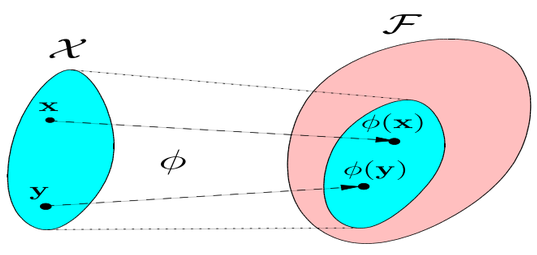
Kernel Springy Discriminant Analysis

ANDRÁS KOCSOR
Research Group on Artificial Intelligence, University of Szeged, Aradi vrt. 1, H-6720, Hungary

Abstract

Making use of the ubiquitous kernel notion, we present a new nonlinear supervised feature extraction technique called Kernel Springy Discriminant Analysis (KSDA). We demonstrate that this method can efficiently reduce the number of features and increase classification performance.

1. INTRODUCTION



The "kernel-idea". $\mathcal{F}$ is the closure of the linear span of the mapped data. The dot product in the kernel feature space $\mathcal{F}$ is defined implicitly. The dot product of $\sum_{i=1}^{n} \alpha_i \phi(\mathbf{x_i})$ and $\sum_{i=1}^{n} \beta_i \phi(\mathbf{x_i})$ is $\sum_{i,j} \alpha_i \beta_j k(\mathbf{x_i}, \mathbf{x_j})$.

Without loss of generality we shall assume that, as a realization of multivariate random variables, there are m-dimensional real attribute vectors in a compact set $\mathcal{X}$ over $\mathbb{R}^m$ describing objects in a certain domain, and that we have a finite $n \times m$ sample matrix $X = [\mathbf{x_1}, \cdots, \mathbf{x_n}]^\top$ containing n random observations. Let us assume as well that we have k classes and an indicator function

$$\mathcal{L} : \{1, \ldots, n\} \to \{1, \ldots, k\}, \tag{1}$$

where $\mathcal{L}(i)$ gives the class label of the sample $\mathbf{x_i}$.
Now let the dot product be implicitly defined (see the fig. on the left) by the kernel function k in some finite or infinite dimensional feature space $\mathcal{F}$ with associated transformation ϕ:

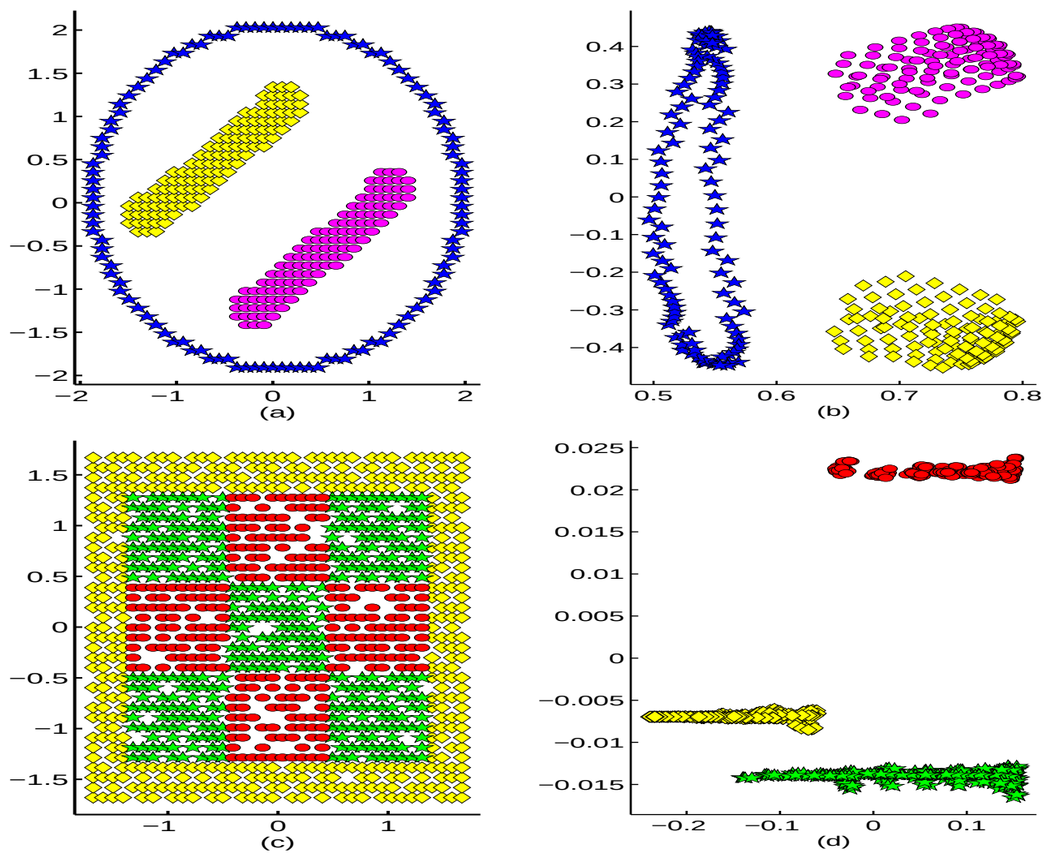
$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y}). \tag{2}$$

Knowing ϕ explicitly – and, consequently, knowing $\mathcal{F}$ – is not necessary. We need only define the kernel function, which then ensures an implicit evaluation. From the functions available, the two most popular are:

$$\text{Polynomial kernel:} \quad k_1(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^\top \mathbf{y}\right)^d, \qquad d \in \mathbb{N}, \tag{3}$$

$$\text{Gaussian RBF kernel:} \quad k_2(\mathbf{x}, \mathbf{y}) = \exp\left(-||\mathbf{x} - \mathbf{y}||^2 / r\right), \qquad r \in \mathbb{R}_+. \tag{4}$$

3. TEST RUN ON ARTIFICIAL DATA



In Figs. (a)-(d) we can see the result of a KSDA transformation using a Gaussian RBF kernel on artificial data with three different class labels. Without a doubt, the classes are well separated.

5. CONCLUSION

- The principal aim was to map the input data to a higher or even infinite dimensional feature space where, based on a physical spring & antispring analogue, we defined a measure for selecting new feature components.
- Since this measure can be extracted only by taking dot products of the mapped data, we overcame the numerical problems using kernel functions. These ensure an implicit access to this extended space in calculations.
- Furthermore, in contrast to [5][2][6][4] (Kernel LDA clones), KSDA can be done by solving a symmetric eigenproblem rather than an unsymmetrical one.
- In order to demonstrate the effectiveness of KSDA, we performed tests on artificial data and on real data as well, which had been prepared especially for a phonological awareness teaching system.
- We found that the applied feature extraction method was rather effective in improving classification performance in spite of a significant dimension reduction.
- But studying the scale-matrix Θ and applying a sparse data representation will be subject of future work.

References

[1] AIZERMAN, M. A., BRAVERMAN, E. M. AND ROZONOER L. I., Theoretical foundation of the potential function method in pattern recognition learning, *Auttomat. Remote Cont.*, vol. 25, pp. 821-837, 1964.

[2] BAUDAT G., ANOUAR F., Generalized discriminant analysis using a kernel approach, *Neural Computation*, Vol. 12, pp. 2385-2404, 2000.

[3] CRISTIANINI, N. AND SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and other kernel-based learning methods,* Cambridge University Press, 2000.

[4] KOCSOR, A., TÓTH, L. AND PACZOLAY, D., A Nonlinearized Discriminant Analysis and its Application to Speech Impediment Therapy, *in V. Matousek et al. (eds.): Proc. of the 4th Int. Conf. on Text, Speech and Dialogue,* LNAI 2166, pp. 249-257, Springer Verlag, 2001.

[5] MIKA, S., RÄTSCH, G., WESTON, J., SCHÖLKOPF B. AND MÜLLER, K.-R. Fisher discriminant analysis with kernels, in: Hu, Y.-H., Larsen J., Wilson, E. and Douglas S. (eds.): *Neural Networks for Signal Processing IX*, pp. 41-48. IEEE, 1999.

[6] TÓTH, L., KOCSOR, A., KOVÁCS, K.: A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, *in: Sojka, P., Kopecek I., Pala K.(eds.): TSD'2000*, LNAI 1902, pp. 307-313, Springer Verlag, 2000.

[7] VAPNIK, V. N., *Statistical Learning Theory,* John Wiley & Sons Inc., 1998.

2. KERNEL SPRINGY DISCRIMINANT ANALYSIS

Kernel Springy Discriminant Analysis (KSDA) searches for a linear transformation in $\mathcal{F}$ having the form $Q\dot{X}$, where Q is a real n by n orthogonal matrix containing variational parameter values obtained by the method, while $\dot{X}$ is a short-hand notation for the image matrix $[\phi(\mathbf{x_1}), \cdots, \phi(\mathbf{x_n})]^\top$. If we have a new attribute vector $\mathbf{y}$, the transformation can be employed by $Q\dot{X}\phi(\mathbf{y}) = Qk(X, \mathbf{y})$, where the column vector $k(X, \mathbf{y})$ is $[k(\mathbf{x_1}, \mathbf{y}), \cdots, k(\mathbf{x_n}, \mathbf{y})]^\top$.
The name Kernel Springy Discriminant Analysis stems from the utilization of a spring & antispring model, which involves searching for directions with optimal potential energy using attractive and repulsive forces. In our case sample pairs in each class are connected by springs, while those of different classes are connected by antisprings. New features can be easily extracted by taking the projection of a new point in those directions where a small spread in each class is attained, while different classes are spaced out as much as possible. Let $\delta(\mathbf{v})$, the potential of the spring model along the direction $\mathbf{v}$ in $\mathcal{F}$, be defined by

$$\sum_{i,j=1}^{n} ((\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j}))^\top \mathbf{v})^2 \Theta_{ij}, \tag{5}$$

where

$$\Theta_{ij} = \begin{cases} -1, & \text{if } \mathcal{L}(i) = \mathcal{L}(j) \\ 1, & \text{otherwise} \end{cases} \qquad i, j = 1, \ldots, n. \tag{6}$$

Technically speaking, KSDA searches for those directions $\mathbf{v}$ of the form $\dot{X}^\top \boldsymbol{\alpha}$ with a variational parameter vector $\boldsymbol{\alpha}$, along which a large potential is obtained. However, instead of the function $\delta(\dot{X}^\top \boldsymbol{\alpha})$, we use its normalized version

$$\gamma(\boldsymbol{\alpha}) = \frac{\delta(\dot{X}^\top \boldsymbol{\alpha})}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}, \tag{7}$$

when selecting the new directions. Intuitively, if larger values of γ indicate better directions and the chosen directions need to provide independent feature information, then choosing stationary points that have large values is a reasonable strategy.
It is easy to prove that $\gamma(\boldsymbol{\alpha})$ is equal to the following Rayleigh quotient formula

$$\gamma(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top \dot{X} A \dot{X}^\top \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}, \tag{8}$$

where

$$A = \sum_{i,j=1}^{n} \left(\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j})\right) \left(\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j})\right)^\top \Theta_{ij}. \tag{9}$$

Moreover, it is also straightforward to prove that (8) takes the following form:

$$\frac{\boldsymbol{\alpha}^\top \left(K\tilde{\Theta}K^\top - K\Theta K^\top\right) \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}, \tag{10}$$

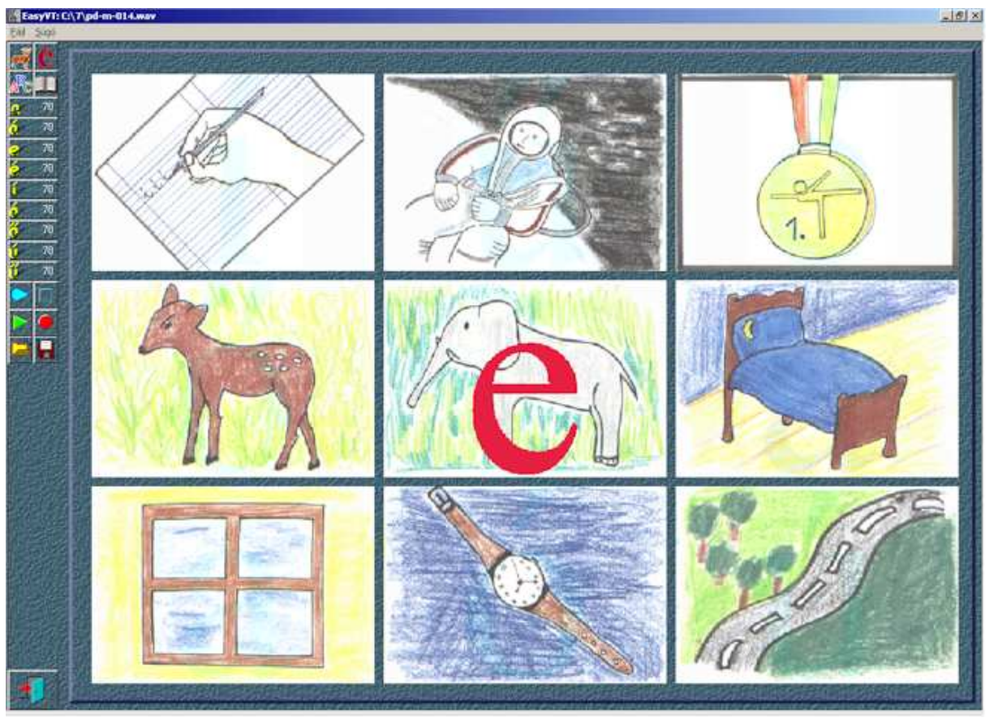
where $K = \dot{X}\dot{X}^\top = [k(\mathbf{x_i}, \mathbf{x_j})]$ and $\tilde{\Theta}$ is a diagonal matrix with the sum of each row of Θ in the diagonal. The stationary points of the above Rayleigh quotient formula will furnish the row vectors of the orthogonal matrix $\mathbf{Q}$.
After taking the derivative of (10) we readily see that the stationary points of $\gamma(\boldsymbol{\alpha})$ can be obtained via an eigenanalysis of the following symmetric eigenproblem:

$$(K\tilde{\Theta}K^\top - K\Theta K^\top)\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}. \tag{11}$$

If we assume that the eigenvectors are $\boldsymbol{\alpha_1}, \cdots, \boldsymbol{\alpha_n}$ then the orthogonal matrix Q is defined by $[\boldsymbol{\alpha_1} c_1, \cdots, \boldsymbol{\alpha_n} c_n]^\top$, where the normalization parameter c_i is equal to $(\boldsymbol{\alpha_i}^\top K \boldsymbol{\alpha_i})^{-1/2}$. This normalization factor ensures that the two-norm of row vectors of the transformation matrix $Q\dot{X}$ is unity.

4. TEST RUN ON REAL DATA

- ***A Phonological Awareness Teaching System.*** The most important clue to the process of learning to read is the ability to separate and identify consecutive sounds that make words and to associate these sounds with its corresponding written form. To learn to read in a fruitful way young learners must, of course, also be aware of the phonemes and be able to manipulate them.



- ***Database.*** For training and testing purposes we also recorded samples from 120 speakers (children of age 6-7) at a sampling rate of 22050 Hz in 16-bit quality. Each speakers uttered all the Hungarian vowels (9 vowels), one after the other, separated by a short pause.
- ***Initial Features.*** Initially the signals were processed in 10 ms frames, from which the log-energies of 24 critical-bands were extracted using FFT and triangular weighting. In our tests we used the filter-bank log-energies from the centermost frame of the steady-state part of each vowel and smoothed the feature trajectories to remove the effect of brief noises and disturbances ("FBLE Smooth" set, 24 features). Afterwards, in a second set of features we extended the smoothed log-energies with the gravity centers of four frequency bands, approximately corresponding to the possible values of the formants ("FBLE+Grav" set, 24+4 features). These gravity centers are supposed to give a crude approximation of the formants.
- ***Feature Extraction using KSDA.*** Naturally, both initial feature sets were transformed by KSDA using the third-order polynomial kernel. Since Q was defined by only those eigenvectors with the largest 16 eigenvalues, we also performed a dimension reduction when applying the KSDA transformation.
- ***Classifiers.*** Then, as a classifier, the well-known Support Vector Machine [7][3] was employed using the same kernel as before. In the trials 80 speakers were used for training and 40 for testing.
- ***Result.*** For the sets "FBLE Smooth" and "FBLE+Grav" the recognition errors were 6.08% and 5.27%, respectively, while after a KSDA transformation they were 3.24% and 2.8%, respectively.