Kernel Springy Discriminant Analysis and its Application to a Phonological Awareness Teaching System

András Kocsor[1] and Kornél Kovács[2]

Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences
and University of Szeged *
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{[1]kocsor, [2]kkornel}@inf.u-szeged.hu

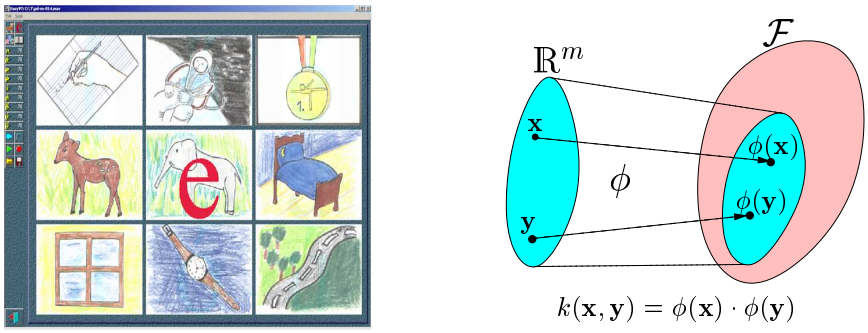
Abstract. Making use of the ubiquitous kernel notion, we present a new nonlinear supervised feature extraction technique called Kernel Springy Discriminant Analysis. We demonstrate that this method can efficiently reduce the number of features and increase classification performance. The improvements obtained admittedly arise from the nonlinear nature of the extraction technique developed here. Since phonological awareness is a great importance in learning to read, a computer-aided training system could be most beneficial in teaching young learners. Naturally, our system employs an effective automatic phoneme recognizer based on the proposed feature extraction technique.

1 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. Many children with learning disabilities have problems in their ability to process phonological information. So we decided to construct a computer-aided training software package which makes use of a very effective automatic phoneme recognizer in the background and provides visual feedback, on a frame-by-frame basis, in the form of flickering letters (see Fig. 1a). So as to make the sound to grapheme association easier a unique picture is attached to each letter. In addition, the transparency of letters is proportional to the output of the speech recognizer. Our experiments and general observations show that young people are more willing to practice with the computer than with traditional drills. To reinforce this point we found we could make impressive progress in a very short training period.

Since a highly efficient automatic phoneme recognizer can make the teaching system reliable, we decided to develop a novel feature extraction technique which proved to be suitable for this task. In the next section we describe this method, followed by results and concluding remarks.

* This work was supported under the contract IKTA No. 2001/055 from the Hungarian Ministry of Education.



(a) (b)

Fig. 1. (a) A phonological awareness teaching system. (b) 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 implicitly defined. 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 $\sum_{i,j} \alpha_i \beta_j k(\mathbf{x_i}, \mathbf{x_j})$.

2 Kernel Springy Discriminant Analysis

The approach of feature extraction could be either linear or nonlinear, but it seems there is a technique (which is most topical nowadays) that is, in some sense, breaking down the barrier between the two types. The key idea was originally presented in [1] and was again applied in connection with the general purpose Support Vector Machine [2][3]. In the following this notion is also used to derive a novel nonlinear feature extractor.

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 $\mathcal{L} : \{1, \ldots, n\} \to \{1, \ldots, k\}$, where $\mathcal{L}(i)$ gives the class label of the sample $\mathbf{x_i}$.

Now let the dot product be implicitly defined (see Fig. 1b) by the kernel function k in some finite or infinite dimensional feature space $\mathcal{F}$ with associated transformation ϕ:

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y}). \tag{1}$$

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 $[k(\mathbf{x_1}, \mathbf{y}), \cdots, k(\mathbf{x_n}, \mathbf{y})]^\top$. Thus in the kernel feature space $\mathcal{F}$ this type of linear transformation can be expressed only by dot products, which requires only kernel function evaluations. Moreover, since ϕ is generally nonlinear the resultant transformation is a nonlinear transformation of the original sample data. Knowing ϕ explicitly – and, consequently, knowing $\mathcal{F}$ – is not necessary. We need only define the kernel

function, which then ensures an implicit evaluation. The construction of an appropriate kernel function (i.e. when such a function ϕ exists) is a non-trivial problem, but there are many good suggestions about the sorts of kernel functions which might be adopted along with some background theory [2], [3]. From the functions available, the two most popular are[1]:

$$\text{Polynomial kernel:} \qquad k_1(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^\top \mathbf{y}\right)^d, \qquad d \in \mathbb{N}, \tag{2}$$

$$\text{Gaussian RBF kernel:} \qquad k_2(\mathbf{x},\mathbf{y}) = \exp\left(-||\mathbf{x}-\mathbf{y}||^2/r\right), \qquad r \in \mathbb{R}_+. \tag{3}$$

The stationary points of the Rayleigh quotient formula will furnish the row vectors of the orthogonal matrix $\mathbf{Q}$ of the KSDA transformation[2]:

$$\gamma(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \dot{X} A \dot{X}^\top \boldsymbol{\alpha} / \boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \tag{4}$$

where

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

and

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

It is straightforward to prove that (4) takes the following form:

$$\gamma(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \left(K\tilde{\Theta}K^\top - K\Theta K^\top\right)\boldsymbol{\alpha} / \boldsymbol{\alpha}^\top \boldsymbol{\alpha}, \tag{7}$$

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. After taking the derivative of (4) we readily see that the stationary points of $\gamma(\boldsymbol{\alpha})$ can be obtained via an eigenanalysis of the following symmetric eigenproblem:[3] $(K\tilde{\Theta}K^\top - K\Theta K^\top)\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$. 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.

[1] For a given kernel function ϕ is not always unique as the kernel k_1, the dimension of the feature space $\mathcal{F}$, is at least $\binom{m+d-1}{d}$ while with k_2 we get infinite dimension feature spaces.

[2] The name of this method 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 $\delta(\mathbf{v})$ be defined by $\sum_{i,j=1}^{n}((\Phi(\mathbf{x_i}) - \Phi(\mathbf{x_j}))^\top \mathbf{v})^2 \Theta_{ij}$. Using this term, which in $\mathcal{F}$ defines the potential of the spring model along the direction $\mathbf{v}$, we find that $\gamma(\boldsymbol{\alpha})$ is equal to $\delta(\dot{X}^\top\boldsymbol{\alpha})/\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$. Technically speaking, KSDA searches for those directions $\mathbf{v}$ of the form $\dot{X}^\top\boldsymbol{\alpha}$ along which a large potential is obtained. 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.

[3] In contrast to [4], KSDA can be performed by solving a symmetric eigenproblem rather than an unsymmetrical one.

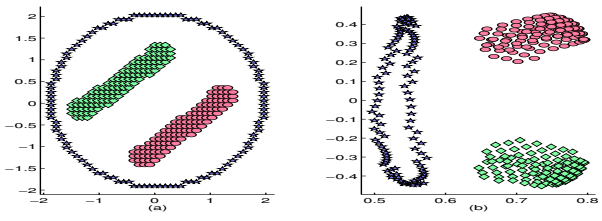


Fig. 2. KSDA transformation of artificial data. (a) is the initial data, and (b) is the result.

3 Results

In Fig. 2 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. 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, one after the other, separated by a short pause. Since we decided not to discriminate their long and short versions, we worked with 9 vowels altogether. 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 centremost 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. 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. Then, as a classifier the well-known Support Vector Machine [2][3] was employed using the same kernel as before. In the trials 80 speakers were used for training and 40 for testing. 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.

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