Multicategory Support Vector Machines

Yoonkyung Lee[a]

Department of Statistics
The Ohio State University

[a]Joint work with Grace Wahba and Yi Lin, Statistics, University of Wisconsin-Madison.

Classification Problems

- Training data set :

  Covariate (attribute) $\mathbf{x}_i \in R^d$, and Class label $y_i \in \{1, \cdots, k\}$,

  $i = 1, \cdots, n$.

  k is the number of classes.

- Task :

  Learn a classification rule matching $\mathbf{x}_i$ with y_i,

  the mapping $\phi(\mathbf{x}) : R^d \to \{1, \cdots, k\}$

Example: Classification Problem

- Fisher (1936): Iris data

  Iris Setosa, Iris Versicolour and Iris Virginica

- Golub et al.(*Science*, 1999): Acute leukemia data

  ALL (acute lymphoblastic leukemia) with subtype B-cell and T-cell lineage, and AML (acute myeloid leukemia)

Training data set: Acute leukemia gene expression profiles

	$gene_1$	$gene_2$	$\cdots$	$gene_{7129}$	$class$
$patient_1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,7129}$	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

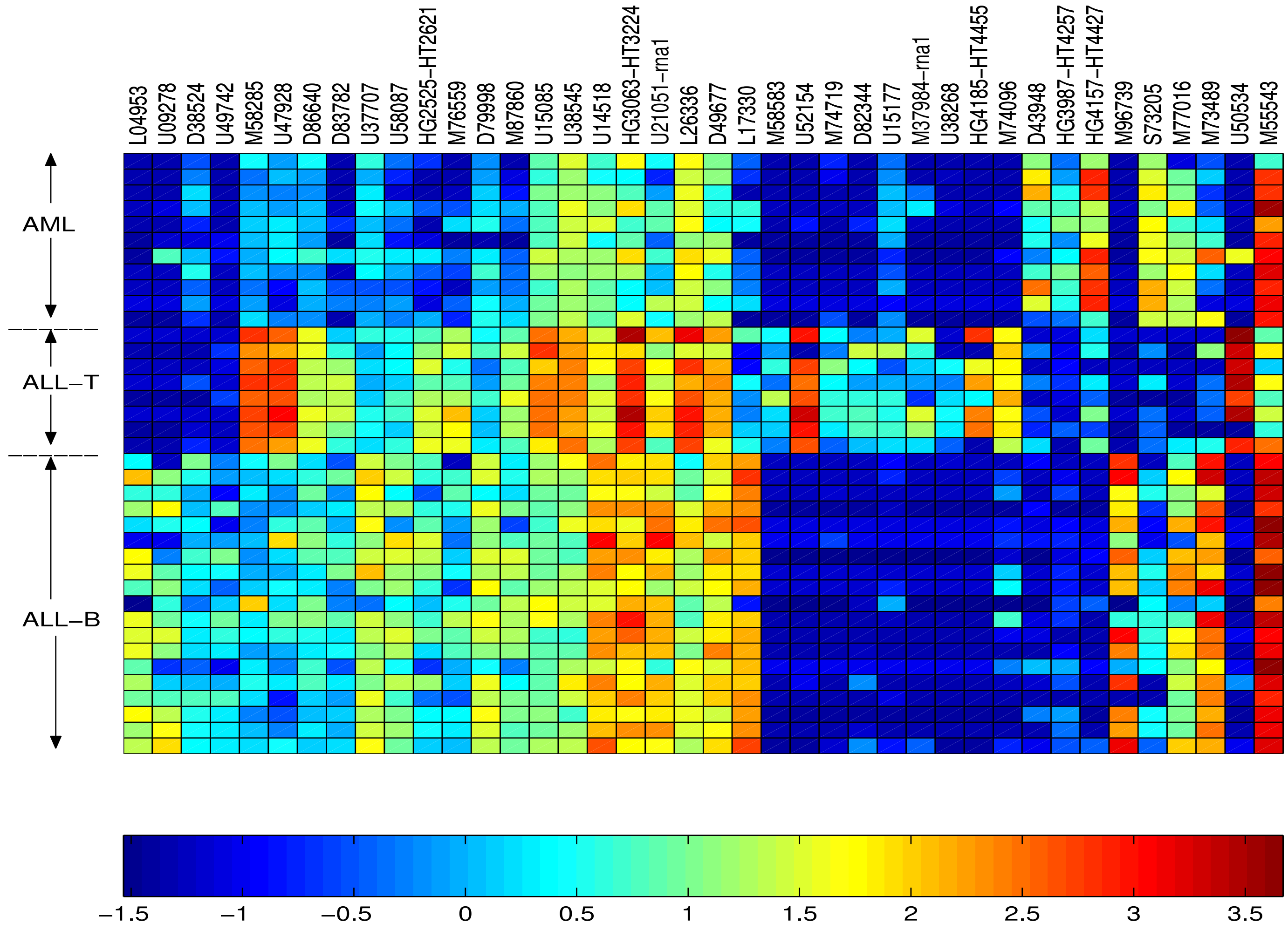


Figure 1: The heat map shows the expression levels of 40 most important genes for the training samples when they are appropriately standardized. Each row corresponds to a sample, which is grouped into the three classes, and the columns represent genes. The 40 genes are clustered in a way the similarity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

The Bayes Classification Rule

- Misclassification Loss function :

$$l(y, \phi(\mathbf{x})) = I(y \neq \phi(\mathbf{x}))$$

- (X, Y) : a random sample from $P(\mathbf{x}, y)$, and $p_j(\mathbf{x}) = P(Y = j | X = \mathbf{x})$

- The Bayes rule minimizes the expected misclassification rate $E[l(Y, \phi(\mathbf{X}))] = P(Y \neq \phi(\mathbf{X}))$

$$\phi_B(\mathbf{x}) = arg \max_{j=1,\cdots,k} p_j(\mathbf{x})$$

Classification Tools

Linear Discriminant Analysis	Quadratic Discriminant Analysis
Logistic Regression	Density Estimation
Classification Tree	Nearest Neighbor
Neural Networks	Support Vector Machine

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 $b \in R$.
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$

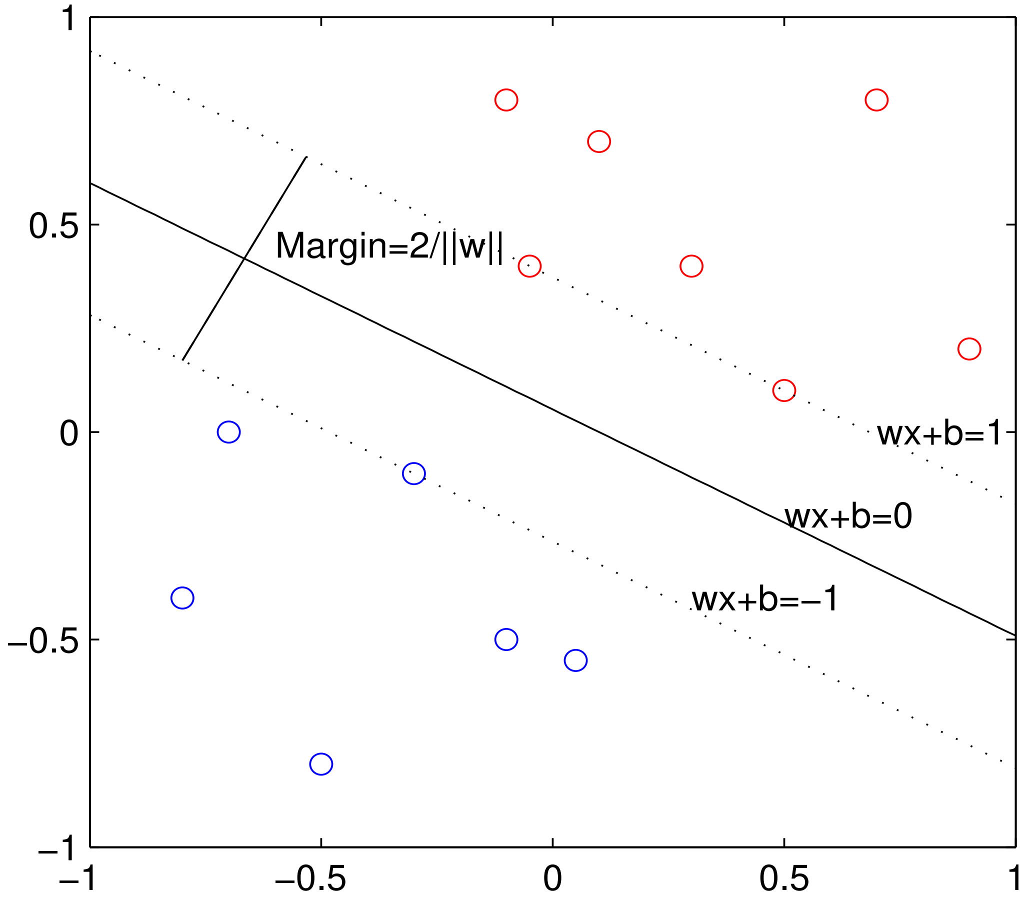


Figure 2: Linear Support Vector Machine example in separable case. Red and blue circles indicate positive and negative examples, respectively. The solid line is the separating hyperplane with the maximum margin in this example.

- 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,

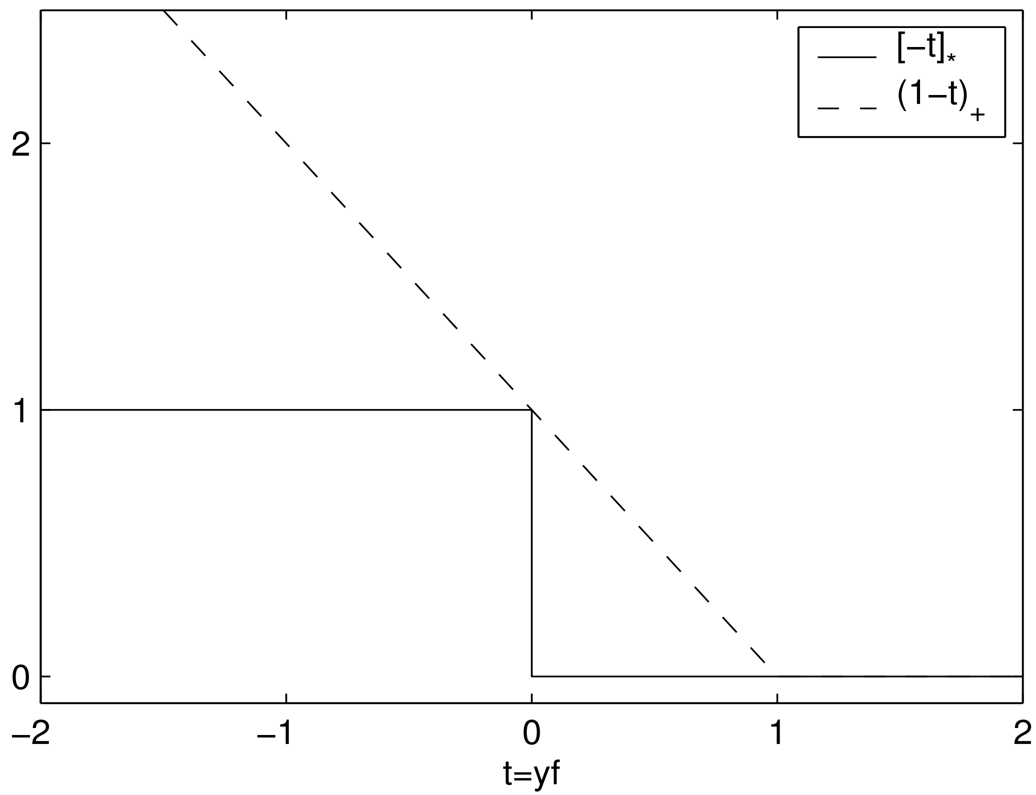
$$y_i f(\mathbf{x}_i) \geq 1 \text{ for } i = 1, \cdots, n.$$

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



Loss function $(1 - y_i f(\mathbf{x}_i))_+$ is an upper bound of the misclassification loss function $I(y_i \neq \phi(\mathbf{x}_i)) = [-y_i f(\mathbf{x}_i)]_* \leq (1 - y_i f(\mathbf{x}_i))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

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

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

- Linear SVM: $\mathcal{H}_K = \{h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} | \mathbf{w} \in R^d\}$

  with $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2$ and $\|h\|^2_{\mathcal{H}_K} = \|\mathbf{w} \cdot \mathbf{x}\|^2_{\mathcal{H}_K} = \|\mathbf{w}\|^2$
- Nonlinear SVM: $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1 \cdot \mathbf{x}_2)^d, \exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$

- By Representer Theorem (Kimeldorf, & Wahba 1971),

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b$$

So, $||h||^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j)$

- Solve a constrained optimization problem through its Primal-Dual relation. Introduce a Lagrange multiplier for each data point.

- Dual problem in terms of Lagrange multipliers:
  Quadratic Programming problem

- If the Lagrange multiplier for the i th example is zero, then $c_i = 0$. Removing such example $(\mathbf{x}_i, y_i)$ would not affect the solution at all. *Support Vectors*: Examples with nonzero coefficient.

SVM and the Bayes Rule for Binary case

- Under the label 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* $f(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$.

Characteristics of Support Vector Machines

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

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 $L(\mathbf{y}_i)\cdot(\mathbf{f}(\mathbf{x}_i)-\mathbf{y}_i)_+ = \sum_{j=1}^{k} L_{cat(i)j}(f_j(\mathbf{x}_i)-y_{ij})_+$,

  $cat(i)$: the category of $\mathbf{y}_i$ and $L_{jj'}$: the cost of misclassifying j as j'.

  When $L_{jj'} = I(j \neq j')$,

  $L(\mathbf{y}_i)\cdot(\mathbf{f}(\mathbf{x}_i)-\mathbf{y}_i)_+ = \sum_{j\neq cat(i)}(f_j(\mathbf{x}_i)+\frac{1}{k-1})_+$.

- 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*

$$f_j(\mathbf{x}) = \begin{cases} 1 & \text{if } j = arg\max_{\ell=1,\cdots,k} p_\ell(\mathbf{x}) \\ -\frac{1}{k-1} & \text{otherwise} \end{cases}$$

Representer theorem for Multicategory SVM

Theorem 1. *To find* $(f_1(\mathbf{x}), \cdots, f_k(\mathbf{x})) \in \prod_1^k(\{1\} + \mathcal{H}_K)$*, with the sum-to-zero constraint, minimizing the MSVM objective function is equivalent to find* $(f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ *of the form*

$$f_j(\mathbf{x}) = b_j + \sum_{i=1}^{n} c_{ij} K(\mathbf{x}_i, \mathbf{x}) \text{ for } j = 1, \cdots, k$$

with the sum-to-zero constraint only at $\mathbf{x}_i$ *for* $i = 1, \cdots, n$*, minimizing the objective function.*

- Solve a constrained optimization problem through its dual problem in terms of Lagrange multipliers.
- Dual problem : Quadratic Programming problem
- If the Lagrange multipliers for the i th example are zero vector, then $(c_{i1}, \cdots, c_{ik}) = 0$.
  Support Vectors: Examples with $(c_{i1}, \cdots, c_{ik}) \neq 0$ for $i = 1, \cdots, n$.
- Multicategory SVM retains the sparsity of solution in the same way as binary SVM!

Numerical Study

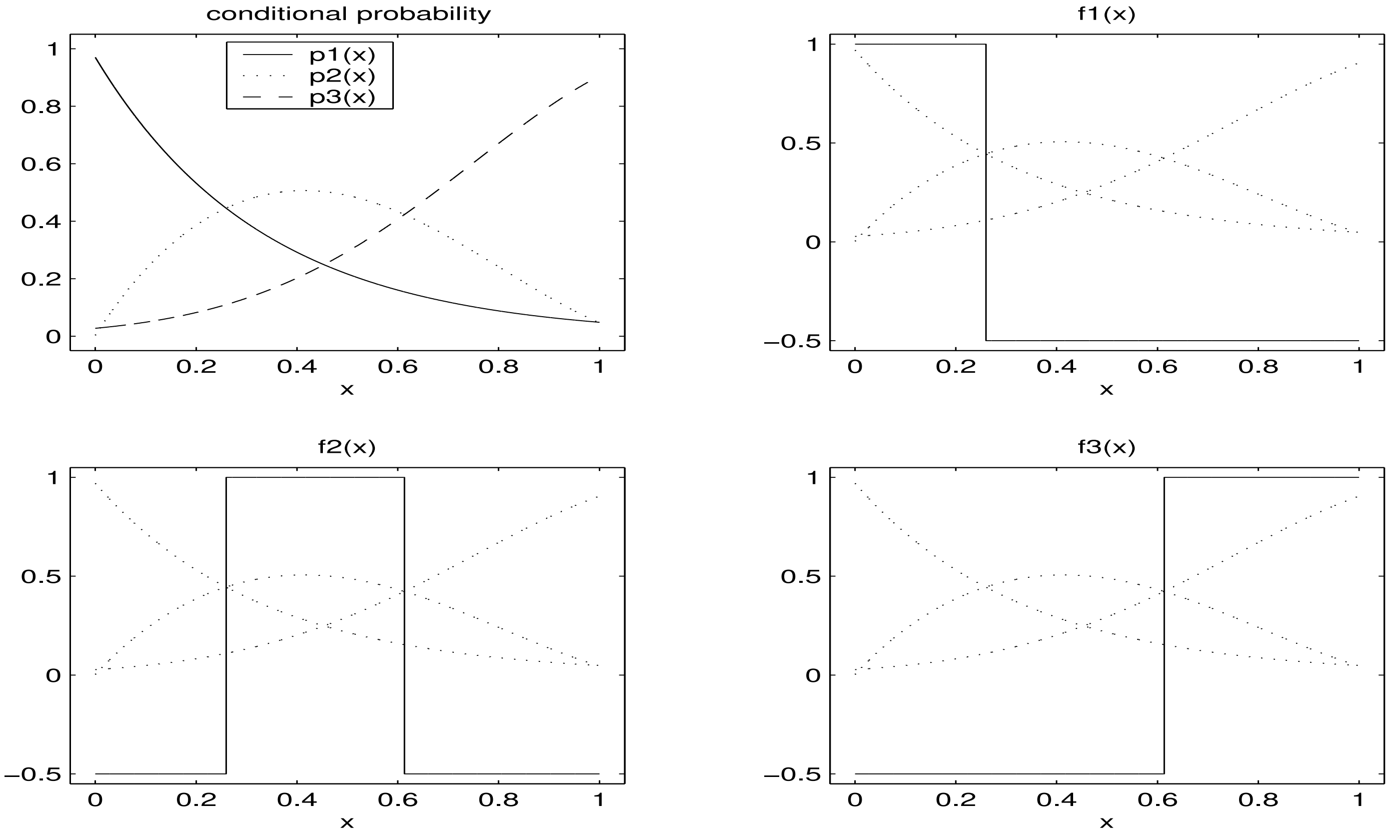


Figure 3: Conditional probabilities and multicategory SVM target functions for three-class example.

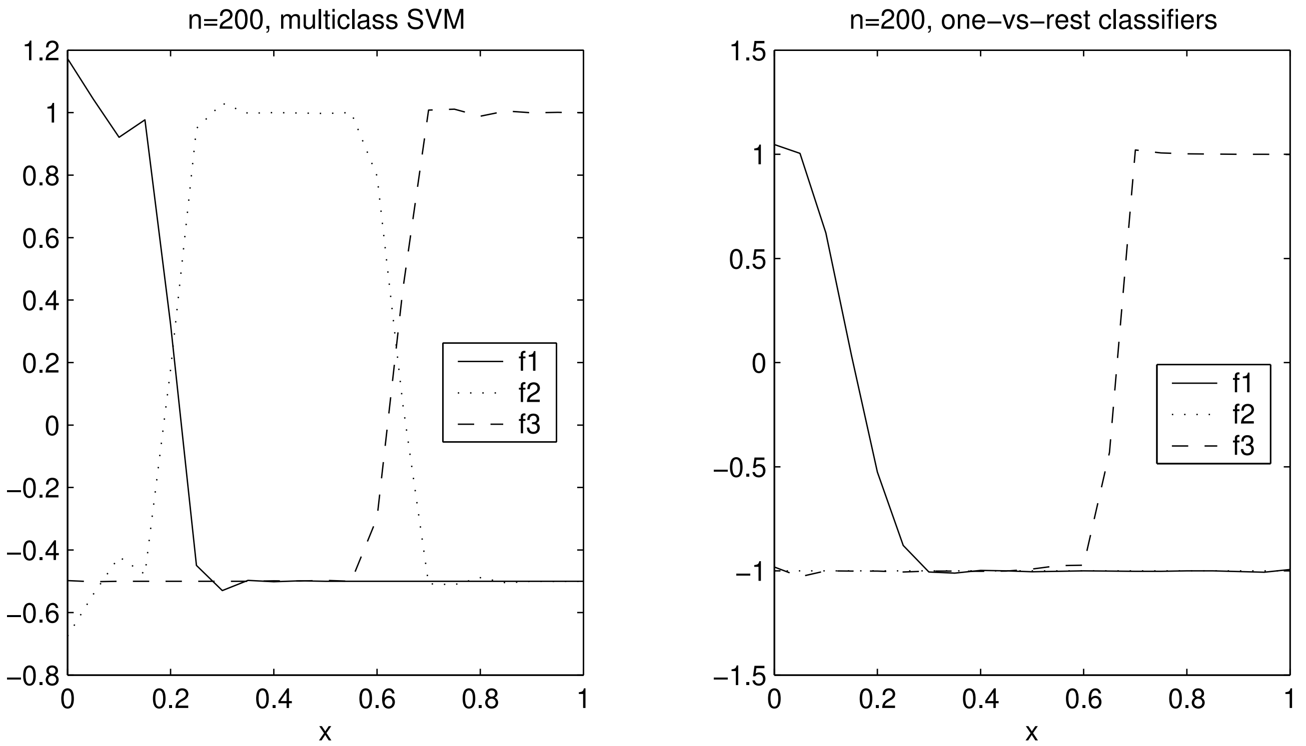


Figure 4: Comparison between multicategory SVM and one-versus-rest method. The Gaussian kernel function is used, and the tuning parameters λ, and σ are simultaneously chosen by $GCKL = E_{true}\frac{1}{n}\sum_{i=1}^{n} L(\mathbf{Y}_i) \cdot (\mathbf{f}_\lambda(x_i) - \mathbf{Y}_i)_+$. Over 10000 newly generated examples, the multicategory SVM has a misclassification rate 0.3890, while that of the one-vs-rest approach is 0.4243

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.

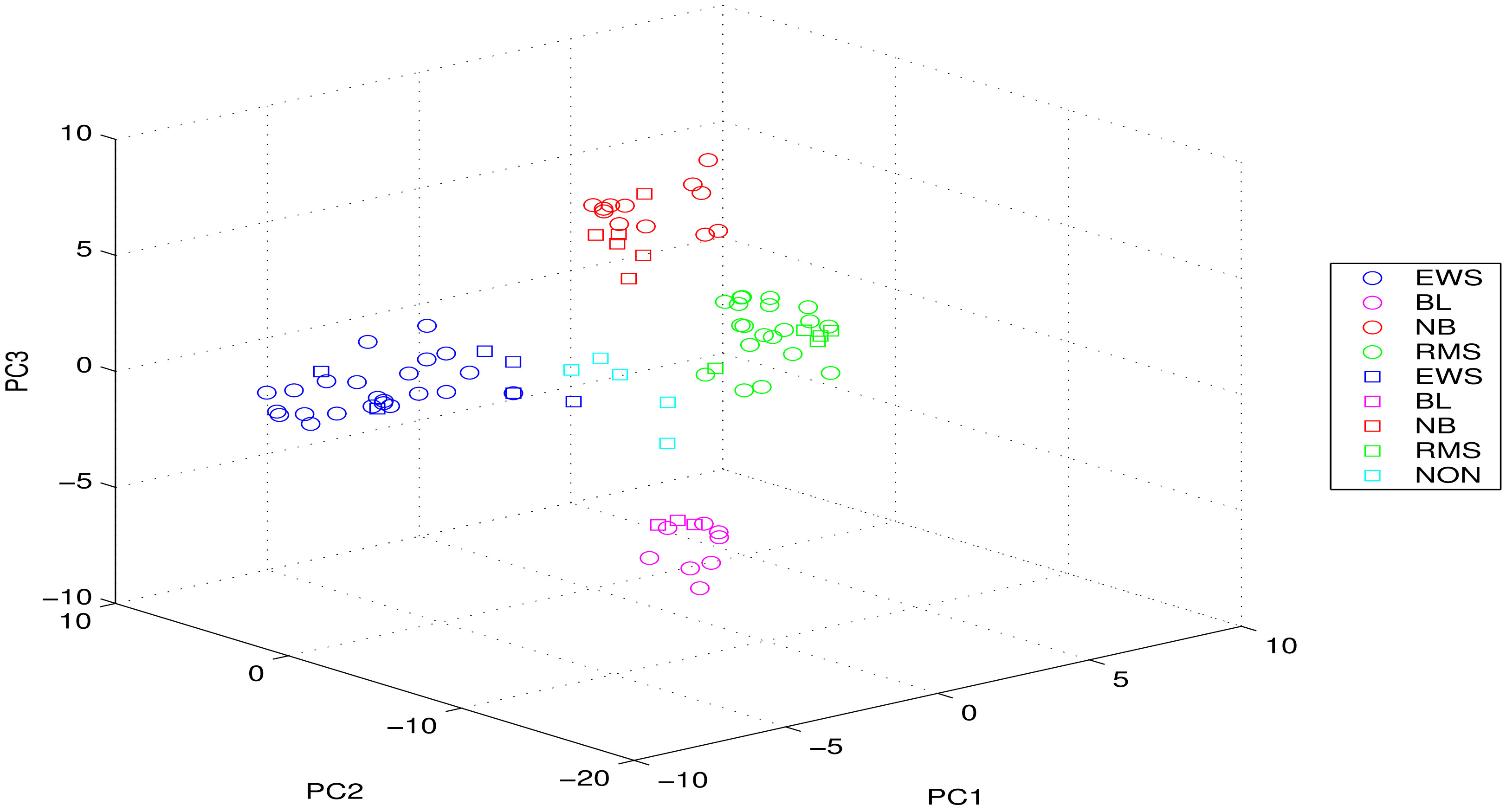


Figure 5: Three principal components of 100 gene expression levels (circles: training samples, squares: test samples including non SRBCT samples). The tumor types are distinguished by colors.

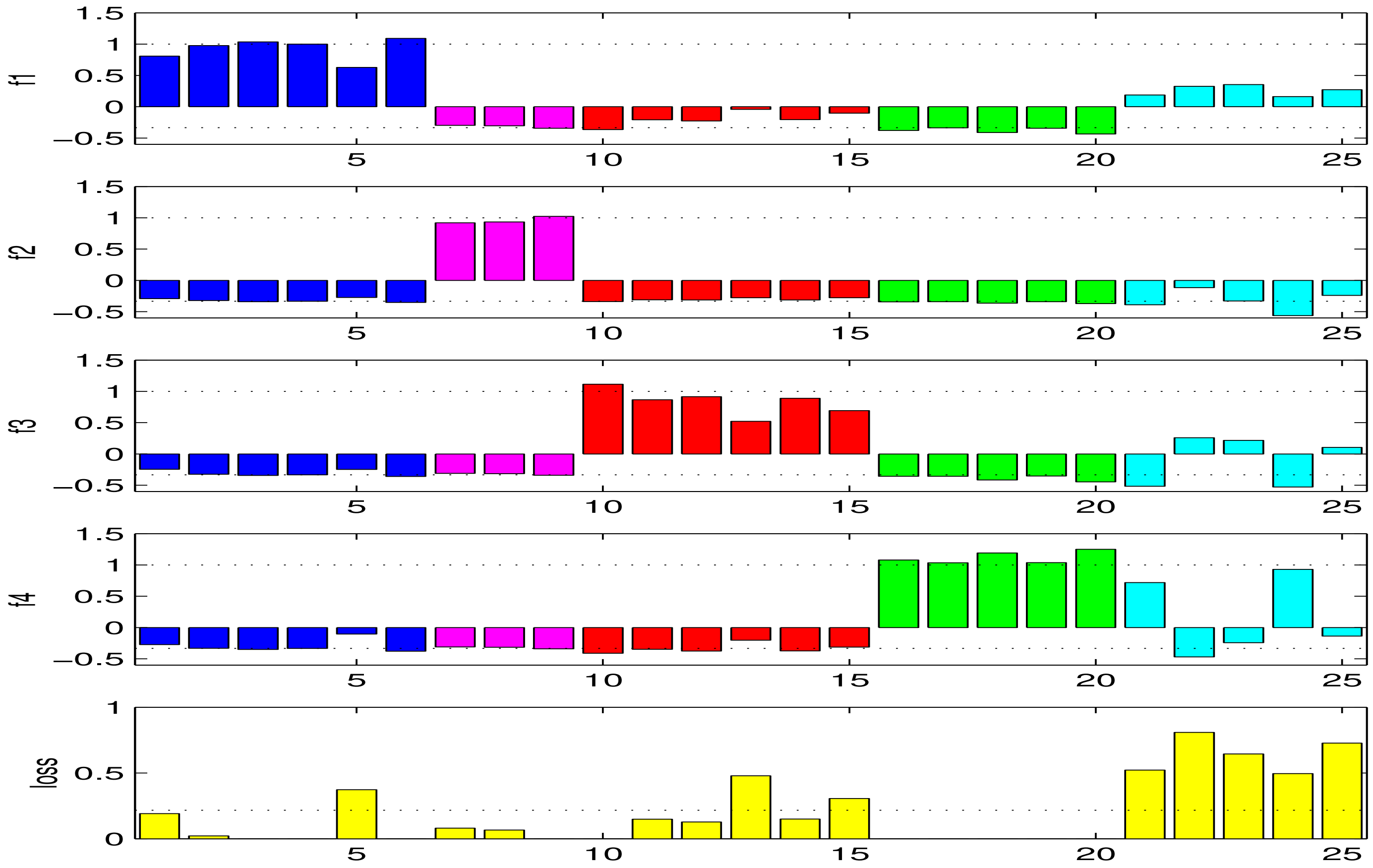


Figure 6: Predicted decision vectors (f_1, f_2, f_3, f_4) at the test samples. EWS: $(1, -1/3, -1/3, -1/3)$, BL: $(-1/3, 1, -1/3, -1/3)$, NB: $(-1/3, -1/3, 1, -1/3)$, and RMS: $(-1/3, -1/3, -1/3, 1)$. The colors indicate the true class identities of the test samples.

Cloud Classification using Radiance Data

- The MODIS (moderate resolution imaging spectroradiometer) measures radiances at 36 wavelengths every 1 to 2 days with spatial resolution 250 m to 1 km.

- Satellite observations at 12 wavelengths (.66, .86, .46, .55, 1.2, 1.6, 2.1, 6.6, 7.3, 8.6, 11, 12 microns or MODIS channels 1, 2, 3, 4, 5, 6, 7, 27, 28, 29, 31, 32) are simulated.

- Total 744 radiance profiles over the ocean (81 clear scenes, 202 liquid clouds and 461 ice clouds). Each simulated radiance profile consists of 7 reflectances and 5 brightness temperatures.

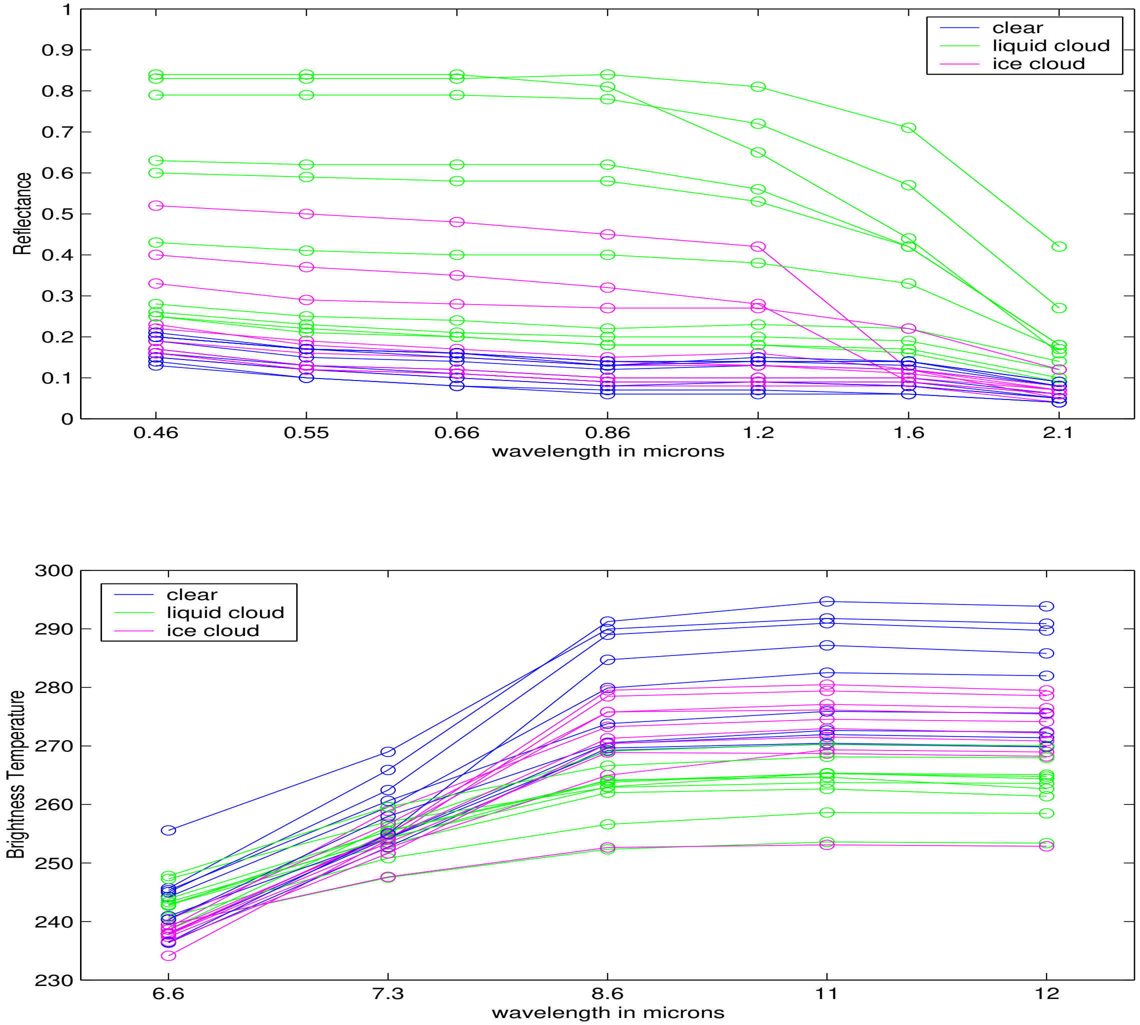


Figure 7: 10 randomly selected radiance profiles from the three scenes.

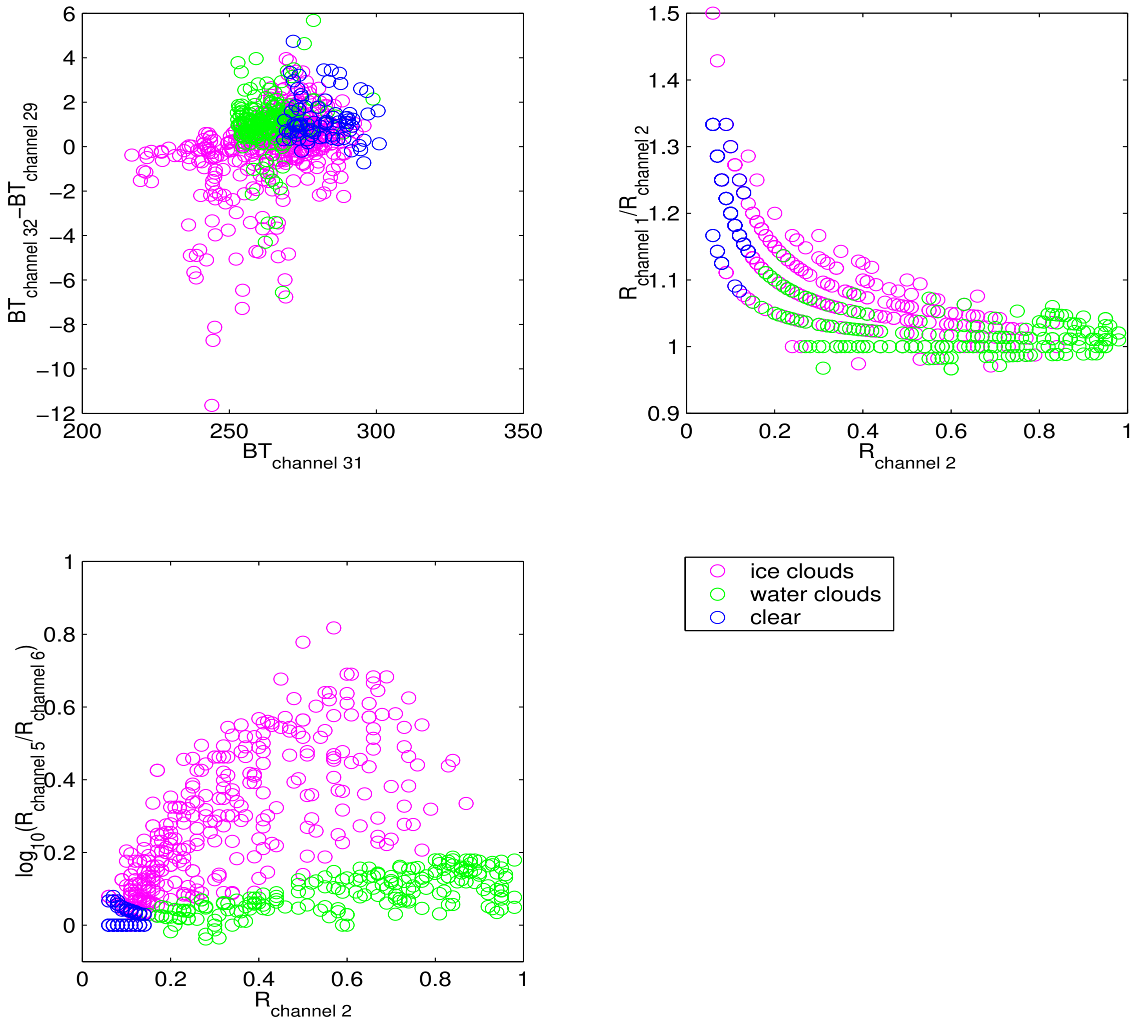


Figure 8: Scatterplots of various features.

Table 2: Test error rates for the combinations of variables and classifiers.

Number of variables	Variable descriptions	Test error rates (%)			
		MSVM	QDA	LDA	TREE
2	(i) $R_2, \log_{10}(R_5/R_6)$	11.50	14.44	36.10	14.97
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	10.16	12.83	27.81	15.24
12	(ii) original 12 variables	12.03	9.36	17.11	16.84
12	log transformed (ii)	9.89	9.63	24.06	16.84

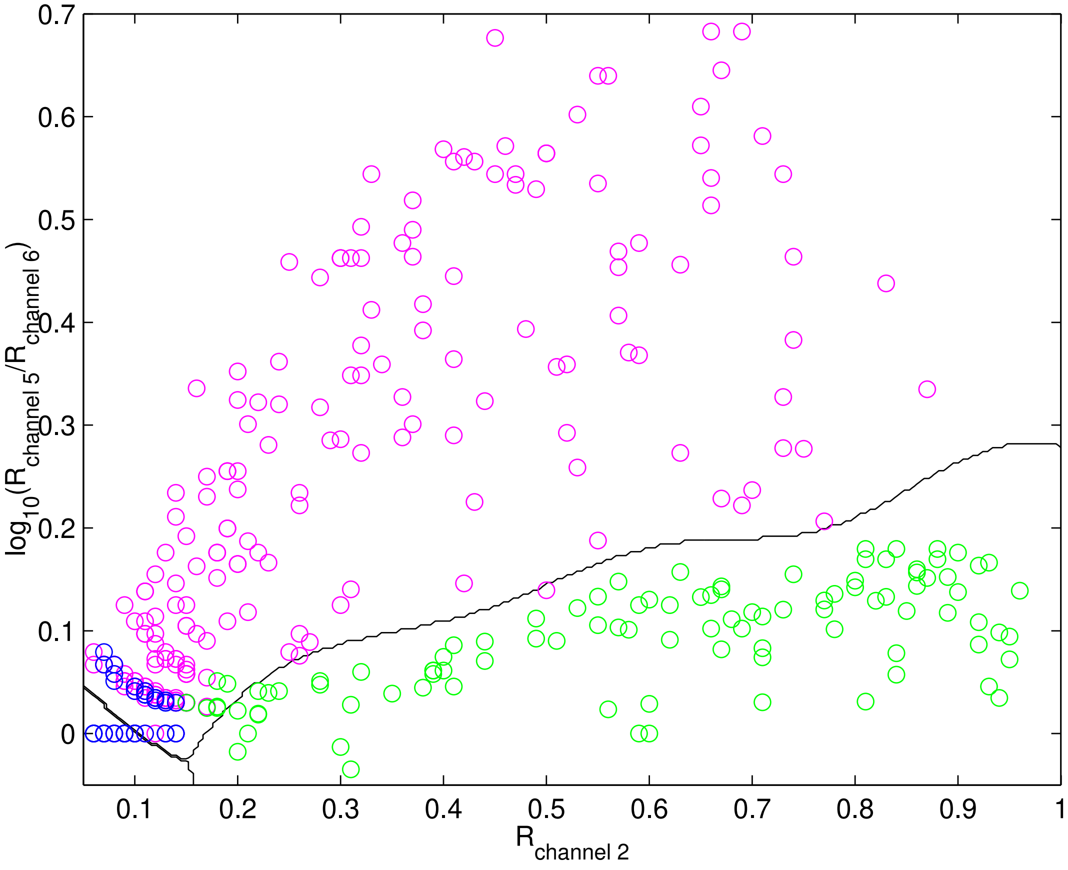


Figure 9: The classification boundaries determined by the MSVM using 370 training samples randomly selected from the bottom left plot in Figure 8.

Summary

- We have extended SVMs for the binary case to the multicategory case which encompass the binary SVM as a special case.
- Multicategory SVM implements the optimal classification rule as the sample size gets large, overcoming the suboptimality of conventional one-versus-rest approach.
- The proposed method deals with the standard and the nonstandard case in a unified way.
- Multicategory SVM has a great potential for real applications.

Future Directions

- Accommodate variable selection.
- Establish theoretical properties.
- Off-the-shelf package.

Slides of this talk and papers are available from

http://www.stat.ohio-state.edu/∼yklee