98.8 ± 0.3	98.3 ± 0.3	97.4	94.9	96.1
	EC	98.4 ± 0.2	95.3	92.8	94.0 ± 0.8	93.1 ± 0.9	97.4	78.9	87.2

Table 4. Comparing with PSORTb v2.0 on singly located proteins from PSORTdb on bacteria. The classes are cytoplasm (C), cytoplasmic membrane (CM), periplasm (P), outer membrane (OM), extracellular (EC) and cell wall (CW).

Data	Class	Our Method				
		Accuracy	Precision	Recall	F1-Score	MCC
PSORT+	C	94.9 ± 0.9	90.6	96.7	93.5 ± 1.1	89.4 ± 1.9
	CM	95.7 ± 0.7	94.7	83.6	88.6 ± 2.1	86.3 ± 2.3
	CW	97.1 ± 0.4	89.9	84.5	86.3 ± 1.9	85.2 ± 2.0
	EC	93.7 ± 0.8	90.9	91.3	91.1 ± 1.1	86.3 ± 1.7
PSORT-	C	95.5 ± 0.4	83.9	96.8	89.9 ± 1.0	87.4 ± 1.2
	CM	97.1 ± 0.2	97.3	89.1	92.9 ± 0.6	91.3 ± 0.8
	P	94.5 ± 0.4	84.1	89.2	86.5 ± 1.0	83.2 ± 1.2
	OM	97.7 ± 0.3	98.2	93.7	95.8 ± 0.6	94.4 ± 0.8
	EC	96.7 ± 0.3	92.5	85.2	88.6 ± 1.1	86.8 ± 1.3

Table 5. Results using all data. The classes are cytoplasm (C), cytoplasmic membrane (CM), periplasm (P), outer membrane (OM), extracellular (EC) and cell wall (CW).

times selected	**mean β_k**	**kernel**
10	41.77%	RBF on BLAST E-value, $\sigma = 10^3$
10	27.32%	RBF on lin phyl. profs, $\sigma = 1$
10	6.23%	motif (●,○,○,○,○) on $[1, Inf]$
10	4.01%	RBF on inv phyl. profs, $\sigma = 300$
10	3.75%	motif (●,○,●,○,●) on $[1, Inf]$
8	3.03%	RBF on log BLAST E-value, $\sigma = 10^5$
9	2.24%	motif (●,○,●,●,●) on $[1, 60]$
7	2.21%	motif (●,○,○,●,●) on $[-15, Inf]$
9	1.81%	linear kernel on BLAST E-value
5	1.64%	motif (●,○,●,●,●) on $[-15, Inf]$
10	1.32%	motif (●,○,○,○,●) on $[1, 15]$
6	1.25%	motif (●,○,●,○,○) on $[1, Inf]$
4	0.92%	motif (●,●,○,○,○) on $[1, Inf]$
8	0.53%	motif (●,○,○,○,○) on $[1, 15]$
4	0.43%	motif (●,○,○,○,○) on $[1, 60]$
6	0.43%	motif (●,○,○,○,○) on $[-15, Inf]$
2	0.32%	motif (●,○,○,●,○) on $[1, 60]$
4	0.25%	motif (●,○,○,○,●) on $[1, Inf]$
2	0.17%	motif (●,○,○,○,●) on $[-15, Inf]$
2	0.16%	motif (●,○,○,○,●) on $[1, 60]$
1	0.11%	motif (●,○,●,○,○) on $[1, 15]$
1	0.07%	motif (●,○,●,○,●) on $[-15, Inf]$
1	0.05%	motif (●,●,○,○,○) on $[1, 15]$

Table 6. Analoguous to Table 2, but for gram-positive bacteria for PSORTb.

times selected	**mean β_k**	**kernel**
10	35.00%	RBF on BLAST E-value, $\sigma = 10^3$
10	26.87%	RBF on log BLAST E-value, $\sigma = 10^5$
10	16.17%	RBF on lin phyl. profs, $\sigma = 1$
10	5.04%	motif (●,○,○,○,○) on $[1, Inf]$
10	2.82%	RBF on inv phyl. profs, $\sigma = 300$
10	2.64%	linear kernel on BLAST E-value
10	1.97%	motif (●,○,○,●,○) on $[1, Inf]$
9	1.57%	motif (●,●,○,○,○) on $[1, Inf]$
10	1.51%	motif (●,○,○,○,○) on $[1, 60]$
10	1.14%	motif (●,○,○,○,○) on $[1, 15]$
7	0.86%	motif (●,○,●,○,○) on $[1, Inf]$
10	0.82%	motif (●,○,○,○,●) on $[-15, Inf]$
5	0.65%	motif (●,○,●,●,●) on $[1, 60]$
5	0.60%	motif (●,○,○,○,●) on $[1, Inf]$
4	0.54%	motif (●,●,○,●,●) on $[1, 60]$
2	0.36%	motif (●,●,●,●,●) on $[1, 60]$
6	0.35%	motif (●,○,●,●,●) on $[1, 15]$
6	0.25%	motif (●,○,○,●,○) on $[-15, Inf]$
3	0.21%	motif (●,○,●,○,●) on $[1, 15]$
3	0.18%	motif (●,○,○,●,●) on $[1, 15]$
2	0.13%	motif (●,○,○,○,●) on $[1, 15]$
5	0.11%	motif (●,○,○,○,○) on $[-15, Inf]$
2	0.10%	motif (●,○,○,●,○) on $[1, 15]$
1	0.09%	motif (●,○,○,●,●) on $[1, 60]$
1	0.03%	motif (●,●,○,○,○) on $[-15, Inf]$
1	0.01%	motif (●,●,●,○,●) on $[1, Inf]$

Table 7. Analoguous to Table 2, but for gram-negative bacteria for PSORTb.

References

1. A. Zien and C.S. Ong. Multiclass multiple kernel learning. In *International Conference on Machine Learning*, 2007.