Comprehensive Analysis of Hyperspectral Data Using Band Selection Based on Sparse Support Vector Machines

Sofya Chepushtanova*, Michael Kirby

Department of Mathematics
Colorado State University

March 2, 2013

Introduction

Feature selection:

Process of selecting a subset of features (variables) according to a certain evaluation criterion.

Motivation:

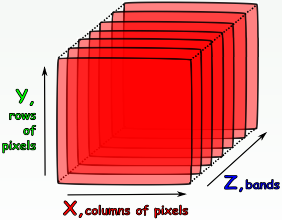
- Improving performance of a predictor
- Dimension reduction
- Data understanding, data storage

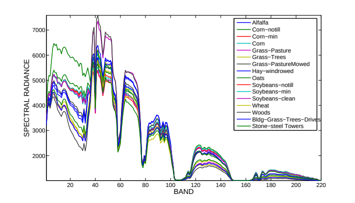
Applications:

- Gene selection problem
- Text categorization
- Hyperspectral imagery (HSI)

Hyperspectral Imagery (HSI)

- Hyperspectral sensors generate imagery in the electromagnetic spectrum, capturing aspects that are imperceptible to the human eye.
- The radiance of materials is measured within each pixel area at a very large number of contiguous spectral wavelength bands.
- Spatial and spectral information is contained in data cubes.
- Each pixel is a vector $x \in \mathbb{R}^n$.





Hyperspectral Imagery (HSI)

Advantage: detailed radiance information
Disadvantage: huge amount of data

Band selection

identify bands that contain the most discriminatory information $\longrightarrow$ use them for further analysis

Methods

1. *Filters:*
   all bands $\longrightarrow$ filter $\longrightarrow$ band subset $\longrightarrow$ predictor
2. *Wrappers:*
   all bands $\longrightarrow$ space of band subsets $\longrightarrow$ predictor (wrapper) $\longrightarrow$ band subset
3. *Embedded algorithms:*
   all bands $\longrightarrow$ predictor $\longrightarrow$ band subset

- **Embedded method:**

  A **S**parse (ℓ_1-norm) linear **S**upport **V**ector **M**achine (SSVM) is a classification method that combines a good performance and automatic variable selection:

  - the ℓ_1**-norm** ($\|w\|_1 = \sum_{i=1}^{n} |w_i|$) supresses many components of the weight vector w in the decision function $f(x) = sgn(w^T x + b)$.

- Nonzero components of w indicate the **relevant bands**.

- **Bagging** (= **B**ootstrap **AGG**regat**ING**) is used to reduce variability in w.

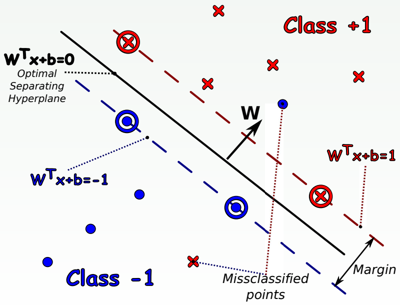
- **Training data** $x_i \in \mathbb{R}^n$ with class labels $d_i \in \{-1, +1\}, i = 1, \ldots, m$ (let $D = diag(d_i)$ and X be the $m \times n$ data matrix)
- **Separating hyperplane** $P = \{x : w^T x + b = 0\}$, $w \in \mathbb{R}^n$ is normal to P
- Points on $w^T x + b = \pm 1$ are **support vectors**
- The optimal P has the largest **margin** $2/\|w\|_1$

SSVM:

$$\min_{w,b,\xi} \quad \|w\|_1 + Ce^T \xi$$
$$\text{s. t.} \quad D(Xw + be) + \xi \geq e,$$
$$\xi \geq 0.$$

Decision function:

$$f(x) = sgn(w^T x + b)$$

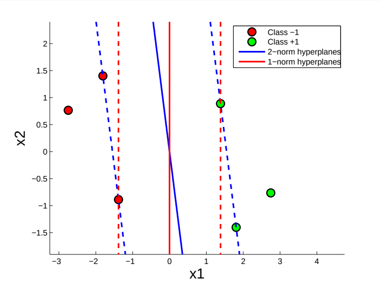


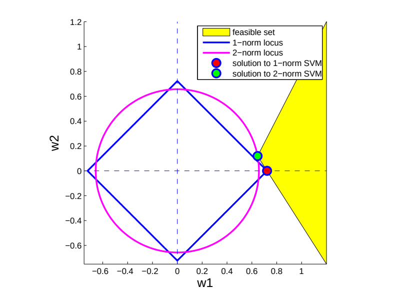
Sparse Support Vector Machines

- SSVM $\Rightarrow$ LP (with $\|w\|_1 = w^+ + w^-$ and $w = w^+ - w^-$):

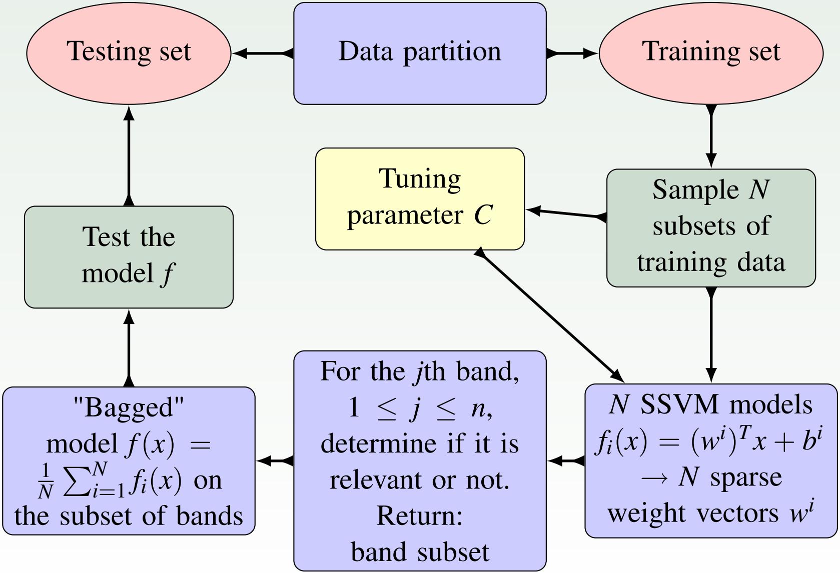
$$\begin{aligned} \min_{w^+,w^-,b,\xi} \quad & e^T(w^+ + w^-) + Ce^T\xi \\ \text{s. t.} \quad & D(X(w^+ - w^-) + be) + \xi \geq e, \\ & w^+, w^-, \xi \geq 0. \end{aligned}$$

- Sparsity of ℓ_1-norm:





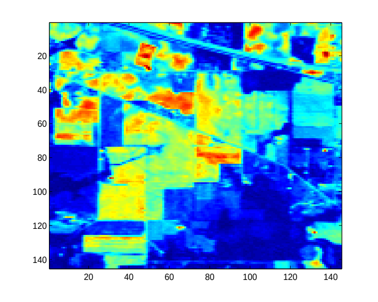
Band Selection Algorithm

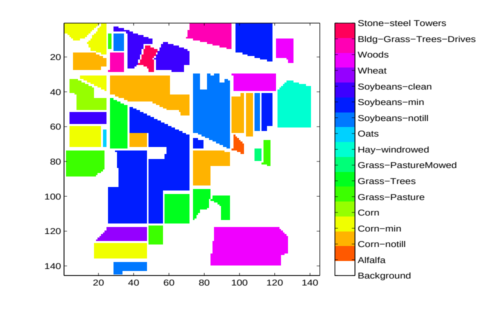


Band Selection Algorithm

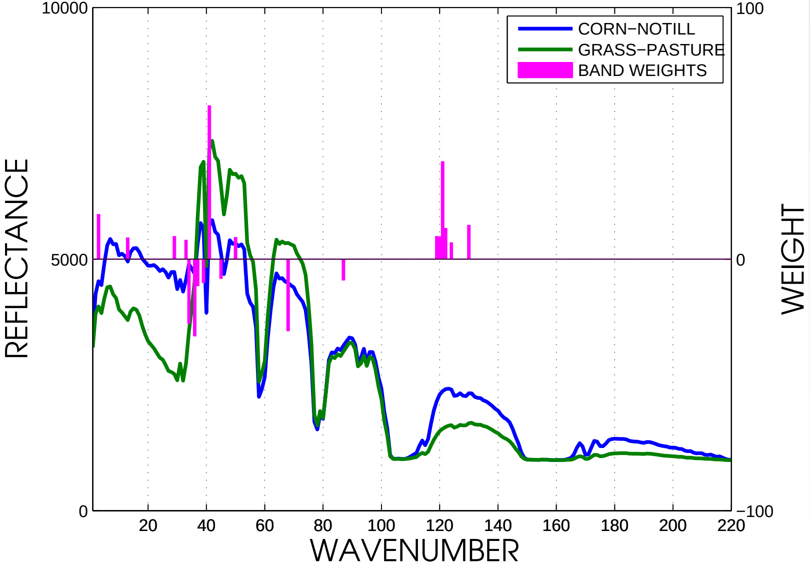
$$
\begin{array}{lcccc}
\text{band } 1: & [w_1^1 & w_1^2 & \cdots & w_1^N] \\
\text{band } 2: & [w_2^1 & w_2^2 & \cdots & w_2^N] \\
 & \vdots & \vdots & \cdots & \vdots \\
\text{band } j: & [w_j^1 & w_j^2 & \cdots & w_j^N] \\
 & \vdots & \vdots & \cdots & \vdots \\
\text{band } n: & [w_n^1 & w_n^2 & \cdots & w_n^N]
\end{array}
$$

- **Reducing variability and band ordering:**
  - Null hypothesis $H_0 : \{ \text{ mean}(j\text{th band weight sample}) = 0\}$
  - Calculate statistic of the sample $t_* = \frac{\text{mean(jth band weight sample)}-0}{\text{std}/\sqrt{N}}$
  - Calculate p-value $= P(t \geq |t_*|)$.
    If p-value $<$ significance level $\rightarrow$ reject H_0
  - Standard deviation of the sample (std) (or variance, e.g. std^2)





- Aiborne Visible/Infrared Imaging Spectrometer (AVIRIS): Indian Pines data set collected in an agricultural area of northern Indiana in 1992
- 145×145 images, 220 spectral bands (ranging from 0.4 to $2.5\mu m$)
- Ground truth is known for 49% of the pixels
- 16 classes ranging from 20 to 2468 pixels



10000
5000
0
100
0
−100
20
40
60
80
100
120
140
160
180
200
220
CORN−NOTILL
GRASS−PASTURE
BAND WEIGHTS
REFLECTANCE
WEIGHT
WAVENUMBER

2 classes: Corn-notill versus Grass/Pasture

- Testing accuracy 98.9% is achieved on both full data and reduced data.

Band number	Mean w value	Standard deviation	P-value
34	-0.0041	0.0024	9.2852e-09
41	0.0037	0.0028	9.0592e-07
3	0.0021	0.0018	6.0899e-06
36	-0.0022	0.0028	3.9240e-04
121	0.0029	0.0037	0.0005
28	0.0018	0.0023	0.0006
68	-0.0024	0.0036	0.0019
29	0.0011	0.0016	0.0034
1	-0.0003	0.0006	0.0039
73	-0.0014	0.0025	0.0080
35	-0.0008	0.0013	0.0111
13	0.0006	0.0012	0.0179
43	0.0007	0.0014	0.0266
33	0.0008	0.0017	0.0285
130	0.0011	0.0025	0.0325
123	0.0012	0.0029	0.0333

One-against-one approach (OAO):
k classes $\rightarrow \frac{k(k-1)}{2}$ binary classifiers
$\rightarrow$ majority voting to assign class to a testing point.
Testing accuracy 87% is achieved on both full and reduced 16-class data.

Problem:
How to find the optimal set of bands in multiclass case?

1. retain the top several bands for each model and take the union of them.
2. find "representative bands" for each model and then take the union of them:
   - Corn-notill and Grass/Pasture:
     1 3 13 28 29 33 34 35 36 41 43 68 73 121 123 130
   - Eliminate more bands in each "cluster":
     1 3 13 [28 29] [33 34 35 36] 41 43 68 73 121 123 130

Spatial Smoothing

Zare and Gader (2008):
After a texting pixel X has been assigned a label, spatial smoothing can be done by summing votes over the eight-connected neighborhood of the pixel X

1	1	1
1	X	1
1	1	1

1	1	1
1	X	2
2	2	2

1	1	1
2	X	2
3	3	3

Accuracy over all classes of the Indian Pines data set:

No. bands	Spatial Smoothing accuracy	Testing accuracy
220	98.69	86.99
137	98.53	82.91
69	97.78	82.66
13	93.42	71.84
7	88.39	65.50
3	71.11	48.25

Conclusions/Future Work

- The band selection algorithm identifies the optimal spectral bands for classification.
- Accuracy classification rates are preserved in the reduced feature space.
- Bagging and statistical analysis allow for the systematic elimination of unimportant bands and to attribute relative significance to the retained bands.
- Multiclass band selection problem: algorithm extension for more than 2 classes of data.
- Spatial smoothing as a method for improving classification rates.

References

- V. N. Vapnik, *The nature of statistical learning theory*, New York: Springer, 1995
- L. Breiman, *Bagging predictors*, Machine learning, 24, pp. 123-140, 1996
- O. L. Mangasarian, *Arbitrary-norm separating plane*, Operations Research Letters, 24, pp. 15-23, 1997
- J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song, *Dimensionality reduction via sparse support vector machines*, Journal of Machine Learning Research, 3, pp. 1229-1243, 2003
- O. L. Mangasarian, *Exact* 1*-norm support vector machines via unconstrained convex differentiable minimization*, Journal of Machine Learning Research, 7, pp. 1517-1530, 2006
- A. Zare and P. Gauder, *Hyperspectral band selection and endmember detection using sparsity promoting priors*, IEEE Geoscience and remote sensing letters, vol. 5, no.2, pp. 256-260, 2008
- S. Chepushtanova, C. Gittins, and M. Kirby, *Band selection for classification of hyperspectral data based on sparsity of ℓ_1-norm linear support vector machine*, preprint 2013