Indefinite Kernel Discriminant Analysis

Bernard Haasdonk[1] and Elżbieta Pękalska[2]

[1] Institute of Applied Analysis and Numerical Simulation
University of Stuttgart, Germany, *haasdonk@mathematik.uni-stuttgart.de*
[2] School of Computer Science
University of Manchester, United Kingdom, *pekalska@cs.man.ac.uk*

Abstract. Kernel methods for data analysis are frequently considered to be restricted to positive definite kernels. In practice, however, indefinite kernels arise e.g. from problem-specific kernel construction or optimized similarity measures. We, therefore, present formal extensions of some kernel discriminant analysis methods which can be used with indefinite kernels. In particular these are the multi-class kernel Fisher discriminant and the kernel Mahalanobis distance. The approaches are empirically evaluated in classification scenarios on indefinite multi-class datasets.

Keywords: Kernel Methods, Indefinite Kernels, Mahalanobis Distance, Fisher Discriminant Analysis

1 Introduction

Kernel methods are powerful statistical learning techniques, widely applied to various data analysis scenarios thanks to their flexibility and good performance, e.g. the support vector machine, kernel principal component analysis (KPCA), kernel Fisher discriminant (KFD), kernel k-means, etc. We refer to the monographs of Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004) for extensive presentations. The class of permissible kernels is often, and frequently wrongly, considered to be limited due to their requirement of being positive definite (pd). In practice, however, many non-pd similarity measures arise, e.g. when invariance or robustness is incorporated into the measure. Naturally, indefinite (dis-)similarities arise from non-Euclidean or non-metric dissimilarities, such as modified Hausdorff distances, or non-pd similarities, such as Kullback-Leibler divergence between probability distributions. Consequently, there is a practical need to handle these measures properly. Apart from embedding into Banach spaces or regularizing indefinite kernels, more general approaches are of high interest. A natural extension of Mercer kernels leads to indefinite kernels and the corresponding learning methods, cf. Ong et al. (2004), Pękalska and Duin (2005), Haasdonk (2005), Pękalska and Haasdonk (2009) and references therein.

In the current presentation, we extend two kernel discriminant methods, known from the positive definite case, to their indefinite counterparts. First, we focus on the generalized discriminant analysis (Baudat and Anouar

(2000)), which is a kernel version of the standard linear discriminant analysis for feature extraction (Duda et al. (2001)). Secondly, we consider the Mahalanobis distance for indefinite kernels, which is an extensions of the pd case of Haasdonk and Pękalska (2008). For completeness, we also want to mention other existing approaches for pd kernels of Ruiz and Lopez-de Teruel (2001) and Wang et al. (2008).

The structure of the paper is as follows. In the next section we provide the basic notation and background on indefinite kernel spaces, which are the geometric framework for indefinite kernels. We derive the kernelized versions of the Fisher discriminant and two versions of the Mahalanobis distance for indefinite kernels in Sec. 3. We perform classification experiments in Sec. 4 and conclude in Sec. 5.

2 Kernels and Feature Space Embedding

The proper frame for indefinite kernel functions, to be used in the sequel, are indefinite vector spaces such as pseudo-Euclidean (Goldfarb (1985), Pękalska and Duin (2005)) or more general Kreĭn spaces (Bognar (1974), Rovnyak (2002)). A *Kreĭn space* over $\mathbb{R}$ is a vector space $\mathcal{K}$ equipped with a non-degenerate indefinite inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}: \mathcal{K}\times\mathcal{K}\to\mathbb{R}$ such that $\mathcal{K}$ admits an orthogonal decomposition as a direct sum, $\mathcal{K}=\mathcal{K}_+\oplus\mathcal{K}_-$, where $(\mathcal{K}_+,\langle\cdot,\cdot\rangle_+)$ and $(\mathcal{K}_-,\langle\cdot,\cdot\rangle_-)$ are separable Hilbert spaces with their corresponding pd inner products. The inner product of $\mathcal{K}$, however, is the difference of $\langle\cdot,\cdot\rangle_+$ and $\langle\cdot,\cdot\rangle_-$, i.e. for any $\xi_+,\xi'_+\in\mathcal{K}_+$ and any $\xi_-,\xi'_-\in\mathcal{K}_-$ holds

$$\left\langle \xi_+ + \xi_-, \xi'_+ + \xi'_-\right\rangle_{\mathcal{K}} := \left\langle \xi_+, \xi'_+\right\rangle_+ - \left\langle \xi_-, \xi'_-\right\rangle_- .$$

The natural projections P_+ onto $\mathcal{K}_+$ and P_- onto $\mathcal{K}_-$ are *fundamental projections*. Any $\xi\in\mathcal{K}$ can be represented as $\xi = P_+\,\xi + P_-\,\xi$, while $I_{\mathcal{K}} = P_+ + P_-$ is the identity operator. The linear operator $\mathcal{J} = P_+ - P_-$ is called the *fundamental symmetry* and is the basic characteristic of a Kreĭn space $\mathcal{K}$, satisfying $\mathcal{J}=\mathcal{J}^{-1}$. The space $\mathcal{K}$ can be turned into its *associated Hilbert space* $|\mathcal{K}|$ by using the positive definite inner product $\langle\xi,\xi'\rangle_{|\mathcal{K}|} := \langle\xi,\mathcal{J}\xi'\rangle_{\mathcal{K}}$. We use the "transposition" abbreviation $\xi^T\xi' := \langle\xi,\xi'\rangle_{|\mathcal{K}|}$ for vectors and now additionally (motivated by $\mathcal{J}$ operating as a sort of "conjugation") a "conjugate-transposition" notation $\xi^*\xi' := \langle\xi,\xi'\rangle_{\mathcal{K}} = \langle\mathcal{J}\xi,\xi'\rangle_{|\mathcal{K}|} = (\mathcal{J}\xi)^T\xi' = \xi^T\mathcal{J}\xi'$.

Finite-dimensional Kreĭn spaces with $\mathcal{K}_+=\mathbb{R}^p$ and $\mathcal{K}_-=\mathbb{R}^q$ are denoted by $\mathbb{R}^{(p,q)}$ and called *pseudo-Euclidean spaces*. They are characterized by the so-called *signature* $(p,q)\in\mathbb{N}^2$. $\mathcal{J}$ becomes the matrix $\mathcal{J}=\mathrm{diag}(\mathbf{1}_p,-\mathbf{1}_q)$ with respect to an orthonormal basis in $\mathbb{R}^{(p,q)}$. Kreĭn spaces are important as they provide feature-space representations of dissimilarity data (Goldfarb (1985)) or indefinite kernels. In analogy to the pd case, an indefinite kernel represents an inner product in an implicitly defined Kreĭn space. Hence algorithms working with indefinite kernels have a geometric interpretation in these spaces.

We assume a c-class problem with n training samples $X := \{x_i\}_{i=1}^n \subset \mathcal{X}$ and integer class labels $\{y_i\}_{i=1}^n \subset \{1, \ldots, c\} \subset \mathbb{N}$, where $\mathcal{X}$ denotes a general set of objects. The class labels induce a partition of the training data $X = \cup_{j=1}^c X^{[j]}$ with $X^{[j]} = \{x_i^{[j]}\}_{i=1}^{n^{[j]}}$, i.e. $n^{[j]}$ denoting the number of samples per class satisfying $\sum_{j=1}^c n^{[j]} = n$. Let $\psi \colon \mathcal{X} \to \mathcal{K}$ be an embedding of the sample set $\mathcal{X}$ into a Kreĭn space $\mathcal{K}$ and $\Psi := [\psi(x_1), \ldots, \psi(x_n)]$ be a sequence of embedded samples. We use natural "sequence-vector-products" to abbreviate linear combinations, so the empirical mean is defined as $\psi_\mu := \frac{1}{n}\sum_{i=1}^n \psi(x_i) = \frac{1}{n}\Psi\mathbf{1}_n$ with $\mathbf{1}_n \in \mathbb{R}^n$ being the vector of all ones. Similarly, we adopt "sequence-sequence-product" notation for expressing matrices, e.g. $K := \Psi^*\Psi = \Psi^T\mathcal{J}\Psi \in \mathbb{R}^{n\times n}$ being the kernel-matrix with respect to the kernel $k(x, x') := \langle\psi(x), \psi(x')\rangle_{\mathcal{K}}$. If $\Psi^{[j]} = [\psi(x_1^{[j]}), \ldots, \psi(x_{n^{[j]}}^{[j]})]$ denotes the sequence of class-wise embedded data, we define the class mean as $\psi_\mu^{[j]} := \frac{1}{n^{[j]}}\Psi^{[j]}\mathbf{1}_{n^{[j]}}$ and we abbreviate the column-blocks of the kernel matrix as $K^{[j]} := \Psi^*\Psi^{[j]} \in \mathbb{R}^{n\times n^{[j]}}$. We finally introduce the kernel-quantities $\mathbf{k}_x := (\psi(x_i)^*\psi(x))_{i=1}^n$. In practice, the embedding ψ will not be given for defining the kernel k. Instead a symmetric kernel function $k(x, x')$ will be chosen in a problem-specific way, which then implicitly represents the inner-product in some Kreĭn space obtained via a suitable embedding ψ. The strength of kernel methods relies in the fact that the computation of this embedding ψ can mostly be avoided, if the analysis algorithm only requires inner-products between embedded samples, as these are provided by the given function k.

3 Kernel Discriminant Analysis

3.1 Indefinite Kernel Fisher Discriminant Analysis

We demonstrated in (Haasdonk and Pękalska (2008b)) how the two-class indefinite Kernel Fisher discriminant classifier can be rigorously derived. This represents an extension of the KFD (Mika et al. (1999)) to indefinite kernels. Here, we generalize this further to multicategory Fisher Discriminant Analysis for feature extraction (Duda et. al. (2001)). The within-class scatter operator $\Sigma_W^{[j]} : \mathcal{K} \to \mathcal{K}$ for the j-th class is defined by

$$\Sigma_W^{[j]}\varphi := \sum_{i=1}^{n^{[j]}} \left(\psi(x_i^{[j]}) - \psi_\mu^{[j]}\right)\left(\psi(x_i^{[j]}) - \psi_\mu^{[j]}\right)^*\varphi, \quad \varphi \in \mathcal{K} \tag{1}$$

which results in the (normalized) overall within-class scatter $\Sigma_W := \frac{1}{n}\sum_{j=1}^c \Sigma_W^{[j]}$. Similarly, the (normalized) between-class scatter operator is defined by

$$\Sigma_B\varphi := \sum_{j=1}^c \frac{n^{[j]}}{n}\left(\psi_\mu^{[j]} - \psi_\mu\right)\left(\psi_\mu^{[j]} - \psi_\mu\right)^*\varphi, \quad \varphi \in \mathcal{K}. \tag{2}$$

The multiple discriminant analysis problem is then solved by searching a sequence of vectors $W = [w_1, \ldots, w_{c-1}] \in \mathcal{K}^{c-1}$ such that

$$J(W) := \frac{\det(W^* \Sigma_B W)}{\det(W^* \Sigma_W W)} = \frac{\det(W^T \mathcal{J} \Sigma_B W)}{\det(W^T \mathcal{J} \Sigma_W W)} \tag{3}$$

is maximized. This is obtained by solving the generalized eigenvalue problem

$$\mathcal{J} \Sigma_B w_j = \lambda_j \mathcal{J} \Sigma_W w_j \tag{4}$$

for the $c-1$ largest eigenvalues λ_j. The practical computation can now be kernelized similarly to Baudat and Anouar (2000) or Haasdonk and Pękalska (2008b). First, we note from the eigenvalue equation (4) that the range of both scatter operators is spanned by embedded training examples, implying $w_j \in \operatorname{span}\{\psi(x_i)\}_{i=1}^n$. Hence, there exists a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{n \times (c-1)}$ such that $W = \Psi \boldsymbol{\alpha}$ and the discriminant quotient (3) becomes

$$J(W) := \frac{\det(\boldsymbol{\alpha}^T \Psi^T \mathcal{J} \Sigma_B \Psi \boldsymbol{\alpha})}{\det(\boldsymbol{\alpha}^T \Psi^T \mathcal{J} \Sigma_W \Psi \boldsymbol{\alpha})} =: \frac{\det(\boldsymbol{\alpha}^T M \boldsymbol{\alpha})}{\det(\boldsymbol{\alpha}^T N \boldsymbol{\alpha})}. \tag{5}$$

The matrices M, N can now be computed based on the kernel data. We define $\mathbf{c} := \frac{1}{n}\mathbf{1}_n$ and $\mathbf{c}^{[j]} := (c_i^{[j]})_{i=1}^n$ with $c_i^{[j]} := 1/n^{[j]}$ for $x_i \in X^{[j]}$ and $c_i^{[j]} := 0$ otherwise. Then the between-class scatter is rewritten as $\Sigma_B = \sum_{j=1}^c \frac{n^{[j]}}{n} \Psi(\mathbf{c}^{[j]} - \mathbf{c})(\mathbf{c}^{[j]} - \mathbf{c})^T \Psi^*$. Setting $D := \sum_{j=1}^c \frac{n^{[j]}}{n}(\mathbf{c}^{[j]} - \mathbf{c})(\mathbf{c}^{[j]} - \mathbf{c})^T$, we obtain

$$M = \Psi^T \mathcal{J} \Sigma_B \Psi = \Psi^T \mathcal{J} \Psi D \Psi^T \mathcal{J} \Psi = KDK. \tag{6}$$

Note that both D and M are positive semidefnite by construction. Further, we introduce the centering matrix $H^{[j]} := I_{n^{[j]}} - \frac{1}{n^{[j]}} \mathbf{1}_{n^{[j]}} \mathbf{1}_{n^{[j]}}^T \in \mathbb{R}^{n^{[j]} \times n^{[j]}}$ and obtain for the class-specific within-class scatter operator $\Sigma_W^{[j]} = \frac{1}{n^{[j]}} \Psi^{[j]} H^{[j]} \Psi^{[j]} \mathcal{J}$. Consequently, $\Sigma_W = \sum_{j=1}^c \frac{n^{[j]}}{n} \Sigma_W^{[j]} = \frac{1}{n} \sum_{j=1}^c \Psi^{[j]} H^{[j]} \Psi^{[j]} \mathcal{J}$. Hence, the denominator matrix of (5) is expressed as

$$N = \Psi^T \mathcal{J} \Sigma_W \Psi = \frac{1}{n} \sum_{j=1}^c K^{[j]} H^{[j]} (K^{[j]})^T. \tag{7}$$

N is positive semidefinite irrespectively of the definiteness of $K^{[j]}$. To see this, it is sufficient to remark that $K^{[j]} H^{[j]} (K^{[j]})^T = (K^{[j]} H^{[j]})(K^{[j]} H^{[j]})^T$ is positive semidefinite as $H^{[j]}$ is idempotent, and N is a positive linear combination of such matrices. Identical to the pd case, the matrix N will be singular and maximizing (5) is not well defined. Therefore, the matrix N is regularized, e.g. by $N_\beta = N + \beta I_{n \times n}$ with $\beta > 0$. The coefficient matrix $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{c-1}]$ is then obtained columnwise by solving the following eigenvalue problem

$$(N_\beta^{-1} M) \boldsymbol{\alpha}_j = \lambda_j \boldsymbol{\alpha}_j,$$

Obviously, thanks to the positive semidefiniteness of $N_\beta^{-1}M$, the eigenvalues λ_j are nonnegative. The normalized eigenvectors $\boldsymbol{\alpha}_j$ define the indefinite kernel Fisher (IKF) feature extractors by suitable projections via the indefinite inner product as follows:

$$f_{IKF}(x) := [\langle w_1, \psi(x)\rangle_{\mathcal{K}}, .., \langle w_{c-1}, \psi(x)\rangle_{\mathcal{K}}]^T = W^* \psi(x) = \boldsymbol{\alpha}^T \Psi^* \psi(x) = \boldsymbol{\alpha}^T \mathbf{k}_x.$$

The above formulation is equivalent to the kernel Fisher discriminant analysis for pd kernels; here, however, K is indefinite.

We now want to note an interesting theoretical fact of the IKF feature extractor: it is equivalent to embedding the data in the associated Hilbert space $|\mathcal{K}|$ and performing a positive definite kernel Fisher discriminant analysis. The latter approach would be in the spirit of "regularizing" the indefinite kernel matrix to a pd matrix, but is algorithmically quite cumbersome. After an eigenvalue decomposition of the possibly huge indefinite kernel matrix the negative eigenvalues are flipped yielding an explicit feature space embedding, which enables the traditional discriminant analysis. If we show the equivalence of IKF to this procedure, then IKF is a simple algorithmical alternative to this regularizing approach. To see this equivalence, we first note with view on (1) and (2) that the between-class and within-class scatter operators in the associated Hilbert space are given by $\Sigma_B^{|\mathcal{K}|} = \Sigma_B \mathcal{J}$ and $\Sigma_W^{|\mathcal{K}|} = \Sigma_W \mathcal{J}$. Then, the corresponding positive definite Fisher discriminant eigenvalue problem $\Sigma_B^{|\mathcal{K}|} w_j^{|\mathcal{K}|} = \lambda_j^{|\mathcal{K}|} \Sigma_W^{|\mathcal{K}|} w_j^{|\mathcal{K}|}$ is solved by $\lambda_j^{|\mathcal{K}|} = \lambda_j$ and $w_j^{|\mathcal{K}|} = \mathcal{J} w_j$ as can be seen by (4). Setting $W^{|\mathcal{K}|} := [w_1^{|\mathcal{K}|}, \ldots, w_{c-1}^{|\mathcal{K}|}]$ allows to define the Fisher discriminant in the associated Hilbert space and gives the equivalence to f_{IKF}:

$$f_{IKF}^{|\mathcal{K}|}(x) := (W^{|\mathcal{K}|})^T \psi(x) = (\mathcal{J} W)^T \psi(x) = W^T \mathcal{J} \psi(x) = W^* \psi(x) = f_{IKM}(x).$$

3.2 Indefinite Kernel Mahalanobis Distances

For simplicity of presentation we describe the computation of the Mahalanobis distance for the complete dataset and assume that it is centered in the embedded Kreĭn space. This can be obtained by explicit centering operations, cf. Shawe-Taylor and Cristianini (2004). As a result, $K = \Psi^* \Psi$ is now a centered kernel matrix. Then, the empirical covariance operator C: $\mathcal{K} \to \mathcal{K}$ acts on $\phi \in \mathcal{K}$ as $C\phi := \frac{1}{n} \sum_{i=1}^n \psi(x_i)\ \langle \psi(x_i), \phi\rangle_{\mathcal{K}} = \frac{1}{n} \Psi \Psi^* \phi$. We will therefore identify the empirical covariance operator as

$$C = \frac{1}{n} \Psi \Psi^* = \frac{1}{n} \Psi \Psi^T \mathcal{J} = C^{|\mathcal{K}|} \mathcal{J},$$

where $C^{|\mathcal{K}|} = \frac{1}{n} \Psi \Psi^T$ is the empirical covariance operator in $|\mathcal{K}|$. The operator C is not pd in the Hilbert sense, but it is pd in the Kreĭn sense. It means that $\langle \xi, C\xi\rangle_{\mathcal{K}} \geq 0$ for $\xi \neq 0$ in agreement with the inner product of that space. In (Haasdonk and Pękalska (2008)) we presented a kernelized version of the

Mahalanobis distance for pd kernels. In the case of an invertible covariance (IC) operator, the derivation directly extends to indefinite kernels resulting in

$$d_{IC}^2(x) := \psi(x)^* C^{-1} \psi(x) = n(\mathbf{k}_x)^T (K^-)^2 \mathbf{k}_x$$

where the superscript $\cdot^- = \mathrm{pinv}(\cdot, \alpha)$ denotes the pseudo-inverse with a threshold $\alpha > 0$. This means that in the computation of the inverse singular values smaller than α are set to zero.

The above distance is evaluated per class and does not involve between-class information. Hence, alternatively, we also proposed a kernel Mahalanobis distance in a full kernel (FK) space, determined via KPCA (Pękalska and Haasdonk (2009)). This distance for class j is obtained as

$$(d_{FK}^{[j]}(x))^2 := \frac{n^{[j]}}{2} (\tilde{\mathbf{k}}_x^{[j]})^T (\tilde{K}_{\mathrm{reg}}^{[j]})^{-1} \tilde{\mathbf{k}}_x^{[j]}, \tag{8}$$

where $\tilde{\mathbf{k}}_x^{[j]} := \mathbf{k}_x - \frac{1}{n^{[j]}} K^{[j]} \mathbf{1}_{n^{[j]}}$ and $\tilde{K}_{\mathrm{reg}}^{[j]} := \tilde{K}^{[j]} + \alpha_j I_n$ for some $\alpha_j > 0$ with $\tilde{K}^{[j]} := K^{[j]} H^{[j]} (K^{[j]})^T \in \mathbb{R}^{n \times n}$. The extension to indefinite kernels is straightforward as the kernel matrices $\tilde{K}^{[j]} := K^{[j]} H^{[j]} (K^{[j]})^T$ still are positive semidefinite. As a result, the indefinite kernel Mahalanobis distance using the full kernel matrix is identical to (8), but based on an indefinite kernel. See Appendix of Pękalska and Haasdonk (2009) for details.

Since we compute kernel Mahalanobis distances per class, we can now define the feature representations of a sample x as a c-dimensional vector by the indefinite kernel Mahalanobis distance with invertible covariance (IKM-IC) as $f_{IKM-IC}(x) := [d_{IC}^{[1]}(x), \ldots, d_{IC}^{[c]}(x)]^T$ and similarly f_{IKM-FK} using the full kernel distance $d_{FK}^{[j]}$.

4 Classification Experiments

Multi-class problems characterized by indefinite proximity data can now be approached via the feature representations f_{IKF}, f_{IKM-IC} and f_{IKM-FK} defined in the previous section. Although the features are extracted from indefinite kernels, the resulting either c- or $(c-1)$-dimensional feature vector spaces are assumed to be equipped with the traditional inner product and Euclidean metric. As a result, different classifiers can now be trained there.

Table 4 lists basic properties of eight multi-class datasets, i.e. the type of dissimilarity measure, the class names and sizes and the fraction of data used for training in the hold-out experiments. Some measures of indefiniteness of the resulting training kernel matrices are also provided: $r_{\mathrm{neg}} \in [0, 1]$ denotes the ratio of negative to overall variance of the centered training kernel matrix and (p, q) indicates the signature of the embedding Kreĭn space. These quantities are averaged over 25 runs based on random drawings of a training subset. For detailed descriptions and references to the single datasets, we refer to Appendix of (Pękalska and Haasdonk (2009)).

Table 1. Characteristics of indefinite datasets, cf. Pękalska and Haasdonk (2009) for details and references.

	Dissimilarity	Kernel	c ($n^{[j]}$)	β	$r_{\text{neg}}(p,q)$
Cat-cortex	Prior knowl.	$-d^2$	4 (10–19)	0.80	0.19 (35, 18)
Protein	Evolutionary	$-d^2$	4 (30–77)	0.80	0.00 (167, 3)
News-COR	Correlation	$-d^2$	4 (102–203)	0.60	0.19 (127,208)
ProDom	Structural	s	4 (271–1051)	0.25	0.01 (518, 90)
Chicken29	Edit-dist.	$-d^2$	5 (61–117)	0.80	0.31 (192,166)
Files	Compression	$-d^2$	5 (60–255)	0.50	0.02 (392, 63)
Pen-ANG	Edit-dist.	$-d^2$	10 (334–363)	0.15	0.24 (261,269)
Zongker	Shape-match.	s	10 (200)	0.25	0.36 (274,226)

In our hold-out experiments, each data set is split into the training and test set of suitable sizes as reported in Table 1. The dissimilarity data set is first scaled such that the average dissimilarity is 1 on the training set. This is done in order to have a consistent choice over a range of crossvalidated parameters. The training kernel matrix based on the kernel $k = -d^2$ is then centered and used to extract features either by f_{IKF}, f_{IKM-IC} or f_{IKM-FK}. Next, four classifiers are constructed in the derived feature spaces, namely the nearest mean classifier (NM), Fisher discriminant (FD), quadratic discriminant (QD) and k-nearest neighbour (KNN) rule. Since we use the same training data both to extract the features and train the classifiers, simple classifiers are preferred to avoid the overuse of the data. The classifiers are then applied to suitably projected test data. This is repeated 20 times and the results are averaged.

As a reference, we use classifiers that directly work with original proximity measures as kernels. These are the indefinite kernel Fisher discriminant (IKFD), indefinite support vector machine (ISVM) and indefinite kernel nearest neighbour (IKNN) classifier. In particular, FD, IKFD and ISVM are binary classifiers, which solve the multi-class problems by the one-versus-all approach. The remaining classifiers are inherently multi-class classifiers.

The free parameters of both feature extractors and classifiers are determined via a (nested, if necessary) 10-fold crossvalidation. The regularization parameters of the kernel discriminant feature extractors are selected from $\{10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 0.1, 0.5, 1, 10, 10^2, 10^3\}$. The number of nearest neighbors is found from 1 to 15. The parameter C for ISVM is found within $\{0.01, 0.1, 0.5, 1, 5, 10, 10^2, 10^3, 10^4, 10^6, 10^8\}$.

The classification results are reported in Table 2. Although strong conclusions cannot be drawn due to high standard deviations, the following observations can be made. The features obtained by f_{IKM-IC} are performing much worse than all other extracted features. This may be due to two facts. First, the assumption of an invertible covariance matrix may be wrong, leading to a degeneration in classification accuracy. Second, this Mahalanobis distance

Table 2. Average classification errors (and standard deviation) over 20 hold-out repetitions of data drawing and cross-validated parameter selection.

Classifier+Features	Cat-cortex	Protein	News-COR	ProDom	Chicken29	Files	Pen-ANG	Zongker
NM+IKM-IC	45.5 (13.2)	21.2 (7.7)	38.6 (2.4)	15.0 (3.5)	36.1 (5.4)	53.1 (4.4)	32.9 (1.9)	37.6 (2.1)
NM+IKM-FK	10.9 (6.3)	2.1 (2.6)	24.9 (2.5)	6.4 (2.4)	20.4 (3.2)	44.4 (3.6)	33.0 (1.6)	14.5 (0.8)
NM+IKF	12.6 (5.7)	0.1 (0.4)	24.1 (1.8)	2.0 (0.6)	6.6 (2.3)	5.3 (1.5)	1.3 (0.5)	5.9 (0.6)
FD+IKM-IC	42.2 (11.3)	25.9 (5.5)	39.7 (2.6)	9.4 (3.0)	36.9 (2.9)	50.4 (5.1)	11.5 (1.1)	33.5 (1.3)
FD+IKM-FK	10.3 (5.4)	1.1 (2.0)	24.2 (2.0)	1.7 (0.6)	8.9 (4.0)	22.3 (3.3)	3.2 (0.9)	6.3 (0.9)
FD+IKF	11.2 (5.2)	0.2 (0.5)	24.2 (3.1)	1.6 (0.6)	5.9 (2.3)	5.4 (1.5)	1.5 (0.4)	7.1 (0.8)
QD+IKM-IC	48.5 (12.4)	11.9 (4.5)	41.4 (3.0)	3.6 (0.9)	30.1 (3.4)	29.3 (4.5)	4.9 (0.9)	39.3 (2.1)
QD+IKM-FK	22.7 (6.7)	0.5 (0.8)	25.5 (2.7)	2.0 (0.7)	6.9 (2.8)	7.7 (2.7)	1.2 (0.3)	6.5 (0.9)
QD+IKF	18.4 (7.4)	0.5 (1.3)	24.4 (3.1)	1.5 (0.5)	5.0 (1.5)	6.7 (1.4)	1.5 (0.4)	5.8 (0.6)
KNN+IKM-IC	43.9 (8.6)	19.8 (7.4)	42.9 (3.1)	5.0 (1.6)	31.8 (5.0)	30.1 (4.6)	5.7 (0.7)	33.9 (1.6)
KNN+IKM-FK	11.3 (6.5)	0.6 (1.7)	25.7 (1.7)	2.2 (0.9)	4.2 (1.3)	8.0 (2.4)	1.8 (0.3)	4.8 (0.5)
KNN+IKF	11.7 (6.5)	0.2 (0.5)	24.7 (2.2)	1.6 (0.7)	5.1 (2.2)	4.9 (1.3)	1.2 (0.3)	5.5 (0.8)
IKFD	10.6 (5.6)	0.3 (0.7)	23.6 (2.4)	2.0 (0.6)	6.4 (2.2)	5.5 (1.4)	1.4 (0.4)	6.3 (0.7)
ISVM	16.5 (5.7)	0.5 (0.8)	24.4 (2.3)	1.6 (0.6)	6.5 (2.2)	11.6 (2.7)	5.0 (0.5)	7.0 (0.5)
IKNN	15.6 (5.8)	4.7 (5.2)	29.6 (2.3)	3.1 (0.8)	5.1 (2.0)	36.7 (2.8)	0.9 (0.3)	11.4 (1.4)

misses the between-class correlations of the datasets, which is increasingly important with higher number of classes. Hence, the full-kernel Mahalanobis distances f_{IKM-FK} are clearly preferable to the former. Still, the $(c-1)$-dimensional feature spaces f_{IKF} mostly lead to better results than the c-dimensional feature spaces obtained by the IKM approaches, hence the IKF-features are overall preferable. Among the chosen classifiers on the extracted features, KNN is overall the best classifier which suggests that these spaces benefit from non-linear classifiers. Among the reference classifiers, IKFD is frequently the best, closely followed by ISVM, but occasionally outperformed by IKNN. Overall, the extracted features f_{IKM-FK} and f_{IKF} in combination with the chosen simple classifiers consistently yield classification results in the range of the reference classifiers.

5 Conclusion

We presented extensions of kernel Fisher discriminant analysis and kernel Mahalanobis distances to indefinite kernels. The natural framework for indefinite kernels are Kreĭn spaces, which give a geometrical interpretation of these indefinite methods. An interesting theoretical finding for the IKF feature extractor is that it correspond to its counterpart in the associated Hilbert space. This implies that the indefinite kernel Fisher discriminant analysis saves the unnecessary preprocessing step of embedding the data, flipping the negative eigenvalues and performing an explicit Fisher discriminant in the embedded space. In particular, the f_{IKF} feature extractor is a real kernel method avoiding the explicit embedding. For the indefinite kernel Mahalanobis distance, we proposed to use two formulations from the pd case for indefinite kernels. We performed experiments on indefinite multi-class classification problems, that demonstrate the applicability of the f_{IKF} and f_{IKM-FK} methods, but clearly discarded the f_{IKM-IC} features. In particular, the successful features yield results comparable to standard indefinite kernel classifiers such as IKFD and ISVM.

6 Acknowledgements

The authors acknowledge funding by the German Research Foundation (DFG) within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart and by the Engineering and Physical Science Research Council in the UK, project no. EP/D066883/1.

References

BAUDAT, G. and ANOUAR, F. (2000): Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404.

BOGNAR, J. (1974): *Indefinite Inner Product Spaces.* Springer Verlag.

DUDA, R.O., HART, P.E. and STORK, D.G. (2001): *Pattern Classification.* John Wiley & Sons, Inc., 2nd edition.

GOLDFARB, L. (1985): A new approach to pattern recognition. In L. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers BV.

HAASDONK, B. (2005): Feature space interpretation of SVMs with indefinite kernels. *IEEE TPAMI*, 27(4):482–492, 2005.

HAASDONK, B. and PEKALSKA, E. (2008): Classification with kernel Mahalanobis distances. In *Proc. of 32nd. GfKl Conference, Advances in Data Analysis, Data Handling and Business Intelligence.*

HAASDONK, B. and PEKALSKA, E. (2008b): Indefinite kernel Fisher discriminant. In *Proc. of ICPR 2008, International Conference on Pattern Recognition*.

MIKA, S., RÄTSCH, G., WESTON, J., SCHÖLKOPF and MÜLLER, K.R. (1999): Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48.

ONG, C.S., MARY, X., CANU, S. and SMOLA, A.J. (2004): Learning with non-positive kernels. In *ICML*, pages 639–646. ACM Press.

PEKALSKA, E. and DUIN, R.P.W. (2005): *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications.* World Scientific.

PEKALSKA, E. and HAASDONK, B. (2009): Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(6):1017–1032.

ROVNYAK, J. (2002): Methods of Krein space operator theory *Operator Theory: Advances and Applications*, 134:31–66.

RUIZ, A. and LOPEZ-DE TERUEL, P.E. (2001): Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 12(1):16–32.

SCHÖLKOPF, B. and SMOLA, A.J. (2002): *Learning with Kernels.* MIT Press, Cambridge.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004): *Kernel Methods for Pattern Analysis.* Cambridge University Press, UK.

WANG, J., PLATANIOTIS, K.N., LU, J. and VENETSANOPOULOS, A.N. (2008): Kernel quadratic discriminant analysis for small sample size problem. *Pattern Recognition*, 41(5):1528–1538.