Adaptive Thresholding for Multi-Label SVM Classification with Application to Protein Subcellular Localization Prediction

Shibiao WAN, Man-Wai MAK

The Hong Kong Polytechnic University, Hong Kong SAR

Sun-Yuan KUNG

Princeton University, USA

Abstract

• Motivation:

Protein subcellular localization is a problem of predicting which part(s) in a cell a protein resides. This information is vitally important for understanding the functions of proteins and for identifying drug targets.

• Proposal:

This paper proposes an adaptive thresholding scheme for multi-label support vector machine (SVM) classifiers.

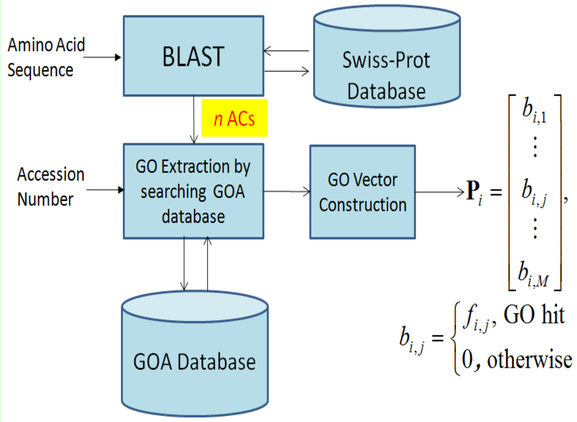
• Findings:

The adaptive thresholding scheme can effectively avoid both over-prediction and under-prediction, resulting in performance significantly better than other gene-ontology based subcellular localization predictors.

Feature Extraction

• We use Gene Ontology information as features.

• A protein accession number (AC) may correspond to 0, 1 or many Gene Ontology (GO) terms.



$$\mathbf{P}_i = \begin{bmatrix} b_{i,1} \\ \vdots \\ b_{i,j} \\ \vdots \\ b_{i,M} \end{bmatrix}, \quad b_{i,j} = \begin{cases} f_{i,j}, \text{ GO hit} \\ 0, \text{ otherwise} \end{cases}$$

ACs: Accession Numbers.
GO: Gene Ontology.
GOA: Gene Ontology Annotation.
BLAST: Basic Local Alignment Search Tool.

Adaptive Thresholding for SVM (AT-SVM)

• Multi-label SVM Scoring:

M one-vs-rest binary SVMs are trained, and the score of the m-th SVM is :

$$s_m(\mathbf{Q}_i) = \sum_{r \in \mathcal{S}_m} \alpha_{m,r} y_{m,r} K(\mathbf{P}_r, \mathbf{q}_{i,k}) + b_m$$

• Adaptive Thresholding:

If $\exists\ s_m(\mathbf{Q}_i) > 0$,

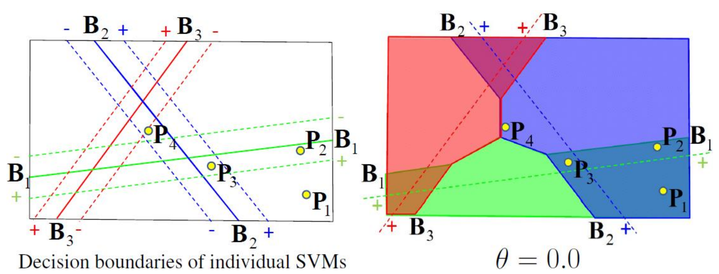
$$\mathcal{M}(\mathbf{Q}_i) = \bigcup_{m=1}^{M} \{\{m : s_m(\mathbf{Q}_i) > 1.0\} \cup \{m : s_m(\mathbf{Q}_i) \geq f(s_{\max}(\mathbf{Q}_i))\}\}$$

otherwise,

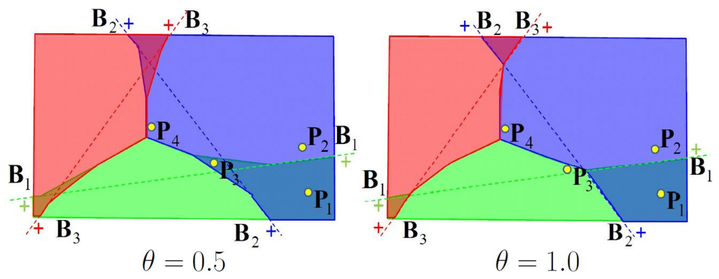
$$\mathcal{M}(\mathbf{Q}_i) = \arg\max_{m=1}^{M} s_m(\mathbf{Q}_i)$$

where $f(s_{\max}(\mathbf{Q}_i)) = \theta s_{\max}(\mathbf{Q}_i)$

A Three-class example:



$$\begin{aligned} &s_{\text{green}}(\mathbf{P}_1) > 1, && s_{\text{blue}}(\mathbf{P}_1) > 1, && s_{\text{red}}(\mathbf{P}_1) < 0; \\ &0 < s_{\text{green}}(\mathbf{P}_2) < 1, && s_{\text{blue}}(\mathbf{P}_2) > 1, && s_{\text{red}}(\mathbf{P}_2) < 0; \\ &0 < s_{\text{green}}(\mathbf{P}_3) < 1, && 0 < s_{\text{blue}}(\mathbf{P}_3) < 1, && s_{\text{red}}(\mathbf{P}_3) < 0; \\ &s_{\text{green}}(\mathbf{P}_4) < 0, && s_{\text{blue}}(\mathbf{P}_4) < 0, && s_{\text{red}}(\mathbf{P}_4) < 0. \end{aligned}$$



Experiments and Results

• Datasets:

Dataset	M	N	TLN
Virus	6	207	252
Plant	12	978	1055

M: number of subcellular locations.
N: number of actual proteins.
TLN: total locative number.

$$N_{\text{loc}}^{v} = 1\times 165 + 2\times 39 + 3\times 3 + \sum_{m=4}^{6} m\times 0 = 252$$

$$N_{\text{loc}}^{p} = 1\times 904 + 2\times 71 + 3\times 3 + \sum_{m=4}^{12} m\times 0 = 1055$$

• Performance Measures:

Denote $\mathcal{L}(\mathbf{P}_i)$ and $\mathcal{M}(\mathbf{P}_i)$ as the true label set and the predicted label set for the i-th protein $\mathbf{P}_i (i = 1, \ldots, N)$, respectively.

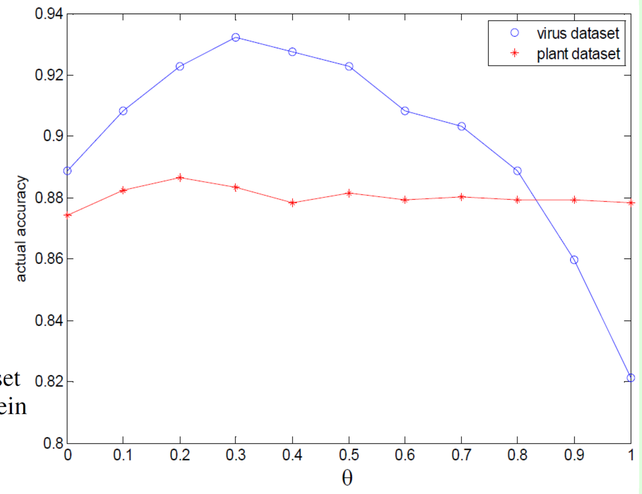
The overall locative accuracy:

$$\Lambda_{\text{loc}} = \frac{1}{\sum_{i=1}^{N} |\mathcal{L}(\mathbf{P}_i)|} \sum_{i=1}^{N} |\mathcal{M}(\mathbf{P}_i) \cap \mathcal{L}(\mathbf{P}_i)|$$

The overall actual accuracy:

$$\Lambda_{\text{act}} = \frac{1}{N} \sum_{i=1}^{N} \Delta[\mathcal{M}(\mathbf{P}_i), \mathcal{L}(\mathbf{P}_i)] \quad \text{where } \Delta[\mathcal{M}(\mathbf{P}_i), \mathcal{L}(\mathbf{P}_i)] = \begin{cases} 1 & \text{if } \mathcal{M}(\mathbf{P}_i) = \mathcal{L}(\mathbf{P}_i) \\ 0 & \text{otherwise.} \end{cases}$$

• Results:



For the virus dataset, AT-SVM peaks (93.2%) at $\theta = 0.3$;
For the plant dataset, AT-SVM peaks (88.7%) at $\theta = 0.2$.

Label	Subcellular Location	LOOCV Locative Accuracy				
		Virus-mPLoc [24]	KNN-SVM [26]	iLoc-Virus [25]	mGOASVM [27]	AT-SVM
1	Viral capsid	8/8 = 100.0%	8/8 = 100.0%	8/8 = 100.0%	8/8 = 100.0%	8/8 = 100.0%
2	Host cell membrane	19/33 = 57.6%	27/33 = 81.8%	25/33 = 75.8%	32/33 = 97.0%	32/33 = 97.0%
3	Host ER	13/20 = 65.0%	15/20 = 75.0%	15/20 = 75.0%	17/20 = 85.0%	17/20 = 85.0%
4	Host cytoplasm	52/87 = 59.8%	86/87 = 98.8%	64/87 = 73.6%	85/87 = 97.7%	83/87 = 95.4%
5	Host nucleus	51/84 = 60.7%	54/84 = 65.1%	70/84 = 83.3%	82/84 = 97.6%	82/84 = 97.6%
6	Secreted	9/20 = 45.0%	13/20 = 65.0%	15/20 = 75.0%	20/20 = 100.0%	20/20 = 100.0%
Overall Locative Accuracy		152/252 = 60.3%	203/252 = 80.7%	197/252 = 78.2%	244/252 = **96.8%**	242/252 = 96.0%
Overall Actual Accuracy		–	–	155/207 =74.8%	184/207 = 88.9%	193/207 = **93.2%**

References:

• S. Wan, M.W. Mak, and S.Y. Kung, "GOASVM: A Subcellular Location Predictor by Incorporating Term-Frequency Gene Ontology into the General Form of Chou's Pseudo-Amino Acid Composition", *J. of Theoretical Biology*, vol. 323 pp. 40–48 2013.

• S. Wan, M.W. Mak, and S.Y. Kung, "mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines", *BMC Bioinformatics*, 2012, 13:290.