PCA-BASED DATA MINING PROBABILISTIC AND FUZZY APPROACHES WITH APPLICATIONS IN PATTERN RECOGNITION

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

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

Panayiotis Vlamos
Ionian University, Corfu, Greece

Viorica Stefanescu
Dept. of Mathematics, Academy of Economic Studies, Calea Dorobantilor #15-17, Bucuresti –1, Romania

Keywords: Principal component analysis, fuzzy clustering, supervised learning, cluster analysis, pattern recognition, data mining.

Abstract: The aim of the paper is to develop a new learning by examples PCA-based algorithm for extracting skeleton information from data to assure both good recognition performances, and generalization capabilities. Here the generalization capabilities are viewed twofold, on one hand to identify the right class for new samples coming from one of the classes taken into account and, on the other hand, to identify the samples coming from a new class. The classes are represented in the measurement/feature space by continuous repartitions, that is the model is given by the family of density functions $(f_h)_{h\in H}$, where H stands for the finite set of hypothesis (classes). The basis of the learning process is represented by samples of possible different sizes coming from the considered classes. The skeleton of each class is given by the principal components obtained for the corresponding sample.

1 PRINCIPAL COMPONENTS

The starting point for PCA is a n-dimensional random vector $\mathbf{X}$. There is available a sample $\mathbf{X}(1)$, $\mathbf{X}(2)$,… ,$\mathbf{X}(T)$ from this random vector. No explicit assumptions on the probability density of the vectors are made in PCA, as long as the first – order and the second –order statistics are known or can be estimated from the sample. Also, no generative model is assumed for vector $\mathbf{X}$.

In the PCA transform, the vector $\mathbf{X}$ is first centered by subtracting its mean, $\mathbf{X} = \mathbf{X} - E(\mathbf{X})$. In practice, the mean of the n-dimensional vector $\mathbf{X}$ is estimated from the available sample. In the following, we assume that the vector $\mathbf{X}$ is centered. Next, $\mathbf{X}$ is linearly transformed to another vector $\mathbf{Y}$ with m elements, $m<n$, so that the redundancy induced by the correlations is removed. The transform consists in obtaining a rotated orthogonal coordinate system such that the elements of $\mathbf{X}$ in the new coordinates become uncorrelated. At the same time, the variances of the projections of $\mathbf{X}$ on the new coordinates become uncorrelated. At the same time, the variances of the projections of $\mathbf{X}$ on the new coordinate axes are maximized so that the first axis corresponds to the maximal variance, the second axis corresponds to the maximal variance in the direction orthogonal to the first axis, and so on (Hyvarinen, 2001).

State L., Cocianu C., Vlamos P. and Stefanescu V. (2006).
PCA-BASED DATA MINING PROBABILISTIC AND FUZZY APPROACHES WITH APPLICATIONS IN PATTERN RECOGNITION.
In *Proceedings of the First International Conference on Software and Data Technologies*, pages 55-60

In mathematical terms, consider a linear combination of the elements $x_1, x_2, ..., x_n$ of the vector $\mathbf{X}$, $y_1 = \sum_{i=1}^{n} w_{i1} x_i = \mathbf{W}_1^T \mathbf{X}$.

We look for a weight vector $\mathbf{W}_1$ maximizing the PCA criterion,

$$\mathrm{E}\left(y_1^2\right) = \mathrm{E}\left(\left(\mathbf{W}_1^T \mathbf{X}\right)^2\right) = \mathbf{W}_1^T \mathbf{S} \mathbf{W}_1 \quad (1)$$

$\|\mathbf{W}_1\| = 1$,

where the matrix $\mathbf{S}$ is the covariance matrix of $\mathbf{X}$ given for the zero-mean vector $\mathbf{X}$ by the correlation matrix.

It is well known from basic linear algebra that the solution of the PCA problem is given in terms of the unit-length eigen vectors of the matrix $\mathbf{S}$, $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_n$. The ordering of the eigen vectors is such that the corresponding eigen values $\lambda_1, \lambda_2, ..., \lambda_n$ satisfy $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. The solution maximizing $\mathrm{E}\left(y_1^2\right) = \mathbf{W}_1^T \mathbf{S} \mathbf{W}_1$ is given by $\mathbf{W}_1 = \boldsymbol{\varphi}_1$ and the first principal component of $\mathbf{X}$ is $y_1 = \boldsymbol{\varphi}_1^T \mathbf{X}$. The PCA criterion can be generalized to m principal components, $1 \leq m \leq n$. Let $y_m = \sum_{i=1}^{n} w_{im} x_i = \mathbf{W}_m^T \mathbf{X}$ be the *m*-th principal component, with $\mathbf{W}_m$ the corresponding unit norm weight vector. The solution maximizing $\mathrm{E}\left(y_m^2\right) = \mathbf{W}_m^T \mathbf{S} \mathbf{W}_m$ under the constrain $E\left(y_m y_k\right) = 0$, for $1 \leq k < m$ is given by, (Hyvarinen, 2001) $\mathbf{W}_m = \boldsymbol{\varphi}_m$ and the *m*-th principal component of $\mathbf{X}$ is $y_m = \boldsymbol{\varphi}_m^T \mathbf{X}$.

2 CLUSTER ANALYSIS

Cluster analysis is a data processing technique aiming to identify the natural grouping trends existing in a data collection, producing a set of overlapping clusters, the elements belonging to the same cluster sharing similar features. So far, there have been proposed a relatively small number of methods for testing the existence/inexistence of a natural grouping tendency in a data collection, most of them being based on arguments coming from mathematical statistics and heuristic graphical techniques (Panayirci and Dubes,1983, Smith and Jain,1984,Jain and Dubes,1988, Tukey,1977, Everitt,1978).

The data are represented by *p*-dimensional vectors, $X = \left(x_1, ..., x_p\right)^t$, whose components are the feature values of a specified attributes and the classification is performed against a certain given label set. The classification of a data collection $\aleph = \left\{X_1, ..., X_n\right\} \subset \Re^p$ corresponds to a labelling strategy of the objects of $\aleph$.

In the fuzzy approaches, the clusters are represented as fuzzy sets $\left(u_i, 1 \leq i \leq c\right)$, $u_i : \aleph \to [0,1]$, where $u_{ik} = u_i\left(X_k\right)$ is the membership degree of X_k to the *i*-th cluster, $1 \leq i \leq c$, $1 \leq k \leq n$. A c-fuzzy partition is represented by the matrix $U = \|u_{ik}\| \in M_{c \times n}$. The number of labels *c* has to be selected in advance, the problem of finding the optimal *c* is usually referred as cluster validation.

The main types of label vectors are *crisp* N_c, *fuzzy* N_p, *and possibilistic* N_{poz}, defined as follows,

$$N_c = \left\{y \mid y \in \Re^c, y = \left(y_1, y_2, ..., y_c\right), y_i \in \{0,1\}, 1 \leq i \leq c, \sum_{i=1}^{c} y_i = 1\right\} = \left\{e_1, e_2, ..., e_c\right\}, \quad (2)$$

where $\left(e_i\right)_j = \delta_{ij} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}$

$$N_p = \left\{y \in \Re^c \mid y = \left(y_1, y_2, ..., y_c\right), \forall i, y_i \in [0,1], \sum_{i=1}^{c} y_i = 1\right\}, \quad (3)$$

$$N_{poz} = \left\{y \in \Re^c \mid y = \left(y_1, y_2, ..., y_c\right), \forall i, y_i \in [0,1], \exists j, y_j \neq 0\right\}, \quad (4)$$

Obviously, $N_{pos} \supset N_p \supset N_c$. If we denote by $U = [U_1, ..., U_n] = \|u_{ij}\|$ a partition of $\aleph$, then, according to the types of label vectors, we get the *c*-partition types M_{pos}, M_p and M_c,

$$M_{pos} = \left\{U \mid U \in M_{c \times n}, U = [U_1, ..., U_n], \forall k, U_k \in N_{pos}, \forall i, \sum_{k=1}^{n} u_{ik} > 0\right\} \quad (5)$$

$$M_p = \left\{U \middle| U \in M_{pos}, \forall k, U_k \in N_p\right\} \quad (6)$$

$$M_c = \left\{U \middle| U \in M_p, \forall k, U_k \in N_c\right\} \quad (7)$$

Note that $M_c \subset M_p \subset M_{pos}$.

3 *C-MEANS* MODEL

The most popular classification methods are the *c-means* algorithms. The variational problem corresponding to *c-means* model is given by

$$\min_{(U,V)}\left\{J_m(U,V;w)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m D_{ik}^2+\sum_{i=1}^{c}w_i\sum_{k=1}^{n}(1-u_{ik})^m\right\}$$

(8) where

$U\in M_c/M_p/M_{pos}$, $V=(v_1,...,v_c)\in M_{c\times p}$, v_i is the centroid of the *i*-th cluster, $w=(w_1,..,w_c)^T$ is the penalties vector corresponding to the cluster system, $m\geq 1$ is the fuzzyfication degree, and $D_{ik}^2=\|x_k-v_i\|^2$.

Let us denote by $(\hat{U},\hat{V})$ a solution of (8). Then,

1. The crisp model:

$(U,V)\in M_c\times M_{c\times p}; w_i=0, 1\leq i\leq c$,

$$\hat{u}_{ik}=\begin{cases}1, D_{ik}\leq D_{ij}, i\neq j\\ 0, \text{otherwise}\end{cases} \quad (9)$$

$$\hat{v}_i=\frac{\sum_{k=1}^{n}\hat{u}_{ik}x_k}{\sum_{k=1}^{n}\hat{u}_{ik}};\quad 1\leq i\leq c, 1\leq k\leq n \quad (10)$$

2. The fuzzy model:

$(U,V)\in M_p\times M_{c\times p}; m>1, w_i=0, 1\leq i\leq c$

$$\hat{u}_{ik}=\left[\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (11)$$

$$\hat{v}_i=\frac{\sum_{k=1}^{n}u_{ik}^m x_k}{\sum_{k=1}^{n}u_{ik}^m};\quad 1\leq i\leq c, 1\leq k\leq n \quad (12)$$

3. The possibilistic model:

$(U,V)\in M_{pos}\times M_{c\times p}; \forall i, w_i>0$

$$\hat{u}_{ik}=\left[1+\left(\frac{D_{ik}^2}{w_i}\right)^{\frac{1}{m-1}}\right]^{-1}$$

$$\hat{v}_i=\frac{\sum_{k=1}^{n}u_{ik}^m x_k}{\sum_{k=1}^{n}u_{ik}^m};\quad 1\leq i\leq c, 1\leq k\leq n \quad (13)$$

The general scheme of a cluster procedure $\wp$ is,

$t\leftarrow 0$

repeat

$t\leftarrow t+1$

$U_t\leftarrow F_\wp(V_{t-1})$

$V_t\leftarrow G_\wp(U_{t-1})$

until$\left(t=Tor\|V_t-V_{t-1}\|\leq\varepsilon\right)$

$(U,V)\leftarrow(U_t,V_t)$

where *c* is the given number of clusters, *T* is upper limit on the number of iterations, *m* is the weight parameter, $1\leq m<\infty$, *C* is the terminal condition, w is the system of weights $\forall i, w_i>0$, $V_0=(v_{1,0},...,v_{c,0})\in M_{c\times p}$ is the initial system of centroids and $F_\wp$, $G_\wp$ are the updating functions.

4 PCA-BASED ALGORITHM FOR EXTRACTING SKELETON INFORMATION

In the following a new learning by examples PCA-based algorithm for extracting skeleton information from data to assure both good recognition performances, and generalization capabilities, is developed. Here the generalization capabilities are viewed twofold, on one hand to identify the right class for new samples coming from one of the classes taken into account and, on the other hand, to identify the samples coming from a new class. The classes are represented in the measurement/feature space by continuous repartitions, that is the model is given by the family of density functions $(f_h)_{h\in H}$, where *H* stands for the finite set of hypothesis (classes).

The basis of the learning process is represented by samples of possible different sizes coming from the considered classes. The skeleton of each class is given by the principal components obtained for the corresponding sample. The recognition algorithm identifies the class whose skeleton is the "nearest" to the tested example, where the closeness degree is expressed in terms of the amount of disturbance determined by the decision of allotting it to the corresponding class. The model is presented as follows. Let $X_1, X_2,..., X_N$ be a series of *n*-dimensional vectors coming from a certain class *C*. The sample covariance matrix is

$$\Sigma_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu_N)(X_i - \mu_N)^T , \quad (14)$$

where $\mu_N = \frac{1}{N}\sum_{i=1}^{N} X_i$.

We denote by $\lambda_1^N \geq \lambda_2^N \geq ... \geq \lambda_n^N$ the eigen values and by $\psi_1^N,...,\psi_n^N$ the corresponding orthonormal eigen vectors of Σ_N.

If X_{N+1} is a new sample, then, for the series $X_1, X_2,..., X_N, X_{N+1}$, we get

$$\Sigma_{N+1} = \Sigma_N + \frac{1}{N+1}(X_{N+1} - \mu_N)(X_{N+1} - \mu_N)^T - \\ - \frac{1}{N}\Sigma_N \quad (15)$$

Lemma. In case the eigen values of $\mathbf{\Sigma}_N$ are pairwise distinct, the following first order approximations hold,

$$\lambda_i^{N+1} = \lambda_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N = \left(\boldsymbol{\psi}_i^N\right)^T \mathbf{\Sigma}_{N+1} \boldsymbol{\psi}_i^N \quad (16)$$

$$\boldsymbol{\psi}_i^{N+1} = \boldsymbol{\psi}_i^N + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N}{\lambda_i^N - \lambda_j^N} \boldsymbol{\psi}_j^N \quad (17)$$

Proof Using the perturbation theory, we get, $\mathbf{\Sigma}_{N+1} = \mathbf{\Sigma}_N + \Delta\mathbf{\Sigma}_N$ and, $\boldsymbol{\psi}_i^{N+1} = \boldsymbol{\psi}_i^N + \Delta\boldsymbol{\psi}_i^N$, $\lambda_i^{N+1} = \lambda_i^N + \Delta\lambda_i^N$, $1 \leq i \leq n$. Then,

$$\Delta\mathbf{\Sigma}_N = \frac{1}{N+1}(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N)(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N)^T - \frac{1}{N}\mathbf{\Sigma}_N \quad (18)$$

$$\left(\mathbf{\Sigma}_N + \Delta\mathbf{\Sigma}_N\right)\left(\boldsymbol{\psi}_i^N + \Delta\boldsymbol{\psi}_i^N\right) = \\ = \left(\lambda_i^N + \Delta\lambda_i^N\right)\left(\boldsymbol{\psi}_i^N + \Delta\boldsymbol{\psi}_i^N\right) \quad (19)$$

Using first order approximations, from (19) we get,

$$\lambda_i^N \boldsymbol{\psi}_i^N + \mathbf{\Sigma}_N \Delta\boldsymbol{\psi}_i^N + \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \boldsymbol{\psi}_i^N + \lambda_i^N \Delta\boldsymbol{\psi}_i^N + \Delta\lambda_i^N \boldsymbol{\psi}_i^N \quad (20)$$

hence,

$$\left(\boldsymbol{\psi}_i^N\right)^T \mathbf{\Sigma}_N \Delta\boldsymbol{\psi}_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \left(\boldsymbol{\psi}_i^N\right)^T \Delta\boldsymbol{\psi}_i^N + \Delta\lambda_i^N \left\|\boldsymbol{\psi}_i^N\right\|^2 \quad (21)$$

Using $\lambda_i^N \left(\boldsymbol{\psi}_i^N\right)^T = \left(\boldsymbol{\psi}_i^N\right)^T \mathbf{\Sigma}_N$ we obtain ,

$$\lambda_i^N \left(\boldsymbol{\psi}_i^N\right)^T \Delta\boldsymbol{\psi}_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \left(\boldsymbol{\psi}_i^N\right)^T \Delta\boldsymbol{\psi}_i^N + \Delta\lambda_i^N \quad (22)$$

hence $\Delta\lambda_i^N = \left(\boldsymbol{\psi}_i^N\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N$ that is,

$$\lambda_i^{N+1} = \lambda_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N = \left(\boldsymbol{\psi}_i^N\right)^T \mathbf{\Sigma}_{N+1} \boldsymbol{\psi}_i^N \quad (23)$$

The first order approximations of the orthonormal eigen vectors of $\mathbf{\Sigma}_{N+1}$ can be derived using the expansion of each vector $\Delta\boldsymbol{\psi}_i^N$ in the basis represented by the orthonormal eigen vectors of $\mathbf{\Sigma}_N$,

$$\Delta\boldsymbol{\psi}_i^N = \sum_{j=1}^{n} b_{i,j} \boldsymbol{\psi}_j^N , \quad (24)$$

where

$$b_{i,j} = \left(\boldsymbol{\psi}_j^N\right)^T \Delta\boldsymbol{\psi}_i^N . \quad (25)$$

Using the orthonormality, we get,

$$1 = \left\|\psi_i^N + \Delta\psi_i^N\right\|^2 \cong \left\|\psi_i^N\right\|^2 + 2\left(\psi_i^N\right)^T \left(\Delta\psi_i^N\right) = \\ = 1 + 2\left(\psi_i^N\right)^T \left(\Delta\psi_i^N\right), \quad (26)$$

that is

$$b_{i,i} = \left(\boldsymbol{\psi}_i^N\right)^T \Delta\boldsymbol{\psi}_i^N = 0 \quad (27)$$

Using (19), the approximation,

$$\mathbf{\Sigma}_N \Delta\boldsymbol{\psi}_i^N + \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \lambda_i^N \Delta\boldsymbol{\psi}_i^N + \Delta\lambda_i^N \boldsymbol{\psi}_i^N . \quad (28)$$

holds for each $1 \leq i \leq n$.

For $1 \leq j \neq i \leq n$, from (28) we obtain the following equations,

$$\left(\boldsymbol{\psi}_N^j\right)^T \mathbf{\Sigma}_N \Delta\boldsymbol{\psi}_i^N + \left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \left(\boldsymbol{\psi}_N^j\right)^T \Delta\boldsymbol{\psi}_i^N + \Delta\lambda_i^N \left(\boldsymbol{\psi}_N^j\right)^T \boldsymbol{\psi}_i^N \quad (29)$$

$$\left(\boldsymbol{\psi}_N^j\right)^T \mathbf{\Sigma}_N \Delta\boldsymbol{\psi}_i^N + \left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \left(\boldsymbol{\psi}_N^j\right)^T \Delta\boldsymbol{\psi}_i^N \quad (30)$$

$$\lambda_j^N \left(\boldsymbol{\psi}_N^j\right)^T \Delta\boldsymbol{\psi}_i^N + \left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \cong \\ \cong \lambda_i^N \left(\boldsymbol{\psi}_N^j\right)^T \Delta\boldsymbol{\psi}_i^N \quad (31)$$

From (31) we get,

$$\left(\lambda_i^N - \lambda_j^N\right)\left(\boldsymbol{\psi}_N^j\right)^T \Delta\boldsymbol{\psi}_i^N = \left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N \quad (32)$$

$$b_{i,j} = \left(\boldsymbol{\psi}_j^N\right)^T \Delta\boldsymbol{\psi}_i^N = \frac{\left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N}{\lambda_i^N - \lambda_j^N} \quad (33)$$

Consequently, the first order approximation of the eigen vectors of $\mathbf{\Sigma}_{N+1}$ are,

$$\boldsymbol{\psi}_i^N + \Delta\boldsymbol{\psi}_i^N = \boldsymbol{\psi}_i^N + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left(\boldsymbol{\psi}_N^j\right)^T \Delta\mathbf{\Sigma}_N \boldsymbol{\psi}_i^N}{\lambda_i^N - \lambda_j^N} \boldsymbol{\psi}_j^N \quad (34)$$

The skeleton of C is represented by the set of estimated principal components $\psi_1^N,...,\psi_n^N$ .When the example X_{N+1} is included in C, then the new skeleton is $\psi_1^{N+1},...,\psi_n^{N+1}$. The skeleton disturbance induced by the decision that X_{N+1} has to be alloted to C is measured by

$$D=\frac{1}{n}\sum_{k=1}^{n}d\left(\psi_k^N,\psi_k^{N+1}\right) \quad (35)$$

The crisp classification procedure identifies for each example the closest cluster in terms of the measure (35). Let $H=\{C_1,C_2,...,C_M\}$. In order to protect against misclassifications of samples coming from new classes not belonging to *H*, a threshold *T*>0 is imposed, that is the example X_{N+1} is alloted to one of C_j for which

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

$$=\min_{1\le p\le M}\frac{1}{n}\sum_{k=1}^{n}d\left(\psi_{k,p}^N,\psi_{k,p}^{N+1}\right) \quad (36)$$

and *D*<*T*, where the skeleton of C_j is $\psi_{1,j}^N,...,\psi_{n,j}^N$.

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*.

In case of fuzzy classification, the value of the membership degree of X_{N+1} to each cluster of $HT=H\cup\{CR\}$ is computed as follows. Let

$$d\left(X_{N+1},C_i\right)=\frac{1}{n}\sum_{k=1}^{n}d\left(\psi_{k,i}^N,\psi_{k,i}^{N+1}\right),\quad 1\le i\le M \quad \text{and}$$

$$d\left(X_{N+1},CR\right)=\begin{cases}\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}d\left(\psi_{k,R}^N,\psi_{k,R}^{N+1}\right), \text{ if } CR\neq\varnothing\\ \infty, \text{ otherwise}\end{cases},$$

where the skeleton of *CR* is represented by the set of estimated principal components $\psi_{1,R}^N,...,\psi_{n,R}^N$.

$$\mu_{C_i}\left(X_{N+1}\right)=1-\frac{d\left(X_{N+1},C_i\right)}{S},\ 1\le i\le M \quad (37)$$

$$\mu_{CR}\left(X_{N+1}\right)=\begin{cases}1-\dfrac{d\left(X_{N+1},CR\right)}{S}, \text{ if } d\left(X_{N+1},CR\right)\neq\infty\\ 0, \text{ otherwise}\end{cases} \quad (38),$$

where

$$S=\begin{cases}\displaystyle\sum_{C\in HT}d\left(X_{N+1},C\right), \text{ if } d\left(X_{N+1},CR\right)\neq\infty\\ \displaystyle\sum_{C\in H}d\left(X_{N+1},C\right), \text{ if } d\left(X_{N+1},CR\right)=\infty\end{cases} \quad (39)$$

5 EXPERIMENTAL RESULTS AND CONCLUDING REMARKS

Several tests were performed on simulated data and they pointed out very successful performance of the proposed classification strategy.

A series of tests were performed on 4-dimensional simulated data coming from 5 clases each of them having 50 examples. The classes are represented by normal repartitions $C_i\sim\left(\mu_i,\Sigma_i\right)$, $1\le i\le 5$, where

μ_1 = [10 11 2 -12]

$$\Sigma_1=\begin{bmatrix}3.5944 & 2.0100 & 1.4720 & 0.6460\\ 2.0100 & 3.7500 & 3.1000 & 1.5350\\ 1.4720 & 3.1000 & 3.0600 & 0.8010\\ 0.6460 & 1.5350 & 0.8010 & 1.8369\end{bmatrix}$$

μ_2 = [12 -5 8 13]

$$\Sigma_2=\begin{bmatrix}1.5566 & 0.7755 & 0.5230 & 0.3745\\ 0.7755 & 1.7766 & 0.9305 & 0.6720\\ 0.5230 & 0.9305 & 2.4586 & 1.0050\\ 0.3745 & 0.6720 & 1.0050 & 2.3961\end{bmatrix}$$

μ_3 = [-10 0 9 11]

$$\Sigma_3=\begin{bmatrix}1.7300 & 1.3740 & 0.0200 & 0.6000\\ 1.3740 & 2.5144 & 0.0120 & 0.4600\\ 0.0200 & 0.0120 & 3.0725 & 0.3250\\ 0.6000 & 0.4600 & 0.3250 & 2.3144\end{bmatrix}$$

μ_4 = [-3 14 3 -11.5]

$$\Sigma_4=\begin{bmatrix}2.3618 & 0.5825 & 0.3814 & 0.9805\\ 0.5825 & 2.4038 & 0.9920 & 0.6029\\ 0.3814 & 0.9920 & 2.3724 & 0.4250\\ 0.9805 & 0.6029 & 0.4250 & 4.1054\end{bmatrix}$$

μ_5 = [-7 -10.5 -14 11.5]

$$\Sigma_5=\begin{bmatrix}1.8017 & 0.2860 & 0.6330 & 0.6540\\ 0.2860 & 2.4436 & 0.0240 & 0.0210\\ 0.6330 & 0.0240 & 1.6301 & 0.3454\\ 0.6540 & 0.0210 & 0.3454 & 2.9441\end{bmatrix}$$

The classification criterion is: allote X_{N+1} to C_{j_i} if

$$D=\min_{1\le l\le t}\frac{1}{m_{j_l}}\sum_{k=1}^{m_{j_l}}d\left(\psi_{j_l}^k,\psi_{j_l,N+1}^k\right) \quad (37)$$

In order to evaluate the generalization capacities, 100 new examples were generated for each distribution. The results are presented in Table 1.

Table 1: Results on new simulated examples.

Class	C_1	C_2	C_3	C_4	C_5
Number of correct classified examples	100	100	96	99	100
Number of missclassified examples	0	0	4 - allotted to C_2	1 – allotted to C_1	0
The mean value of D in case of correct classifications	0.08	0.05	0.75	0.21	0.14
The maximum value of D in case of correct classified examples	0.41	0.19	1.85	0.55	0.53

The evaluation of the generalization capacities in case of examples coming from new classes was performed on 1000 samples generated from $N(\mu, \Sigma)$, where $\mu = [0\ 11\ -9\ -9.5]$ and

$$\Sigma = \begin{bmatrix} 8.2725 & 3.1080 & 1.7925 & 1.3680 \\ 3.1080 & 6.8986 & 1.8390 & 2.5561 \\ 1.7925 & 1.8390 & 6.0422 & 1.6410 \\ 1.3680 & 2.5561 & 1.6410 & 5.2261 \end{bmatrix}.$$

The admissibility criterion for allotting a sample to a certain class is given by the maximum value of D corresponding to correct classifications. The results showed that about 975 examples were classified in CR, that is the algorithm managed to detect the intruded examples.

REFERENCES

Al Sultan K.S., Selim,S.Z., 1993. Global Algorithm for Fuzzy Clustering Problem, *Patt.Recogn*. 26,1375-1361

Bensaid,A., Hall,L.O.,Bezdek,J.C., Clarke,L.P., 1996 Partially supervised clustering for image segmentation, *Patt.Recog.*,29(5),859-871.

Bezdek,J.C., Keller,J., Krisnapuram,R., Pal,N.K., 2005. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing,* Springer Verlag

Bezdek, J.C., 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press

Bensaid, H., Bezdek,J.C., Clarke, L.P., Silbiger, M.L., Arrington, J.A., Murtagh, R.F., 1996. Validity-Guided (Re)clustering with applications to image segmentation, *IEEE Trans. Fuzzy Systems,* 4, 112-123

Clark,M., Hall,L.O., Goldgof,D.B., Clarke,L.P., Velthuizen,R.P., Silbiger,M.S., 1994. MRI Segmentation using Fuzzy Clustering Techniques, *IEEE Engineering in Medicine and Biology*, 1994

Everitt, B. S., 1978. *Graphical Techniques for Multivariate Data*, North Holland, NY

Gath,J., Geva,A.B., 1989. Unsupervised optimal fuzzy clustering, *IEEE Trans. Pattern.Anal.Machine Intell*, 11, 773-781

A. Hyvarinen, J. Karhunen, E. Oja, 2001. *Independent Component Analysis*, John Wiley &Sons

Huang,C., Shi,Y. 2002. *Towards Efficient Fuzzy Information Processing*, Physica-Verlag, Heidelberg

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

Jin,Y., 2003. *Advanced Fuzzy Systems Design and Applications*, Physica-Verlag, Heidelberg

Krishnapuram, R., Keller,J.M., 1993. A possibilistic approach to clustering, *IEEE Trans. Fuzzy Syst.,* 1(2)

Li,C., Goldgof,D.B., Hall, L.O., 1993. Automatic segmentation and tissue labelling of MR brain images, *IEEE Transactions on Medical Imaging*, 12(4),1033-1048

Pal,N.,R., Bezdek,J.C., 1995. On Cluster validity for the Fuzzy c-Means Model, *IEEE Trans. On Fuzzy Syst.* ,Vol. 3,no.3

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

Wu,K-L., Yang,M-S, 2005. A Cluster validity index for fuzzy clustrering, *Patt.Recog.Lett.* 26, 1275-1291

Zahid,N., Abouelala,O., Limouri,M., Essaid,A., 1999. Unsupervised fuzzy clustering, *Patt.Recog.Lett*., 20,1