Sparse Classification Methods for High Dimensional Data

Panos M. Pardalos

Distinguished Professor
Center For Applied Optimization
Industrial and Systems Engineering, University of Florida, USA.

Leading Scientist
National Research University Higher School of Economics
Laboratory of Algorithms and Technologies for Network Analysis (LATNA), Russia
http://www.ise.ufl.edu/pardalos/
http://nnov.hse.ru/en/latna/

December 7, 2012

Introduction - High Dimensional Data

- Massive amounts of high-throughput data can be collected simultaneously due to technological advances.
- Each observation is characterized with **thousands** of features (**p**).
  - MRI and FMRI images
  - Gene-expression microarrays
  - Spectroscopic studies
  - Web documents
- Expensive measurement costs limit the size (**n**) of most datasets to **tens** or **low hundreds**.
- **High Dimension Low Sample Sizes (HDLSS)** - **p** $>>$ **n**.

Introduction - Classification

- Classification is a **supervised** machine learning technique that maps some combination of input variables into pre-defined classes.
- Classification models estimate a **decision rule** from training data that helps to predict the class of an unknown sample.
- Classification problems appear in several applications:
  - Discrimination of cancer cells from non-cancer cells
  - Web-document classification
  - Categorization of images in Remote-Sensing applications
- Several classification methods exist in literature like,
  - Support Vector Machines
  - Neural Networks
  - Logistic Regression
  - Linear Discriminant Analysis
  - Random Forests
  - Adaboost

Classification on HDLSS datasets

- The high-dimensional data poses significant challenges to standard classification methods:
  - **Poor generalization ability** - *curse of dimensionality*
  - **Geometric distortion** - *equidistant points*
  - **Unreliable parameter estimation** - *class covariance*

G.V. Trunk. *A Problem of Dimensionality: A Simple Example* - IEEE Transactions on Pattern Analysis and Machine Intelligence (1979)

Motivation & Significance

- Poor performance of standard classification methods.
- Continued technological advances.
- **Biomarker-type** information in biomedical applications.

Scalable and **efficient** classification models with good **generalization ability** along with **model interpretability** for high dimensional data problems.

Dimensionality Reduction

- The dimensionality reduction techniques decrease the complexity of the classification model and thus improve the classification performance.
- Dimensionality reduction techniques can be categorized as:
  - **Feature Extraction**
    - Transform the input data into a set of *meta*-features that extract relevant information from the input data for classification.
    - Limited model interpretability.
  - **Feature Selection**
    - Select a subset of features based on some *optimality* criteria
    - Advantage of model interpretability by a domain expert.
    - *Biomarker-type* information in biomedical applications.
    - Combinatorial optimization.

Feature Selection

- Feature Selection can be broadly classified as:
  - **Filter methods**
  - **Wrapper methods**
  - **Embedded methods**

Y. Saeys, I. Inza, & P. Larranaga. *A review of feature selection techniques in bioinformatics* - Bioinformatics (2007)

Filter Methods

- Feature subsets are ranked using a *feature relevance* score and low-ranking features are removed.
- Filter methods are independent of the classification method.
- Filter methods can be broadly categorized as:
  - **Univariate techniques**
    - Computationally efficient
    - Scalability
    - Ignore feature dependencies
  - **Multivariate techniques**
    - Feature dependencies
    - **NP-hard** problem
    - Higher computational complexity
    - Prone to over-fitting

Wrapper Methods

- Wrapper methods integrate the classifier hypothesis search within the feature subset search.
- A search procedure is defined in the feature space to select subsets of features.
- A specific feature subset is evaluated by training and testing a **specific** classification model.
- Advantages:
  - Feature dependencies
  - Interaction between feature subset selection and model selection
- Disadvantages:
  - Over-fitting
  - Computationally intensive

Embedded Methods

- Embedded methods also integrate the classifier hypothesis search within the feature subset search.
- Feature selection is part of model building and is generally achieved by **regularization** techniques.
- **Specific** to a classification model.
- Selects **common** subset of features for all classes. - *global sparsity*

Current Research

- **Sparse Proximal Support Vector Machines (sPSVMs)**
- **Fisher-based Feature Selection Combined with Support Vector Machines to Characterize Breast Cell Lines using Raman Spectroscopy.**

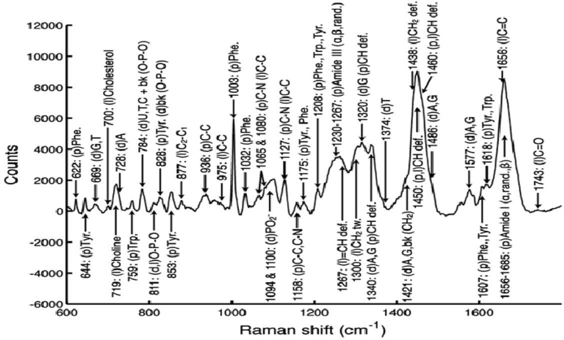
Fisher-based Feature Selection Combined with Support Vector Machines to Characterize Breast Cell Lines using Raman Spectroscopy

Introduction - Cancer

- **Cancer** remains one the leading causes of death throughout the world.
- **Breast cancer** is the most common type of cancer in women, excluding skin cancers.
- In 2009, approximately 40,107 women died from breast cancer, and over 250,000 new cases were diagnosed.
- Lack of cell and tumor specific treatments - **personalized medicine**.
- **Classify** and **characterize** cell types for the selection of therapies for use *in-vivo*.
- Extract **biomarker-type** information that contribute to the differences between cell-types.

Introduction - Raman Spectroscopy

- Raman Spectroscopy has demonstrated the potential to significantly aid in the research, diagnosis and treatment of various cancers.
- Raman spectroscopic analysis of biological specimens provides a **spectral fingerprint** rich in molecular compositional information without disrupting the biological environment.



http://w4.phys.uvic.ca/medphys/people/AJ/jirasek.html

Research Objective

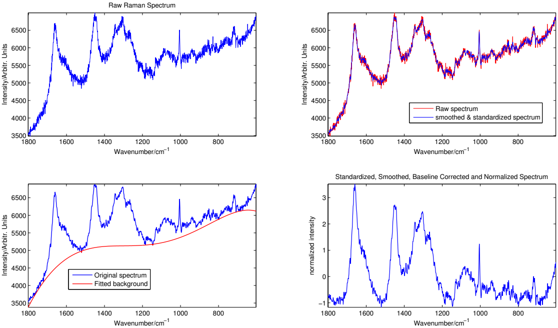
Construct a **classification framework** that would combine **feature selection** and **classification** to characterize Breast cell lines using Raman Spectroscopy.

Data Collection

- Raman spectra of five breast cell lines **MCF7, BT474, MDA-MB-231** (cancer cell lines) and **MCF10A,MCF12A** (non-cancer cell lines) are collected by Renishaw 2000 InVia Spectrometer System coupled to a Leica Microscope.
- **25-40** spectra (**n**) were collected from each cell line.
- Apparent outliers were removed by visual inspection.

Data Preprocessing

- X-axis standardization
- Savitsky-Golay Smoothing
- Background Subtraction
- Normalization



Each spectrum is characterized by **1200** measurements (**p**) between wavenumbers 601 cm^{-1} and 1800 cm^{-1}

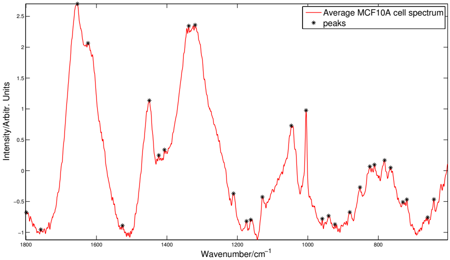
Raman spectral datasets (**p** $>>$ **n**) can be characterized as **HDLSS datasets**.

Fisher-based Feature Selection (FFS)

- Several comparative studies have been performed on univariate and multivariate filter techniques for gene expression datasets.
- Surprisingly, it has been shown that the univariate selection techniques yield consistently better results than multivariate techniques.
- The differences are attributed to the difficulty in extracting the feature dependencies from limited sample sizes.
- In a Raman spectrum, most biologically relevant molecular species correspond to the **peaks**.
- A univariate filter-based technique based on Fisher Criterion called **Fisher-based Feature Selection (FFS)** is developed and involves the following stages:
    - **Peak finding**
    - **Peak coalescing**
    - **Feature ranking**

FFS - Peak Finding

- The set of peaks S for a specific cell line are defined as local maxima given by:
$$S = \{x^* | f(x^*) \geq f(x) \quad \forall x \in \mathcal{N}_\epsilon(x^*)\}, \tag{1}$$
where x^* represents the peak location, $f(x^*)$ is the corresponding intensity value of the average spectrum and $\mathcal{N}_\epsilon(x^*)$ represents an ϵ-neighborhood around x^*.

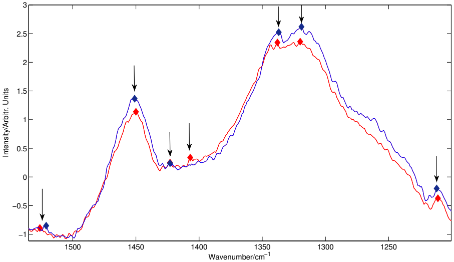


FFS - Peak Coalescing

- The number of clusters N_C is defined as:

$$N_C = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{c} \sum_{x_j \in C_i} (x_j - \mu_i)^2 \quad i = 1, 2, \ldots, c \tag{2}$$

C_i represents the cluster i, μ_i is the mean of cluster i, x_j is the peak j assigned to cluster i.



FFS - Feature Ranking

- The features are ranked based on **Fisher Criterion**.
- For a given feature i, the fisher score is defined as:

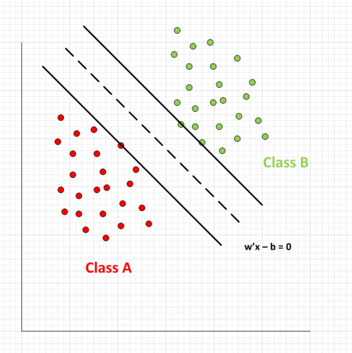
$$J_i = \frac{(\mu_1^i - \mu_2^i)^2}{\frac{(s_1^i)^2}{n_1} + \frac{(s_2^i)^2}{n_2}} \quad \forall i \in S, \tag{3}$$

  where, μ_j^i, $(s_j^i)^2$ and n_j are the sample mean, variance and the number of data samples in class j and S is the set of selected peaks.
- Fisher scores would be high for features having high **mean inter-class** separation while the total **within-class variance** is small.

Fisher-based Feature Selection (FFS)

Support Vector Machines (SVMs)

- Binary classifier
- Linearly separable datasets
- Margin maximization



V. Vapnik. *The Nature of Statistical Learning Theory* - Data Mining and Knowledge Discovery (1995).

SVMs

- Consider binary classification problem with the training set S defined as:

$$S = \{(\mathbf{x_i}, y_i) | \mathbf{x_i} \in \Re^p, y_i \in \{-1, 1\}\}, \quad i = 1, 2, ...n \tag{4}$$

- Let the separating hyperplane P that maximizes the margin be defined as:

$$P = \{x \in \Re^p \quad | \quad \langle \mathbf{w}, \mathbf{x} \rangle - b = 0\} \tag{5}$$

- The optimal (**w**,b) is found by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} ||w||^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle - b) \geq 1 \quad \forall i = 1, 2, \ldots, n \end{aligned} \tag{6}$$

C-SVMs

- SVMs are susceptible to the presence of outliers.
- Linear separation in real-world datasets.
- SVMs are modified as:

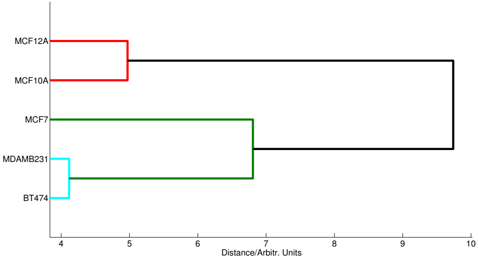
$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle - b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, 2, \ldots, n \end{aligned} \tag{7}$$

C-Support Vector Machines (C-SVMs)

Multi-class SVMs

- Two general approaches to extend SVMs to multi-class problems:
  - One-against-One (OAO) - $n(n-1)/2$ binary classification tasks
  - One-against-All (OAA) - n binary classification tasks
- Instead, SVMs is extended using **hierarchical clustering**.
- An **agglomerative** hierarchical cluster tree is generated from the pairwise **euclidean** distances of the average spectra of cell lines.

- Four binary classification tasks:
  - Cancer Vs. Non-Cancer
  - MCF7 Vs. Rest Cancer
  - MCF10A Vs. MCF12A
  - MDA-MB-231 Vs. BT474



FFS-SVMs Classification framework

Given any two cell lines, the classification framework is built as:

- Spectral Preprocessing
- Fisher-based Feature Selection
  - Peak Finding
  - Peak Coalescing
  - Feature Ranking
- C-SVMs Classification
- Cross Validation using repeated random sub-sampling (100 repetitions).

FFS-SVMs Classification framework

Classification Accuracies

Classification Task	# of selected features	Accuracy (%)	Sensitivity (%)	Specificity (%)
Cancer Vs Non-Cancer	38	99.5	99.8	98.6
MCF7 Vs Rest-Cancer	32	99.3	96.6	100
BT474 Vs MDA-MB231	42	97.4	91.7	100
MCF10A Vs MCF12A	42	91	97.1	62

Table: Sensitivity, Specificity and average classification accuracy for the four binary classification tasks obtained from C-SVMs and validated using random sub-sampling(100 repetitions).

Accuracy Comparison

	Cancer vs. Non-Cancer	MCF7 vs. Rest-Cancer	BT474 vs. MDA-MB231	MCF10A vs. MCF12A
SVMs				
Accuracy(%)	99.2	100	97.6	93.4
Sensitivity(%)	100	100	94.8	100
Specificity(%)	99.4	100	99.5	80.6
PCA-SVMs				
Accuracy(%)	99.4	98.4	98.6	92.8
Sensitivity(%)	100	95.1	96.4	99.3
Specificity(%)	98.2	100	99.5	72.9
PCA-LDA				
Accuracy(%)	99.5	98.3	96.4	85.8
Sensitivity(%)	99.9	98.9	88.2	82.8
Specificity(%)	98.6	97.6	99.3	96.6
FFS-SVMs				
Accuracy(%)	97.3	98.9	98.0	89.0
Sensitivity(%)	100	96.7	93.4	97.6
Specificity(%)	93.3	100	100	62.3

Table: Sensitivity, Specificity and average classification accuracies of four frameworks SVMs, PCA-SVMs, PCA-LDA and FFS-SVMs for the four binary classification tasks. The classification accuracies are obtained from cross-validation using random subsampling(100 repetitions).

Selected Features

Cancer vs. Non-Cancer	MCF7 vs. Rest-Cancer	BT474 vs. MDA-MB231	MCF10A vs. MCF12A
1047	1341	1049	1047
811	986	1063	1320
823	1322	760	1156
765	1658	830	1174
1450	1405	1085	1211
1660	1066	1318	941
829	622	1518	811
1086	1159	604	1338
1621	1799	1129	719
785	1316	1661	967

Table: The top 10 features selected by FFS for the four binary classification tasks.

Biological Relevance of Selected Features

- **Cancer Vs. Non-Cancer**
  - Five of the top ten discriminative features (**811**, 823, 765, 829, and **785** cm^{-1}) all correlate to **DNA** and **RNA** vibrational modes.
  - The features 1086, 1450, 1621, and 1660 cm^{-1} indicate differences in **cell membrane composition** and **cell morphology**.
- **MCF7 Vs. Rest-Cancer**
  - The majority of the features correlate to vibrations observed from **structural proteins** and the **secondary protein structure**.
- **MCF10A Vs. MCF12A**
  - The analysis of features reveal that the most significant differences may be related to **lipid composition**.
- **MDA-MB-231 Vs. BT474**
  - Several of the features listed have assignments related to **fatty acids** and **lipids**.

Sparse Proximal Support Vector Machines (sPSVMs)

Motivation

- Several embedded methods like Regularized Logistic Regression(RLRs), Sparse Support Vector Machines (S-SVMs) etc., induce *global* sparsity.
- Class-specific features - *local* sparsity.
- *Biomarker-type* information in biomedical applications.

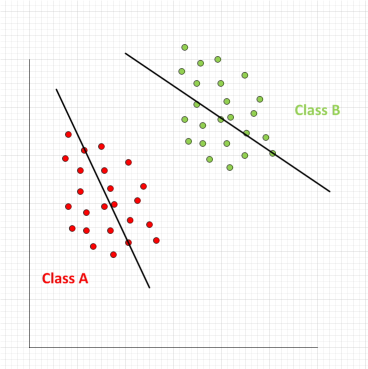
Research Objective:

Construct a new binary classifier that incorporates **class-specific** feature selection.

Sparse Proximal Support Vector Machines (sPSVMs)

Proximal Support Vector Machines (PSVMs)

- Binary Classifier
- Non-parallel hyperplanes
- Closest to one class and farthest from the other class
- Two **generalized eigenvalue problems**



O. L. Mangasarian & E. W. Wild, *Multisurface Proximal Support Vector Machine. Classification via Generalized Eigenvalues* - IEEE Transactions on Pattern Analysis and Machine Intelligence (2005)
M. R. Guarracino, C. Cifarelli, O. Seref & P. M. Pardalos, *A Classification Method based on Generalized Eigenvalue Problems* - Optimization methods and Software (2005)

PSVMs formulation

Let A $\in \Re^{m \times p}$ and B $\in \Re^{n \times p}$ represent the two classes. The hyperplane close to class A is given by:

$$P_A = \{x \in \Re^p \quad | \quad \langle \mathbf{w_A}, \mathbf{x} \rangle - b_A = 0\} \tag{8}$$

The hyperplane P_A is found by solving the following optimization problem:

$$\min_{\mathbf{w_A} \in \Re^p, b_A \in \Re} \quad \frac{\|Aw_A - eb_A\|^2}{\|Bw_A - eb_A\|^2} \tag{9}$$

PSVMs formulation

Adding Tikhonov regularization term to (9),

$$\min_{\mathbf{w_A}\in\Re^p, b_A\in\Re} \quad \frac{\|Aw_A - eb_A\|^2 + \nu\|[w_A' \quad b_A]\|^2}{\|Bw_A - eb_A\|^2} \tag{10}$$

ν is the regularization term.
Let,

$$G_A = [A \quad -e]'[A \quad -e] + \nu I, H_B = [B \quad -e]'[B \quad -e], z' = [w_A' \quad b_A] \tag{11}$$

Re-writing (10),

$$\min_{\mathbf{z}\in\Re^{p+1}} r(z) = \quad \frac{\mathbf{z}' G_A \mathbf{z}}{\mathbf{z}' H_B \mathbf{z}} \tag{12}$$

Rayleigh Quotient Problem

Rayleigh Quotient Properties

$$\min_{\mathbf{z}\in\Re^{p+1}} r(z) = \frac{\mathbf{z}'G\mathbf{z}}{\mathbf{z}'H\mathbf{z}} \tag{13}$$

- **Boundedness:**
  Assuming H is positive definite, $r(z)$ is bounded between $[\lambda_1, \lambda_{p+1}]$, where λ_1 and λ_{p+1} are the minimum and maximum eigenvalues of the following **generalized eigenvalue problem GEV(G,H)**:

$$G\mathbf{z} = \lambda H\mathbf{z} \tag{14}$$

- **Stationarity:**

$$\bigtriangledown r(z) = \frac{G\mathbf{z} - r(\mathbf{z})H\mathbf{z}}{\mathbf{z}'H\mathbf{z}} \tag{15}$$

  The stationary points are given by the eigenvectors of the generalized eigenvalue problem (14).

PSVMs Solution - Hyperplane P_A

$$\min_{\mathbf{z} \in \Re^{p+1}} r(z) = \frac{\mathbf{z}' G_A \mathbf{z}}{\mathbf{z}' H_B \mathbf{z}} \tag{16}$$

or,

$$\max_{\mathbf{z} \in \Re^{p+1}} r(z) = \frac{\mathbf{z}' H_B \mathbf{z}}{\mathbf{z}' G_A \mathbf{z}} \tag{17}$$

The solution is given by the eigenvector corresponding to the maximum eigenvalue of the following **generalized eigenvalue problem GEV(H_B, G_A):**

$$H_B \mathbf{z} = \lambda G_A \mathbf{z} \tag{18}$$

PSVMs Solution - Hyperplane P_B

Similarly, the hyperplane P_B (closest to class B and farthest from class A) given by:

$$P_B = \{x \in \Re^p \quad | \quad \langle \mathbf{w_B}, \mathbf{x} \rangle - b_B = 0\} \tag{19}$$

can be found by solving for the eigenvector corresponding to maximum eigenvalue of the following **generalized eigenvalue problem GEV(H_A, G_B)**:

$$H_A \mathbf{z} = \lambda G_B \mathbf{z} \tag{20}$$

$$G_B = [B \quad -e]'[B \quad -e] + \nu I, H_A = [A \quad -e]'[A \quad -e], z' = [w_B' \quad b_B] \tag{21}$$

Sparse Proximal Support Vector Machines (sPSVMs)

- sPSVMs are constructed by inducing **sparsity** in the hyperplanes obtained from PSVMs.
- Sparsity is defined as the optimal vectors $\mathbf{z}_{\mathbf{A}}^*$ and $\mathbf{z}_{\mathbf{B}}^*$ having only *few* non-zero components.
- The non-zero coefficients of optimal **sparse** vectors $\hat{\mathbf{z}}_{\mathbf{A}}^*$ and $\hat{\mathbf{z}}_{\mathbf{B}}^*$ may be interpreted as **class-specific** features.

Regularization in Linear Regression (LR)

- **Sparsity** via regularization has been well studied in the context of linear regression.
- Given a dataset S defined as:

$$S = \{(\mathbf{x_i}, y_i) \quad | \quad \mathbf{x_i} \in \Re^p, y_i \in \Re\}, \quad i = 1, 2, ...n \tag{22}$$

  the linear regression problem finds a coefficient vector $\mathbf{w}$ that *best* maps the input vector $\mathbf{x}$ to the output y.
- The following **least squares (LS)** problem is solved to obtain $\mathbf{w}$:

$$\min_{\mathbf{w}} \quad ||\mathbf{y} - X\mathbf{w}||^2 \tag{23}$$

$X \in \Re^{n \times p}, y \in \Re^n, w \in \Re^p$

Regularization in Linear Regression

- **Sparsity** is induced in linear regression problems via l_1*-norm*

$$\min_{\mathbf{w}} \quad ||\mathbf{y} - X\mathbf{w}||_2^2 + \lambda\|\mathbf{w}\|_1 \tag{24}$$

- Well known efficient algorithms like **Least Angle Regression (LARS)** exist in literature to solve (24)

B. Efron, T. Hastie, I. Johnstone & R. Tibshirani, *Least angle regression*. - The Annals of statistics (2004)

sPSVMs - Idea

Idea:

Transform **PSVMs** to an equivalent **least-squares (LS)** problem and induce **sparsity** via l_1-norm

Equivalence between Eigendecomposition and Linear Regression

Theorem 1: Consider a real matrix $X \in \Re^{n \times p}$ with rank $r \leq min(n, p)$. Let matrices $V \in \Re^{p \times p}$ and $D \in \Re^{p \times p}$ satisfy the following relation:

$$V^T(X^TX)V = D \tag{25}$$

where, $D = diag(\sigma_1^2, \sigma_2^2, \ldots \sigma_r^2, 0, 0, \ldots, 0)_{p \times p}$. Assume $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2$. For the following least-squares problem,

$$\begin{aligned} \min_{\alpha,\beta} \sum_{i=1}^{n} \quad & ||X_i - \alpha\beta^T X_i||^2 + \lambda\beta^T\beta \\ \text{subject to} \quad & \alpha^T\alpha = 1 \end{aligned} \tag{26}$$

$\beta_{opt} \propto V_1$, where X_i is the $ith - row$ of matrix X and V_1 is the eigenvector corresponding to the largest eigenvalue σ_1^2.

H. Zou, T. Hastie, & R. Tibshirani. *Sparse Principal Component Analysis.* - Journal of computational and graphical statistics (2006).

PSVMs via Least-Squares Approach

- Consider the **generalized eigenvalue problem** in PSVMs given by:

$$H_B\mathbf{z} = \lambda G_A\mathbf{z} \tag{27}$$

$$G_A = [A \quad -e]'[A \quad -e] + \nu I, H_B = [B \quad -e]'[B \quad -e], z' = [w'_A \quad b_A] \tag{28}$$

- Assuming G_A and H_B are **positive-definite**, the *cholesky decomposition* of the matrices give:

$$G_A = L_A L_A^T \quad = U_A^T U_A \tag{29}$$

$$H_B = L_B L_B^T \quad = U_B^T U_B \tag{30}$$

L_A, L_B are lower triangular matrices, and U_A, U_B are upper triangular matrices.

Relation between **generalized** eigenvalue problems and SVD

- Substituting (29) and (30) in (27),

$$H_B z = \lambda G_A z \tag{31}$$
$$L_B L_B^T z = \lambda U_A^T U_A z \tag{32}$$
$$U_A^{-T} L_B L_B^T z = \lambda U_A z \tag{33}$$
$$U_A^{-T} L_B L_B^T U_A^{-1} U_A z = \lambda U_A z \tag{34}$$
$$(L_B^T U_A^{-1})^T (L_B^T U_A^{-1}) U_A z = \lambda U_A z \tag{35}$$

Let, $\hat{X} = L_B^T U_A^{-1}$ and $v = U_A z$

$$(\hat{X}^T \hat{X}) v = \lambda v \tag{36}$$

PSVMs via Least-Squares Approach

- **PSVMs** can now be solved by an equivalent **least squares problem**.
- Using **Theorem 1** and substituting $X = L_B^T U_A^{-1}$, $\beta = U_A \hat{\beta}$ in (26),

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \sum_{i=1}^{n} \quad & ||(L_B^T U_A^{-1})_i - \alpha \hat{\beta}^T U_A^T (L_B^T U_A^{-1})_i||^2 + \lambda \hat{\beta}^T U_A^T U_A \hat{\beta} \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{37}$$

- Substituting $U_A^T U_A = G_A$ and $(L_B^T U_A^{-1})_i = U_A^{-T} U_{B,i}$,

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \sum_{i=1}^{n} \quad & ||U_A^{-T} U_{B,i} - \alpha \hat{\beta}^T U_{B,i}||^2 + \lambda \hat{\beta}^T G_A \hat{\beta} \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{38}$$

PSVMs via Least-Squares Approach

Re-writing in a nicer way,

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B\hat{\beta}\alpha^T||^2 + \lambda\hat{\beta}^T G_A\hat{\beta} \\ \text{s.t.} \quad & \alpha^T\alpha = 1 \end{aligned} \tag{39}$$

$\hat{\beta}_{opt}$ is proportional to z_A^* representing the hyperplane P_A in PSVMs.

PSVMs-via-LS

Solution Strategy

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B\hat{\beta}\alpha^T||^2 + \lambda\hat{\beta}^T G_A\hat{\beta} \\ \text{s.t.} \quad & \alpha^T\alpha = 1 \end{aligned} \tag{40}$$

Strategy:

The optimization problem is solved by alternating over α and $\hat{\beta}$.

Solving for α

- The **PSVMs-via-LS** is given by:

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & \|U_B U_A^{-1} - U_B \hat{\beta}\alpha^T\|^2 + \lambda \hat{\beta}^T G_A \hat{\beta} \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{41}$$

- For a fixed $\hat{\beta}$, the following optimization problem is solved to obtain α.

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & \|U_B U_A^{-1} - U_B \hat{\beta}\alpha^T\|^2 \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{42}$$

Solving for α

- Expanding the objective function,

$$\begin{aligned}(U_B U_A^{-1} - U_B\hat{\beta}\alpha^T)^T(U_B U_A^{-1} - U_B\hat{\beta}\alpha^T) \quad (43)\\ \approx -2\alpha^T U_A^{-T} H_B\hat{\beta} + \alpha^T\alpha\hat{\beta}H_B\hat{\beta} \quad (44)\end{aligned}$$

Subsituting $\alpha^T\alpha = 1$, the optimization problem in (42) reduces to:

$$\begin{aligned}\max_{\alpha} \quad & \alpha^T U_A^{-T} H_B\hat{\beta}\\ \text{s.t.} \quad & \alpha^T\alpha = 1\end{aligned} \quad (45)$$

- An analytical solution for this problem exists and the α_{opt} is given by,

$$\alpha_{opt} = \frac{U_A^{-T} H_B\hat{\beta}}{\|U_A^{-T} H_B\hat{\beta}\|} \quad (46)$$

Solving for $\hat{\beta}$

- The PSVMs-via-LS is given by:

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B \hat{\beta}\alpha^T||^2 + \lambda \hat{\beta}^T G_A \hat{\beta} \\ \text{s.t.} \quad & \alpha^T\alpha = 1 \end{aligned} \tag{47}$$

- Let $\hat{A}$ be an orthogonal matrix such that $[\alpha; \hat{A}]$ is $p \times p$ orthogonal. Then the objective function can be written as,

$$\begin{aligned} & ||U_B U_A^{-1} - U_B \hat{\beta}\alpha^T||^2 + \lambda \hat{\beta}^T G_A \hat{\beta} && (48) \\ \approx\; & tr(U_B U_A^{-1} - U_B \hat{\beta}\alpha^T)^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T) && (49) \\ \approx\; & tr([\alpha; \hat{A}][\alpha; \hat{A}]^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T)^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T) && (50) \\ \approx\; & tr([\alpha; \hat{A}]^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T)^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T)[\alpha; \hat{A}]) && (51) \\ \approx\; & tr((U_B U_A^{-1} - U_B \hat{\beta}\alpha^T[\alpha; \hat{A}])^T (U_B U_A^{-1} - U_B \hat{\beta}\alpha^T[\alpha; \hat{A}])) && (52) \end{aligned}$$

Solving for $\hat{\beta}$

$$
\begin{aligned}
&\approx tr([\alpha^T; \hat{A}^T]U_A^{-T} U_B^T U_B U_A^{-1}[\alpha; \hat{A}] - [\alpha^T; \hat{A}^T]U_A^{-T} U_B^T U_B \hat{\beta}\alpha^T [\alpha; \hat{A}] \\
&\qquad - [\alpha^T; \hat{A}^T]\alpha\hat{\beta}^T U_B^T U_B U_A^{-1}[\alpha; \hat{A}] + [\alpha^T; \hat{A}^T]\alpha\hat{\beta}^T U_B^T U_B \hat{\beta}\alpha^T [\alpha; \hat{A}]) \\
&\approx tr([\alpha^T; \hat{A}^T]U_A^{-T} U_B^T U_B U_A^{-1}[\alpha; \hat{A}] - [\alpha^T; \hat{A}^T]U_A^{-T} U_B^T U_B \hat{\beta} \\
&\qquad - \hat{\beta}^T U_B^T U_B U_A^{-1}[\alpha; \hat{A}] + \hat{\beta}^T U_B^T U_B \hat{\beta}) \\
&\approx tr((U_B U_A^{-1}[\alpha; \hat{A}])^T (U_B U_A^{-1}[\alpha; \hat{A}]) + (U_B\hat{\beta})^T (U_B\hat{\beta}) \\
&\qquad - 2(U_B\hat{\beta})^T U_B U_A^{-1}[\alpha; \hat{A}]) \\
&\approx tr((U_B U_A^{-1}[\alpha; \hat{A}] - U_B\hat{\beta})^T (U_B U_A^{-1}[\alpha; \hat{A}] - U_B\hat{\beta})) \\
&\approx ||U_B U_A^{-1}[\alpha; \hat{A}] - U_B\hat{\beta}||^2 \\
&\approx ||U_B U_A^{-1}\alpha - U_B\hat{\beta}||^2 + ||U_B U_A^{-1}\hat{A}||^2
\end{aligned}
\tag{53}
$$

Solving for $\hat{\beta}$

- For a fixed α, utilizing (53), the optimization problem in (47) reduces to **ridge-regression**:

$$\min_{\beta} \quad ||U_B U_A^{-1}\alpha - U_B\hat{\beta}||^2 + \lambda\hat{\beta}^T G_A\hat{\beta} \tag{54}$$

- An analytical solution exists and $\hat{\beta}_{opt}$ can be found by:

$$\hat{\beta}_{opt} = (H_B + \lambda G_A)^{-1} H_B U_A^{-1}\alpha \tag{55}$$

Algorithm

Algorithm 1 PSVMs-via-LS (H_B, G_A)

1. Initialize $\hat{\beta}$.
2. Find the upper triangular matrix U_A from the cholesky decomposition of G_A.
3. Find α from the following relation:

$$\alpha = \frac{U_A^{-T} H_B \hat{\beta}}{\| U_A^{-T} H_B \hat{\beta} \|} \tag{56}$$

4. Find $\hat{\beta}$ as follows:

$$\hat{\beta} = (H_B + \lambda G_A)^{-1} H_B U_A^{-1} \alpha \tag{57}$$

5. Alternate between 3 and 4 until convergence.

Sparse Proximal Support Vector Machines (sPSVMs)

- The PSVMs-via-LS is given by:

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B \hat{\beta} \alpha^T||^2 + \lambda \hat{\beta}^T G_A \hat{\beta} \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{58}$$

- **Sparsity** is introduced by adding l_1-**norm** in the above problem.

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B \hat{\beta} \alpha^T||^2 + \lambda \hat{\beta}^T G_A \hat{\beta} + \delta ||\hat{\beta}||_1 \\ \text{s.t.} \quad & \alpha^T \alpha = 1 \end{aligned} \tag{59}$$

Sparse Proximal Support Vector Machines (sPSVMs)

- The sPSVMs (59) is again solved by alternating over α and $\hat{\beta}$.

Solving for α

- The sPSVMs is given by:

$$\begin{aligned} \min_{\alpha,\hat{\beta}} \quad & ||U_B U_A^{-1} - U_B\hat{\beta}\alpha^T||^2 + \lambda\hat{\beta}^T G_A\hat{\beta} + \delta||\hat{\beta}||_1 \\ \text{s.t.} \quad & \alpha^T\alpha = 1 \end{aligned} \tag{60}$$

- For a fixed $\hat{\beta}$, an analytical solution exists for α and is given by,

$$\alpha_{opt} = \frac{U_A^{-T} H_B\hat{\beta}}{\|U_A^{-T} H_B\hat{\beta}\|} \tag{61}$$

Solving for $\hat{\beta}$

- For a fixed α, utilizing (52), **sPSVMs** in (59) can be written as:

$$\min_{\beta} \quad ||U_B U_A^{-1}\alpha - U_B\hat{\beta}||^2 + \lambda\hat{\beta}^T G_A\hat{\beta} + \delta||\hat{\beta}||_1 \tag{62}$$

- Expanding (62),

$$\begin{aligned}
&\min_{\hat{\beta}} \quad (U_B U_A^{-1}\alpha - U_B\hat{\beta})^T(U_B U_A^{-1}\alpha - U_B\hat{\beta}) + \lambda\hat{\beta}^T G_A\hat{\beta} + \delta||\hat{\beta}||_1 \\
&\min_{\hat{\beta}} \quad -\alpha^T U_A^{-T} H_B^T\hat{\beta} - \hat{\beta}^T H_B U_A^{-1}\alpha + \hat{\beta}^T H_B\hat{\beta} + \lambda\hat{\beta}^T G_A\hat{\beta} + \delta||\hat{\beta}||_1 \\
&\min_{\hat{\beta}} \quad \hat{\beta}^T(H_B + \lambda G_A)\hat{\beta} - \alpha^T U_A^{-T} H_B^T\hat{\beta} - \hat{\beta}^T H_B U_A^{-1}\alpha + +\delta||\hat{\beta}||_1 \\
&\min_{\hat{\beta}} \quad \hat{\beta}^T(H_B + \lambda G_A)\hat{\beta} - 2\alpha^T U_A^{-T} H_B\hat{\beta} + \delta||\hat{\beta}||_1
\end{aligned} \tag*{(63}$$

Solving for $\hat{\beta}$

- Assuming,
  $W^T = [U_B \quad \sqrt(\lambda) U_A], y^T = [U_B U_A^{-1}\alpha \quad 0],$

$$\min_{\hat{\beta}} \quad \hat{\beta}^T W^T W \hat{\beta} - 2y^T W \hat{\beta} + \delta ||\hat{\beta}||_1 \tag{64}$$

LASSO Regression

- Efficient algorithms like **Least Angle Regression (LARS)** exist to solve (64).

Algorithm

Algorithm 2 sPSVMs (H_B, G_A)

1. Initialize $\hat{\beta}$
2. Find the upper triangular matrix U_A and U_B from the cholesky decomposition of G_A and H_B.
3. Find α from the following equation:

$$\alpha = \frac{U_A^{-T} H_B \hat{\beta}}{\|U_A^{-T} H_B \hat{\beta}\|} \tag{65}$$

4. Construct W and y as follows:

$$W = [U_B \quad \sqrt(\lambda) U_A]^T, y = [U_B U_A^{-1} \alpha \quad 0]^T \tag{66}$$

and solve the following **LASSO** regression to obtain $\hat{\beta}$:

$$\min_{\hat{\beta}} \quad \hat{\beta}^T W^T W \hat{\beta} - 2y^T W \hat{\beta} + \delta ||\hat{\beta}||_1 \tag{67}$$

5. Alternate between 3 and 4 until convergence.

Results

- sPSVMs is compared with other classification methods SVMs, LDA and PSVMs on publicly available datasets.
- 10-fold cross validation is performed and the average accuracies are reported.
- For each fold, λ is chosen as zero and a grid search is performed over different values of ν and δ to choose the best values that yield the highest classification accuracy.
- Final model for testing is chosen as the one that yields the highest accuracy among the 10 folds.

Results - Example (Spambase dataset)

- The spambase dataset consists of 4601 samples and 57 features with 1813 samples in class 1 and 2788 samples in class 2.
- ν and δ are varied in logspace between $10^{-3} - 10^{4}$ and $10^{-5} - 1$ respectively.

Fold	Nu	Delta	Accuracy*	# $Features_A$	# $Features_B$
1	100	0.1	72.6%	14	4
2	10^{-3}	10^{-5}	69.8%	58	55
3	0.01	0.1	71.5%	17	7
4	100	10^{-5}	73.9%	55	37
5	**0.01**	**0.1**	**80.9%**	**14**	**4**
6	10^{-3}	0.1	76.3%	15	6
7	10^{-3}	0.1	75%	18	4
8	100	0.1	73.5%	15	4
9	100	0.01	73.7%	28	11
10	100	0.01	74.6%	28	11

Table: Classification accuracies of the 10-folds for Spambase dataset

Results

- All the classification methods have been implemented in MATLAB.
- LibSVM package is used for SVMs.
- LDA is solved using the 'classify' function in MATLAB.
- PSVMs are solved using the 'eig' function in MATLAB.
- LARS package provided by the authors on their website is used for sPSVMs.

Dataset	**Dimensions**	**SVMs**	**LDA**	**PSVMs**	**sPSVMs**	$\# \ features_A$	$\# \ features_B$
WDBC	569*30	96.8%	95.6%	95.4%	**97.5%**	6	11
Spambase	4601*57	77.1%	**90.7%**	71.6%	78.6%	14	4
Ionosphere	351*34	**91.2%**	88.3%	84.6%	85.5%	2	2
WPBC	198*33	**84.9%**	72.2%	77.7%	82.8%	7	9
Mushroom	8124*126	98.6%	99.2%	**100%**	**100%**	42	41
German	1000*20	**76.4%**	71.8%	68.7%	71.7%	1	2
Waveform	5000*21	**88.6%**	82.9%	78.5%	78%	8	14

Table: Classification accuracies for different classification methods on publicly available datasets.

Results - HDLSS datasets

- **sPSVMs** has been tested on three publicly available HDLSS datasets.
- The results are compared with other classification frameworks that combine dimensionality reduction techniques with a standard classification model.
- The chosen dimensionality reduction techniques are **Principal Component Analysis (PCA), Fisher-based Feature Selection (FFS)** and **Correlation-based Feature Selection (CFS)**.
- The number of principal components in **PCA** are chosen such that they account for 80% of the total variance in data.
- The standard classification methods tested are **SVMs, LDA** and **PSVMs**.
- Classification accuracies are obtained using *10-fold* cross validation.

Results - HDLSS datasets

Dataset	Dimensions	SVMs	PSVMs	sPSVMs	# $features_A$	# $features_B$
Colon	62*2000	75.9%	87.1%	**89%**	13	8
DBWorld	64*4702	88.1%	90.7%	**92.4%**	1	2
DLBCL	77*5469	**94.8%**	81.8%	81.8%	2	7

Table: Classification accuracies for publicly available HDLSS datasets using SVMs, PSVMs, and sPSVMs.

Colon dataset

	SVMs	PSVMs
FFS	92.4%	97%
CFS	83.9%	88.8%
PCA	90.7%	87.4%

DBWorld dataset

	SVMs	PSVMs
FFS	94.1%	97.1%
CFS	97.1%	97.1%
PCA	89.5%	82%

DLBCL dataset

	SVMs	PSVMs
FFS	96.3%	91.1%
CFS	98.8%	79.3%
PCA	96.3%	83.4%

Results - HDLSS datasets

- **sPSVMs** is compared with other embedded methods **Regularized Logistic Regression (RLR)** and **Sparse SVMs (S-SVMs)** on the HDLSS datasets.
- Classification accuracies are obtained using a *10-fold* cross validation.

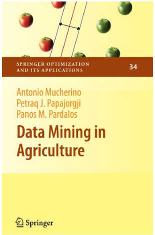
Dataset	**RLR**	**# features**	**S-SVMs**	**# features**	**sPSVMs**	# $features_A$	# $features_B$
Colon	83.9%	12	69.5%	16	**89%**	13	8
DBWorld	82.8%	9	82.6%	14	**92.4%**	1	2
DLBCL	**96.1%**	25	88.2%	12	81.8%	2	7

Table: Classification accuracies for different classification methods on publicly available HDLSS datasets.

Publications

- G. Pyrgiotakis, E. Kundakcioglu, K. Finton, K. Powers, B. M. Moudgil & P.M. Pardalos, *Cell death discrimination with Raman spectroscopy and support vector machines* - Annals of Biomedical Engineering (2009)
- G. Pyrgiotakis, E. Kundakcioglu, B. M. Moudgil & P.M. Pardalos, *Raman spectroscopy and Support Vector Machines for quick toxicological evaluation of Titania nanoparticles* - Journal of Raman Spectroscopy (2011)
- M. B. Fenn, P. Xanthopoulos, L. Hench, S.R. Grobmyer, G. Pyrgiotakis & P.M. Pardalos, *Raman spectroscopy for Clinical Oncology* - Advances in Optical Technologies (2011)
- M. Fenn, V. Pappu, P. Xanthopoulos & P.M. Pardalos, *Data Mining and Optimization Applied to Raman Spectroscopy for Oncology Applications* - BIOMAT (2011)
- P. Xanthopoulos, R. De Asmudis, M.R. Guarracino, G. Pyrgiotakis & P.M. Pardalos, *Supervised classification methods for mining cell differences as depicted by Raman spectroscopy* - Lecture Notes in Bioinformatics (2011)
- M.B. Fenn, & V. Pappu, *Data Mining for Cancer Biomarkers with Raman Spectroscopy* - Data Mining for Biomarker Discovery (2012)

Books

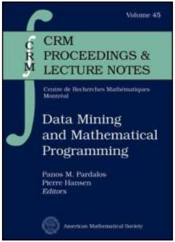


SPRINGER OPTIMIZATION AND ITS APPLICATIONS 34

Antonio Mucherino
Petraq J. Papajorgji
Panos M. Pardalos

Data Mining in Agriculture

Springer



Volume 45

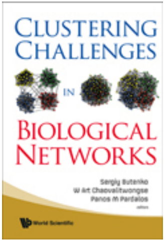
CRM PROCEEDINGS & LECTURE NOTES

Centre de Recherches Mathématiques Montréal

Data Mining and Mathematical Programming

Panos M. Pardalos
Pierre Hansen
Editors

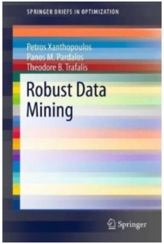
American Mathematical Society



CLUSTERING CHALLENGES IN BIOLOGICAL NETWORKS

Sergiy Butenko
W Art Chaovalitwongse
Panos M Pardalos
editors

World Scientific



SPRINGER BRIEFS IN OPTIMIZATION

Petros Xanthopoulos
Panos M. Pardalos
Theodore B. Trafalis

Robust Data Mining

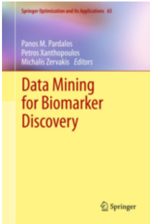
Springer

Springer Optimization and Its Applications

DATA MINING IN BIOMEDICINE

PANOS M. PARDALOS
VLADIMIR L. BOGINSKI
ALKIS VAZACOPOULOS

Springer



Springer Optimization and Its Applications 65

Panos M. Pardalos
Petros Xanthopoulos
Michalis Zervakis *Editors*

Data Mining for Biomarker Discovery

Springer