Classification of Multiple Cancer Types by Multicategory Support Vector Machines using Gene Expression Profiles

Yoonkyung Lee[a]

Department of Statistics
The Ohio State University

[a]Research partly supported by NSF and NIH Grants to Grace Wahba. Joint work with Cheol-Koo Lee, Genetics and Medical Genetics, UW-Madison.

Gene expression profiles from Microarray

- Relative amount of mRNAs of thousands of genes.
- Fingerprints at the molecular level.
- Potential for diagnosis of cancer types.

Acute leukemia data set

- Golub et al.(1999) in *Science*.
- ALL (acute lymphoblastic leukemia) with subtype B-cell and T-cell lineage, and AML (acute myeloid leukemia)
- Total of 72 samples (38 in training set, and 34 in test set).

Training data set

	$gene_1$	$gene_2$	$\cdots$	$gene_{7129}$	$class$
$patient_1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,7129}$	ALL − B
$patient_2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,7129}$	ALL − B
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{20}$	$x_{20,1}$	$x_{20,2}$	$\cdots$	$x_{20,7129}$	ALL − T
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{28}$	$x_{28,1}$	$x_{28,2}$	$\cdots$	$x_{28,7129}$	AML
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{38}$	$x_{38,1}$	$x_{38,2}$	$\cdots$	$x_{38,7129}$	AML

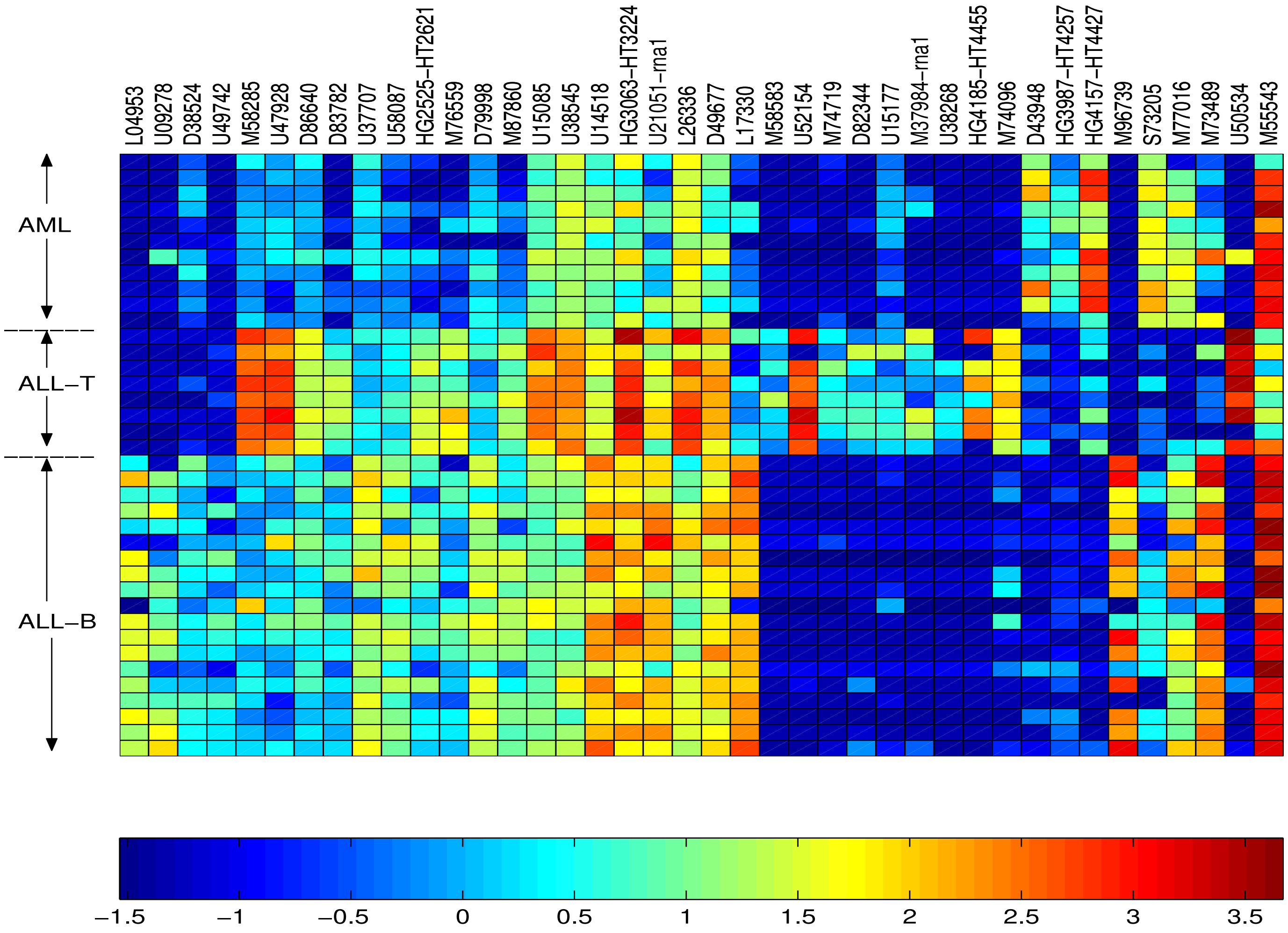

Figure 1: The heat map shows the expression levels of 40 most important genes for the training samples when they are appropriately standardized. Each row corresponds to a sample, which is grouped into the three classes, and the columns represent genes. The 40 genes are clustered in a way the similarity within each class and the dissimilarity between classes are easily recognized.

Challenges in classification by microarrays

- Relatively small number of samples.
- High dimensional attributes (large number of genes).
- Identify genes that contribute most to class distinction.

Cancer classification with microarray data

Data set	Methods	References
Leukemia	Weighted voting scheme	Golub et al. (*Science* 1999)
(2-class)	Support Vector Machine	Furey et al. (*Bioinformatics* 2000)
	LDA/QDA/CART/Nearest Neighbor	Dudoit et al. (*JASA* 2002)
SRBCT	Neural Networks	Khan et al. (*Nature Medicine* 2001)
(4-class)	Nearest shrunken centroids	Tibshirani et al. (*PNAS* 2002)
	SVM/NN (one-vs-rest)	Yeo & Poggio (2001)
Multiple	SVM/k-NN/weighted voting	Ramaswamy et al. (*PNAS* 2001)
(14-class)	(one-vs-rest/all-pairs/partitioning)	

Characteristics of Support Vector Machines

- Classification accuracy
- Flexibility
- Handle high dimensional data
- Sparsity - Quadratic programming problem
- No probability estimates

Support Vector Machine for Binary Case

Boser, Guyon, & Vapnik (1992), Vapnik (1995), Burges (1998), Cristianini & Shawe-Taylor (2000), http://www.kernel-machines.org

- Training data set :
  $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in R^d, y_i \in \{-1, 1\},$ for $i = 1, \cdots, n\}$
- Linear SVM: $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ where $\mathbf{w} \in R^d$ and $b \in R$.
- Nonlinear SVM: $f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b$
  where K is a positive definite function.
  Popular choices of $K(\mathbf{x}_1, \mathbf{x}_2)$: $(1 + \mathbf{x}_1^t \mathbf{x}_2)^d$, $\exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$

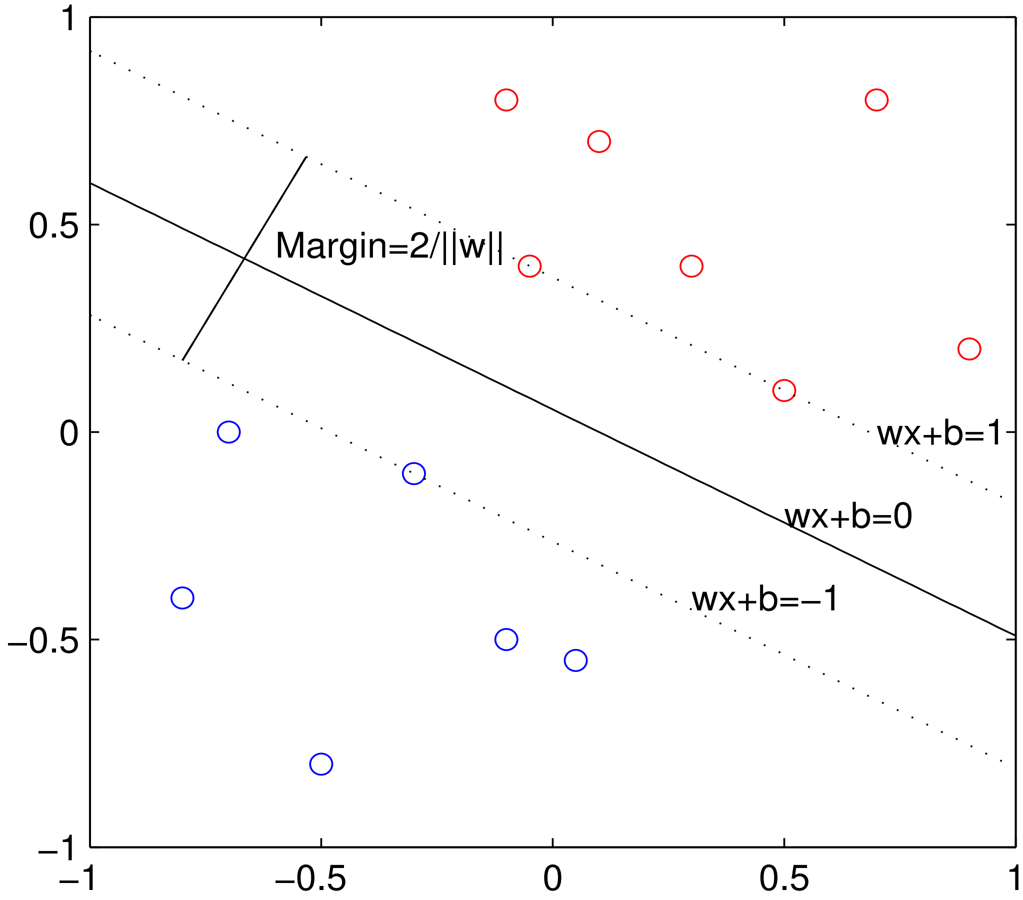

Figure 2: Linear Support Vector Machine example in separable case. Red and blue circles indicate positive and negative examples, respectively. The solid line is the separating hyperplane with the maximum margin in this example.

- Separable case :

  Find $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ minimizing $\|\mathbf{w}\|^2$

$$\text{subject to} \qquad \begin{array}{ll} f(\mathbf{x}_i) \geq 1 & \text{if } y_i = 1 \\ f(\mathbf{x}_i) \leq -1 & \text{if } y_i = -1. \end{array}$$

  Or more succinctly,

$$y_i f(\mathbf{x}_i) \geq 1 \text{ for } i = 1, \cdots, n.$$

- Nonseparable case :

  Find $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|\mathbf{w}\|^2$$

  where $(x)_+ = \max(x, 0)$.

SVM in Regularization Framework

Wahba (1998)

- $f(\mathbf{x}) = h(\mathbf{x}) + b$ with $h \in \mathcal{H}_K$ a Reproducing Kernel Hilbert space (RKHS) with reproducing kernel K.
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$
- SVM scheme : to find f minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+ + \lambda\|h\|^2_{\mathcal{H}_K}$$

λ is a tuning parameter which balances the data fit and the complexity of $f(\mathbf{x})$.

SVM and the Bayes Rule for Binary case

- Under the representation of $Y = 1$ or -1,
  Let $p^+(\mathbf{x}) = P(Y = 1|X = \mathbf{x})$.

- the Bayes rule : $\phi_B(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$

Lemma 1 (Lin 1999). *The minimizer of* $E[(1 - Yf(X))_+]$ *is* $f(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$.

Approach to Multiclass by Binary SVM

- Combine binary classifiers to solve multiclass problems.
- Pairwise approach.
- One-vs-rest approach : very common for SVMs.

  (i) Solve k binary problems.

  (ii) Assign a test sample the class giving the largest $f_j(\mathbf{x})$, where $f_j(\mathbf{x})$ is the SVM solution from training class j versus the rest.

  One-vs-rest approach is not optimal!
- Some multiclass extensions:

  Weston & Watkins(1998), Vapnik(1998), Crammer & Singer(2000)

Multicategory SVM

- Class codes :

  $\mathbf{y}_i$ is represented by a k-dimensional vector with 1 in the jth coordinate and $-\frac{1}{k-1}$ elsewhere if example i falls into class j. For example, when $k = 3$,

$$\mathbf{y}_i = \begin{cases} (1, -\frac{1}{2}, -\frac{1}{2}) & \text{for class 1} \\ (-\frac{1}{2}, 1, -\frac{1}{2}) & \text{for class 2} \\ (-\frac{1}{2}, -\frac{1}{2}, 1) & \text{for class 3} \end{cases}$$

- Separating functions :

  $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ with sum-to-zero constraint, $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$ for any $\mathbf{x} \in R^d$, and $f_j(\mathbf{x}) = h_j(\mathbf{x}) + b_j$ with $h_j \in \mathcal{H}_K$.

- Multicategory SVM formulation :

  Find $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$, with sum-to-zero constraint, minimizing

$$\frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i) \cdot (\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i)_+ + \frac{\lambda}{2}\sum_{j=1}^{k} \|h_j\|^2_{\mathcal{H}_K}$$

  where $L(\mathbf{y}_i)$ = the jth row of the cost matrix C if $\mathbf{y}_i$ indicates class j. C is defined as k by k matrix with 0 on the diagonal, and 1 elsewhere. For example, when $k = 3$,

$$C = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

- A classification rule induced from $\mathbf{f}(\mathbf{x})$:

$$\phi(\mathbf{x}) = arg\max_j f_j(\mathbf{x})$$

- Binary SVM is a special case of Multicategory SVM when $k = 2$.

- Let $p_j(\mathbf{x}) = P(Y = j|X = \mathbf{x})$.

- The Bayes rule: $\phi_B(\mathbf{x}) = arg\max_{j=1,\cdots,k} p_j(\mathbf{x})$

> **Lemma 2 (Lee, Lin & Wahba 2001).** *The minimizer of* $E[L(Y) \cdot (\mathbf{f}(X) - Y)_+]$ *under sum-to-zero constraint is* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ *with*
>
> $$f_j(\mathbf{x}) = \begin{cases} 1 & \text{if } j = arg\max_{\ell=1,\cdots,k} p_\ell(\mathbf{x}) \\ -\frac{1}{k-1} & \text{otherwise} \end{cases}$$

Small Round Blue Cell Tumors of Childhood

- Khan et al. (2001) in *Nature Medicine*
- Tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS)
- Number of genes : 2308
- Class distribution of data set

Data set	NB	RMS	BL(NHL)	EWS	total
Training set	12	20	8	23	63
Test set	6	5	3	6	20
Total	18	25	11	29	83

- Gene selection : Dudoit et al. (2000)

  For gene ℓ, the ratio of between classes sum of squares to within class sum of squares is defined as

$$\frac{BSS(\ell)}{WSS(\ell)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j)(\bar{x}_{\cdot\ell}^{(j)} - \bar{x}_{\cdot\ell})^2}{\sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j)(x_{i\ell} - \bar{x}_{\cdot\ell}^{(j)})^2}$$

  Pick genes with the largest ratios.

Table 1: Leaving-out-one cross validation (LOOCV) error and Test error by the MSVMs with the Gaussian kernel, $K(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$.

Number of genes	LOOCV error	Test error
20	0	0
60	0	0
100	0	0
all	1	0 to 3
3 PCs (100)*	0	0

*: three principal components (PCs) from 100 genes.

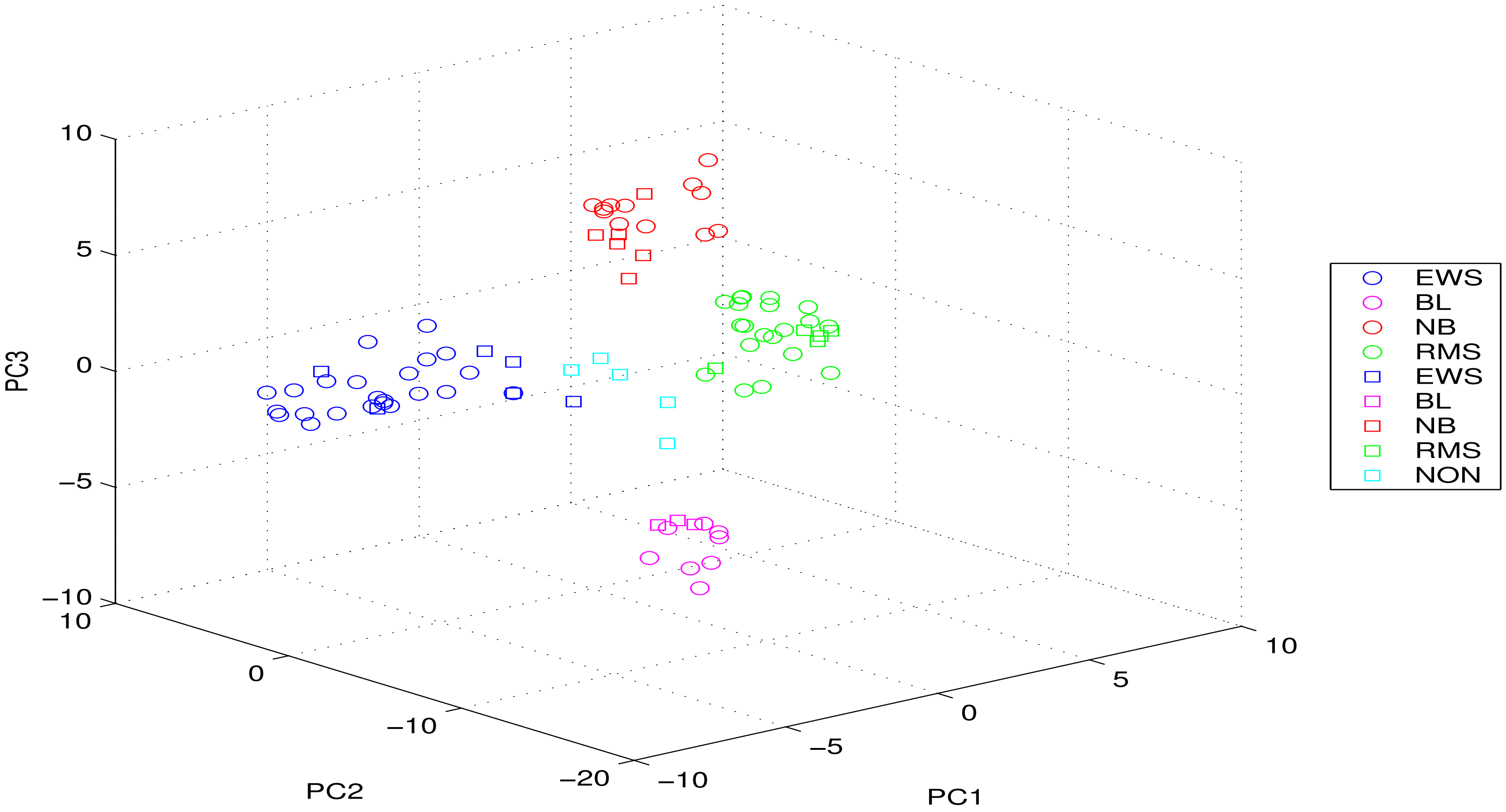

Figure 3: Three principal components of 100 gene expression levels (circles: training samples, squares: test samples including non SRBCT samples). The tumor types are distinguished by colors.

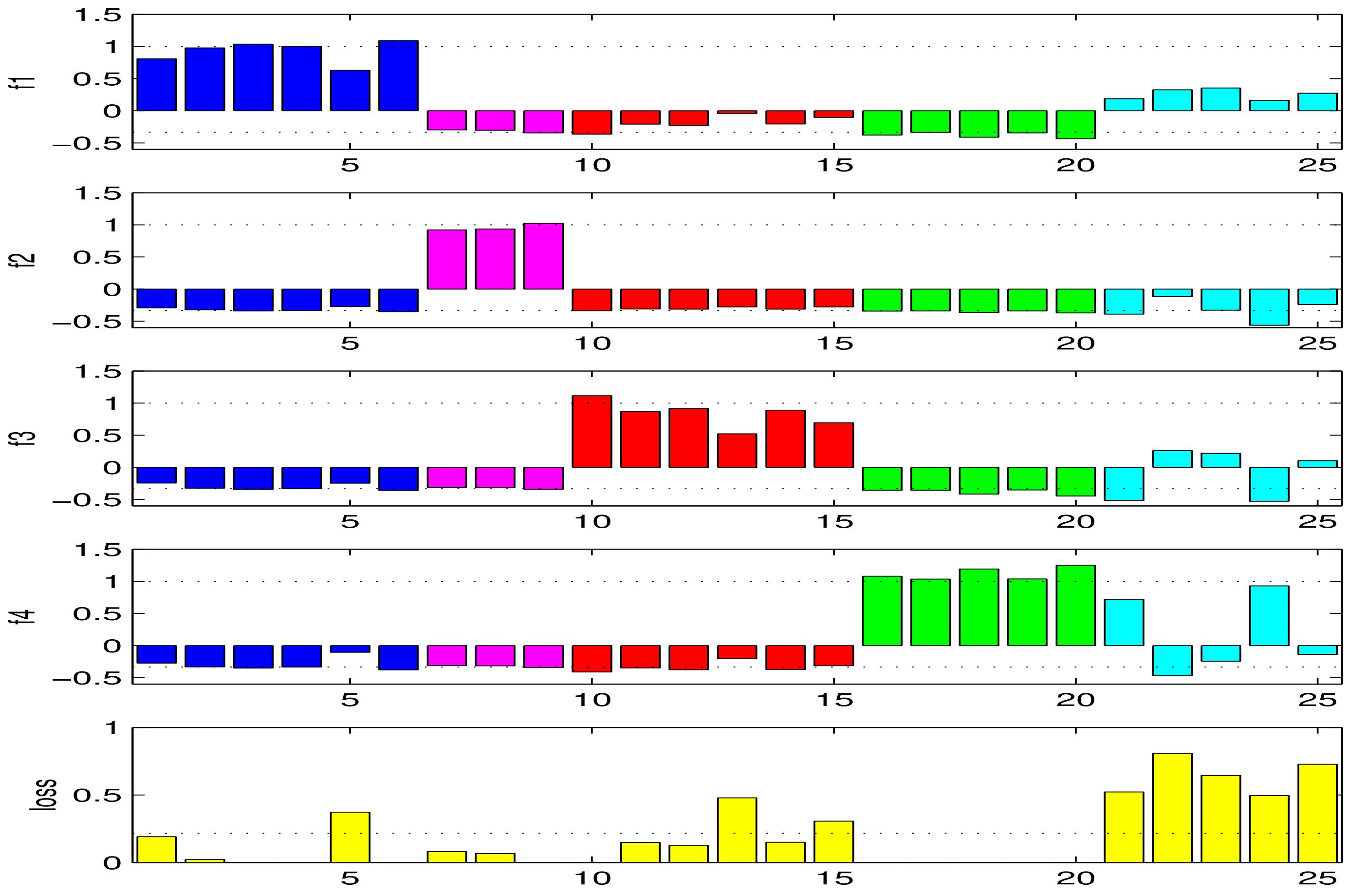

Figure 4: Predicted decision vectors (f_1, f_2, f_3, f_4) at the test samples. EWS: $(1, -1/3, -1/3, -1/3)$, BL: $(-1/3, 1, -1/3, -1/3)$, NB: $(-1/3, -1/3, 1, -1/3)$, and RMS: $(-1/3, -1/3, -1/3, 1)$. The colors indicate the true class identities of the test samples.

Summary

- Multicategory SVM is an optimal extension of the binary SVM to the multiclass case.
- MSVM has a great potential for medical diagnosis.
- Cancer classification with gene expression profiles

  - feasible and multitude of equally good answers.

Future Directions

- Integrate gene selection with learning classification rule.

  Q: Is the number of genes involved in classifier critical?

  What about redundancy from correlated genes?

- Assess statistical significance of the findings.

- Customized mini-arrays for cancer diagnosis.

Slides of this talk and papers are available from

`http://www.stat.ohio-state.edu/~yklee`