Minimum Classification Error Principal Component Analysis

T. B. A. de Carvalho[1], M. A. A. Sibaldo[1], Tsang I. R.[2], G. D. C. Cavalcanti[2]

[1] Universidade Federal Rural de Pernambuco, Brasil
{tiago.buarque, mariaaparecida.sibaldo}@ufrpe.br
[2] Universidade Federal de Pernambuco, Brasil
{tir,gdcc}@cin.ufpe.br

Abstract. We present an alternative method to use Principal Component Analysis (PCA) for supervised learning. The proposed method extract features similarly to PCA but the features are selected by minimizing the Bayes error rate for classification. Experiments using two real datasets shows that the recognition accuracy of the proposed technique is improved compared to PCA.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Principal component analysis, Dimensionality reduction and manifold learning, Supervised learning by classification, Data mining

1. INTRODUCTION

Principal Component Analysis (PCA) is a technique used to reduce data dimensionality. It projects the points into the directions of maximal variance within data space. These directions are the eigenvectors of data covariance matrix. In most of the cases, only some few eigenvectors are selected, normally the ones that have the highest eigenvalues. The eigenvalue is equivalent to the variance of a new variable, that is obtained by projecting the data into an eigenvector. The new variables not only have maximal variance, but they are also uncorrelated [Bishop 2006]. PCA is a very well-known technique that is used in several different applications such as face recognition [Turk and Pentland 1991] and text classification [Alencar et al. 2014].

From the perspective of machine learning, PCA is an unsupervised feature extraction technique. Nonetheless, it is also used in supervised tasks such as in classification and regression. Some versions of supervised PCA have been proposed, for example, Barshan et al. [Barshan et al. 2011] proposed a version of supervised PCA for classification. The method defines class representatives and computes PCA for these points. Directions with maximal variances for those points are also the directions that best separate the classes. Another version of supervised PCA was proposed by Bair et al. [Bair et al. 2006] for regression. The technique selects features that have high predictive power and compute PCA using only those features. Therefore, avoiding the interference of features that have high variance but low predictive power.

The Bayesian approach for classification is very robust and, similar to PCA it depends on the data covariance matrix [Duda et al. 2000]. Here, we propose a supervised version of PCA that minimizes the Bayes error rate for classification. The method projects the same features as PCA but selects the ones that minimize the Bayes error rate, while PCA select the features with maximal variance. Therefore, it can be more suitable for classification task than standard PCA. Since projections of

This work was partially supported by CAPES, CNPq and FACEPE.

maximal variance might not be the best way to separate (discriminate) data form different classes [Bishop 2006].

2. FEATURE EXTRACTION WITH PCA

Suppose a dataset matrix $\mathbf{X}'_{n\times d}$ with n points and d features. Each row of $\mathbf{X}'$ is a point and each column is a feature.

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}. \tag{1}$$

The j-th point is defined as a d dimensional column vector $\mathbf{x}_j$,

$$\mathbf{x}_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jd} \end{bmatrix}, \tag{2}$$

for $j = 1, \ldots, n$ and the data mean vector is

$$\bar{\mathbf{x}} \;=\; n^{-1}\sum_{j=1}^{n} \mathbf{x}_j. \tag{3}$$

The centered matrix is $\mathbf{X}$ having the j-th row equal to $(\mathbf{x}_j - \bar{\mathbf{x}})^T$:

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \end{bmatrix}. \tag{4}$$

The covariance matrix of $\mathbf{X}$ is defined as

$$\mathbf{\Sigma}_{\mathbf{X}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}. \tag{5}$$

Each column $\boldsymbol{\xi}_i$, for $i = 1, \ldots, k$, of the matrix

$$\mathbf{E}_k = [\boldsymbol{\xi}_1 \ldots \boldsymbol{\xi}_k], \tag{6}$$

is an eigenvector of $\mathbf{\Sigma}_{\mathbf{X}}$. $\mathbf{E}_k$ have up to d eigenvectors, for $k = 1, \ldots, d$. Each eigenvector $\boldsymbol{\xi}_i$ have an associated eigenvalue λ_i, which is the variance of the extracted feature

$$\mathbf{f}_i = \mathbf{X}\boldsymbol{\xi}_i. \tag{7}$$

The value of the i-th extracted feature for the j-th point is w_{ij}, where $\mathbf{f}_i = [w_{1i} \ldots w_{ni}]^T$.

The projection of the point $\mathbf{x}_j^T = [x_{j1} \ldots x_{jd}]$ for the space of projected features is $\mathbf{w}_j^T = [w_{j1} \ldots w_{jk}]$, given by

$$\mathbf{w}_j^T = \mathbf{x}_j^T\mathbf{E}_k. \tag{8}$$

The eigenvectors in $\mathbf{E}_k$ are sorted, so that $\lambda_1 > \ldots > \lambda_k$. In PCA, the points are projected in the directions of maximal variances, these directions are the eigenvectors of the covariance matrix that has the greatest eigenvalues. The new data matrix $\mathbf{W}_{n\times k}$ is defined as:

$$\mathbf{W} = \mathbf{X}\mathbf{E}_k, \tag{9}$$

Each row of this matrix is a point and each column an extracted feature.

The covariance matrix of $\mathbf{W}$ is $\mathbf{\Sigma_W} = n^{-1}\mathbf{W}^T\mathbf{W}$, so that $\mathbf{\Sigma_W} = diag(\lambda_1, \ldots, \lambda_k)$. The variables are uncorrelated since the off-diagonal elements of $\mathbf{\Sigma_W}$ are equal to 0. This property is very relevant for supervised learning, because it allows the selection of any subset of the projected variables by ignoring their interaction. However, selecting eigenvectors of highest eigenvalues may not be the best strategy for classification problems. Therefore, we propose a method of selecting the eigenvectors by minimizing the Bayes error rate for classification.

3. BAYES ERROR RATE

The Bayes error rate for classification is defined as the probability of the classification error, *i.e.*, the expected error rate. This error estimation can have a simplified form by imposing some restrictions [Duda et al. 2000]. Here, we consider the following five restrictions. (1) The data presents a multivariate normal distribution. (2) The problem has only two classes. (3) The prior probabilities of both classes are equal. (4) Both classes have the same covariance matrix, the same assumption is used for PCA. Finally, (5) the features are statistically independent, similarly to PCA. Then the Bayes error rate is given by

$$P(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du. \tag{10}$$

The Bayes error rate decreases as r increases. We define r^2 as the Mahalanobis distance between the mean vectors of the classes ($\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$):

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{11}$$

For independent features, the covariance matrix is diagonal. The off-diagonal elements are the features covariances, which have values equal to zero. This means that each feature is uncorrelated so r have a special form:

$$r = \sqrt{\sum_{i=1}^{d} \left(\frac{\mu_{1i} - \mu_{2i}}{\sigma_i} \right)^2}, \tag{12}$$

where $i = 1, \ldots, d$ are the indexes for the features. The variables μ_{1i} and μ_{2i} are the mean of the feature i for classes 1 and 2, respectively. And σ_i is the variance of the feature i that is the same for both classes.

We emphasize that the probability of classification error decreases as r increases. From Equation 12, we can conclude that each feature contributes for minimizing the probability of classification error. In fact, some feature contribute more than others. The larger the difference between the means of the two classes related to the feature variance, the higher is the contribution of this feature to minimize Bayes error.

4. PROPOSED METHOD

Since PCA generate uncorrelated features and considering that the covariance matrix is the same for every class in the dataset (because it computes direction of maximal variance for a covariance matrix estimated for all data), then the Bayes error rate can be minimized, proportionally to r as defined in (12), for features extracted using (9). The proposed method considers these equations to choose the PCA projected variables. However, instead of selecting the directions of maximal variance for the classification task, we select the directions that minimizes Bayes error rate.

The problem continues to be restricted to two classes, setting $\mathbf{W} = \mathbf{X}\mathbf{E}_d$, as in (9). However, now the features are extracted for d eigenvectors. We define w_{ij} as the value of the i-th new feature

$(i = 1, \ldots, d)$ for the j-th point $(j = 1, \ldots, n)$. The mean of the i-th feature for the c-th class $(c = 1, 2)$ is

$$\bar{w}_{ci} = \frac{\sum_{j=1}^{n} w_{ij}\delta_{jc}}{\sum_{j=1}^{n} \delta_{jc}}, \tag{13}$$

where δ_{jc} is the Dirac delta function $\delta_{jc} = 1$ if the j-point belongs to the c-th class, and $\delta_{jc} = 0$, otherwise.

We propose a score of the relevance for the classification of a feature extracted with PCA, s_i is the score for the i-th extracted feature:

$$s_i = \begin{cases} |\bar{w}_{1i} - \bar{w}_{2i}|/\lambda_i, & \text{if } \lambda_i \neq 0 \\ 0, & \text{if } \lambda_i = 0 \end{cases}, \tag{14}$$

where λ_i is the eigenvalue of the eigenvector from which the i-th feature were computed, and $\bar{w}_{ci}$ is the mean of the i-th feature for the c-th class $(c = 1, 2)$. If $\lambda_i = 0$ the variance of the i-th extracted feature is zero, which means that the variable has the same value for all points. Therefore it is not useful for classification and its score is set as $s_i = 0$. Otherwise the score is positive and is defined as the absolute value of the difference between the mean of each class divided by the variance of the feature. Features selected according to this score minimize the Bayes error rate. The proposed method consists in the following steps:

(1) Project the data as $\mathbf{W} = \mathbf{X}\mathbf{E}_d$, similar to Equation (9).

(2) Compute the mean of each feature for each class $\bar{w}_{ci}$, Equation (13).

(3) Compute the score s_i of each feature, Equation (14).

(4) Select k features with the highest score.

(5) Define the projection matrix as:

$$\mathbf{S}_k = [\boldsymbol{\xi}_1 \ldots \boldsymbol{\xi}_k] \tag{15}$$

with the eigenvectors that have the highest scores s_i, such that $s_i \geq s_j$ if $\boldsymbol{\xi}_i \in \mathbf{S}$ and $\boldsymbol{\xi}_j \notin \mathbf{S}$.

(6) Project the data as:

$$\mathbf{V} = \mathbf{X}\mathbf{S}_k, \tag{16}$$

where $\mathbf{V}_{n \times k}$ is the projected data matrix with n points and k discriminant features.

The difference between standard PCA and the proposed method is that the selected features in PCA are the ones of highest eigenvalues (λ_i) and the selected features in the proposed method are the ones with the highest discriminant score (s_i).

5. EXPERIMENTS

The experiments were performed using two datasets from the UCI Machine Learning Repository [Lichman 2013]. The Climate Model Simulation Crashes Data Set that has 540 points, 18 features and the Banknote Authentication Data Set that has 1,372 points, 4 features, both datasets have two classes. Accuracy, the rate of corrected classified points, is the metric used to evaluated the methods. Each point in the plot is the average accuracy for 100 holdout experiments. In each holdout experiment, 50% of the points from each class were randomly chosen for training and the remaining points were used for testing. The training set were used for both PCA and the proposed method. Both training and test sets were projected using k selected eigenvector, $k = 1, \ldots, d$. The 1-NN (Nearest Neighbor) with Euclidean distance, Naive Bayes with normal kernel smoothing density estimate, pruned Decision Tree with Gini's diversity index and a minimum of 10 nodes per leaf, and Fisher's Linear Discriminant were used for classification. The experiment were performed using Matlab 2015b

Statistics and Machine Learning Toolbox[1]. The result are shown in Figures 1, 2, 3, and 4. The results are also detailed in Tables I, II, III, and IV.

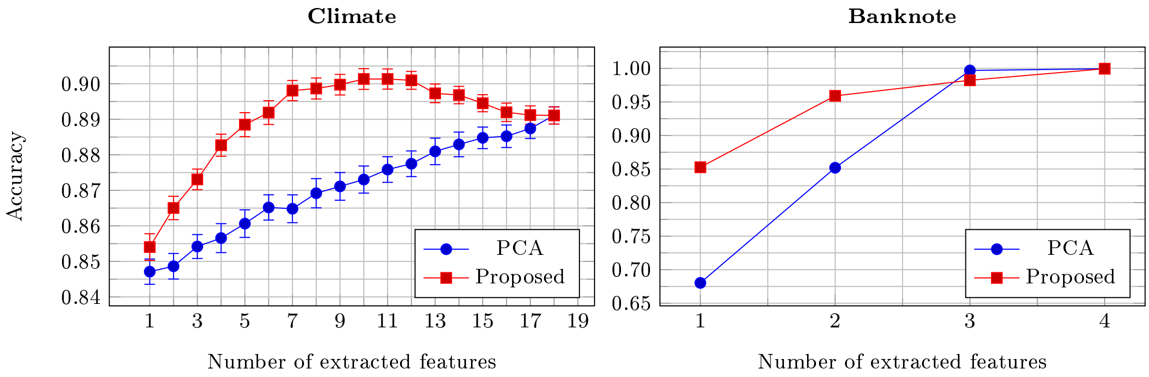

Fig. 1. Accuracy per number of extracted features for Climate and Banknote datasets using features extracted with PCA and the proposed method using **1-NN** Classifier.

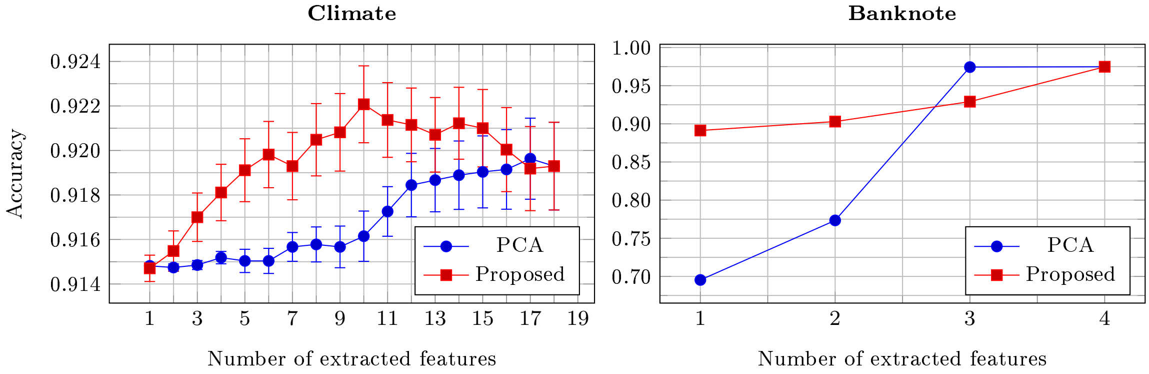

Fig. 2. Accuracy per number of extracted features for Climate and Banknote datasets using features extracted with PCA and the proposed method using **Naive Bayes** Classifier.

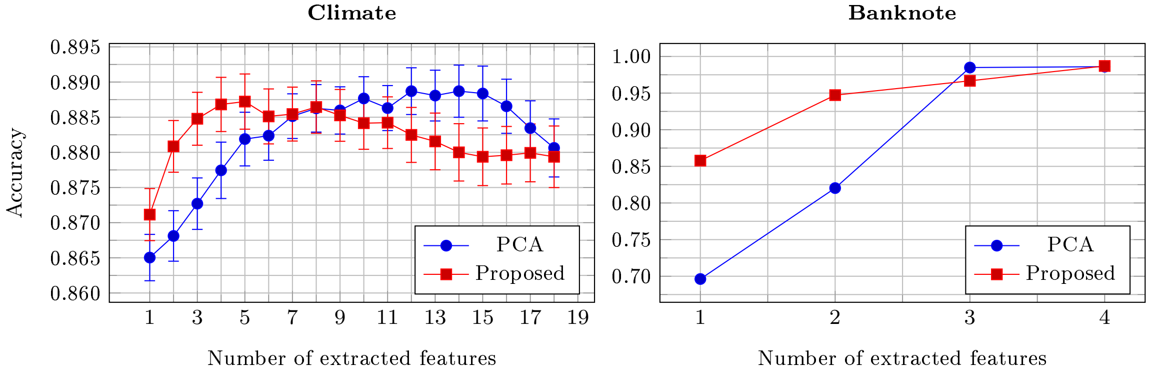

Fig. 3. Accuracy per number of extracted features for Climate and Banknote datasets using features extracted with PCA and the proposed method using **Decision Tree** Classifier.

[1] http://www.mathworks.com/help/stats/index.html.

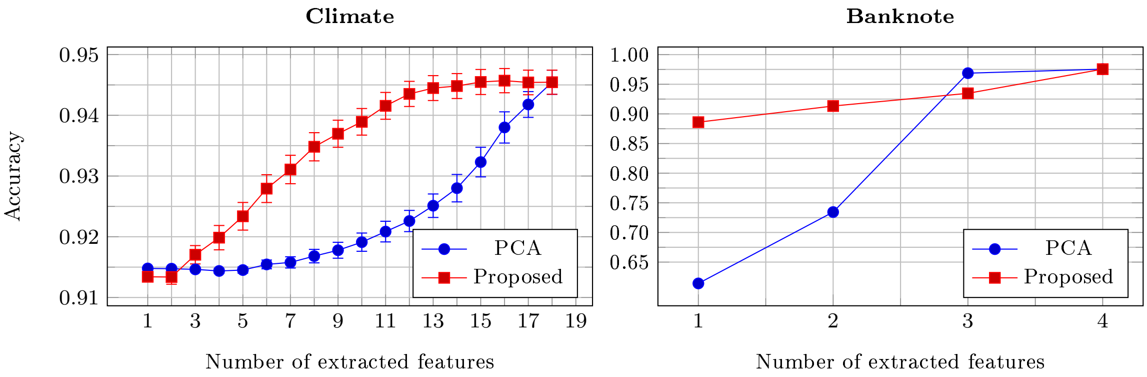

Fig. 4. Accuracy per number of extracted features for Climate and Banknote datasets using features extracted with PCA and the proposed method using **Linear Discriminant** Classifier.

We calculated confidence intervals assuming that each mean follows a Student's t distribution. For a 95% confidence level this interval is $[\bar{a} - E, \bar{a} + E]$, where $\bar{a}$ is the mean accuracy, b is the accuracy standard deviation, and $E = 1.984b/\sqrt{100}$. If there is no overlap between the confidence intervals of PCA and the proposed method the difference is considered significant [Schenker and Gentleman 2001]. The error bars shown for the Climate dataset in the Figures, represent the confidence intervals. For Banknote dataset, the values are too small to appear in the plots. The results for each classifier are discussed in the following subsections.

Analysis for the 1-NN classifier. The results show that the proposed method present accuracy significantly higher than PCA from 2 to 16 extracted features, for the Climate dataset; and for 1 and 2 extracted features for the Banknote dataset. For the Climate dataset (Table I), the maximum mean accuracy obtained using the proposed method was 0.901, for 10 extracted features. PCA presented a smaller mean accuracy 0.873 for the same number of features. The maximum mean accuracy obtained using PCA was 0.891 with 18 extracted features. For the Banknote dataset (Table III), for 1 and 2 extracted features the difference are quite significant. The obtained values were 0.680 (PCA), 0.852 (proposed) and 0.851 (PCA), 0.959 (proposed) respectively.

Analysis for the Naive Bayes classifier. The results show that the proposed method present accuracy significantly higher than PCA from 3 to 11 extracted features, for the Climate dataset; and for 1 and 2 extracted features for the Banknote dataset. For the Climate dataset (Table I), the maximum mean accuracy obtained using the proposed method was 0.922, for 10 extracted features. PCA presented a smaller mean accuracy 0.916 for the same number of features. The maximum mean accuracy obtained using PCA was 0.920 with 17 extracted features. For the Banknote dataset (Table III), for 1 and 2 extracted features the difference are quite significant. The obtained values were 0.695 (PCA), 0.891 (proposed) and 0.773 (PCA), 0.903 (proposed) respectively.

Analysis for the Decision Tree classifier. The results show that the proposed method present accuracy significantly higher than PCA from 2 to 4 extracted features, for the Climate dataset; and for 1 and 2 extracted features for the Banknote dataset. For the Climate dataset (Table II), the maximum mean accuracy obtained using the proposed method was 0.887, for 4 extracted features. PCA presented a smaller mean accuracy 0.877 for the same number of features. The maximum mean accuracy obtained using PCA was 0.889 with 12 extracted features. for 1 and 2 extracted features the difference are quite significant. The obtained values were 0.696 (PCA), 0.858 (proposed) and 0.820 (PCA), 0.947 (proposed) respectively.

Analysis for the Linear Discriminant classifier. The results show that the proposed method present accuracy significantly higher than PCA from 4 to 16 extracted features, for the Climate dataset; and for 1 and 2 extracted features for the Banknote dataset. For the Climate dataset (Table

Table I. The results of the **Climate** database showing the Mean Accuracy (M.A.), Standard Deviation (S.D.) and the number of Extracted Features (E.F.) for each method using 1-NN and Naive Bayes classifiers.

	1-NN		**Naive Bayes**	
	PCA	Proposed	PCA	Proposed
E.F.	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)
1	0.847 (0.021)	0.854 (0.022)	0.915 (0.000)	0.915 (0.003)
2	0.849 (0.021)	0.865 (0.019)	0.915 (0.001)	0.915 (0.005)
3	0.854 (0.020)	0.873 (0.017)	0.915 (0.001)	0.917 (0.006)
4	0.857 (0.024)	0.883 (0.018)	0.915 (0.002)	0.918 (0.007)
5	0.861 (0.023)	0.888 (0.019)	0.915 (0.003)	0.919 (0.008)
6	0.865 (0.021)	0.892 (0.019)	0.915 (0.003)	0.920 (0.008)
7	0.865 (0.023)	0.898 (0.016)	0.916 (0.004)	0.919 (0.008)
8	0.869 (0.024)	0.899 (0.017)	0.916 (0.004)	0.920 (0.009)
9	0.871 (0.023)	0.900 (0.016)	0.916 (0.005)	0.921 (0.010)
10	0.873 (0.022)	0.901 (0.016)	0.916 (0.006)	0.922 (0.009)
11	0.876 (0.021)	0.901 (0.016)	0.917 (0.006)	0.921 (0.009)
12	0.877 (0.021)	0.901 (0.014)	0.918 (0.008)	0.921 (0.009)
13	0.881 (0.021)	0.897 (0.015)	0.919 (0.008)	0.921 (0.009)
14	0.883 (0.020)	0.897 (0.014)	0.919 (0.008)	0.921 (0.009)
15	0.885 (0.017)	0.895 (0.014)	0.919 (0.009)	0.921 (0.010)
16	0.885 (0.018)	0.892 (0.015)	0.919 (0.010)	0.920 (0.010)
17	0.887 (0.016)	0.891 (0.015)	0.920 (0.010)	0.919 (0.010)
18	0.891 (0.014)	0.891 (0.014)	0.919 (0.011)	0.919 (0.011)

II) with 11 extracted features the proposed method have accuracy 0.942, and PCA 0.92. For the Banknote dataset (Table IV), for 1 and 2 extracted features the difference are quite significant. The obtained values were 0.614 (PCA), 0.886 (proposed) and 0.734 (PCA), 0.913 (proposed) respectively.

Classifier independent analysis. The proposed method presents greater accuracy for fewer extracted features if compared to PCA. For the Climate dataset and Decision Tree classifier, the maximum accuracy is achieve using only 5 of 18 features, using other classifiers more than 10 features are needed. For the Banknote dataset, if the accuracy for 1 extracted feature is about 0.9 the accuracy is similar for 2 extracted features (Naive Bayes and Linear Discriminant). If the accuracy for 1 extracted feature is about 0.85 the accuracy is about 0.95 for 2 extracted features (1-NN and Decision Tree).

6. CONCLUSION

We proposed a feature extraction technique that is similar to PCA but selects features that minimizes the Bayes error rate instead of features that maximizes the variance. The method presented a higher mean accuracy compared to PCA in two datasets using a small number of features. For future work, we will evaluate more databases, extend the proposed method to problems with more than two classes and to test in other PCA-based techniques [de Carvalho et al. 2015], [de Carvalho et al. 2014].

REFERENCES

Alencar, A. S. C., Gomes, J. P. P., Souza, A. H., Freire, L. A. M., Silva, J. W. F., Andrade, R. M. C., and Castro, M. F. Regularized supervised distance preserving projections for short-text classification. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on.* pp. 216–221, 2014.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. Prediction by supervised principal components. *Journal of the American Statistical Association* vol. 101, pp. 119–137, 2006.

Barshan, E., Ghodsi, A., Azimifar, Z., and Jahromi, M. Z. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition* 44 (7): 1357 – 1371, 2011.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Table II. The results of the **Climate** database showing the Mean Accuracy (M.A.), Standard Deviation (S.D.) and the number of Extracted Features (E.F.) for each method using 1-NN and Naive Bayes classifiers.

	Decision Tree		**Linear Discriminant**	
	PCA	Proposed	PCA	Proposed
E.F.	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)
1	0.865 (0.0192)	0.871 (0.0214)	0.915 (0.000)	0.913 (0.004)
2	0.868 (0.0208)	0.881 (0.0210)	0.915 (0.001)	0.913 (0.007)
3	0.873 (0.0211)	0.885 (0.0214)	0.915 (0.002)	0.917 (0.008)
4	0.877 (0.0230)	0.887 (0.0219)	0.914 (0.002)	0.920 (0.011)
5	0.882 (0.0218)	0.887 (0.0223)	0.915 (0.003)	0.923 (0.012)
6	0.882 (0.0199)	0.885 (0.0223)	0.915 (0.004)	0.928 (0.012)
7	0.885 (0.0181)	0.885 (0.0218)	0.916 (0.005)	0.931 (0.013)
8	0.886 (0.0192)	0.886 (0.0212)	0.917 (0.006)	0.935 (0.013)
9	0.886 (0.0191)	0.885 (0.0210)	0.918 (0.007)	0.937 (0.012)
10	0.888 (0.0176)	0.884 (0.0211)	0.919 (0.008)	0.939 (0.012)
11	0.886 (0.0182)	0.884 (0.0209)	0.921 (0.009)	0.942 (0.012)
12	0.889 (0.0189)	0.882 (0.0223)	0.923 (0.010)	0.944 (0.011)
13	0.888 (0.0205)	0.882 (0.0230)	0.925 (0.011)	0.944 (0.011)
14	0.889 (0.0210)	0.880 (0.0233)	0.928 (0.012)	0.945 (0.011)
15	0.888 (0.0221)	0.879 (0.0235)	0.932 (0.013)	0.945 (0.011)
16	0.887 (0.0219)	0.880 (0.0235)	0.938 (0.014)	0.946 (0.011)
17	0.883 (0.0222)	0.880 (0.0236)	0.942 (0.011)	0.945 (0.011)
18	0.881 (0.0236)	0.879 (0.0251)	0.945 (0.011)	0.945 (0.011)

Table III. The results of the **Banknote** database showing the Mean Accuracy (M.A.), Standard Deviation (S.D.) and the number of Extracted Features (E.F.) for each method.

	1-NN		**Naive Bayes**	
	PCA	Proposed	PCA	Proposed
E.F.	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)
1	0.680 (0.015)	0.853 (0.014)	0.695 (0.015)	0.891 (0.013)
2	0.852 (0.012)	0.959 (0.008)	0.773 (0.013)	0.903 (0.011)
3	0.997 (0.002)	0.982 (0.013)	0.974 (0.006)	0.929 (0.036)
4	0.999 (0.001)	0.999 (0.001)	0.975 (0.006)	0.975 (0.006)

Table IV. The results of the **Banknote** database showing the Mean Accuracy (M.A.), Standard Deviation (S.D.) and the number of Extracted Features (E.F.) for each method.

	Decision Tree		**Linear Discriminant**	
	PCA	Proposed	PCA	Proposed
E.F.	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)	M.A. (S.D.)
1	0.696 (0.017)	0.858 (0.017)	0.614 (0.011)	0.886 (0.012)
2	0.820 (0.017)	0.947 (0.011)	0.734 (0.011)	0.913 (0.009)
3	0.985 (0.006)	0.967 (0.019)	0.969 (0.006)	0.935 (0.029)
4	0.986 (0.006)	0.987 (0.007)	0.975 (0.005)	0.975 (0.005)

de Carvalho, T. B. A., Costa, A. M., Sibaldo, M. A. A., Tsang, I. R., and Cavalcanti, G. D. C. Supervised fractional eigenfaces. In *Image Processing (ICIP), 2015 IEEE International Conference on.* pp. 552–555, 2015.

de Carvalho, T. B. A., Sibaldo, M. A. A., Tsang, I. R., Cavalcanti, G. D. C., Tsang, I. J., and Sijbers, J. Fractional eigenfaces. In *2014 IEEE International Conference on Image Processing (ICIP).* pp. 258–262, 2014.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification (2Nd Edition).* Wiley-Interscience, 2000.

Lichman, M. UCI machine learning repository, 2013.

Schenker, N. and Gentleman, J. F. Statistical practice: On judging the significance of differences by examining the overlap between confidence intervals. 55 (3): 182–186, Aug., 2001.

Turk, M. and Pentland, A. Eigenfaces for recognition. *J. Cognitive Neuroscience* 3 (1): 71–86, Jan., 1991.