Constrained Subspace Classifier for High Dimensional Datasets

Panos M. Pardalos

Joint work with Orestis P. Panagopoulos & Petros Xanthopoulos

Work of Panos M. Pardalos was conducted at National Research University Higher School of Economics and supported by RSF Grant 14-41-00039

Information Gathered

- Sample/Data point
  - $\mathcal{X} \in \Re^D$
  - $D = \#$ of features
- Data Space
  - $\mathcal{S}$ with dimension equal to number of features

Binary Classification

- Classifying the samples of set $\mathcal{S}$ into two groups according to a classification rule

High Dimensional Datasets

- What are they?
  - High number of features
  - Relative small number of samples
- Examples of high dimensional datasets

Dataset	Reference
Customer Relationship Management data	(Tseng and Huang, 2007)
Covariation information of stocks	(Campbell and Lo, 1997)
Text datasets for classification	(Hassell and Arpinar, 2006)
Data collected from Surveys	(Belloni and Hansen, 2014)
Netflix dataset	(Bennett and Lanning, 2007)
MRI data	(Kampa et al., 2014)
Mass Spectroscopy data	(Fenn and Pappu, 2012)

Difficulties with High Dimensional Datasets

- *Curse of dimensionality*
  - Can cause model overfitting and estimation instability
  - Common classifiers fail

Clarke, R. et al. *The properties of high-dimensional data spaces: implications for exploring gene and protein expression data* - Nature Publishing Group (2008)

Difficulties with High Dimensional Datasets

- Volume increases exponentially as dimensionality increases
  - Points tend to become equidistant
  - Metric functions fail

Beyer, K. and Goldstein, J. and Ramakrishnan, R. and Shaft, U. *When is nearest neighbor meaningful?* - Springer Database Theory ICDT (1999)

Difficulties with High Dimensional Datasets

- Estimation of class covariance matrix unreliable
  - Most statistical classifiers require knowing class covariances *apriori*
  - Statistical classifiers fail
- *How do we deal with the aforementioned issues?*
  - Reducing the dimensionality of the dataset prior to classification
    - *Feature Selection*
    - *Feature Extraction*

Feature Selection

- Select only a subset of relevant features to use for classification
- Good for removing irrelevant data, increasing learning accuracy, and improving result comprehensibility

Yu, Lei, and Huan Liu *Feature selection for high-dimensional data: A fast correlation-based filter solution* - ICML (2003)

Feature Selection

- Categories of feature selection techniques:
  - **Filter methods**
  - **Wrapper methods**
  - **Embedded methods**

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

Feature Selection

- **Filter methods**
  - Access features during a separate process prior to classification
  - Variables are given a score according to a filtering function and are ordered accordingly
  - Features with the lowest scores are discarded while the rest are used from the classifier
- **Hypothesis testing, t-test**

Feature Selection

- **Wrapper methods**
  - Use the classifier structure itself to evaluate the importance of features
  - Based on the idea that the classifier can provide a better estimate of accuracy than a separate independent process
  - Increased computational power is often required - the classification process has to be repeated for each feature set considered
- **Metaheuristics**

Feature Selection

- **Embedded methods**
  - Perform feature selection in a way so that the classification algorithm is executed while variables are evaluated and selected
- **Recursive feature elimination in SVMs**
- **Random forests for feature evaluation**

Feature Extraction

- Feature extraction techniques transform the input data into a set of *meta*-features that extract the relevant information from the input data for classification
- *Principal Component Analysis (PCA)*

Rene Vidal,Yi Ma,S. Shankar Sastry *Generalized Principal Component Analysis* - ERL, UC (2006)

PCA

- Removes redundancy by transforming the data from a higher dimensional space into an **orthogonal lower dimensional space**
- First principal component captures as much variation in the data as possible - each succeeding component accounts for a decreasing amount of variance
- Number of retained principal components is less than or equal to the number of original variables
  - Criteria: **eigenvalue-one criterion, scree test, proportion of variance accounted for**

Local Subspace Classifier

- Local Subspace Classifier (LSC) utilizes PCA to perform classification
- Training phase
  - A lower dimensional subspace is found for each class that approximates the data
- Testing phase
  - A new data point is classified by calculating the distance of the point to each subspace and choosing the class with minimal distance

Laaksonen, Jorma *Local subspace classifier* - Artificial Neural Networks ICANN (1997)

Local Subspace Classifier

- Consider a binary classification problem
  - Let the matrices $\mathcal{X}_1 \in \Re^{p \times m}$ and $\mathcal{X}_2 \in \Re^{p \times l}$ be given, whose columns represent the training examples of two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively
- LSC attempts to find two subspaces separately, one for each class that *best* approximates the data

Laaksonen, Jorma *Local subspace classifier* - Artificial Neural Networks ICANN (1997)

Local Subspace Classifier

- Let

$$\boldsymbol{U}_1 = [\boldsymbol{u}_1^{(1)}, \boldsymbol{u}_2^{(1)}, \ldots, \boldsymbol{u}_k^{(1)}]_{p\times k} \tag{1}$$

and

$$\boldsymbol{U}_2 = [\boldsymbol{u}_1^{(2)}, \boldsymbol{u}_2^{(2)}, \ldots, \boldsymbol{u}_k^{(2)}]_{p\times k} \tag{2}$$

represent orthonormal bases of two k-dimensional linear subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ that approximate classes $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively

  - We assume the dimensionality of subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ to be the same and equal to k

Training Phase

- $\mathcal{S}_1$ and $\mathcal{S}_2$ attempt to capture *maximal* variance in classes $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively by solving the following optimization problems:
  - $$\begin{aligned} \underset{\boldsymbol{U}_1 \in \Re^{p \times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_1^T \mathcal{X}_1 \mathcal{X}_1^T \boldsymbol{U}_1) \\ \text{subject to} \quad & \boldsymbol{U}_1^T \boldsymbol{U}_1 = \boldsymbol{I}_k \end{aligned} \tag{3}$$

    The solution to (3) is given by the eigenvectors corresponding to k largest eigenvalues of matrix $\mathcal{X}_1 \mathcal{X}_1^T$
  - $$\begin{aligned} \underset{\boldsymbol{U}_2 \in \Re^{p \times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_2^T \mathcal{X}_2 \mathcal{X}_2^T \boldsymbol{U}_2) \\ \text{subject to} \quad & \boldsymbol{U}_2^T \boldsymbol{U}_2 = \boldsymbol{I}_k \end{aligned} \tag{4}$$

    Similarly the orthonormal basis $\boldsymbol{U}_2$ is obtained by choosing eigenvectors corresponding to k largest eigenvalues of matrix $\mathcal{X}_2 \mathcal{X}_2^T$

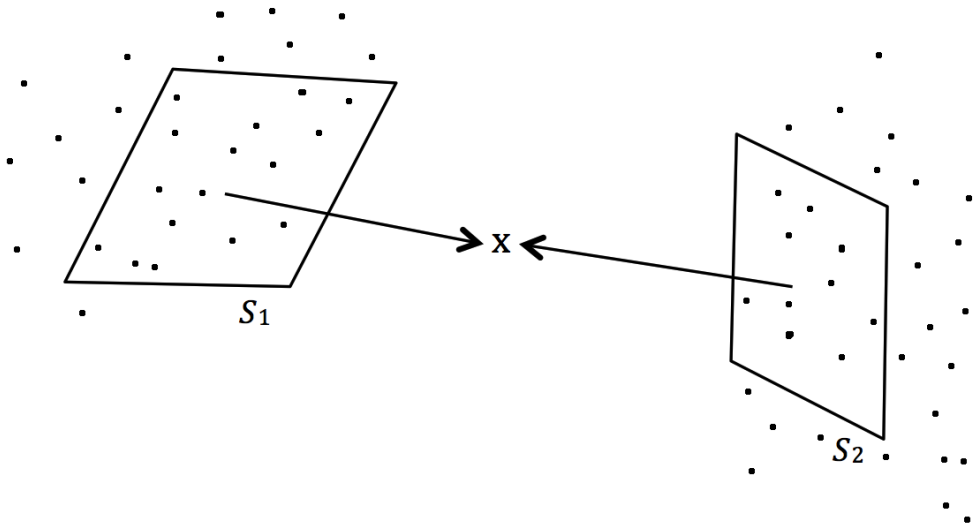
Testing Phase

- A new point $\boldsymbol{x}$ is classified by computing the distance from subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$:

$$\text{dist}(\boldsymbol{x}, \mathcal{S}_i) = \text{tr}(\boldsymbol{U}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{U}_i) \tag{5}$$

and the class of $\boldsymbol{x}$ is determined as:

$$\text{class}(\boldsymbol{x}) = argmin_{i \in \{1,2\}}\{\text{dist}(\boldsymbol{x}, \mathcal{S}_i)\} \tag{6}$$



Motivation

- Though the subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ approximate the classes well, these projections may not be *ideal* for classification tasks as each of them are obtained *without* the knowledge of another class/subspace

In order to account for the presence of another subspace, we consider the *relative orientation* of the subspaces

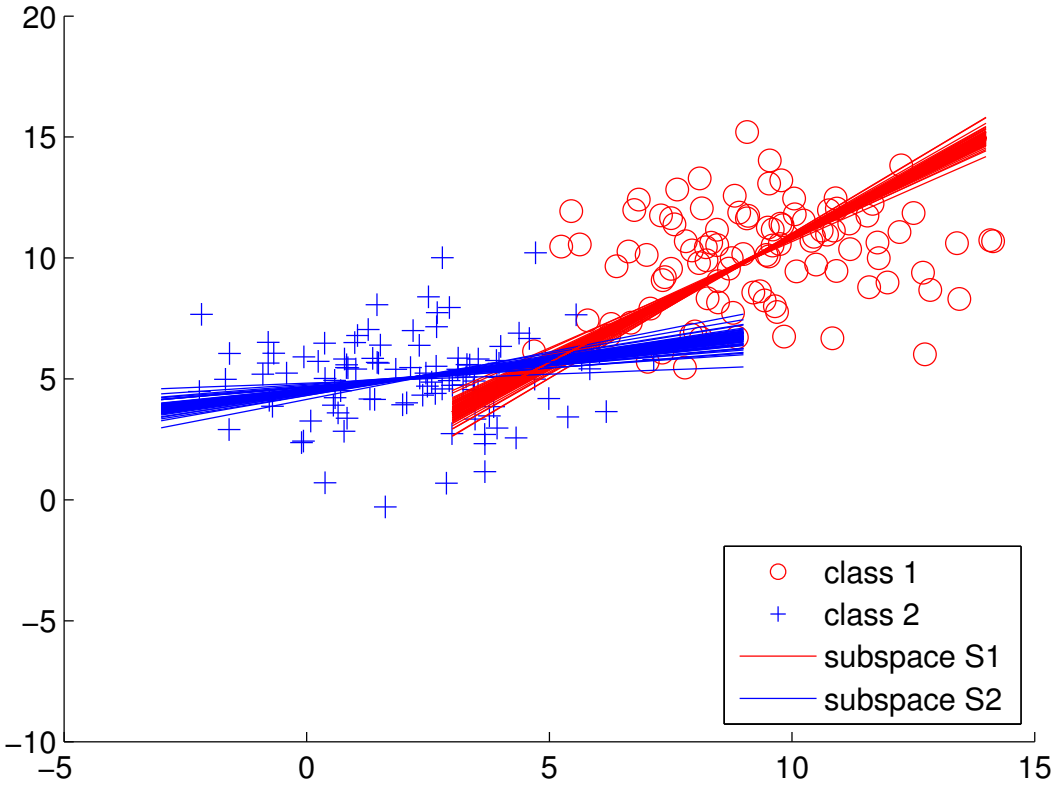
Motivating Example

- Datasets are generated from two bivariate normal distributions $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ representing classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Each class consists of 100 randomly generated points from $\mathcal{N}_1$ and $\mathcal{N}_2$ respectively

Datasets	$\mathcal{N}_1$		$\mathcal{N}_2$		LSC		CSC	
	$\boldsymbol{\mu}_1$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}_2$	Acc(%)	Angle(θ)	Acc(%)	Angle(θ)
Example 1	$\begin{bmatrix} 9 \\ 10 \end{bmatrix}$	$\begin{bmatrix} 4 & 1.1 \\ 1.1 & 4 \end{bmatrix}$	$\begin{bmatrix} 2 \\ 5 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$	74	0.54	**92**	0.99
Example 2	$\begin{bmatrix} 3 \\ 5 \end{bmatrix}$	$\begin{bmatrix} 4 & -2 \\ -2 & 6 \end{bmatrix}$	$\begin{bmatrix} 10 \\ 10 \end{bmatrix}$	$\begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}$	87	0.92	**97**	0.16

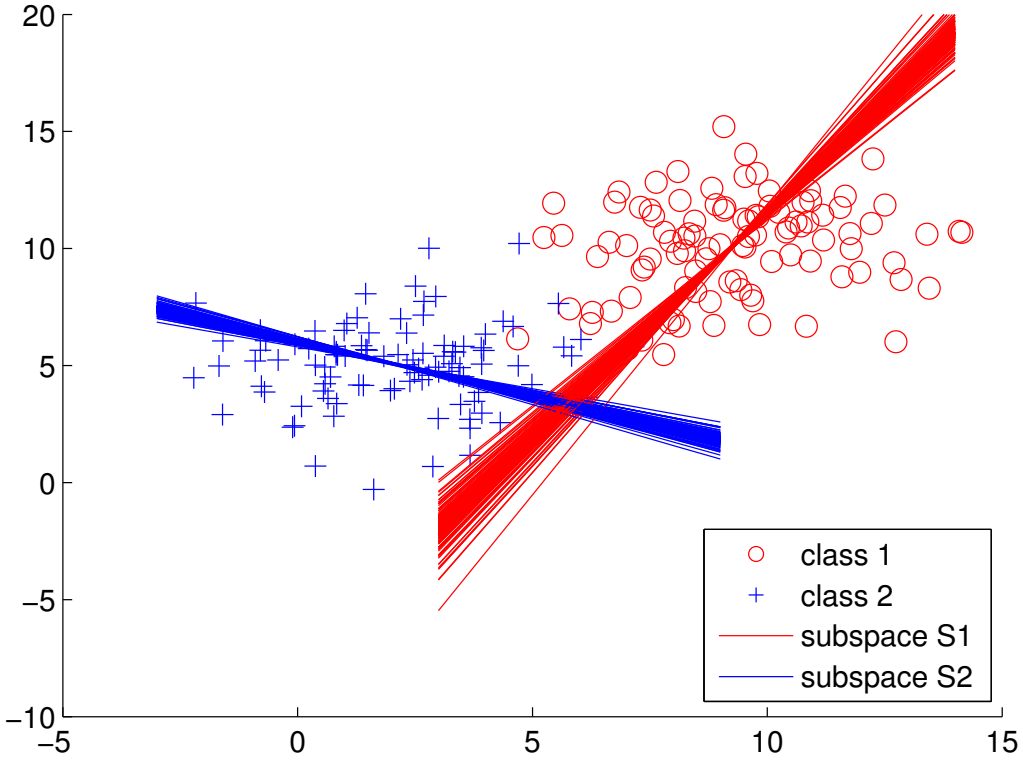
Motivating Example

- LSC and CSC are trained on the data with $k = 1$ and the classification accuracies are obtained via 10-fold cross validation
  - Example 1:
    - LSC



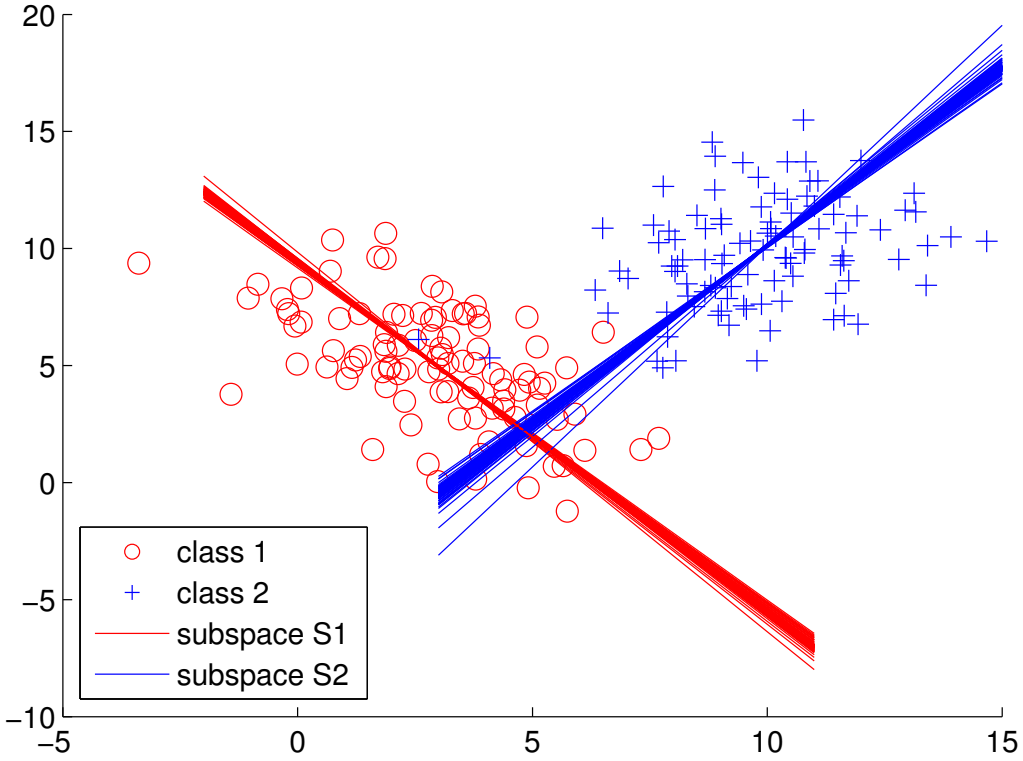
Motivating Example

- Example 1:
  - CSC

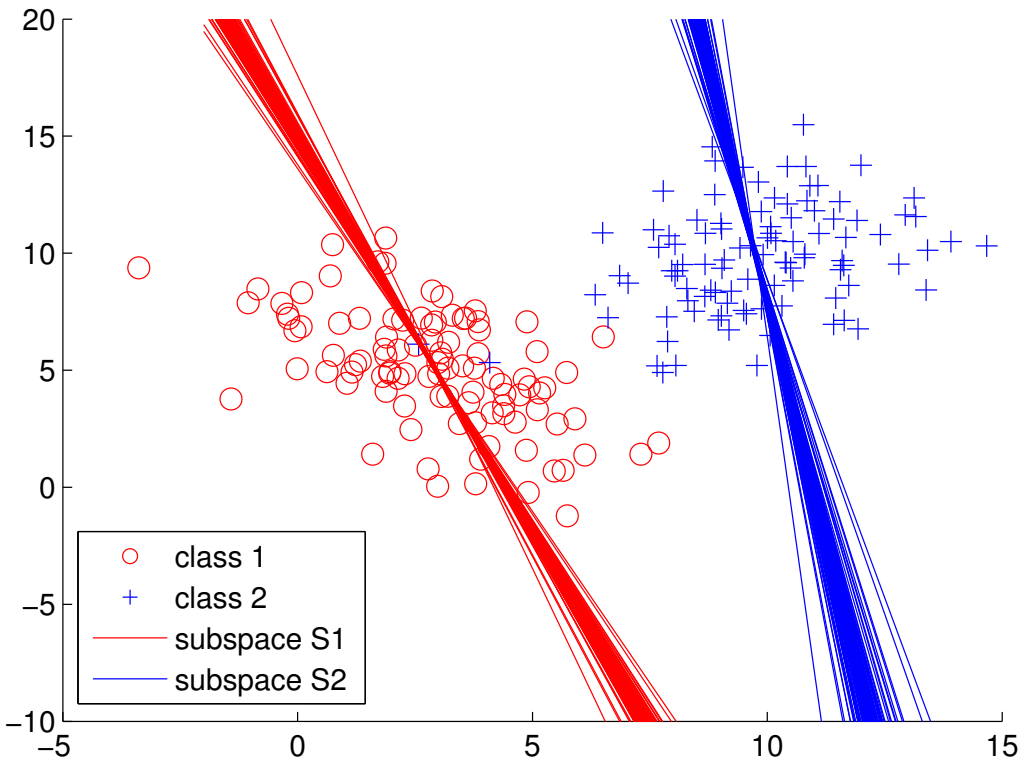


Motivating Example

- Example 2:
  - LSC



- Example 2:
  - CSC



- These examples show that *relative orientation* of the subspaces should also be considered in addition to capturing *maximal* variance in data

Constrained Subspace Classifier (CSC)

- Constrained subspace classifier (CSC) finds two subspaces *simultaneously*, one for each class, such that each subspace accounts for maximal variance in the data in the *presence* of the other class/subspace

Relative orientation in terms of principal angles

Definition 1: *Let* $\boldsymbol{U}_1 \in \Re^{p\times k}$ *and* $\boldsymbol{U}_2 \in \Re^{p\times k}$ *be two orthonormal matrices spanning subspaces* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*. The principal angles* $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq \cdots \leq \theta_k \leq \pi/2$ *between subspaces* $\mathcal{S}_1$ *and* $\mathcal{S}_2$*, are defined recursively by:*

$$\begin{aligned} \cos\theta_i = \max_{\boldsymbol{x}_m \in \mathcal{S}_1} \max_{\boldsymbol{y}_n \in \mathcal{S}_2} \quad & \boldsymbol{x}_m^{\mathsf{T}} \boldsymbol{y}_n \\ \text{subject to} \quad & \boldsymbol{x}_m^{\mathsf{T}} \boldsymbol{x}_n = 1, \quad \boldsymbol{y}_m^{\mathsf{T}} \boldsymbol{y}_n = 1, \quad \textit{for } m = n \\ & \boldsymbol{x}_m^{\mathsf{T}} \boldsymbol{x}_n = 0, \quad \boldsymbol{y}_m^{\mathsf{T}} \boldsymbol{y}_n = 0, \quad \textit{for } m \neq n \\ & \forall m, n = 1, 2, \ldots, k \end{aligned} \tag{7}$$

Hamm, Jihun and Lee, Daniel D *Grassmann discriminant analysis: a unifying view on subspace-based learning* -ACM (2008)

Finding the canonical correlations

Theorem 1: Let $U_1 \in \Re^{p \times k}$ and $U_2 \in \Re^{p \times k}$ be rectangular matrices whose column vectors span the subspaces $S_1 \in \Re^k$ and $S_2 \in \Re^k$ respectively. Let $M = U_1^\top U_2 \in \Re^{k \times k}$, using SVD we can express M by:

$$M = YCZ^\top \tag{8}$$

where $Y^\top Y = I_k,\ Z^\top Z = I_k$ and $C = diag(\sigma_1, \sigma_2, \ldots, \sigma_k)$

- If we assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ then the principal angles are given by $\cos\theta_k \ = \sigma_k(M) \quad \forall i = 1, 2, \ldots, k$

Bjorck, Ake, and Gene H. Golub *Numerical methods for computing angles between linear subspaces* -Mathematics of computation (1973)

Seeking a Distance Metric

We consider the metric that defines the relative orientation between the two subspaces S_1 and S_2 spanned by U_1 and U_2 respectively to be the *projection F-norm* defined by:

$$d_{pF}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \frac{1}{\sqrt{2}} \| \boldsymbol{U}_1 \boldsymbol{U}_1^\top - \boldsymbol{U}_2 \boldsymbol{U}_2^\top \|_F \tag{9}$$

Edelman, Alan, Toms A. Arias, and Steven T. Smith *The geometry of algorithms with orthogonality constraints* -SIAM journal on Matrix Analysis and Applications (1998)

In terms of principal angles

$$
\begin{aligned}
&\|\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top\|_F^2 \\
&= \mathrm{tr}((\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top)^\top(\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top)) \\
&= \mathrm{tr}(\boldsymbol{U}_1\boldsymbol{U}_1^\top\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_1\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{U}_2^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top\boldsymbol{U}_1\boldsymbol{U}_1^\top + \boldsymbol{U}_2\boldsymbol{U}_2^\top\boldsymbol{U}_2\boldsymbol{U}_2^\top) \\
&= \mathrm{tr}(\boldsymbol{U}_1\boldsymbol{U}_1^\top) + \mathrm{tr}(\boldsymbol{U}_2\boldsymbol{U}_2^\top) - 2\mathrm{tr}(\boldsymbol{U}_1\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{U}_2^\top) \\
&= \mathrm{tr}(\boldsymbol{U}_1^\top\boldsymbol{U}_1) + \mathrm{tr}(\boldsymbol{U}_2^\top\boldsymbol{U}_2) - 2\mathrm{tr}(\boldsymbol{U}_2^\top\boldsymbol{U}_1\boldsymbol{U}_1^\top\boldsymbol{U}_2) \\
&= \|\boldsymbol{U}_1\|_F^2 + \|\boldsymbol{U}_2\|_F^2 - 2\|\boldsymbol{U}_2^\top\boldsymbol{U}_1\|_F^2
\end{aligned}
\tag{10}
$$

According to **Theorem 1**:

$$
\|\boldsymbol{U}_2^\top\boldsymbol{U}_1\|_F^2 = \sum_{i=1}^{k}\sigma_i^2 = \sum_{i=1}^{k}\cos^2\theta_i \tag{11}
$$

In terms of principal angles

Using (11) on (10) becomes:

$$
\begin{aligned}
&= \sum_{i=1}^{k} \lambda_i + \sum_{i=1}^{k} \lambda_i - 2\sum_{i=1}^{k} \cos^2\theta_i \\
&= k + k - 2\sum_{i=1}^{k} \cos^2\theta_i \\
&= 2\left[k - \sum_{i=1}^{k} \cos^2\theta_i\right] \qquad (12) \\
&= 2\left[(1-\cos^2\theta_1) + (1-\cos^2\theta_2) + \cdots + (1-\cos^2\theta_k)\right] \\
&= 2\sum_{i=1}^{k} \sin^2\theta_i
\end{aligned}
$$

In terms of principal angles

- Hence the projection F-norm becomes:

$$d_{pF}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \frac{1}{\sqrt{2}} \| \boldsymbol{U}_1 \boldsymbol{U}_1^\top - \boldsymbol{U}_2 \boldsymbol{U}_2^\top \|_F = \sqrt{\sum_{i=1}^{k} \sin^2 \theta_i} \tag{13}$$

Formulating CSC

- The projection metric is utilized to incorporate the relative orientation between subspaces in LSC
- The formulation of LSC is modified as shown below to obtain the *Constrained Subspace Classifier (CSC)*:

$$\begin{aligned} \underset{\boldsymbol{U}_1,\boldsymbol{U}_2\in\Re^{p\times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_1^T\mathcal{X}_1\mathcal{X}_1^T\boldsymbol{U}_1)+\text{tr}(\boldsymbol{U}_2^T\mathcal{X}_2\mathcal{X}_2^T\boldsymbol{U}_2)-C\|\boldsymbol{U}_1\boldsymbol{U}_1^\top-\boldsymbol{U}_2\boldsymbol{U}_2^\top\|_F^2 \\ \text{subject to} \quad & \boldsymbol{U}_1^T\boldsymbol{U}_1=\boldsymbol{I}_k, \quad \boldsymbol{U}_2^T\boldsymbol{U}_2=\boldsymbol{I}_k \end{aligned} \tag{14}$$

where the parameter C controls the tradeoff between the relative orientation of the subspaces and the approximation of the data

Formulating CSC

- Using (11) and (12) :
  - $\|\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top\|_F^2 = 2k - 2\text{tr}(\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{U}_2^\top\boldsymbol{U}_1)$
- Hence the optimization problem becomes:

$$\begin{aligned} \underset{\boldsymbol{U}_1,\boldsymbol{U}_2\in\Re^{p\times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_1^T\mathcal{X}_1\mathcal{X}_1^T\boldsymbol{U}_1) + \text{tr}(\boldsymbol{U}_2^T\mathcal{X}_2\mathcal{X}_2^T\boldsymbol{U}_2) + C\text{tr}(\boldsymbol{U}_1^T\boldsymbol{U}_2\boldsymbol{U}_2^T\boldsymbol{U}_1) \\ \text{subject to} \quad & \boldsymbol{U}_1^T\boldsymbol{U}_1 = \boldsymbol{I}_k, \quad \boldsymbol{U}_2^T\boldsymbol{U}_2 = \boldsymbol{I}_k \end{aligned} \tag{15}$$

Algorithm

- We introduce an alternating optimization algorithm to solve (15)
- For a fixed $\boldsymbol{U}_2$, (15) reduces to:

$$\begin{aligned} \underset{\boldsymbol{U}_1 \in \Re^{p \times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_1^T(\mathcal{X}_1\mathcal{X}_1^T + C\boldsymbol{U}_2\boldsymbol{U}_2^T)\boldsymbol{U}_1) \\ \text{subject to} \quad & \boldsymbol{U}_1^T\boldsymbol{U}_1 = \boldsymbol{I}_k \end{aligned} \tag{16}$$

The solution to (16) is obtained by choosing eigenvectors corresponding to k largest eigenvalues of symmetric matrix $\mathcal{X}_1\mathcal{X}_1^T + C\boldsymbol{U}_2\boldsymbol{U}_2^T$

Algorithm

- Similarly, for a fixed $\boldsymbol{U}_1$, (15) reduces to:

$$\begin{aligned} \underset{\boldsymbol{U}_2 \in \Re^{p \times k}}{\text{maximize}} \quad & \text{tr}(\boldsymbol{U}_2^T(\mathcal{X}_2\mathcal{X}_2^T + C\boldsymbol{U}_1\boldsymbol{U}_1^T)\boldsymbol{U}_2) \\ \text{subject to} \quad & \boldsymbol{U}_2^T\boldsymbol{U}_2 = \boldsymbol{I}_k \end{aligned} \tag{17}$$

where the solution to (17) is again obtained by choosing eigenvectors corresponding to k largest eigenvalues of symmetric matrix $\mathcal{X}_2\mathcal{X}_2^T + C\boldsymbol{U}_1\boldsymbol{U}_1^T$

Termination Rules

We define the following three termination rules:

- Maximum limit Z on the number of iterations,
- Relative change in $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ at iteration m and m+1,

$$\text{tol}_{\boldsymbol{U}_1}^m = \frac{\|\boldsymbol{U}_1^{(m+1)} - \boldsymbol{U}_1^{(m)}\|_F}{\sqrt{q}}, \quad \text{tol}_{\boldsymbol{U}_2}^m = \frac{\|\boldsymbol{U}_2^{(m+1)} - \boldsymbol{U}_2^{(m)}\|_F}{\sqrt{q}} \tag{18}$$

  where $q = pk$

- Relative change in objective function value of (14) at iteration m and m+1,

$$\text{tol}_f^m = \frac{F^{(m+1)} - F^{(m)}}{|F^{(m)}| + 1} \tag{19}$$

CSC Algorithm

The algorithm for CSC can be summarized as follows:

Algorithm 1 CSC $(\mathcal{X}_1, \mathcal{X}_2,\ k,\ C)$

1. Initialize $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ such that $\boldsymbol{U}_1^T\boldsymbol{U}_1 = \boldsymbol{I}_k$, $\boldsymbol{U}_2^T\boldsymbol{U}_2 = \boldsymbol{I}_k$
2. Find eigenvectors corresponding to the k largest eigenvalues of symmetric matrix $\mathcal{X}_1\mathcal{X}_1^T + C\boldsymbol{U}_2\boldsymbol{U}_2^T$
3. Find eigenvectors corresponding to the k largest eigenvalues of symmetric matrix $\mathcal{X}_2\mathcal{X}_2^T + C\boldsymbol{U}_1\boldsymbol{U}_1^T$
4. Alternate between 2 and 3 until one of the termination rules is satisfied

Algorithm 1 converges. For proof of convergence see:

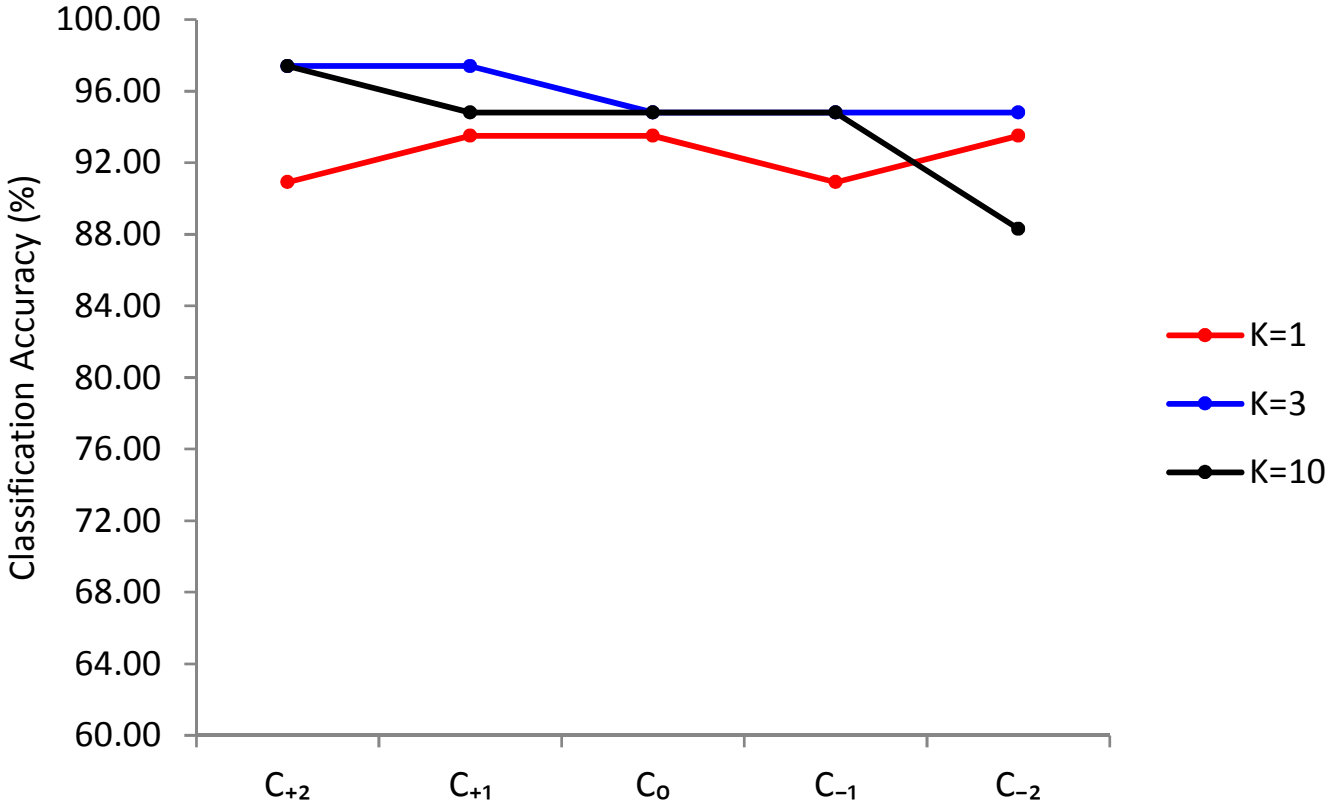
Panagopoulos, O. P., Pappu, V., Xanthopoulos, P., Pardalos, P. M. *Constrained subspace classifier for high dimensional datasets* - Omega (Volume 59, Part A, March 2016, Pages 40–46)
http://www.sciencedirect.com/science/article/pii/S0305048315001188

Numerical Experiments

- The performance of CSC is evaluated on four high dimensional publicly available datasets
- CSC is also tested on two lower dimensional datasets
- The performance of CSC is evaluated for different values of C, and compared to that of LSC
- The values of C are chosen in such a way that they vary *uniformly*
- The classification performance is evaluated using *leave-one-out cross validation (LOOCV)* technique
- *The value of k is chosen as $\{1, 3, 10\}$*
- Experiments are performed with a 2.60GHz Intel Core i5 CPU running OS X with 8.0 GB of main memory

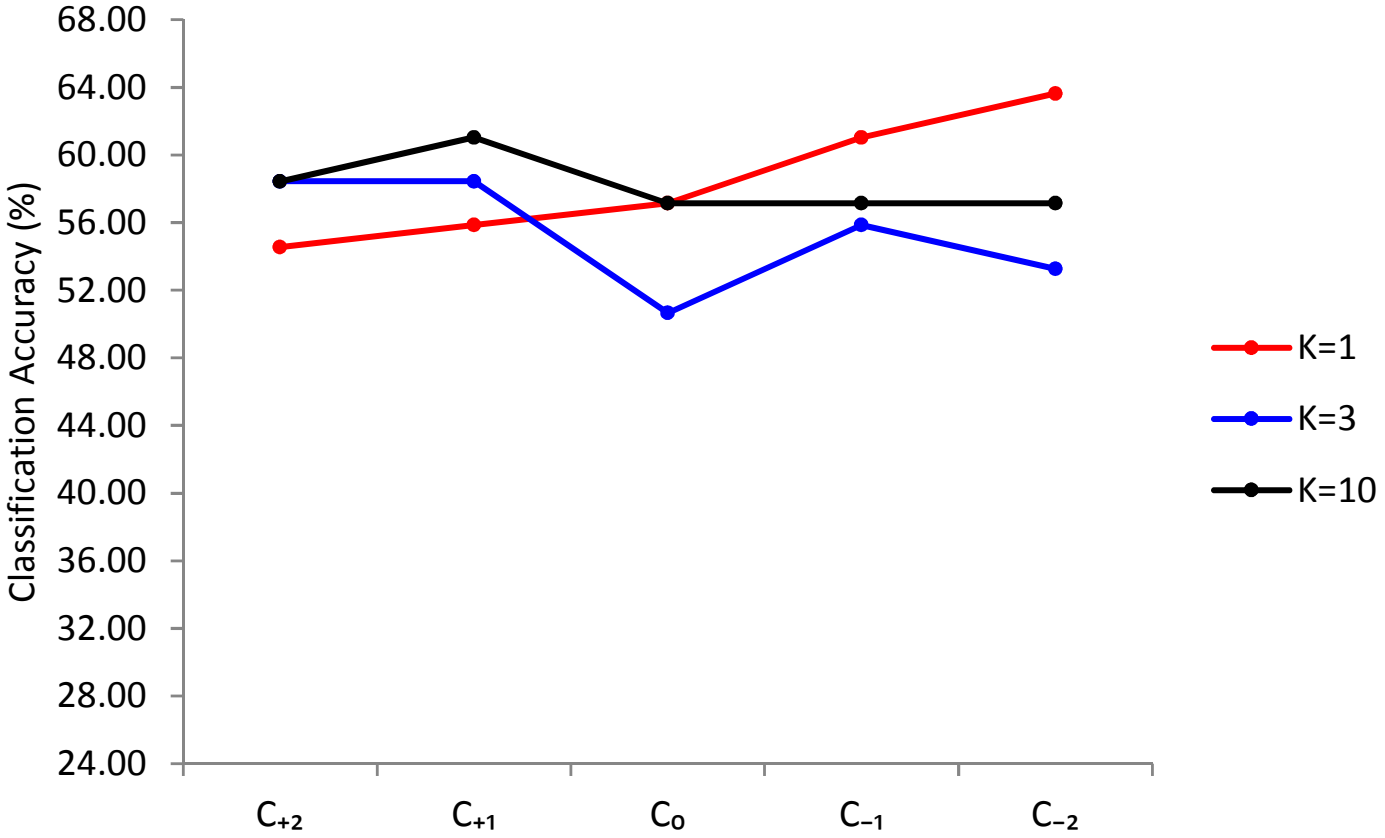
DLBCL

Diffuse large B-cell lymphoma DLBCL, the most common lymphoid malignancy in adults, is curable in less than 50% of patients. The DLBCL dataset consists of 77 samples with 5469 features. CSC was used to identify cured versus fatal or refractory disease



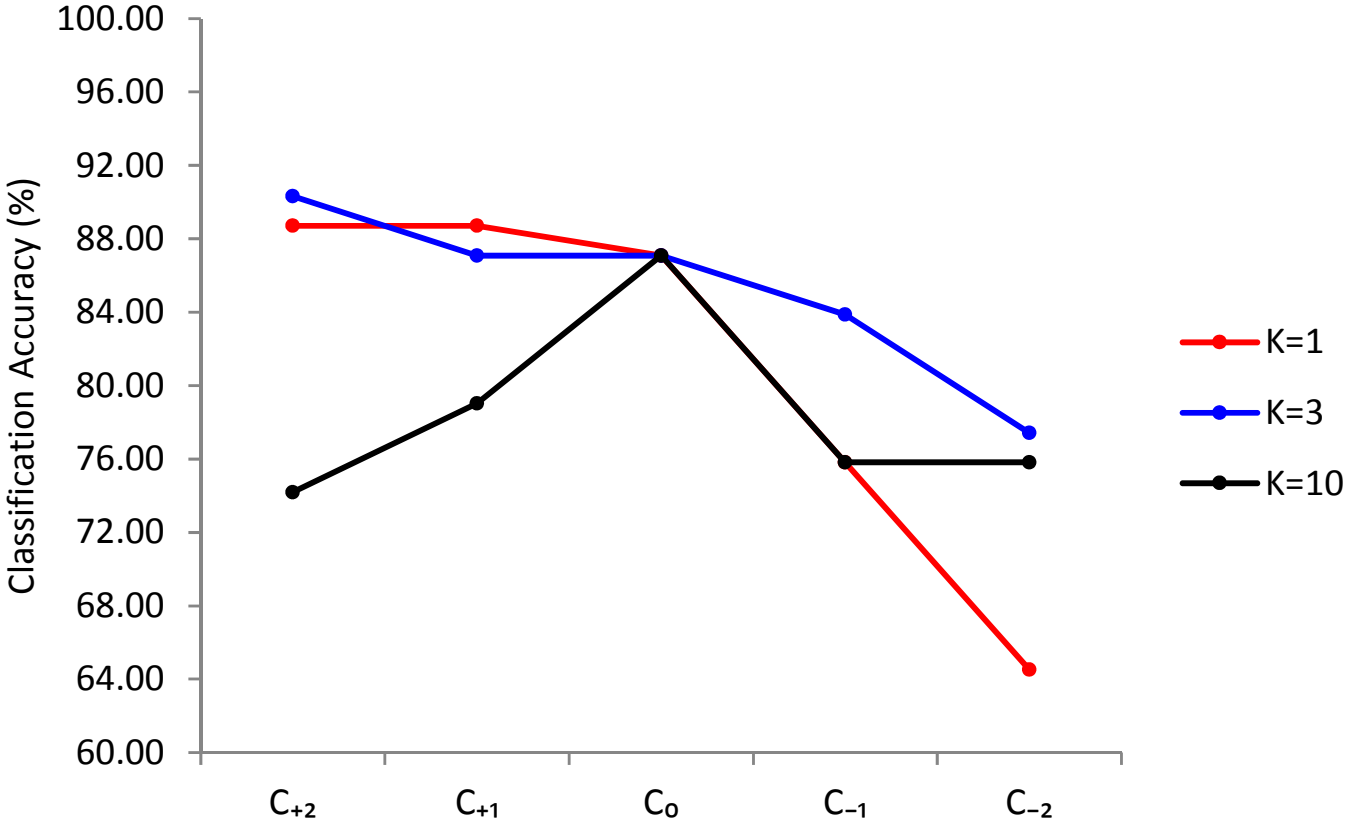
Breast Cancer

Breast Cancer dataset consists of 77 samples of breast tumors. 4869 features describe each one of those tumors. CSC classified the tumors as recurring or non-recurring



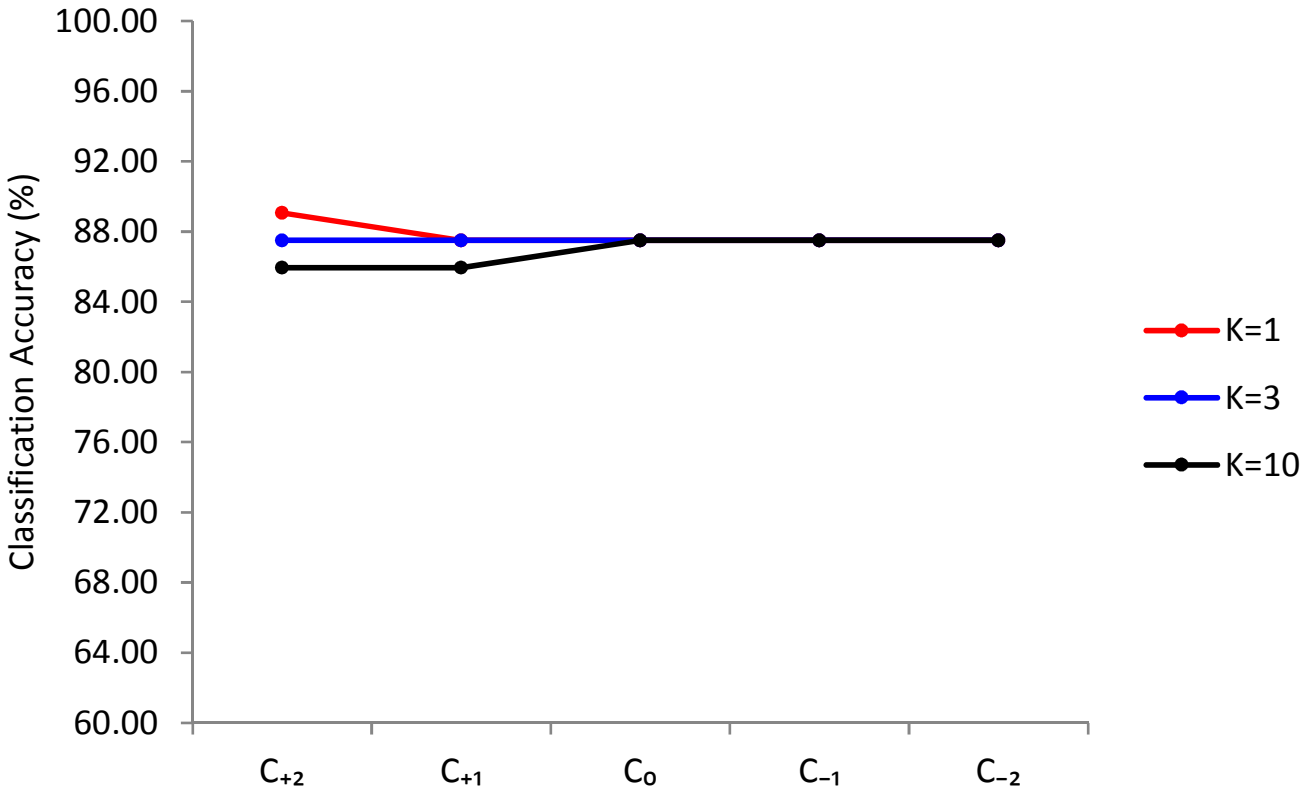
Colon

40 tumor and 22 normal colon tissue samples make up Colon dataset. 2000 features describe each one of those samples. CSC classified the samples as tumorous or not



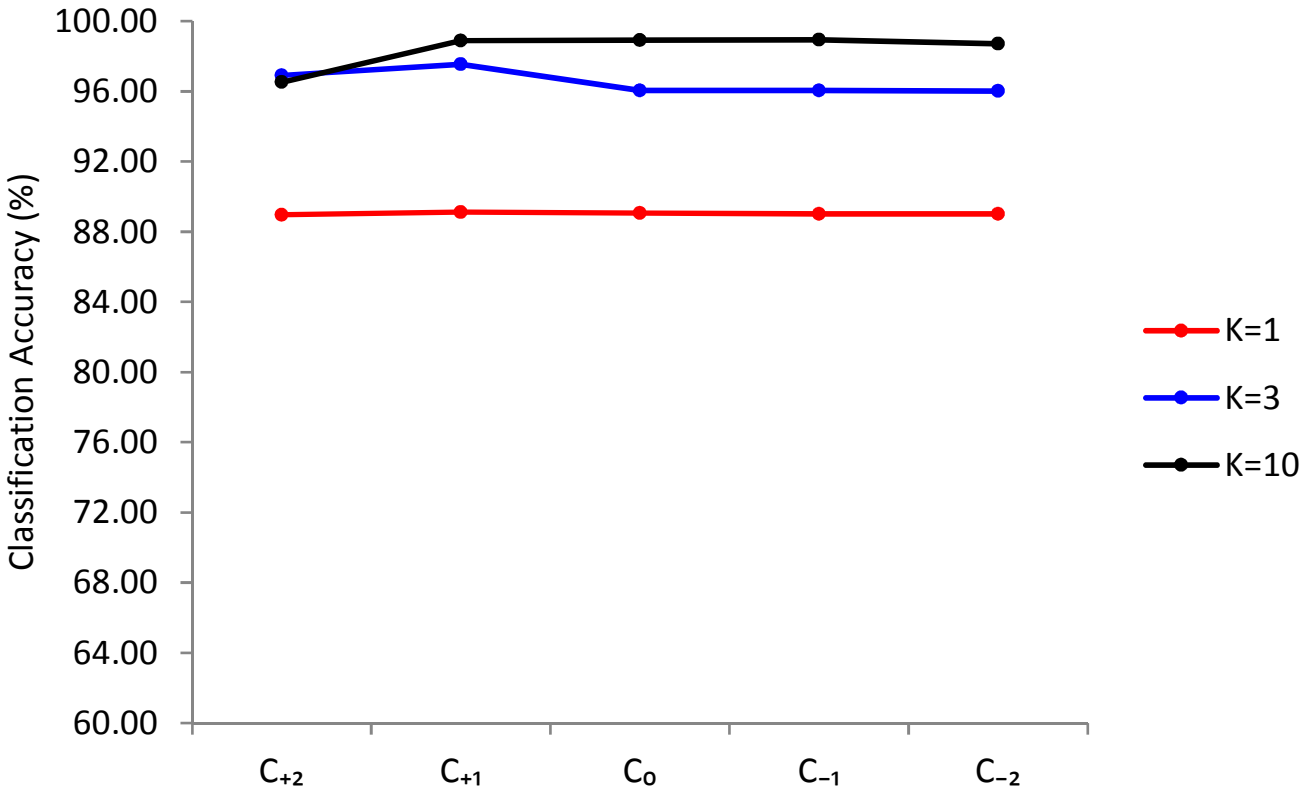
DBWorld

DBWorld dataset consists of 64 e-mails (samples) divided in two classes. The first one consists of only subject lines, while the second consists of only bodies. 4702 features describe each one of those samples. CSC classified the samples as subjects or bodies



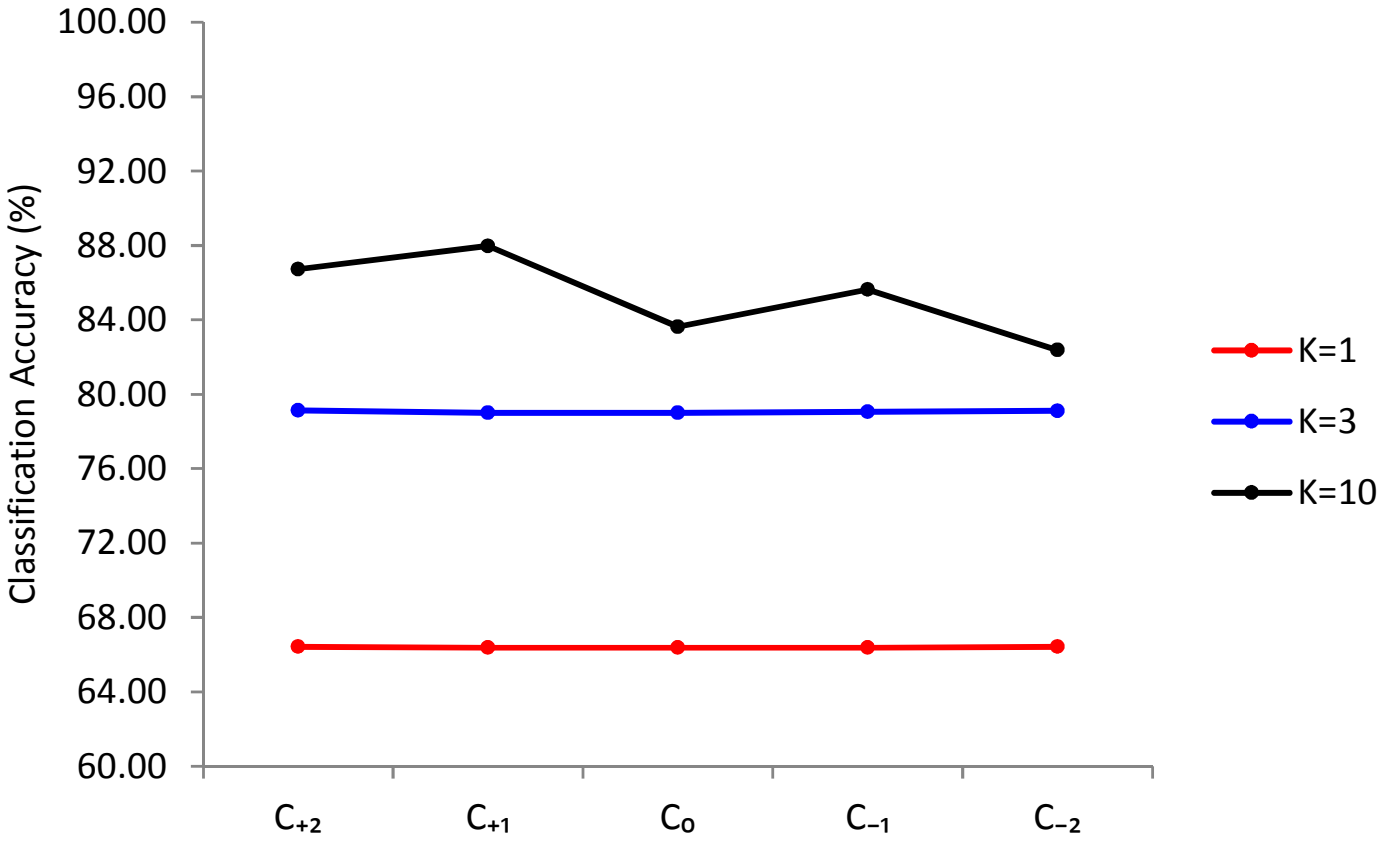
Mushroom

Mushroom dataset describes characteristics of gilled mushrooms. It consists of 8124 samples with 126 features. CSC classified the samples onto two categories, edible and non-edible



Spambase

Spambase dataset consists of 4601 samples(emails) with 57 features. CSC was used to find weather an email is spam or not



Results

- For DLBCL and Colon datasets, classification *accuracy is improved* by reducing the relative angle between subspaces for $k = 3$, $k = 10$ and $k = 1$, $k = 3$ respectively
- In the case of Breast dataset, increasing the relative angle for $k = 1$ *considerably improves the classification accuracy*
- The classification accuracy of CSC was almost identical to that of LSC for the DBWorld dataset
- With respect to the lower dimensional datasets, CSC performed *at least as good as* LSC
  - In the case of Spambase dataset, CSC was able to slightly increase the accuracy of classification for positive values of C

Comparative computational results

Dataset	SVM	PCA/SVM	Naive Bayes	CSC
DLBCL	94.8	97.5	75	97.4
Breast	68	68	62.5	63.6
Colon	75.9	92.1	71.4	90.3
DBWorld	88	88	57.1	89
Mushroom	100	100	88.1	98.9
Spambase	91	66	56.3	87.9

- CSC demonstrates competitive behavior with respect to dataset dimensionality
- CSC remains robust

Conclusion

- A new classification algorithm, CSC, was proposed and designed for high dimensional datasets
- CSC *improves* upon local subspace classifier
- The *improvement in classification accuracy* shows the importance of considering the relative angle between subspaces while approximating the classes
- The robust nature of CSC reveals that it can serve as a *one-step method* for preprocessing-free classification

Future Research

- A *cost sensitive* version for *imbalanced classification* problems
- A stream mining version that will incrementally retrain as new training data samples arrive in the form of a data stream
- A robust optimization version for handling datasets that are inexact or uncertain

Panagopoulos, O. P., Pappu, V., Xanthopoulos, P., Pardalos, P. M.
Constrained subspace classifier for high dimensional datasets - Omega
(Volume 59, Part A, March 2016, Pages 40–46)
http://www.sciencedirect.com/science/article/pii/S0305048315001188

THANK YOU!

- Saeys, Y. and Inza, I. and Larrañaga, P. A review of feature selection techniques in bioinformatics. 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), pp 4–8, 2000.
- Johnstone, I.M. and Titterington, D.M.,Statistical challenges of high-dimensional data, Johnstone, I.M. and Titterington, D.M.,Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367, 4237–4253, 2009
- Köppen, M.,The curse of dimensionality,5th Online World Conference on Soft Computing in Industrial Applications (WSC5),4–8, 2000
- Beyer, K. and Goldstein, J. and Ramakrishnan, R. and Shaft, U., When is nearest neighbor meaningful?, Database Theory ICDT 99, 217–235, 1999
- Liu, H. and Motoda, H. Feature extraction, construction and selection: A data mining perspective, Springer, 1998

- Alon, U. and Barkai, N. and Notterman, D.A. and Gish, K. and Ybarra, S. and Mack, D. and Levine, A.J., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences, 6745–6750, 1999
- van't Veer, L.J. and Dai, H. and Van De Vijver, M.J. and He, Y.D. and Hart, A.A.M. and Mao, M. and Peterse, H.L. and van der Kooy, K. and Marton, M.J. and Witteveen, A.T. and others, Gene expression profiling predicts clinical outcome of breast cancer, nature, 530–536, 2002
- Laaksonen, Jorma, Local subspace classifier, Artificial Neural Networks ICANN'97, 637–642, 1997
- Golub, Gene H and Van Loan, Charles F, Matrix computations, 2012
- Hamm, Jihun and Lee, Daniel D, Grassmann discriminant analysis: a unifying view on subspace-based learning, Proceedings of the 25th international conference on Machine learning, 376–383, 2008

- Shipp, Margaret A and Ross, Ken N and Tamayo, Pablo and Weng, Andrew P and Kutok, Jeffery L and Aguiar, Ricardo CT and Gaasenbeek, Michelle and Angelo, Michael and Reich, Michael and Pinkus, Geraldine S and others, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, 68–74, 2002
- Petros Xanthopoulos, Panos M. Pardalos, Theodore B. Trafalis, Robust Data Mining, 2010
- Zhang, L. and Lin, X., Some considerations of classification for high dimension low-sample size data, Statistical Methods in Medical Research, 2011
- Edelman, Alan, Toms A. Arias, and Steven T. Smith, The geometry of algorithms with orthogonality constraints, SIAM journal on Matrix Analysis and Applications, 303–353 ,1998

- Hassell, Joseph, Boanerges Aleman-Meza, and I. Budak Arpinar, Ontology-driven automatic entity disambiguation in unstructured text, 2006
- Fenn, M.B. and Pappu, V., Data Mining for Cancer Biomarkers with Raman Spectroscopy, Data Mining for Biomarker Discovery, pages 143–168, 2012
- Golub, T.R. and Slonim, D.K. and Tamayo, P. and Huard, C. and Gaasenbeek, M. and Mesirov, J.P. and Coller, H. and Loh, M.L. and Downing, J.R. and Caligiuri, M.A. and others, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science, 531–537, 1999
- Lee, Sang Min, Dong Seong Kim, Ji Ho Kim, and Jong Sou Park, Spam detection using feature selection and parameters optimization, Intelligent and Software Intensive Systems (CISIS), 883–888, 2010

- Jiang, Sheng-Yi, and Xia Li, A hybrid clustering algorithm, Fuzzy Systems and Knowledge Discovery, 2009
- Satsangi, Amit, and Osmar R. Zaiane, Contrasting the contrast sets: An alternative approach, Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International, 2007
- Yu, Lei, and Huan Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, 856–863, 2003
- Kohavi, Ron., A study of cross-validation and bootstrap for accuracy estimation and model selection, 1137–1145, 1995
- Tseng T-LB, Huang C-C. Rough set-based approach to feature selection in customer relationship management. Omega 35(4):365–83, 2007
- Campbell J, Lo AW, M. A. The econometrics of financial markets. Princeton, New Jersey, USA: Princeton University Press, 1997

- Joseph Hassell BA-M, Arpinar IB. Ontology-driven automatic entity disambi- guation in unstructured text, vol. 4273. Springer, 2006
- Belloni Alexandre VC, Hansen C. High-dimensional methods and inference on structural and treatment effects. The Journal of Economic Perspectives, 28:29–50, 2014
- Bennett J, Lanning S. The netflix prize. In: Proceedings of the KDD cup and workshop, 2007
- Kampa K, Mehta S, Chou C, Chaovalitwongse W, Grabowski T. Sparse optimization in feature selection: application in neuroimaging. Journal of Global Optimization, 59(2-3):439–57, 2014
- Clarke, R. and Ressom, H.W. and Wang, A. and Xuan, J. and Liu, M.C. and Gehan, E.A. and Wang, Y., The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature Reviews Cancer, 37–49, 2008