The (Nonstandard) Multicategory Suport Vector Machine, with Application to Classification of Satellite-Observed Radiance Profiles

Grace Wahba
Department of Statistics
University of Wisconsin-Madison
`http://www.stat.wisc.edu/~wahba`
$\rightarrow$ *TRLIST*

SYSID 2003
Rotterdam
August 27, 2003

We describe the Bayes rule for multicategory classification with unequal costs. Then we make some remarks about the two category SVM and other (standard) large margin classifiers. We describe the non-standard multicategory SVM, and show how it has been applied to classification of satellite-observed radiance profiles, to classify the profiles as coming from clear sky, water clouds or ice clouds.

OUTLINE

1. Multicategory Bayes risk.

2. Two category (standard) SVM's and other large margin classifi ers.

3. Multicategory penalized likelihood.

4. The (nonstandard) multicategory SVM (MSVM).

5. Application to classifi cation of satellite-observed radiance profi les.

6. Concluding remarks.

References

www.kernel-machines.org

Y. Lee, Y. Lin and G. Wahba. Multicategory Support Vector Machines, Theory, and Application to the Classifi cation of Microarray Data and Satellite Radiance Data, TR 1064(2002), revised as (OSU)TR714(2003), subm.

Y. Lee, G. Wahba and S. Ackerman. Classifi cation of Satellite Radiance Data by Multicategory Support Vector Machines, TR 1075, to appear *J. Atmos. Ocean Technology*.

Y. Lee and C.-K. Lee. Classifi cation of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19:1132–1139, 2003.

Y. Lin. Support vector machines and the Bayes rule in classifi cation. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

Y. Lee *Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data* PhD. thesis, also UW-Madison TR 1063. 2002.

X. Lin. *Smoothing Spline Analysis of Variance for Polychotomous Response Data*. PhD thesis, Department of Statistics, Uiversity of Wisconsin, Madison WI, 1998. Also TR 1003, available via G. Wahba home page.

G. Wahba. Soft and hard classifi cation by reproducing kernel Hilbert space methods. *Proceedings, National Academy of Sciences* 2002, 99, 16524-16503.

G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. In ‘Advances in Kernel Methods - Support Vector Learning’, Schölkopf, Burges and Smola (eds.), MIT Press 1999, 69-88.

G. Wahba, Y. Lin, Y. Lee, and H. Zhang. Optimal properties and adaptive tuning of standard and non-standard support vector machines. In D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, pages 129–148. Springer, 2002.

♣♣ 1. Multicategory Bayes risk.

Conskder k populations, each member of which has a predictor (attribute) variable $t \in \mathcal{T}$, (prior) density of t in the jth population is $h_j(t)$, suppose that the prior probability (relative frequency) of the jth population is π_j. Let $p_j(t)$ be the probability that the next observation is from population j:

$$p_j(t) = \frac{\pi_j h_j(t)}{\sum_{j=1}^{k} \pi_j h_j(t)}$$

Let C_{jr} be the cost of misclassifying a j as an r. Then the Bayes rule, to minimize the expected cost is to choose j to minimize

$$\sum_{\ell=1}^{k} C_{\ell j} p_j(t).$$

Problem: Build a classifi er which is targeted at the Bayes rule, from an unrepresentative training set. This has been done with the (nonstandard) multicategory support vector machine (MSVM) of [LeeLinWahba(2002)] [LeeWahbaAckerman(2003)] [LeeLee(2003)] [Lee(2002)] [Wahba(2002)] [WahbaLinLeeZhang(2002)].

♣♣ 2. Two category (standard) SVM's and other large margin classifi ers.

Standard, two-category large margin classifi ers can be described as follows: The classifi er is obtained by constructing a function $f(t)$ such that $f(t) > 0$ labels a subject with attribute vector t as being in the $'' + ''$ class, and $f(t) < 0$ as being in in the $'' - ''$ class. Given a training set $\{y_i, t_i, i = 1, \cdots, n, y_i = \pm 1\}$. f is obtained as the minimizer in $\mathcal{H}_K$ of

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|f\|^2_{\mathcal{H}_K}$$

where $\mathcal{H}_K$ is some reproducing kernel space (whose RK may contain some parameters) and

$$\mathcal{C}(y_i, f(t_i))) \equiv c(y_i f(t_i)) = c(\tau),$$

say. For the (original) SVM, $c(\tau) = (1-\tau)_+$. The penalized log likelihood estimate corresponds to $c(\tau) = log(1 + e^{-\tau})$. Many other $c's$ have been proposed: $(1 - \tau)^p, p \geq 1$, which for $p = 2$ is equivalent to penalized least squares a.k.a ridge regression, $e^{-\tau}$, ${(1 - \tau)_+}^p$ and others (some noted in [Wahba2002]).

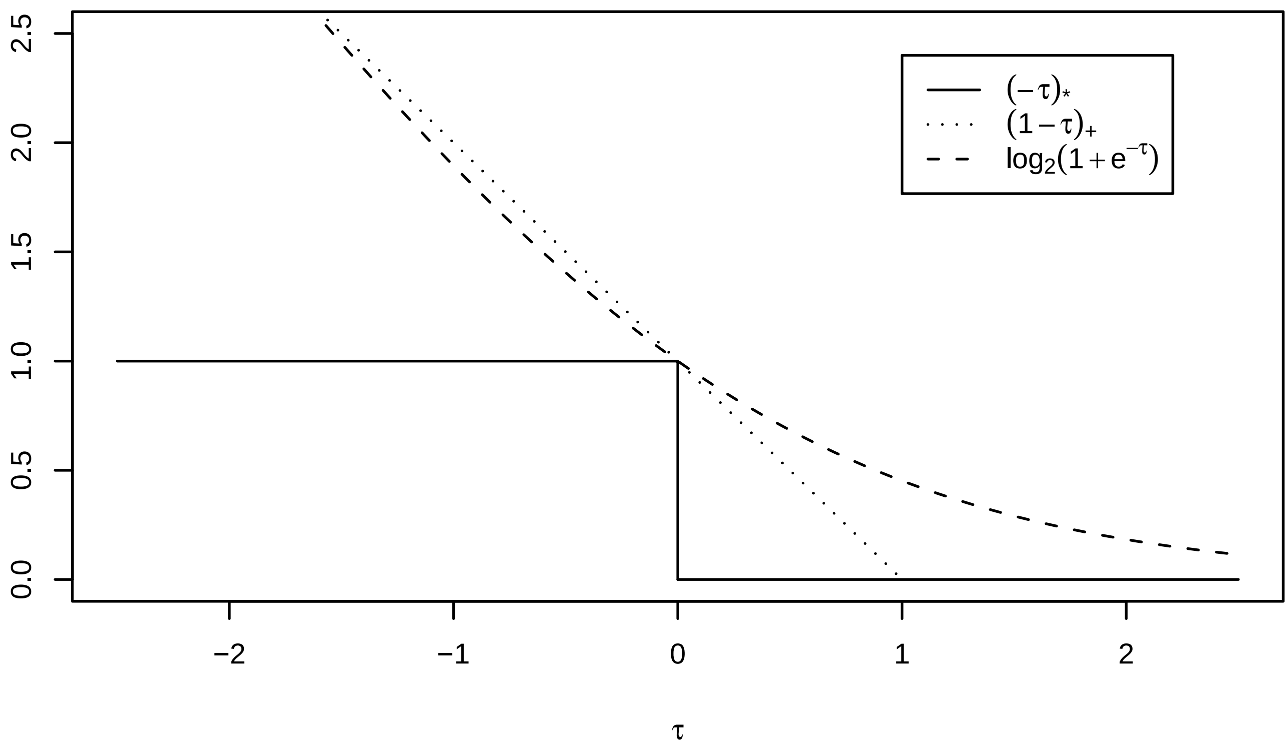


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$. Comparison of the misclassifi cation counter $c(\tau) = (-\tau)_*$, the c for the SVM $(1-\tau)_+$, and the penalized log likelihood $log_2(1+e^{-\tau})$. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on $(-\tau_*)$ and will be a looser bound than some SVM (hinge) function $(1-\theta\tau)_+$. Many other "large margin" classifi ers.

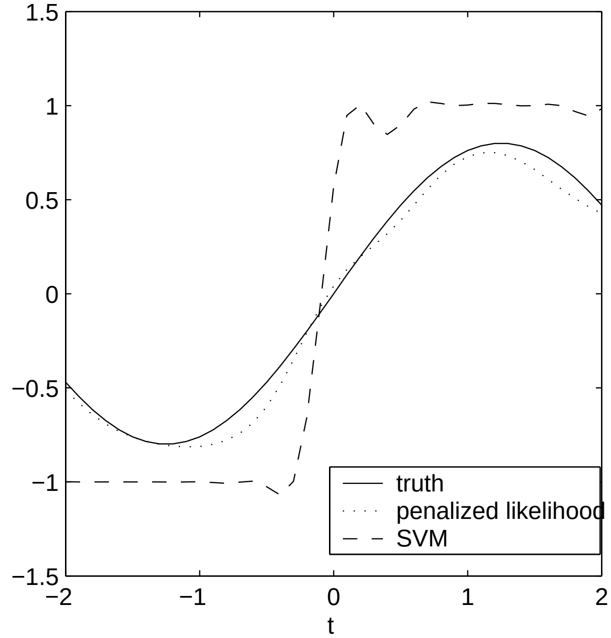
♣♣ 2. Two category (standard) SVM's and other large margin classifi ers (cont.).

The standard, two-category SVM estimates the $sign$ of the log odds ratio $f(t)$, $sign\ log[p_1(t)/p_2(t)] \equiv sign\ f(t)$. The penalized likelihood ($c(\tau) = log(1 + e^{-\tau})$ estimates the log odds ratio $f(t)$ itself, and hence estimates $p_1(t) \equiv e^{f(t)}/(1 + e^{f(t)})$.

All of the reasonable large margin classifi ers will estimate some function $f(\hat{t})$ sucn that $sign\ f(\hat{t})$ approximates $sign\ log[p_1(t)/p_2(t)]$ (or they could be considered not reasonable).

The various suggestions have differing computational demands on differing examples, and, if the classes are easily separable, various classifi ers have been shown to behave similarly, although their behavior on on overlapping classes may be different. The various proponents of the different suggestions generally have reasons why their classifi er is good, but claims of a universal best classifi er probably will not withstand scrutiny. Good tuning is at least as important as the particular choice of c.

♣♣ 2. Two category (standard) SVM's and other large margin classifi ers (cont.).



Demonstration of Yi Lin's lemma:(Lin2002). 300 Bernoulli random variables were generated, equally spaced t from $p(t) = 0.4sin(0.4\pi t) + 0.5$ Solid line: $(2p(t) - 1)$. Dotted line:$(2p_\lambda - 1)$, where p_λ is (optimally tuned) penalized likelihood estimate of p. Dashed line: $f_{svm\ \lambda}$, is (optimally tuned) SVM. Observe $f_{svm\ \lambda} \sim \pm 1$, thus p_λ is estimating $p(t)$, whereas $f_{svm\ \lambda}$ is estimating $sign(2p-1)$ = $sign(p-1/2)$= $sign\ f$. (based on Gaussian K) (plot: Yoonkyung Lee)

♣♣ 3. Multicategory penalized likelihood estimates.

[X. Lin 1998]

♣♣ 4. Multicategory support vector machines (MSVMs).

From [LeeLinWahba02] [LeeWahbaAckerman03][LeeLee03] [Lee02] [WahbaLinLeeZhang02]. $k > 2$ categories. In the papers above, the data is coded in a special way:

$$y_i = (y_{i1}, \cdots, y_{ik}), \sum_{j=1}^{k} y_{ij} = 0,$$

with $y_{ij} = 1$ if the ith subject is in category j and $y_{ij} = -\frac{1}{k-1}$ otherwise. $y_i = (1, -\frac{1}{k-1}, \cdots, -\frac{1}{k-1})$ indicates y_i is from category 1. The MSVM produces $f(t) = (f^1(t), \cdots f^k(t))$, with each $f^j = d^j + h^j$ with $h^j \in \mathcal{H}_K$, *required to satisfy a sum-to-zero constraint*

$$\sum_{j=1}^{k} f^j(t) = 0,$$

for all t in $\mathcal{T}$. The largest component of f indicates the classifi cation.

♣♣ 4. Multicategory support vector machines (MSVMs) (cont.).

Standard case: representative samples, equal miss-classifi cation costs:

Let $L_{jr} = 1$ for $j \neq r$ and 0 otherwise. The MSVM is defi ned as the vector of functions $f_\lambda = (f_\lambda^1, \cdots, f_\lambda^k)$, with each h^k in $\mathcal{H}_K$ satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k} L_{cat(i)r}(f^r(t_i) - y_{ir})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

equivalently

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r \neq cat(i)} (f^r(t_i) + \frac{1}{k-1})_+ + \lambda \sum_{j=1}^{k} \|h^j\|^2_{\mathcal{H}_K}$$

where $cat(i)$ is the category of y_i (i.e. a ”charge” on $f^r(t_i)$ if y_i is not category r.)

The $k = 2$ case reduces to the usual 2-category SVM.

The target for the MSVM has been shown to be $f(t) = (f^1(t), \cdots, f^k(t))$ with $f^j(t) = 1$ if $p_j(t)$ is bigger than the other $p_l(t)$ and $f^j(t) = -\frac{1}{k-1}$ otherwise-that is, it targets the code for the correct classifi cation.

♣♣ 4. Multicategory support vector machines(MSVMs)(cont.).

The nonstandard MSVM:

More generally, suppose the sample is not representative, and misclassifi cation costs are not equal. Let

$$L_{jr} = (\pi_j/\pi_j^s)C_{jr}, \quad j \neq r, \quad = 0, j = r.$$

where C_{jr} is the cost of misclassifying a j as an r, π_j is the prior probability of category j, and π_j^s is the fraction of samples from category j in the training set. The nonstandard MSVM minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)} L_{cat(i)r}(f^r(t_i)+\frac{1}{k-1})_+ + \lambda\sum_{j=1}^{k}\|h^j\|^2_{\mathcal{H}_K}$$

subject to the sum-to-zero constraint. As before the largest component determines the classifi cation. Then the nonstandard MSVM has as its target the Bayes rule, which is to choose the j which minimizes

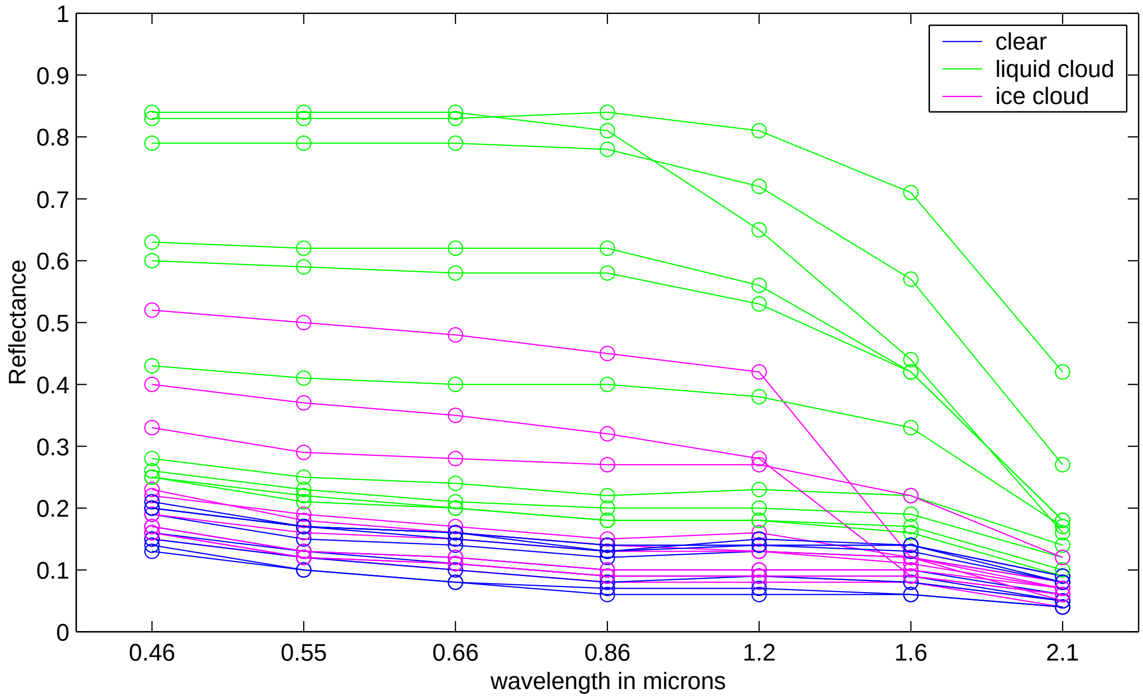
$$\sum_{\ell=1}^{k} C_{\ell j}p_\ell(x).$$

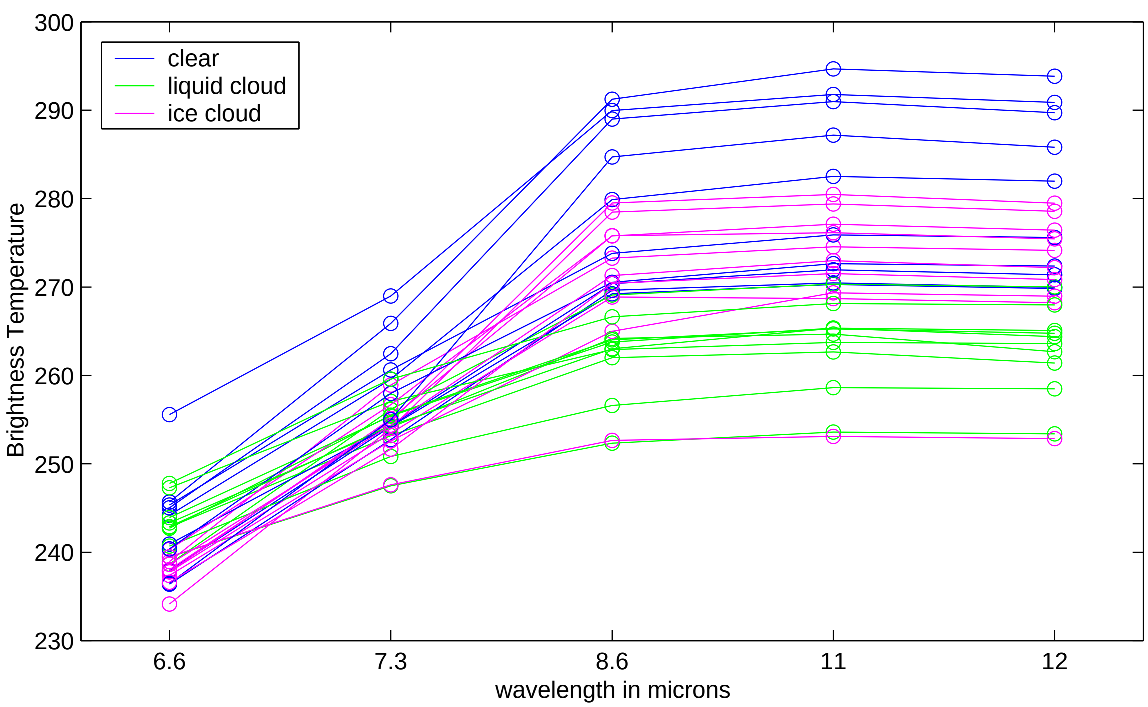
♣♣ 5. The classifi cation of upwelling MODIS radiance data to clear sky, water clouds or ice clouds.

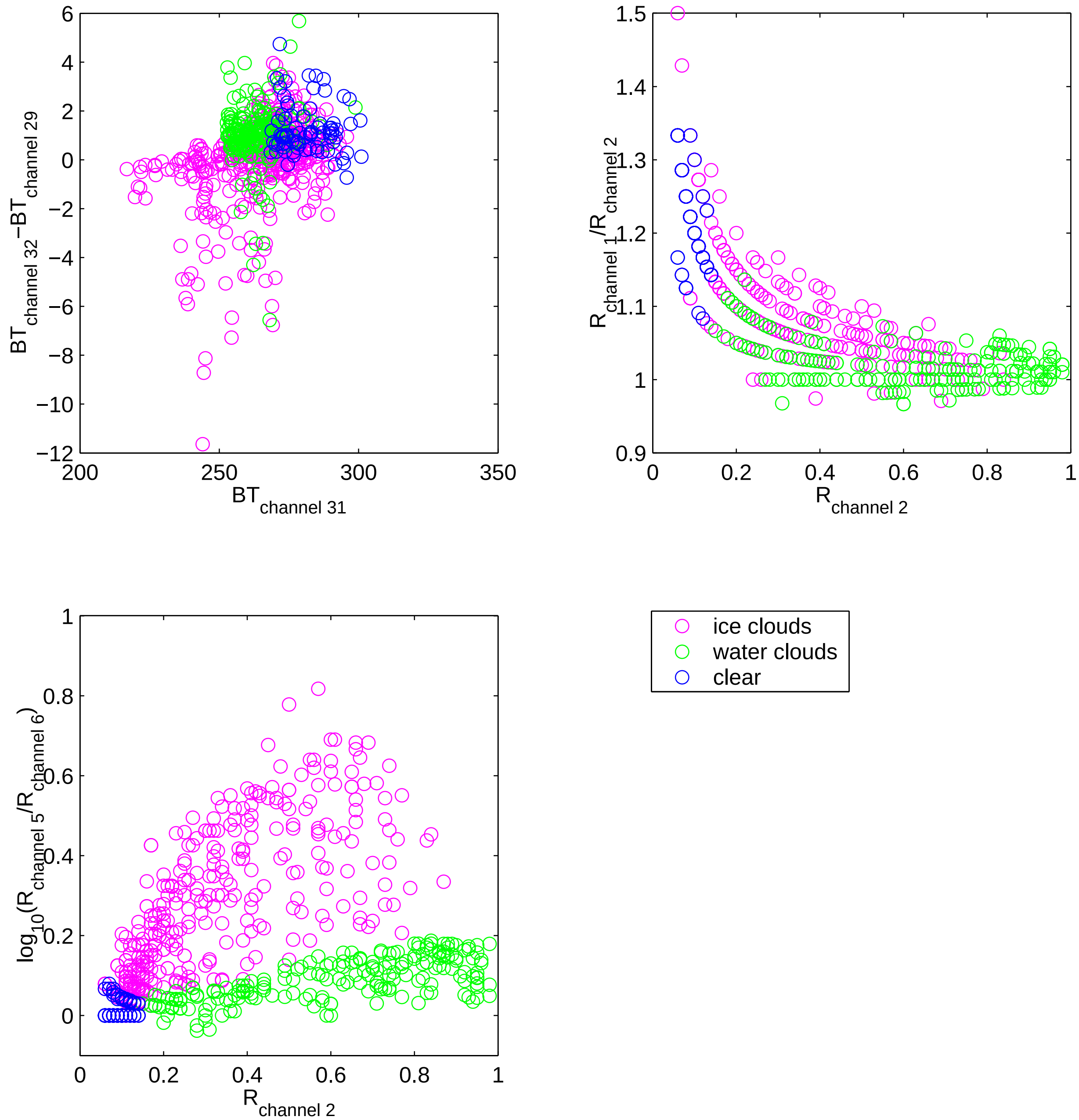
From [LeeWahbaAckerman03].Classifi cation of 12 channels of upwelling radiance data from the satellite- borne MODIS instrument. MODIS is a key part of the Earth Observing System (EOS).

Classify each vertical profi le as coming from clear sky, water clouds, or ice clouds.

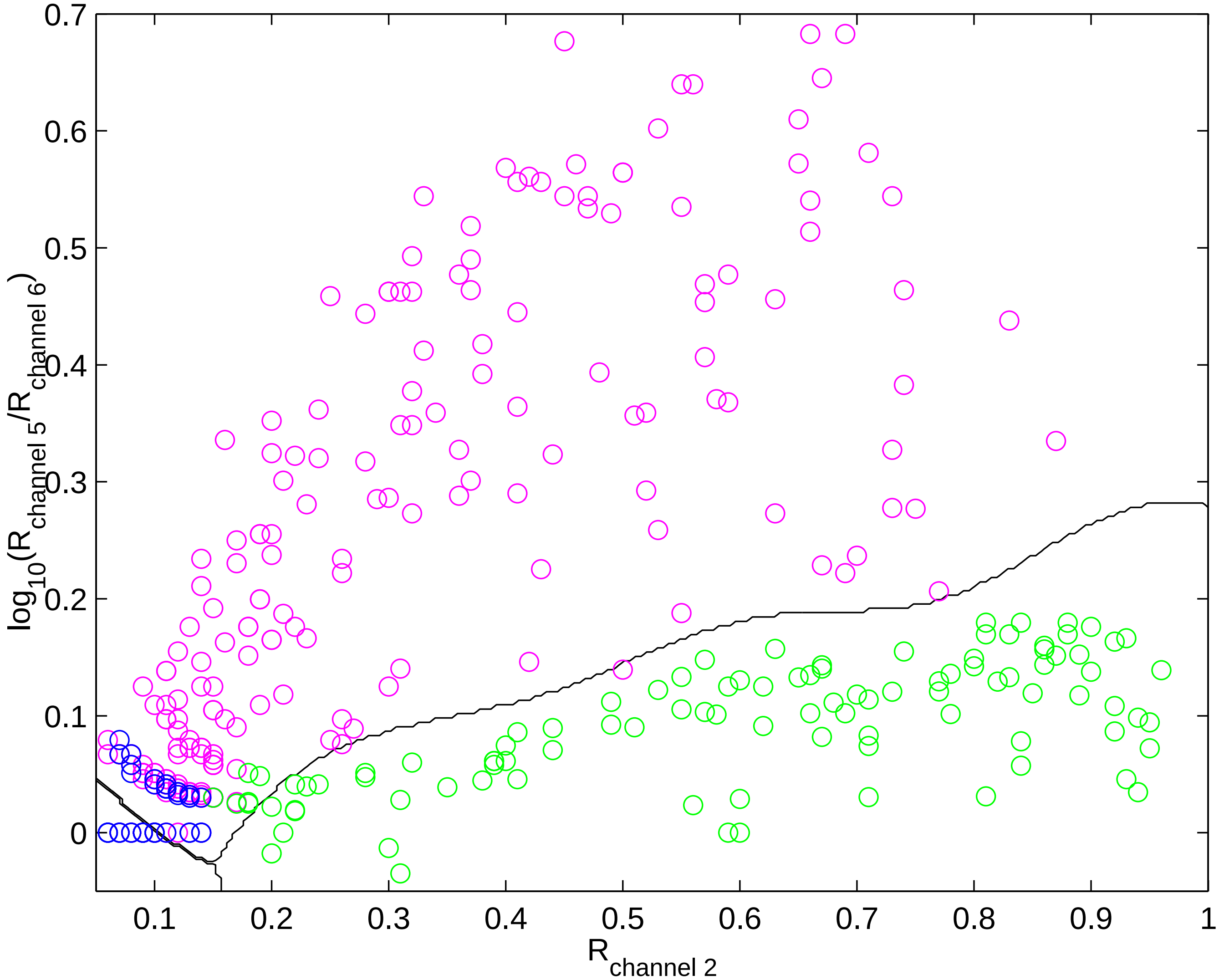
Next page: 744 simulated radiance profi les (81 clear-blue, 202 water clouds-green, 461 ice clouds-purple). 10 samples from clear, from water and from ice:


Reflectance
wavelength in microns
clear
liquid cloud
ice cloud
0.46
0.55
0.66
0.86
1.2
1.6
2.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

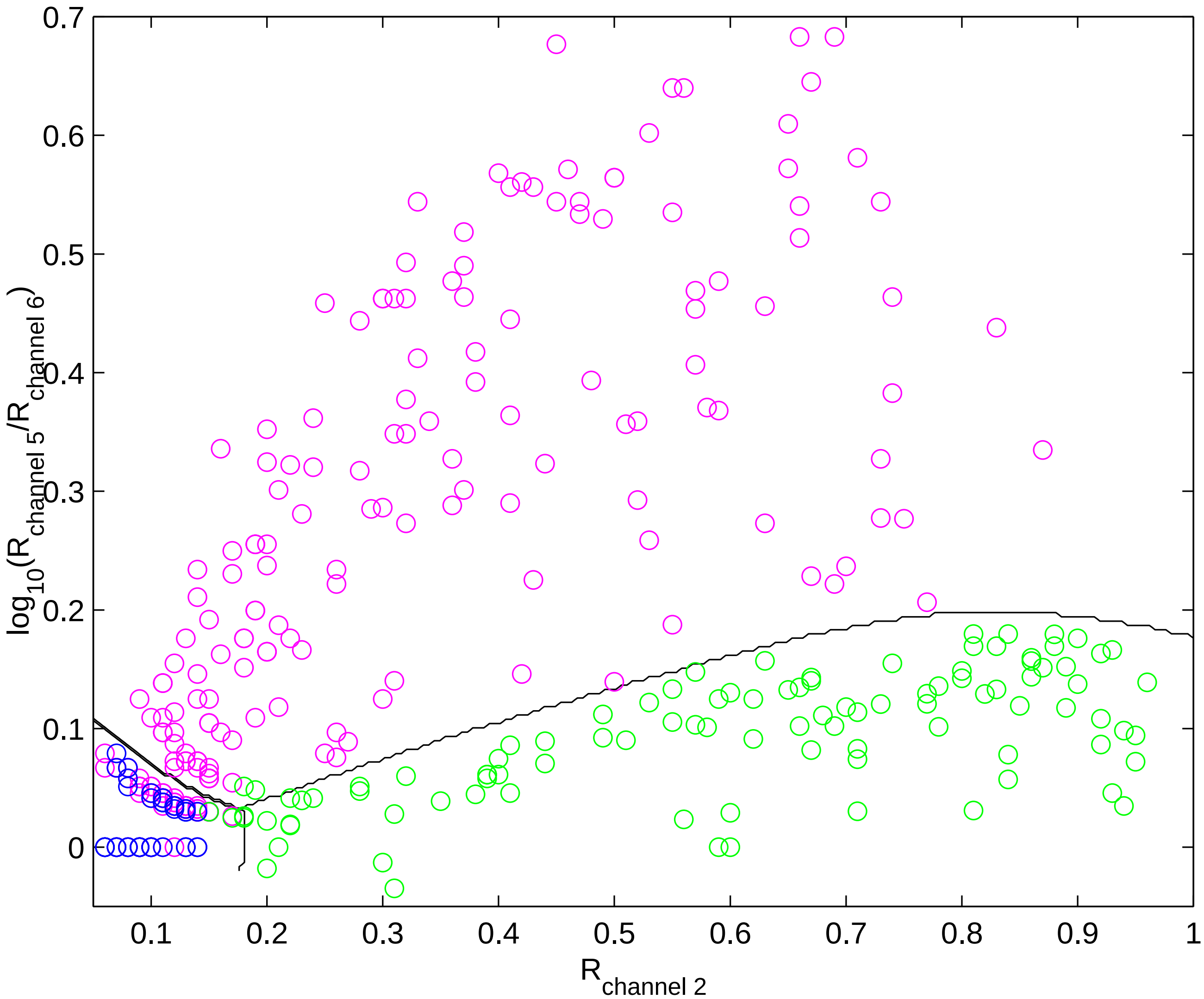

Brightness Temperature
wavelength in microns
clear
liquid cloud
ice cloud
6.6
7.3
8.6
11
12
230
240
250
260
270
280
290
300



Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



Classifi cation boundaries on the 374 test set determined by the MSVM using 370 training examples, two variables, one is composite. Y. K. Lee Student poster prize AMetSoc Satellite Meteorology and Oceanography session.



Classifi cation boundaries determined by the nonstandard MSVM when the cost of misclassifying clear clouds is 4 times higher than other types of misclassifi cations.