TOWARD A SEMI-SUPERVISED APPROACH IN CLASSIFICATION BASED ON PRINCIPAL DIRECTIONS

Luminita State
Dept. of Computer Science, University of Pitesti
Caderea Bastliei #45, Bucuresti – 1, Pitesti,Romania

Catalina Cocianu
Dept. of Computer Science, Academy of Economic Studies
Calea Dorobantilor #15-17, Bucuresti –1, Bucharest, Romania

Doru Constantin, Corina Sararu
Dept. of Computer Science, University of Pitesti, Pitesti, Romania

Panayiotis Vlamos
Ionian University, Corfu, Greece

Keywords: Principal component analysis, Data compression/decompression, Statistical signal classification, First order approximation for eigen values and eigen vectors.

Abstract: Since similarity plays a key role for both clustering and classification purposes, the problem of finding a relevant indicators to measure the similarity between two patterns drawn from the same feature space became of major importance. The advantages of using principal components reside from the fact that bands are uncorrelated and no information contained in one band can be predicted by the knowledge of the other bands. The semi-supervised learning (SSL) problem has recently drawn large attention in the machine learning community, mainly due to its significant importance in practical applications. The aims of the research reported in this paper are to report experimentally derived conclusions on the performance of a PCA-based supervised technique in a semi-supervised environment. A series of conclusions experimentally established by tests performed on samples of signals coming from two classes are exposed in the final section of the paper.

1 INTRODUCTION

In supervised learning, the basis is represented by a training set of examples (inputs) with associated labels (output values). Usually, the examples are in the form of attribute vectors, so that the input space is a subset of $\mathbf{R}^n$. Once the attribute vectors are available, a number of sets of hypothesis could be chosen for the problem.

Traditional statistics and the classical neural network literature have developed many methods for discriminating between two classes of instances using linear functions, as well as methods for interpolation using linear functions. These techniques, which include both efficient iterative procedures and theoretical analysis of their generalization properties, provide a suitable framework within which the construction of more complex systems are usually developed.

The semi-supervised learning (SSL) problem has recently drawn large attention in the machine learning community, mainly due to its significant importance in practical applications.

In statistical machine learning, there is a sharp distinction between unsupervised and supervised learning. In the former scenario we are given a

State L., Cocianu C., Constantin D., Sararu C. and Vlamos P. (2009).
TOWARD A SEMI-SUPERVISED APPROACH IN CLASSIFICATION BASED ON PRINCIPAL DIRECTIONS.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 68-73
DOI: 10.5220/0002233200680073

sample $\{x_i\}$ of patterns in $\aleph$ drawn independently and identically distributed (i.i.d.) from some unknown data distribution with density P(x), the goal being to estimate either the density or a functional thereof. Supervised learning consists of estimating a functional relationship x → y between a covariate $x \in \aleph$ and a class variable $y \in \{1,2,...,M\}$, with the goal of minimizing a functional of the joint data distribution P(x, y) such
as the probability of classification error.

The terminology "unsupervised learning" is a bit unfortunate: the term density estimation should probably suit better. Traditionally, many techniques for density estimation propose a latent (unobserved) class variable y and estimate P(x) as mixture distribution $\sum_{y=1}^{M} P(x|y)P(y)$. Note that y has a fundamentally different role than in classification, in that its existence and range c is a modeling choice rather than observable reality.

The semi-supervised learning problem belongs to the supervised category, since the goal is to minimize the classification error, and an estimate of P(x) is not sought after. The difference from a standard classification setting is that along with a labeled sample $D_l = \{(x_i, y_i) | i = 1,...,n\}$ drawn i.i.d. from P(x, y) we also have access to an additional unlabeled sample $D_u = \{x_{n+j} | j = 1,...,m\}$from the marginal P(x). We are especially interested in cases where *n«m* which may arise in situations where obtaining an unlabeled sample is cheap and easy, while labeling the sample is expensive or difficult.

Principal Component Analysis, also called Karhunen-Loeve transform is a well-known statistical method for feature extraction, data compression and multivariate data projection and so far it has been broadly used in a large series of signal and image processing, pattern recognition and data analysis applications.

The advantages of using principal components reside from the fact that bands are uncorrelated and no information contained in one band can be predicted by the knowledge of the other bands, therefore the information contained by each band is maximum for the whole set of bits (Diamantaras, 1996).

Recently, alternative methods as discriminant common vectors, neighborhood components analysis and Laplacianfaces have been proposed allowing the learning of linear projection matrices for dimensionality reduction. (Liu, Chen, 2006; Goldberger, Roweis, Hinton, Salakhutdinov, 2004)

The aims of the research reported in this paper are to report experimentally derived conclusions on the performance of a PCA-based supervised technique in a semi-supervised environment.

The structure of a class is represented in terms of the estimates of its principal directions computed from data, the overall dissimilarity of a particular object with a given class being given by the "disturbance" of the structure, when the object is identified as a member of this class. In case of unsupervised framework, the clusters are computed using the estimates of the principal directions, that is the clusters are represented in terms of skeletons given by sets of orthogonal and unit eigen vectors (principal directions) of each cluster sample covariance matrix. The reason for adopting this representation relies on the property that a set of principal directions corresponds to the maximum variability of each class.

A series of conclusions experimentally established by tests performed on samples of signals coming from two classes are exposed in the final section of the paper.

2 THE MATHEMATICS BEHIND THE PROPOSED ATTEMPT

The classes are represented in terms of multivariate density functions, and an object coming from a certain class is modeled as a random vector whose repartition has the density function corresponding to this class. In cases when there is no statistical information concerning the set of density functions corresponding to the classes involved in the recognition process, usually estimates based on the information extracted from available data are used instead.

The principal directions of a class are given by a set of unit orthogonal eigen vectors of the covariance matrix. When the available data is represented by a set of objects $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$, belonging to a certain class C, the covariance matrix is estimated by the sample covariance matrix,

$$\hat{\Sigma}_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_N)(X_i - \hat{\mu}_N)^T , \qquad (1)$$

where $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$.

Let us denote by $\lambda_1^N \geq \lambda_2^N \geq ... \geq \lambda_n^N$ the eigen values and by $\boldsymbol{\psi}_1^N,...,\boldsymbol{\psi}_n^N$ a set of orthonormal eigen vectors of $\hat{\Sigma}_N$.

If a new example $\mathbf{X}_{N+1}$ coming from the same class has to be included in the sample, the new estimate of the covariance matrix can be recomputed as,

$$\hat{\mathbf{\Sigma}}_{N+1} = \hat{\mathbf{\Sigma}}_N + \frac{1}{N+1}\left(\mathbf{X}_{N+1} - \hat{\boldsymbol{\mu}}_N\right)\left(\mathbf{X}_{N+1} - \hat{\mu}_N\right)^T - -\frac{1}{N}\hat{\Sigma}_N \qquad (2)$$

Using first order approximations (State, Cocianu, 2006), the estimates of the eigen values and eigen vectors respectively are given by,

$$\lambda_i^{N+1} = \lambda_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\hat{\mathbf{\Sigma}}_N \boldsymbol{\psi}_i^N = \left(\boldsymbol{\psi}_i^N\right)^T \hat{\mathbf{\Sigma}}_{N+1}\boldsymbol{\psi}_i^N \qquad (3)$$

$$\boldsymbol{\psi}_i^{N+1} = \boldsymbol{\psi}_i^N + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left(\boldsymbol{\psi}_N^j\right)^T \Delta\hat{\mathbf{\Sigma}}_N \boldsymbol{\psi}_i^N}{\lambda_i^N - \lambda_j^N}\boldsymbol{\psi}_j^N \qquad (4)$$

On the other hand, when an object has to be removed from the sample, then the estimate of the covariance matrix can be computed as,

$$\hat{\mathbf{\Sigma}}_N = \hat{\mathbf{\Sigma}}_{N+1} + \Delta\mathbf{\Sigma}_{N+1}, \qquad (5)$$

where

$$\Delta\mathbf{\Sigma}_{N+1} = \frac{1}{N-1}\hat{\mathbf{\Sigma}}_{N+1} - \\ -\frac{N}{(N-1)(N+1)}\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)^T$$

and

$$\boldsymbol{\mu}_N = \frac{(N+1)\boldsymbol{\mu}_{N+1}}{N} - \frac{\mathbf{X}_{N+1}}{N}$$

The conclusion formulated in the next lemma can be proved by straightforward computation.

Lemma. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K$ be an *n*-dimensional Bernoullian sample. We denote by $\hat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$, $\hat{\mathbf{\Sigma}}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)^T$, and let $\left\{\lambda_i^N\right\}_{1\le i\le n}$ be the eigen values and $\left\{\boldsymbol{\psi}_i^N\right\}_{1\le i\le n}$ a set of orthogonal unit eigen vectors of $\hat{\mathbf{\Sigma}}_N$, $2 \le N \le K-1$. In case the eigen values of $\hat{\mathbf{\Sigma}}_{N+1}$ are pairwise distinct, the following first order approximations hold,

$$\lambda_i^N = \lambda_i^{N+1} + \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \qquad (6)$$

$$\boldsymbol{\psi}_i^N = \boldsymbol{\psi}_i^{N+1} + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1}}{\lambda_i^{N+1} - \lambda_j^{N+1}}\boldsymbol{\psi}_j^{N+1} \qquad (7)$$

where $\Delta\mathbf{\Sigma}_{N+1} = \hat{\mathbf{\Sigma}}_N - \hat{\mathbf{\Sigma}}_{N+1}$

Let $\psi_1^N, ..., \psi_n^N$ be set of principal directions of the class C computed using $\hat{\mathbf{\Sigma}}_N$. When the example $\mathbf{X}_{N+1}$ is identified as a member of the class *C*, then the disturbance implied by extending C is expressed as,

$$D = \frac{1}{n}\sum_{k=1}^{n} d\left(\psi_k^N, \psi_k^{N+1}\right) \qquad (8)$$

where *d* is the Euclidian distance and $\boldsymbol{\psi}_1^{N+1}, ..., \boldsymbol{\psi}_n^{N+1}$ are the principal directions computed using $\hat{\mathbf{\Sigma}}_{N+1}$.

Let $H = \left\{C_1, C_2, ..., C_M\right\}$ be a set of classes, where the class C_j contens N_j elements. The new object X is alloted to C_j, one of the classes for which

$$D = \frac{1}{n}\sum_{k=1}^{n} d\left(\psi_{k,j}^{N_j}, \psi_{k,j}^{N_j+1}\right) = \\ = \min_{1\le p\le M}\frac{1}{n}\sum_{k=1}^{n} d\left(\psi_{k,p}^{N_p}, \psi_{k,p}^{N_p+1}\right) \qquad (9)$$

In order to protect against misclassifications, due to insufficient "closeness" to any class, we implement this recognition technique using a threshold *T*>0 such that the example **X** is allotted to C_j only if relation (8) holds and *D*<*T*.

The classification of samples for which the resulted value of *D* is larger than *T* is postponed and the samples are kept in a new possible class CR. The reclassification of elements of CR is then performed followed by the decision concerning to either reconfigure the class system or to add CR as a new class in *H*.

For each new sample allotted to a class, the class characteristics (the covariance matrix and the principal axes) are re-computed using (2), (3) and (4). The skeleton of each class is computed using an exact method, **M**, in case PN samples have been already classified in $H = \left\{C_1, C_2, ..., C_M\right\}$.

Briefly, the recognition procedure, P1, is described below (Cocianu, State, 2007).

Input: $H = \left\{C_1, C_2, ..., C_M\right\}$ the set of samples coming from M classes respectively

Step 1: For each class, compute a set of orthogonal unit eigen vectors (characteristics of the classes)

Repeat

i←1

Step 2: Generate **X** a new test example and classify **X** according to (8)

Step 3: If $\exists j, 1 \le j \le M$ such that **X** is allotted to C_j, then

```
    3.1.re-compute the characteristics of $C_j$ using (3), (4) and (5)
    3.2. i←i+1
Step 4: If i<PN goto Step 2
    Else
        4.1. For i=$\overline{1,M}$, compute the characteristics of the class $C_i$ using M.
        4.2. goto Step 2.
Until all test examples are classified
Output: The new set $\{C_1, C_2, ..., C_M\} \cup CR$
```

3 EXPERIMENTAL ANALYSIS ON THE PERFORMANCE OF THE PROPOSED CLASSIFICATION METHOD

In this section, we present the results in testing the performance of the proposed approach evaluated in terms of the recognition error. The tests were performed in discriminating between two classes of signals, with known statistical properties. The evaluation of the error is computed on new test examples. The approach can be taken as a semi-supervised approach because each new test example is included in the class established by the decision rule (not necessarily being the true provenance class) therefore becoming involved in the re-actualization of the new characteristics.

The classes are represented by NP examples coming from each class.

Test 1. The evaluation of error using the leaving one out method. Sequentially, one of the given examples is removed from the sample. The classifier is designed using the rest of 2NP-1 examples (that is the characteristics of the classes are computed in terms of the NP, NP-1 remaining examples) and the removed example is classified into one of resulted classes. The error is evaluated as $\frac{F}{2NP}$, where F is the number of misclassified examples.

Let $\left\{\psi_i^{1,NP}\right\}_{1\le i\le n}$, $\left\{\psi_i^{2,NP}\right\}_{1\le i\le n}$, $\left\{\lambda_i^{1,NP}\right\}_{1\le i\le n}$, $\left\{\lambda_i^{2,NP}\right\}_{1\le i\le n}$ be the characteristics of the classes and the corresponding eigen values at the initial moment and $\mu_{NP}^1, \mu_{NP}^2, \Sigma_{NP}^1, \Sigma_{NP}^2$ the sample means and the sample covariance matrices respectivelly. Let X be the removed example. In case X comes from the first class, then the new characteristics are,

$$\psi_i^{1,NP-1} = \psi_i^{1,NP} + \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{\left(\psi_j^{1,NP}\right)^T \Delta\Sigma_{NP}^1 \psi_i^{1,NP}}{\lambda_i^{1,NP} - \lambda_j^{1,NP}} \psi_j^{1,NP}$$

for the first class and remains unchanged for the second one, where

$$\Delta\Sigma_{NP}^1 = \Sigma_{NP-1}^1 - \Sigma_{NP}^1$$

$$\Sigma_{NP-1}^1 = \frac{NP-1}{NP-2}\Sigma_{NP} - $$
$$- \frac{NP-1}{(NP-2)NP}\left(X - \mu_{NP-1}^1\right)\left(X - \mu_{NP-1}^1\right)^T$$

$$\mu_{NP-1}^1 = \frac{NP\mu_{NP}^1}{NP-1} - \frac{X}{NP-1}$$

In case X comes from the second class, similar formula are used.

The evaluation of the error is performed for $NP = 10,20,30,40,50$. Several tests were performed on samples generated from 3 repartitions, Gaussian, Rayleigh and geometric, each class corresponding to one of them. All tests reported to a surprising conclusion, that is the misclassification error is very closed to 0.

a) The classes correspond to the Gaussian repartition and Rayleigh repartition respectively, NP=150, n=50, e=50, where n is the data dimensionality and e is the number of epochs, the resulted empirical error is 0.0327. The variation of the empirical error in terms of e is presented in Figure 1.

b) The classes correspond to the geometric repartition and Rayleigh repartition respectively, NP=150, n=50, e=50, where n is the data dimensionality and e is the number of epochs, the resulted empirical error is 0.0112. The variation of the empirical error in terms of e is presented in Figure 2.

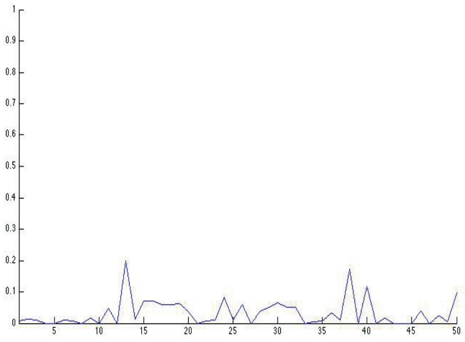

Figure 1.

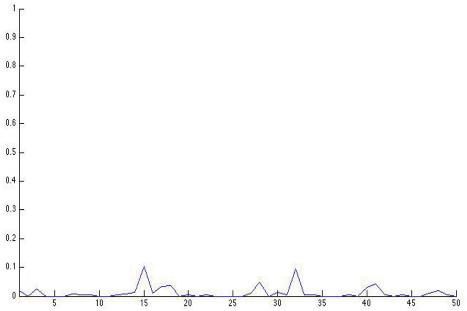

Figure 2.

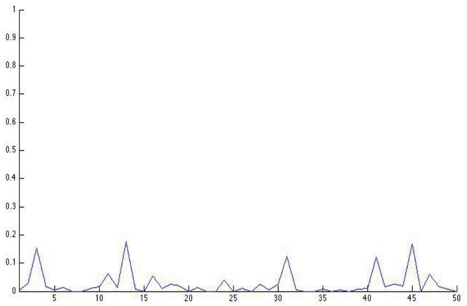

Figure 3.

c) The classes correspond to the Gaussian repartition, NP=150, n=50, e=50, where n is the data dimensionality and e is the number of epochs, the resulted empirical error is 0.0261. The variation of the empirical error in terms of e is presented in Figure 3.

Test 2. The evaluation of the error by counting the misclassified examples from a set of NC new test samples coming from the given classes of the same repartitions.

In this case, the learning is performed in a non-adaptive way, that is first order approximations of the characteristics for each class are used for classification purposes only (the characteristics of the classes are the initial computed characteristics during the classification process).

The tests were performed for NP=10,20,30,40,50, NC=10,20,30,40,50, n=50, e=50, where n is the data dimensionality and e is the number of epochs.

a) The classes correspond to the Gaussian repartition and Rayleigh repartition respectively.

b) The classes correspond to the geometric repartition and Rayleigh repartition respectively.

c) The classes correspond to the Gaussian repartition.

The values of the empirical error in terms of e lie in the interval [0.02,0.15] in case a), [0.32,0.4] in case b), and [0.04,0.4] in case c) respectively. In all cases, a decreasing tendency is identified while the number of epochs increases.

Test 3. The evaluation of the error by counting the misclassified examples from a set of NC new test samples coming from the given classes of the same repartitions.

In this case, the learning is performed in an adaptive way, that is, each new classified example contributes to the new characteristics of the class the exampled is assigned to, the new characteristics being computed using first order approximations in terms of the previous ones. Besides, after each iteration, the characteristics of the new resulted classes are re-computed using an exact method **M**.

The tests were performed for NP=150, NC=10,20,30,40,50, n=50, e=100, where n is the data dimensionality and e is the number of epochs.

a) The classes correspond to the Gaussian repartition and Rayleigh repartition respectively, The empirical error stabilises in few epochs at the value 0.015 and remains unchanged while the number of epochs increases.

b) The classes correspond to the geometric repartition and Rayleigh repartition respectively. The variation of the empirical error in terms of e is presented in Figure 4.

c) The classes correspond to the Gaussian repartition. The variation of the empirical error in terms of e is presented in Figure 5.

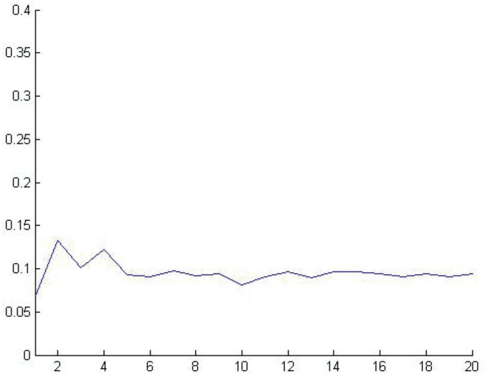

Figure 4.

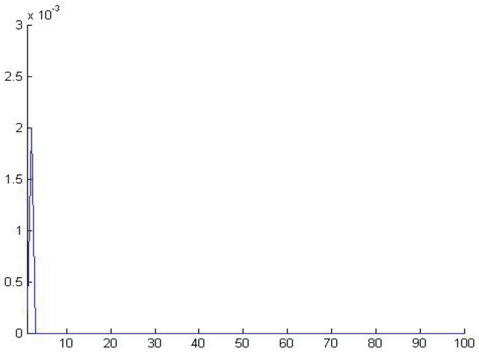

Figure 5.

Finally, we conclude that the long series of tests on the proposed classification procedure pointed out very good performance in terms of the misclassification error. In spite of the apparent complex structure, using first order approximations for the class characteristics, its complexity significantly decreases without degrading the classification accuracy.

REFERENCES

Chapelle, O., Scholkopf, B., Zien, A. (Eds.), 2006. *Semi-Supervised Learning*, MIT Press

Cocianu, C., State, L., Rosca, I., Vlamos, P. 2007. A New Adaptive Classification Scheme Based on Skeleton Information, *Proceedings of 2nd International Conference on Signal Processing and Multimedia Applications* 2007

Diamantaras, K.I., Kung, S.Y., 1996. Pr*incipal Component Neural Networks: theory and applications*, John Wiley &Sons

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighbourhood Component Analysis. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*

Gordon, A.D. 1999. *Classification*, Chapman&Hall/CRC, 2nd Edition

Hastie, T., Tibshirani, R., Friedman, J. 2001. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer-Verlag

Jain,A.K., Dubes,R., 1988. *Algorithms for Clustering Data*, Prentice Hall,Englewood Cliffs, NJ.

Liu, J., and Chen, S. 2006. Discriminant common vectors versus neighbourhood components analysis and Laplacianfaces: A comparative study in small sample size problem. *Image and Vision Computing* 24 (2006) 249-262

Smith,S.P., Jain,A.K., 1984. Testing for uniformity in multidimensional data, In *IEEE Trans.`Patt. Anal.` and Machine Intell.*, 6(1),73-81

State, L., Cocianu, C., Vlamos, P, Stefanescu, V., 2006. PCA-Based Data Mining Probabilistic and Fuzzy Approaches with Applications in Pattern Recognition. In *Proceedings of ICSOFT 2006*, Portugal, pp. 55-60**.**

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge