O-IPCAC and its application to EEG classification

A. Rozza, G. Lombardi, M. Rosa, and E. Casiraghi
{rozza,lombardi,rosa,casiragh}@dsi.unimi.it

Università degli Studi di Milano

Outline

1. Introduction
   - EEG classification
   - IPCAC
2. O-IPCAC
   - Theoretical Problems
   - The Algorithm
3. Experimental Evaluation
   - EEG Dataset
   - Results
4. Conclusions
   - References
5. Appendix

EEG classification

- Recently this problem is raising a wide interest since it is the fundamental step of Brain to Computer Interface (BCI) systems: the translatation of the brain activity into commands for computers;
- The task of EEG classification is a hard problem:
  - The data are high dimensional;
  - The classes to be discriminated are often highly unbalanced;
  - The selection of discriminative information is difficult;
  - The cardinality of the training set is often lower than the space dimensionality.

Existing Approaches

- Feature extraction/selection techniques are generally used;
- This approach causes loss of discriminative information, and might affect the classification accuracy.

Different Approach

- Develop an efficient classifier that deals with high dimensional datasets whose cardinality is lower than the space dimensionality.
  - Apply it to the raw data.

Existing Approaches

- Feature extraction/selection techniques are generally used;
- This approach causes loss of discriminative information, and might affect the classification accuracy.

Different Approach

- Develop an efficient classifier that deals with high dimensional datasets whose cardinality is lower than the space dimensionality.
  - Apply it to the raw data.

Isotropic Principal Component Analysis Classifier [5]

IPCAC

A linear two-class classification algorithm, based on a new estimation of the Fisher Subspace [1], assuming points drawn by an isotropic Mixture of two Gaussian Functions.

- The Fisher subspace is spanned by the one-dimensional vector defined as follows:

$$\mathbf{F} = \frac{\boldsymbol{\mu}_A - \boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|} \tag{1}$$

Training task: In this phase the classifier exploits the training set to estimate the Fisher subspace $\mathbf{F}$ and the thresholding value γ.

Classification task: An unknown test point $\mathbf{p}$ is classified by projecting it on $\mathbf{F}$ and then thresholding with γ.

Isotropic Principal Component Analysis Classifier [5]

IPCAC

A linear two-class classification algorithm, based on a new estimation of the Fisher Subspace [1], assuming points drawn by an isotropic Mixture of two Gaussian Functions.

- The Fisher subspace is spanned by the one-dimensional vector defined as follows:

$$\mathbf{F} = \frac{\boldsymbol{\mu}_A - \boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|} \tag{1}$$

Training task: In this phase the classifier exploits the training set to estimate the Fisher subspace **F** and the thresholding value γ.

Classification task: An unknown test point **p** is classified by projecting it on **F** and then thresholding with γ.

IPCA-based Classifier - Training phase

Data whitening

- The probability distribution related to several classification tasks is not mean-centered, and its random variables are often correlated; To avoid this problem data whitening is performed ($\mathbf{W}$ is the whitening matrix).

Fisher subspace estimation

- The whitened training points are employed to compute the class means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$, and $\mathbf{F}$ (see Equation (1)).

Thresholding value

$$\gamma = \left\langle \underset{\{\bar{\gamma}\} \subseteq \{\mathbf{w} \cdot (\mathbf{p}_i - \tilde{\boldsymbol{\mu}})\}}{\operatorname{argmax}} \; Score(\bar{\gamma}) \right\rangle$$

Theoretical Problems in High Dimensionality

Covariance Matrix Estimation Problem

- Given the matrix $\mathbf{P} \in \Re^{D \times N}$, representing a training dataset $\mathcal{P} = \mathcal{P}_A \cup \mathcal{P}_B, |\mathcal{P}| = N = N_A + N_B$, let α be the ratio D/N;

If $\alpha \approx 1$, the sample covariance matrix $\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\mathbf{P}\mathbf{P}^T$ is not a consistent estimator of the population covariance matrix $\boldsymbol{\Sigma}$ [3].

Theoretical Problems (2)

Noise Problem

- Assuming that $\mathbf{\Sigma} = \mathbf{\Sigma}^* + \sigma^2\mathbf{I}$, where $\mathbf{\Sigma}^*$ has rank $k < D$ and $\sigma^2 I$ represents the contribution of a zero mean Gaussian noise affecting the data;
- Calling $\sigma^2 = \lambda_1 = \ldots = \lambda_{D-k-1} < \ldots < \lambda_D$ the ordered eigenvalues of $\mathbf{\Sigma}$;

Only the portion of the spectrum of $\mathbf{\Sigma}$ above $\sigma^2 + \sqrt{\alpha}$ can be correctly estimated from the sample [4].

- Denoting with $\tilde{\lambda}_1 < \ldots < \tilde{\lambda}_D$ the ordered eigenvalues of $\tilde{\Sigma}$;

If $\alpha \approx 1$ the estimates of the smallest eigenvalues $\tilde{\lambda}_i$ can be much larger than the real ones, and the corresponding estimated eigenvectors are uncorrelated with the real ones.

Theoretical Problems (2)

Noise Problem

- Assuming that $\mathbf{\Sigma} = \mathbf{\Sigma}^* + \sigma^2\mathbf{I}$, where $\mathbf{\Sigma}^*$ has rank $k < D$ and $\sigma^2 I$ represents the contribution of a zero mean Gaussian noise affecting the data;
- Calling $\sigma^2 = \lambda_1 = \ldots = \lambda_{D-k-1} < \ldots < \lambda_D$ the ordered eigenvalues of $\mathbf{\Sigma}$;

Only the portion of the spectrum of $\mathbf{\Sigma}$ above $\sigma^2 + \sqrt{\alpha}$ can be correctly estimated from the sample [4].

- Denoting with $\tilde{\boldsymbol{\lambda}}_1 < \ldots < \tilde{\boldsymbol{\lambda}}_D$ the ordered eigenvalues of $\tilde{\mathbf{\Sigma}}$;

If $\alpha \approx 1$ the estimates of the smallest eigenvalues $\tilde{\boldsymbol{\lambda}}_i$ can be much larger than the real ones, and the corresponding estimated eigenvectors are uncorrelated with the real ones.

Problems with dimensionality reduction

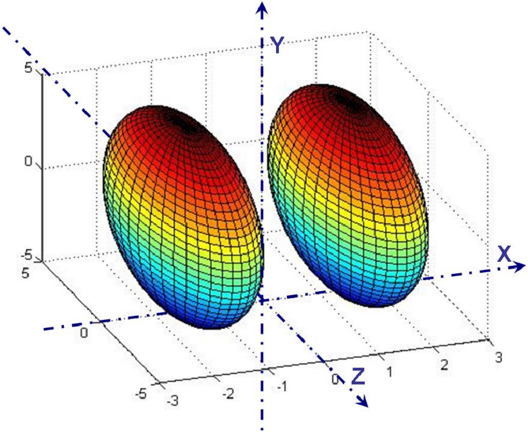
- Dimensionality reduction might delete discriminative information, decreasing the classification performance;

Consider two classes with the shape of parallel pancakes in $\Re^D$:

1. if the direction defined by the Fisher subspace in the original space is orthogonal to the subspace π_d defined by the first $d \leq D$ principal components, the dimensionality reduction process projects the data on π_d, obtaining an isotropic mixture of two completely overlapped Gaussian distributions.

Problems with dimensionality reduction (2)

Figure: Parallel Pancakes



O-IPCAC: the algorithm (1)

- To estimate the linear transformation $\mathbf{W}$, which represents the partial whitening operator, we apply the Truncated Singular Value Decomposition;
- The d largest singular values on the diagonal of $\mathbf{Q}_d$, and the associated left singular vectors, are employed to project the points in $\mathbf{P}$ on the subspace $\mathcal{SP}_d$ spanned by the columns of $\mathbf{U}_d$, and to perform the whitening, as follows:

$$\bar{\mathbf{P}}_{\mathbf{W}_d} = q_d \mathbf{Q}_d^{-1} \mathbf{P}_{\perp \mathcal{SP}_d} = q_d \mathbf{Q}_d^{-1} \mathbf{U}_d^T \mathbf{P} = \mathbf{W}_d \mathbf{P}$$

O-IPCAC: the algorithm (2)

- To avoid this information loss, we add to the partially whitened data the residuals ($\mathbf{R}$) of the points in $\mathbf{P}$ with respect to their projections on $\mathcal{SP}_d$:

$$\begin{aligned}
\mathbf{R} &= \mathbf{P} - \mathbf{U}_d \mathbf{P}_{\perp \mathcal{SP}_d} = \mathbf{P} - \mathbf{U}_d \mathbf{U}_d^T \mathbf{P} \\
\bar{\mathbf{P}}_{\mathbf{W}_D} &= \mathbf{U}_d \bar{\mathbf{P}}_{\mathbf{W}_d} + \mathbf{R} = \mathbf{U}_d \mathbf{W}_d \mathbf{P} + \mathbf{P} - \mathbf{U}_d \mathbf{U}_d^T \mathbf{P} \\
&= \left(q_d \mathbf{U}_d \mathbf{Q}_d^{-1} \mathbf{U}_d^T + \mathbf{I} - \mathbf{U}_d \mathbf{U}_d^T \right) \mathbf{P} \\
&= \mathbf{W}\mathbf{P}
\end{aligned}$$

  where $\mathbf{W} \in \Re^{D \times D}$ represents the linear transformation that whitens the data along the first d principal components, while keeping unaltered the information along the remaining components.

O-IPCAC: the algorithm (3)

- The Fisher subspace is estimated by exploiting the whitened class means, $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$, obtained by the class means in the original space $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$ as follows:

$$\begin{aligned}
\boldsymbol{\mu}_A &= \mathbf{W}\hat{\boldsymbol{\mu}}_A \\
&= \left(q_d \mathbf{U}_d \mathbf{Q}_d^{-1} \mathbf{U}_d^T + \mathbf{I} - \mathbf{U}_d \mathbf{U}_d^T\right) \hat{\boldsymbol{\mu}}_A \\
&= q_d \mathbf{U}_d \mathbf{Q}_d^{-1} \mathbf{U}_d^T \hat{\boldsymbol{\mu}}_A + \hat{\boldsymbol{\mu}}_A - \mathbf{U}_d \mathbf{U}_d^T \hat{\boldsymbol{\mu}}_A
\end{aligned}$$

- Using these quantities we estimate $\mathbf{f} = \frac{\boldsymbol{\mu}_A - \boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\|}$.
- We process an unknown point $\mathbf{p}$ by transforming it with $\mathbf{W}$, and projecting it on $\mathbf{f}$;

$$\mathbf{w} = \mathbf{W}^T \mathbf{f} = q_d \mathbf{U}_d^T \mathbf{Q}_d^{-1} \mathbf{U}_d \mathbf{f} + \mathbf{f} - \mathbf{U}_d^T \mathbf{U}_d \mathbf{f}$$

- Given a thresholding value γ, $\mathbf{p}$ is assigned to class A if $\mathbf{w} \cdot \mathbf{p} < \gamma$, to class B otherwise.

O-IPCAC: the algorithm (4)

- We never explicitly compute the matrix $\mathbf{W}$, but we perform the matrix times vector operations thus preventing a quadratic time/space complexity.

The Online algorithm

- With training sets of high cardinality, or when mini-batches of training data are dynamically supplied, subsequent training phases must be applied to update the classification model.
- To this aim, the algorithm has been extended to perform **online/incremental** training by updating:

$N_k, N_{A,k}, N_{B,k}$: number of training points seen until the k-th training phase;

$\boldsymbol{\mu}_k, \hat{\boldsymbol{\mu}}_{A,k}, \hat{\boldsymbol{\mu}}_{B,k}$: the means employed to obtain the centered sets $\mathcal{P}_k$, $\mathcal{P}_{A,k}$, and $\mathcal{P}_{B,k}$ respectively;

$\mathbf{U}_{d_k}, \mathbf{Q}_{d_k}, \mathbf{V}_{d_k}$: the SVD matrices related to $\mathcal{P}_k$, truncated to d_k principal components;

σ_A, σ_B : the standard deviations of the projections $\mathbf{w}_k^T \mathbf{P}_{A,k}$ and $\mathbf{w}_k^T \mathbf{P}_{B,k}$.

Data Description

- The data used in our tests have been distributed by the organizers of the MLSP 2010 [2] competition and consist of EEG brain signals collected while the subject viewed satellite images and tried to detect those containing a predefined target:
  - 64 channels of EEG data;
  - The total number of samples is 176378, and the sampling rate is 256Hz;
  - During the EEG recording 2775 satellite images were shown, partitioned in 75 activation blocks with 37 images per block;
  - **The classifier must analyze the brain activity to recognize those images containing the target.**

Pre-processing

- We pre-processed each channel with a Gaussian filter with cut-frequency of 2.2Hz, and we subtracted the filtered data from the original one to obtain high-pass filtered signals.
- These signals were then used to extract 64×97 image blocks, where each image block starts exactly 65 time samples ($\approx$ 250ms) after the corresponding image trigger.
- The extracted blocks are serialized in 2775 vectors in $\Re^{6208}$, of which only 58 points represent images with target.

Performance evalutation

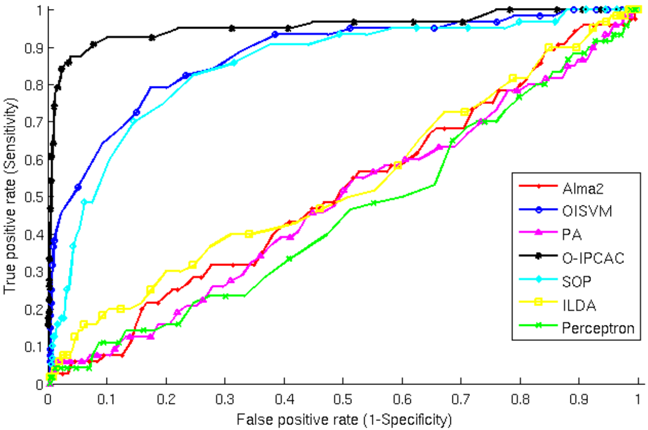
To evaluate the performance of our classifier:

- We computed the Receiver Operating Characteristic (ROC) curve;
- We estimated the Area Under the Curve (AUC).

To obtain an unbiased evaluation, we performed ten-fold cross validation, and we averaged the computed sensitivity and specificity values.

Results

Figure: ROC curves



Results and Comparison

Table: AUC per classifier

Classifier	AUC
O-IPCAC	**0.9541**
OISVM	0.8766
SOP	0.8479
ILDA	0.5315
Alma	0.5110
PA	0.4835
Perceptron	0.4507

Conclusions and Future Works

Conclusions

We proposes an online/incremental linear binary classifier that has been developed to deal with:

1. High dimensional data;
2. Classification problems where the cardinality of the point set is high;
3. Data dynamically supplied;
4. Highly unbalanced training sets whose cardinality is lower than the space dimensionality.

These peculiarities allow to manage EEG classification problem:

1. Without focusing on complex features extraction/selection techniques;
2. Dealing with the raw data;
3. Achieving good results.

Conclusions and Future Works

Future Works

- Apply our method to biological data (such as Microarray) where the datasets are characterized by a very large ratio between dimension and training points.
- Develop an adaptive version of O-IPCAC, to cope with classification problems where the probability distribution underlying the data changes with time.

References I

S. C. Brubaker and S. Vempala.
Isotropic pca and affine-invariant clustering.
CoRR, abs/0804.3575, 2008.

K. Hild, M. Kurimo, and V. Calhoun.
The sixth annual mlsp competition.
In *MLSP '10*, Sept. 2010.

I. M. Johnstone and A. Y. Lu.
Sparse principal components analysis.
Journal of the American Statistical Association, 2004.

References II

D. Paul.
Asymptotics of sample eigenstructure for a large dimensional spiked covariance model.
Statistica Sinica, 2007.

A. Rozza, G. Lombardi, and E. Casiraghi.
Novel ipca-based classiers and their application to spam filtering.
In *Proceedings of the 9th International Conference on Intelligent System Design and Applications (ISDA09)*. IEEE CS, 2009.

Any questions?

Whitening Process

1. estimate the expectation $\tilde{\boldsymbol{\mu}} = N^{-1} \sum_i \mathbf{p}_i$, and the covariance matrix $\tilde{\boldsymbol{\Sigma}} = N^{-1} \sum_i (\mathbf{p}_i - \tilde{\boldsymbol{\mu}})(\mathbf{p}_i - \tilde{\boldsymbol{\mu}})^T$;
2. estimate the principal components through the covariance matrix Eigen-decomposition $\mathbf{X} \boldsymbol{\Lambda} \mathbf{X}^T = \tilde{\boldsymbol{\Sigma}}$;
3. estimate the whitening matrix as $\mathbf{W} = \mathbf{X} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{X}^T$.

Back