Visualization of Some Multi-Class Erosion Data Using GDA and Supervised SOM

Anna Bartkowiak[1], Niki Evelpidou[2]

[1] Institute of Computer Science, University of Wrocław,
Przesmyckiego 20, 51-151 Wroclaw Poland, aba@ii.uni.wroc.pl
[2] Remote Sensing Laboratory, University of Athens
Panepistimiou Zoografou, Athens, Greece, evelpidou@geol.uoa.gr

Abstract. We present our experience in visualization multivariate data when the data vectors have class assignment. The goal is then to visualize the data in such a way that data vectors belonging to different classes (subgroups) appear differentiated as much as possible. We consider for this purpose the traditional CDA (Canonical Discriminant Functions), the GDA (Generalized Discriminant Analysis, Baudat and Anouar, 2000) and the Supervised SOM (Kohonen, Makivasara, Saramaki 1984). The methods are applied to a set of 3-dimensional erosion data containing N=3420 data vectors subdivided into 5 classes of erosion risk. By performing the mapping of these data to a plane, we hope to gain some experience how the mentioned methods work in practice and what kind of visualization is obtained. The final conclusion is that the traditional CDA is the best both in speed (time) of the calculations and in the ability of generalization.

1 Introduction

We consider the problem of multivariate data visualization when each data vector has a class (group) assignment. Generally, methods of data visualization perform linear or nonlinear mapping to a manifold of lower dimension. Say, this lower dimension is q. The most common visualization uses $q = 2$. Generally, it is expected that the projection gives us an idea on the shape of the data cloud. Here, we want more: Using the information about crisp group assignment ('crisp' is used here in the opposite meaning of 'soft'), we seek for such a projection (mapping), which shows distinctly differentiation between various groups of the data.

When intending a graphical visualization of the data, we should ask in first step about the intrinsic dimensionality of the data. It could happen that all the observed variables are generated by some unobserved variables, so called 'latent variables' located in a manifold of lower dimension – and we should know it. Thus, we should ask about the intrinsic dimensionality of the analyzed data. We will use for this purpose the correlation integral C(r) and the correlation dimension D introduced by Grassberger and Procaccia [6], [4].

The next question is: What kind of projection or mapping should we use? The most simple method is the classical one using Fishers' criterion based on the between and

within class variance and yielding so called 'canonical discriminant variates' or canonical discriminant functions [2], [5], [7], [12]. The method belongs to the class of linear methods and is referred to as CDA or Fisherian LDA. The method is extendable to the class of nonlinear methods – by use of appropriate transformation of the data. In particular one may use kernel transformations [8], [9], [5], [11].
Using the kernel approach, Baudat and Anouar [2] proposed a non trivial generalization of the canonical discriminant analysis. They called their algorithm GDA (generalized discriminant analysis). It represents the nonlinear discriminant functions.
As an alternative to the nonlinear GDA we will consider also quite a different algorithm, called SOM supervised (SOM_s) and based on a modification of Kohonens' self organizing map .
In the following, we will show how the mentioned methods work when analyzing a real data set of a considerable size, i.e. about 3 thousands of data vectors. The data set is subdivided into 5 erosion classes. We take for our illustration only 3 variables known as predictors for the erosion risk. In the case of 3 variables it is possible to visualize the data in a 3D plot. For the considered erosion data, the 3D plot shows plainly that the relations between the variables are highly nonlinear; thus nonlinear projection methods might show a more distinctive differentiation among the erosion classes.
The paper is organized as follows. In Sect. 2 we describe the data and their correlation dimension. Sect. 3 explains the accepted Fishers' criterion of separation between classes and the principles of building canonical discriminant functions (CDA alias LDA). Sect. 4 shows the nonlinear extension of LDA using the kernel approach proposed by Baudat and Anouar. In Sect. 5 we describe briefly the supervised SOM. Finally, Sect. 6 contains some concluding remarks.

2 The Erosion Data

Our interest in a trustful visualisation of subgroups of data originated from the research of erosion risk observed in the Greek island Kefallinia. The entire island was covered by a grid containing 3422 cells. The area covered by each cell of the grid was characterized by several variables. For our purpose, to illustrate some visualization concepts, we will consider in the following only 3 variables: drainage density, slope and vulnerability of the soil (rocks). The values of the variables were rescaled to belong to the interval [0, 1]. Thus, for our analysis, we got a data set containing N=3422 data vectors, each vector characterized by 3 variables. Using an expert GIS system, each data vector was assigned to one of five erosion classes: 1, very high (vH), 2. high (H), 3. medium (Me), 4. low (L) and 5. very low (vL). A 3D plot of the data is shown in Fig. 1. The data set contains a few outliers, which are strongly atypical observations. Two of them will be removed from further analysis.

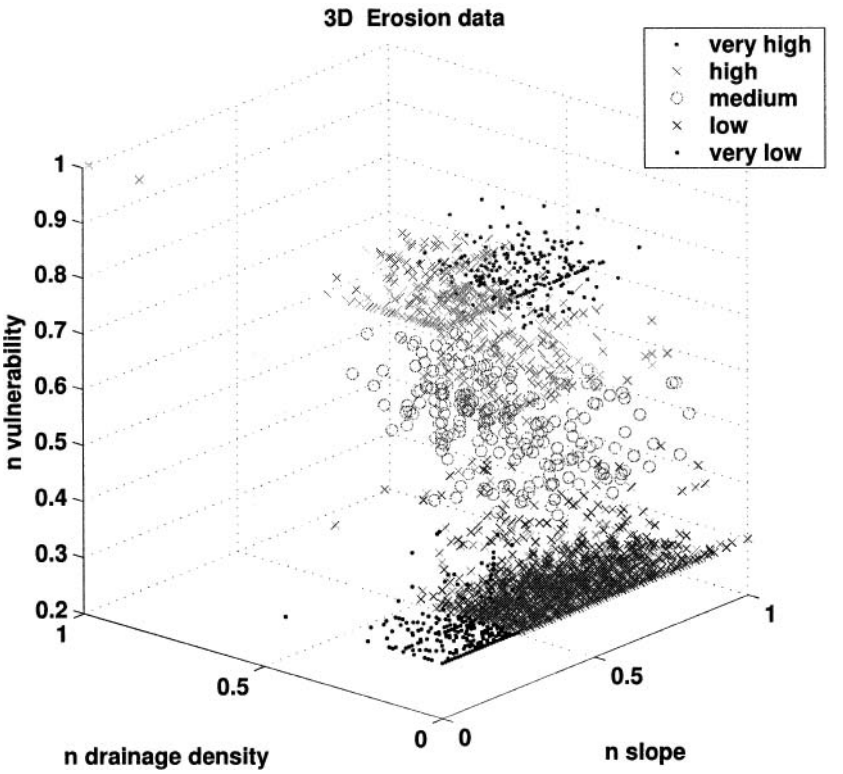

Fig.1. Visualization of the Kefallinia erosion data containing N=3422 data points, subdivided into 5 classes of progressing erosion risk. In some parts of the space the data points are much condensed. Two severe outliers are visible top left – they will be dropped in further analysis

The different classes of the data set are marked by different symbols and/or colours. Looking at the plot in Fig. 1 one may state that, generally, the distribution of the data is far from normality, also far from the ellipsoidal shape. The hierarchy of the classes exhibits a nonlinear pattern. Some parts of the space show a great concentration of the data points, while some other parts are sparsely populated.

The fractal correlation dimension calculated using the Grassberger-Proccacia index [6], [4] equals D = 1.6039. This is the estimate of the intrinsic dimension for the considered erosion data (For comparison, we have performed analogous calculations for two synthetic data sets of the same size, generated from the 3D and 5D normal distributions; as expected, we obtained the values $D_3 = 3.0971$ and $D_5 = 4.9781$ appropriately). Thus – the intrinsic dimension of the considered data set is less then 2 and a planar representation of the data is justified.

The data set contained two big outliers. To not confound the effects of the outliers and the effects of the methods, we have removed the outliers from the analyzed set. Next we subdivided the remaining 3400 data vectors into two parts (halves), each

counting N = 1710 data items. The first part (labelled samp1) was destined for learning (establishing the parameters of the models), and the second part (samp2) as test. In the next three sections we will show mapping of the data to a 2D plane using three different methods: canonical discriminant functions (CDA alias LDA), kernel discriminant functions (GDA) and the supervised SOM (SOM_s).

3 Canonical Discriminant Functions

We show now the canonical discriminant functions derived from Fishers' criterion. The method is called sometimes also LDA [5], [2].
The case of ***the two-class problem.*** R. A. Fisher proposed to seek for the linear combination (**a**) of the variables, which separates the two indicated classes as much as possible. The criterion of separateness, proposed by Fisher, is the ratio of between-class to within-class variances. Formally, the criterion is defined as the ratio (see, e.g. Duda [6] or Webb [11])

$$JF2 = [aT(m1-m2)]2 / [aTSw\ a], \quad \text{(2-class problem)}$$

where **a** is the sought linear form, $\mathbf{m}_1$ and $\mathbf{m}_2$ denote the sample group means, and $\mathbf{S}_w$ is the pooled within-class sample covariance matrix, in its bias-corrected form given by

$$\mathbf{S}_w = (n_1\mathbf{S}_1+n_2\mathbf{S}_2) / (n_1+n_2-2).$$

Maximizing the J_{F2} criterion yields as solution the sought linear combination **a** for the two-class problem.
In the case of ***the multi-class problem***, – when we have k classes, k≥2, with sample sizes $n_1, \ldots, n_k$ totaling N, and the overall mean $\mathbf{m}_*$ – the criterion J_{F2} is rewritten as the criterion J_{Fk}, which accommodates the between class and within class variances:

$$J_{Fk} = \sum_j n_j \mathbf{a}^T(\mathbf{m}_j-\mathbf{m}_*)^2 / [\mathbf{a}^T\mathbf{S}_w\ \mathbf{a}],\ j=1,\ldots k \quad (k\text{-class problem})$$

where $\mathbf{m}_j$ denotes the mean of the j-th class and $\mathbf{m}_*$ stands for the overall sample mean. The within class variance $\mathbf{S}_w$ is evaluated as ($\mathbf{S}_j$ denotes the covariance matrix in the jth class, j=1, … ,k):

$$\mathbf{S}_w = (\sum_j n_j \mathbf{S}_j)/(N-k).$$

Maximizing the criterion J_{Fk} with respect to **a** we obtain, with accuracy to the sign, h solutions, i.e. h vectors $\mathbf{a}_1, \ldots, \mathbf{a}_h$, h = min (k-1, rank of **X**), with **X** being the data matrix. From these we obtain h canonical variates: $y_j = \mathbf{X}\mathbf{a}_j$, j = 1, … , h, called also ***canonical discriminant functions***. The separateness of the subgroups, attained when considering the transformation yielded by subsequent canonical discriminant variates, is measured by the criterion J_{Fk} evaluated as $J_{Fk}(\mathbf{a}_j)$ and called also ***lambda*** $_j$. For each

vector $\mathbf{a}_j$ we obtain its corresponding value $lambda_j = lambda(\mathbf{a}_j)$ denoting the ratio of the between to the within class variance of the respective canonical variate derived from the vector $\mathbf{a}_j$. Thus a big value of $lambda_j$ indicates a high discriminative power of the derived canonical variate.

For the analyzed erosion data we got h=3 vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ and corresponding to them 3 values of *lambda* equal to [22.1995 0.7982 0.0003]. One may notice that the first canonical variate – compared to the remaining ones - has a very big discriminative power, while the contribution of the third canonical variate is practically none.

The projection of the data, when using the first two canonical variates, is shown in Fig. 2. One may notice that the subgroups are quite well separated. One may notice also that the second canonical variate, which – taken alone – has practically no discriminative power, however, when combined with the first variate, helps much in the display, i.e. in distinguishing the classes of erosion risk. We got very similar values of *lambda* and very similar displays both for the learning and the testing data sets (i.e. for samp1 and samp2) – thus the method has good generalization abilities.

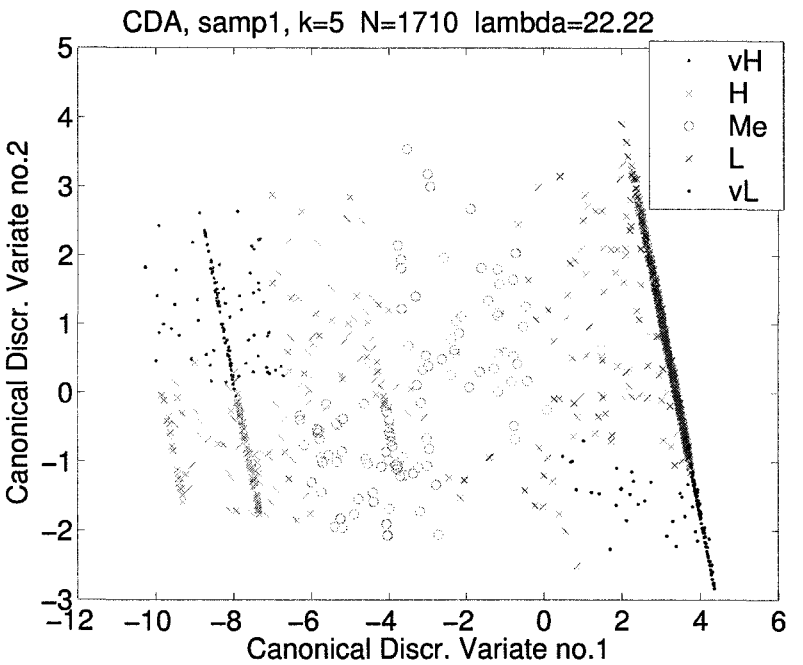

Fig.2. Projection of the samp1 data using the first two canonical discriminant functions derived from Fisher's criterion. The very low and very high erosion points keep opposite position, right and left, in the exhibit. The display for the samp2 data looks identical

4 Nonlinear Projection Using the Kernel Approach

The CDA, described in previous section, considers only linear functions of the variables and is proper when the groups (classes) are distributed elliptically. For our data this is not the case. Therefore, some nonlinear methods might be better for visualizing the class differentiation. A kind of non-linear discriminant analysis, called GDA (***Generalized Discriminant Analysis***) was proposed by Baudat and Anouar [2]. Their algorithm maps the input space into an extended high dimensional feature space. In the extended space, one can solve the original nonlinear problem in a classical way, e.g., using the CDA. Speaking in other words, the main idea is to map the input space into a convenient feature space in which variables are nonlinearly related to the input space. The fact of mapping original data in a nonlinear way into an extended feature space was met in the context of support vector machines (SVM) see e.g., [5], [8], [9], [11]. The mapping uses predominantly kernel functions. Direct coordinates – in the extended space – are not necessary, because the kernel approach needs only computations of so called 'dot products' formed from the original features.
Generally, the mapping reads

$$\Phi: X \rightarrow F,$$

with X denoting the input space (original data), and F the extended feature space, usually of higher dimensionality as the original data space. The mapping Φ transforms elements $\mathbf{x} \in X$ from the original data space into elements $\Phi(\mathbf{x}) \in F$, i.e. elements of the feature space.
Statistical and/or pattern recognition problems use extensively cross products (inner products), e.g. for obtaining the within and between group covariance. To calculate them, a special notation of kernel products was invented.
Let $\mathbf{x}_i$ and $\mathbf{x}_j$ denote two elements (row data vectors) of the input data matrix $\boldsymbol{X}$. The ***kernel function*** $k(\mathbf{x}_i,\mathbf{x}_j)$ returns the inner product $\Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_j)$ between the images of these inputs (located in the feature space). It was proved that for kernel functions satisfying some general analytical conditions (possessing so called Mercer properties) the kernel functions $k(\mathbf{x}_i,\mathbf{x}_j)$ can be expressed as simple functions of the inner product $<\mathbf{x}_i, \mathbf{x}_j>$ of the original vectors. In such a case, we can compute the inner product between the projections of two points into the feature space *without evaluating explicitly* their coordinates (N denotes the number of data vectors, i.e. the number of rows in the data matrix $\mathbf{X}$):

$$k(\mathbf{x}_i,\mathbf{x}_j) = \Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_j) = k(<\mathbf{x}_i, \mathbf{x}_j>), \quad \text{for } i,j = 1, \ldots, N.$$

The GDA algorithm operates on the kernel dot product matrix $\mathbf{K} = \{k(<\mathbf{x}_i, \mathbf{x}_j>\}$ of size N x N, evaluated from the learning data set. The most commonly used kernels are Gaussians (RBFs) and polynomials.
Let $\boldsymbol{x}$, $\boldsymbol{y}$ be the two (row) input vectors. Let $\boldsymbol{d} = \boldsymbol{x}-\boldsymbol{y}$. Using Gaussian kernels, the element $z = k(<\mathbf{x}_i, \mathbf{x}_j>)$ is evaluated as: $z = \exp\{-(\boldsymbol{d}*\boldsymbol{d}^T) / \sigma\}$. The constant σ, called kernel width, is a parameter of the model; its value has to be declared by the user.

Baudat and Anouar use as the index of separateness of the constructed projection a criterion, which they call **inertia**. This criterion is defined as the ratio of the between class to the total variance of the constructed discriminant variates. The inertia criterion takes values from the interval [0, 1]. High values of inertia indicate a good separation of the displayed classes.

For our evaluation we have used Matlab software implemented by Baudat and Anouar. For k = 5 classes we got 4 discriminative variates. The largest values of inertia were noted, as expected, for the first two GDA variates. What concerns the kernel width σ, we have tried several values: σ = 0.0005, 0.005, 0.05, 0.5, 1, 4, 6.5, 9, 14. For each value of σ, the system has been learning using the samp1 data, next the established model was tested using the samp2 data. Each run (i.e. calculations for one value of σ) needed about 12 minutes of computer time (PC, XPHome, Intel® Pentium® 4, Mobile CPU 1.80GHz, 512 MB RAM). The samp1 and samp2 data were of size [1710, 3]. Thus the computations were quite lengthy.

Generally, it was stated that for decreasing values of σ the classes appeared more and more separated (for values σ = 0.5 to 14, the displays were quite similar). As an exemplary exhibit we show here Fig. 3, obtained for σ = 1. The resulting inertias for variates no. 1-4 are: [0.968650 0.705236 0.550547 0.257248].

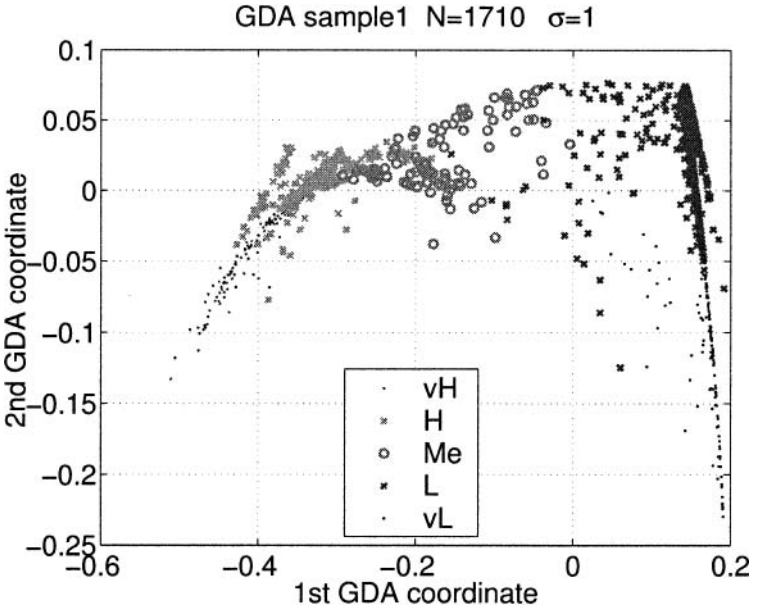

Fig. 3. GDA using Gaussian kernels with σ = 1 applied to the samp1 data. Horizontal and vertical axes denote first and second GDA coordinates. Five classes of data points corresponding to decreasing erosion risk – appearing from left (very high risk) to right (very low risk) – are marked differently. Generally, the topology of the subgroups is preserved and the groups appear fairly separated and condensed

The overall pattern of the point configuration in Fig. 3 is the same as in Fig. 2. From left to right we see groups of points corresponding to areas with very high (vH), high

(H), medium (Me), low (L), and very low (vL) erosion risk. Generally, the topology of the subgroups is preserved. Both variates contribute significantly to the differentiation of the risk classes. Unfortunately, the model when applied to the test set, yields projections appearing in quite different areas; thus it is not able to make the generalization.

5 Supervised SOM

Kohonen's self-organizing maps are a popular tool for visualization of multivariate data. The method was successfully applied to the Kefallinia erosion data [1].
The SOM method uses a general purpose methodology without accounting specially for the additional information on class membership of the data points. However, after constructing the map, we may indicate by so called 'hits', what is the distribution (location) of the different classes. Map with hits of the classes is shown in Fig. 4 below.

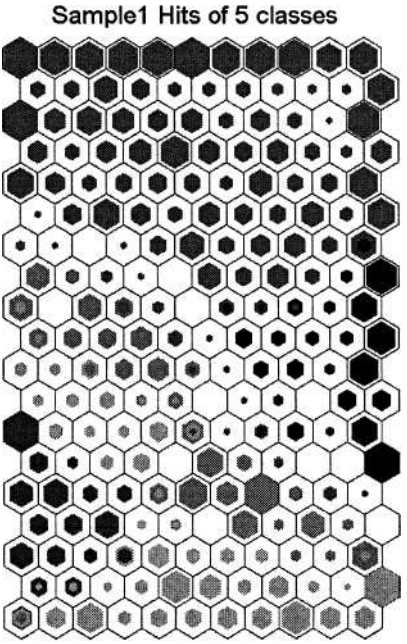

Fig.4. Ordinary self-organizing map SOM of size 19 x 11 constructed from the samp1 learning data set using the Matlab SOM Toolbox by Vesanto *et al.* [10]. The erosion risk classes are neatly separated, with single overlapping hexagons. The erosion risk is progressing from the north (low risk) to the south (high risk)

Similarly as Fig. 3, also Fig. 4 was obtained using the data set samp1. When constructing the map, the class membership information was not used, The map was

created and graphed using the Matlab SOM Toolbox [10]. The same toolbox contains also another procedure, called 'som_supervised' (SOM_s), and based on a proposal by Kohonen et al. [7], how to include during the process of training the information on class membership. The procedure was added to the Matlab SOM Toolbox by Juha Parhankangas, who keeps the copyright of that procedure [10].
We have applied the 'som_supervised' technique to our data with the hope that it will ameliorate the already good differentiation among the classes. The result was negative: we got even a deterioration of the display.
We considered the idea that perhaps we should normalize our data in a different way, say statistically, to have the data with mean=0 and variance=1. Result: We got even more mixed classes.
The **quality of a SOM** is usually measured by two indices: the quantization error q_e and the topographical error t_e [10]. They are:

	q_e	t_e
Ordinary map:	0.0362	0.0327
Supervised map:	0.0453	0.1544
Ordinary normalized:	0.2203	0.0246
Supervised normalized:	0.2247	0.0596

Thus our conclusion: the best **SOM quality** is attained for the ordinary SOM.

6 Concluding Remarks

We compared in detail three methods serving for visualization of multivariate data, whose intrinsic dimension - as evaluated by the correlation fractal dimension - equals 1.60. This justifies the mapping of the data to a plane. The data were subdivided into 5 erosion risk classes and we wanted the mapping algorithm to take into account the class membership.
From the 3 investigated methods, the first one uses classical canonical discriminant functions (CDA alias LDA), which provide linear projections. The other two applied methods were: Generalized Discriminant Analysis (GDA) based on Gaussian kernels, and the som supervised SOM (SOM_s), a variant of Kohonen's self-organizing map. All the 3 considered methods yielded similar results. In all projections, the erosion risk subgroups appeared fairly separated, as it should be. The GDA, by a proper tuning of the parameter 'sigma', yielded the classes more and more condensed and separated, however without generalization to other samples.
All the 3 methods preserved roughly the topology of the data, although the GDA has twisted sometimes the planar representation of the high and very high erosion group. The SOM_s appeared worse than the ordinary SOM, both in som quality and in differentiation of the risk classes. This is to a certain degree justified, because the ordinary SOM is trained to be optimal in the som quality, which means to be optimal in achieving both small quantization error and small topographic error. A change in conditions of the training may cause a deviation from optimality.

What concerns time of computing, the classical CDA and the SOM (also SOM_s) worked extremely fast (several seconds), while the kernel GDA needed about 12 minutes. This happens not only for the GDA. Let us mention that lengthy calculations do happen also for some other nonlinear methods, especially, when the complexity of calculations depends essentially from the cardinality of the analyzed data set.

References

1. Bartkowiak, A., Szustalewicz, A., Evelpidou, N., Vassilopoulos, A.: Choosing data vectors representing a huge data set: a comparison of Kohonen's maps and the neural gas method. Proc. of the First Int. Conf. on Environmental Research and Assessment, Bucharest, Romania, March 23–27, 2003. Ars Docendi Publishing House, Bucharest (2003) 5–20
2. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12 (2000) 2385–2404
3. Baudat G., Anouar F., Feature vector selection and projection using kernels. Neurocomputing 55 (2003) 21–38
4. Camastra, F., Vinciarelli, A.: Estimating the intrinsic dimension of data with a fractal-based method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002) 10 1404–1407
5. Duda, R.O., Hart, P.E., Stork D.E.: Pattern Classification, 2nd Edition, Wiley (2001)
6. Grassberger, P., Procaccia, I.: Measuring the strangeness of strange attractors. Physica D9 (1983) 189–208
7. Kohonen, T., Makisavara, K., Saramaki, T. :Phonetic maps – insightful representation of phonological features for speech recognition. ICPR Montreal, Canada (1984) 182–185
8. Shawe-Taylor, J., Christianini, N.: Kernel Methods for Pattern Recognition, Cambridge University Press (2004)
9. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10 (1998) 1291–1398
10. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. : SOM Toolbox for Matlab 5. Som Toolbox Team, Helsinki, HUT, Finland, Libella Oy, Espoo (2000), http://www.cis.hut.fi/projects/somtoolbox (2005)
11. Webb, A., Statistical Pattern Recognition. 2nd Edition, Wiley (2002) reprint (2004)
12. Yang, J., Jin, Z., et. al. : Essence of kernel Fisher discriminant: KPCA plus LDA. Pattern Recognition 37 (2004) 2097–2100