On the Equivalence of Kernel Fisher Discriminant Analysis and Kernel Quadratic Programming Feature Selection

Irene Rodriguez-Lujan, Ramon Huerta, Carlos Santa Cruz

Departamento de Ingeniería Informática and
Instituto de Ingeniería del Conocimiento
Universidad Autónoma de Madrid

BioCircuits Institute
University of California, San Diego

25 February 2011

Outline

Outline

Feature Selection and Extraction Methods (I)

- Increasing size and dimensionality of real-world datasets.

Linear Feature Selection and Extraction Methods

- Fast and simple.
- Do not handle nonlinear relationships in the data.
- Principal Component Analysis (PCA).
- Canonical Correlation Analysis (CCA).
- Fisher Discriminant Analysis (FDA).
- ...

Feature Selection and Extraction Methods (II)

Kernelized Feature Selection and Extraction Methods

- Capture nonlinear dependences in the data.
- Maps the data from an original space to a *feature space* $\mathcal{F}$ via a (nonlinear) mapping $\Phi : \mathbb{R}^d \longrightarrow \mathcal{F}$.
- The dot-product in the feature space $\mathcal{F}$ is defined by a Mercer kernel $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$.
- Reformulation of traditional linear methods using only dot-products of training samples $\Rightarrow$ nonlinear method in the input space.
  - Kernel Principal Component Analysis (KPCA).
  - Kernel Canonical Correlation Analysis (KCCA).
  - Kernel Fisher Discriminant Analysis (KFDA).
  - ...

Outline

Kernel Fisher Discriminant Analysis (KFDA) (I)

Notation

- $\mathcal{X}_1 = \{x_1^1, \dots, x_{l_1}^1\}$ and $\mathcal{X}_2 = \{x_1^2, \dots, x_{l_2}^2\}$ samples from two different classes ($x_i \in \mathbb{R}^d$).
- $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$.
- $y \in \{-1, 1\}^l$ be the target vector.
- Mapping function to the kernel space: $\Phi : \mathbb{R}^d \longrightarrow \mathcal{F}$.
- Mercer Kernel: $K(x, y) =< \Phi(x), \Phi(y) >$.

Kernel Fisher Discriminant Analysis (KFDA) (II)

- Mapping the data nonlinearly into the feature space $\mathcal{F}$ and computing Fisher's linear discriminant there.

KFDA Objective Function

$$\max_{w \in \mathcal{F}} J(w) = \max_{w \in \mathcal{F}} \frac{w^T S_B^\Phi w}{w^T S_W^\Phi w}$$

$\mathcal{F}$ Mean Vector

$m_i^\Phi = \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi(x_j^i)$

$\mathcal{F}$ Between Scatter Matrix

$S_B^\Phi = (m_1^\Phi - m_2^\Phi)(m_1^\Phi - m_2^\Phi)^T$

$\mathcal{F}$ Within Scatter Matrix

$S_W^\Phi = \sum_{i=1,2} \sum_{x \in \chi_i} (\Phi(x) - m_i^\Phi)(\Phi(x) - m_i^\Phi)^T$

Kernel Fisher Discriminant Analysis (KFDA) (III)

- $\Phi(x)$ is not known in general.
- Finding a solution in $\mathcal{F} \Rightarrow$ reformulate it in terms of only dot products of the input patterns.

KFDA Objective Function

$$\max_{\alpha} J(\alpha) = \max_{\alpha} \frac{\alpha^T M \alpha}{\alpha^T N \alpha}$$

Reproducing Kernels

$w = \sum_{i=1}^{l} \alpha_i \Phi(x_i)$

Kernelized Between Scatter Matrix

$M = (M_1 - M_2)(M_1 - M_2)^T$

$(M_i)_j = \frac{1}{l_i} \sum_{k=1}^{l_i} K(x_j, x_k^i)$

Kernelized Within Scatter Matrix

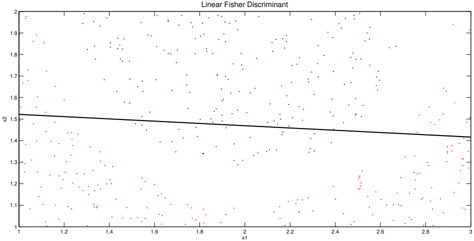
$N = \sum_{j=1,2} K_j (I - 1_{l_j}) K_j^T$

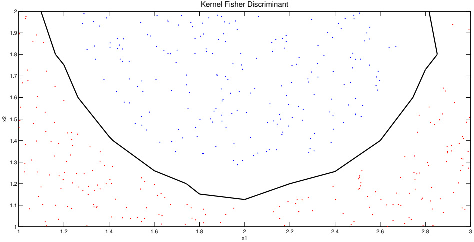
Kernel Fisher Discriminant Analysis (KFDA) (IV)

Kernel Fisher Coefficients (α^*_{KFD})

- $\alpha^*_{\text{KFD}} \equiv$ the leading eigenvector of $N^{-1}M$.
- $\alpha^*_{\text{KFD}} = N^{-1}(M_2 - M_1)$.
  - Ill-posed problem: N matrix not positive.
  - Some kind of regularization in needed ($\|\alpha\|^2$, $\|w\|^2$, ...).
  - Used regularization: $N_\mu = N + \mu_N I$.

Kernel Fisher Discriminant Analysis (KFDA) (V)





Outline

Quadratic Programming Feature Selection (QPFS)

- Select those features which provide a good tradeoff between relevance maximization and redundancy minimization for the classification task.

QPFS Objective Function

$$
\begin{aligned}
\min_{x} \quad & \tfrac{1}{2}x^T Q x - F^T x \\
\text{s.t.} \quad & x_i \geq 0 \qquad \forall i = 1 \dots M \\
& \|x\|_1 = 1.
\end{aligned}
$$

- **Q**: similarity among variables (redundancy).
- **F**: how correlated each feature is with the target class (relevance).
- Components of solution vector $\mathbf{x}^*$: weight of each feature.

Kernel QPFS (KQPFS) (I)

- For some kernels, it is not possible to give a weight to each feature in the kernel space due to its potential infinite dimension.
- QPFS objective function can be adapted to find an optimal direction w to project the data into the kernel space $\mathcal{F}$.
- KQPFS represents a **feature extraction** method.

KQPFS Objective Function

$$\min_{w} \frac{1}{2} w^T Q^{\Phi} w - \left(F^{\Phi}\right)^T w$$

Kernel QPFS (KQPFS) (II)

Similarity Measures

- QPFS: Correlation and Mutual Information.
  - The mapping function Φ is usually *implicit*.
  - The dimension of the kernel space $\mathcal{F}$ may be infinite.
  - Basis set in the kernel space is needed.
- KQPFS: **Covariance** $\Rightarrow$ KQPFS formulation does not require an explicit basis in the kernel space.

Kernel QPFS (KQPFS) (III)

KQPFS Redundacy-Relevance Matrices

$$
\begin{aligned}
Q^{\Phi} &= \sum_{x\in\mathcal{X}} \left(\Phi(x) - m^{\Phi}\right)\left(\Phi(x) - m^{\Phi}\right)^{T} \\
F^{\Phi} &= \sum_{x\in\mathcal{X}} (y_x - m^{y})\left(\Phi(x) - m^{\Phi}\right) \\
m^{\Phi} &= \frac{1}{l}\sum_{x\in\mathcal{X}} \Phi(x) \\
m^{y} &= \frac{1}{l}\sum_{i=1}^{l} y_i \; .
\end{aligned}
$$

Kernel QPFS (KQPFS) (IV)

- Theory of Reproducing Kernels: $w = \sum_{i=1}^{l} \alpha_i \Phi(x_i)$.
- $Q_K = K(I - 1_l) K$.
- $F_K = K(I - 1_l) y$.

KQPFS Objective Function

$$\min_{\alpha} G(\alpha) = \min_{\alpha} \frac{1}{2}\alpha^T Q_K \alpha - F_K^T \alpha$$

Kernel QPFS Coefficients

- $\nabla_\alpha G(\alpha) = 0 \Rightarrow Q_K \cdot \alpha = F_K$.
  - matrix Q_K is always singular.
  - Again, some kind of regularization in needed ($\|\alpha\|^2$, $\|w\|^2$, ...).
  - Used regularization: $Q_\mu = Q_K + \mu_Q I$.

Kernel QPFS (KQPFS) (V)

Regularized KQPFS Objective Function

$$G_\mu(\alpha) = \frac{1}{2}\alpha^T \left(Q_K + \mu_Q I\right) \alpha - F_K^T \alpha$$

Regularized Kernel QPFS Coefficients (α^*_{KQPFS})

$$\alpha^*_{\text{KQPFS}} = \left(Q_K + \mu_Q I\right)^{-1} F_K$$

- α^*_{KQPFS}: minimizes the covariance among features in the kernel space + maximizes the covariance of each feature in the kernel space with the target class.
- α^*_{KQPFS} only depends on the kernel matrix K and the class labels y.

Outline

Kernel Fisher Discriminant Analysis as Quadratic Programming Problem (I)

Proposition Mika et al.

KFD is equivalent to the quadratic programming problem:

$$\begin{aligned} \min_{\alpha} \quad & \alpha^T N\alpha + CP(\alpha) \\ \text{s.t.} \quad & \alpha^T (M_1 - M_2) = 2 \end{aligned} \tag{1}$$

Regularization: $N_\mu = N + \mu_N I$

- $C = \mu_N$.
- $P(\alpha) = \|\alpha\|^2$.

Kernel Fisher Discriminant Analysis as Quadratic Programming Problem (II)

Proposition Mika et al.

For given $C \in \mathbb{R}$, any optimal solution α to the optimization problem (1) is also optimal for

$$\begin{array}{ll} \min_{\alpha,b,\xi} & \|\xi\|^2 + \mu_N \|\alpha\|^2 \\ \text{s.t.} & K\alpha + \overrightarrow{1} b = y + \xi \\ & \overrightarrow{1}_i^T \xi = 0 \ \text{ for } i = 1,2 \end{array}$$

and vice versa

Proposition

Given $\mu_N \in \mathbb{R}$ and let $\mu_N = \mu_Q$, any optimal solution (α^*, b^*, ξ^*) to the optimization problem (2) is also optimal for the Regularized KQPFS and vice versa.

Proof (I)

- It is straightforward to show that the following proof is also valid for other regularization functions.

1. Working out ξ in (3):

$$\xi\,(\alpha, b) = K\alpha + \overrightarrow{1}b - y$$

$$\min_{\alpha,b,\xi} \quad \|\xi\|^2 + \mu_N\|\alpha\|^2 \tag{2}$$

$$\text{s.t.} \quad K\alpha + \overrightarrow{1}b = y + \xi \tag{3}$$

$$\overrightarrow{1}_i^T\xi = 0 \;\text{ for } i = 1, 2 \tag{4}$$

Proof (I)

- It is straightforward to show that the following proof is also valid for other regularization functions.

$$
\begin{array}{lll}
\min\limits_{\alpha,b,\xi} & \|\xi\|^2 + \mu_N \|\alpha\|^2 & (2) \\
\text{s.t.} & K\alpha + \overrightarrow{1} b = y + \xi & (3) \\
 & \overrightarrow{1}_i^T \xi = 0 \;\; \text{for } i = 1,2 & (4)
\end{array}
$$

1. Working out ξ in (3):

$$\xi\,(\alpha, b) = K\alpha + \overrightarrow{1} b - y$$

2. Optimization problem (2):

$$
\begin{array}{ll}
\min\limits_{\alpha,b} & \{\alpha^T KK\alpha - lb^2 - 2y^T K\alpha + \\
 & +y^T y + \mu_N \|\alpha\|^2\} \\
\text{s.t.} & \overrightarrow{1}_i^T \xi(\alpha, b) = 0 \text{ for } i = 1,2
\end{array}
$$

Proof (I)

- It is straightforward to show that the following proof is also valid for other regularization functions.

$$\begin{aligned} \min_{\alpha,b,\xi} \quad & \|\xi\|^2 + \mu_N \|\alpha\|^2 & (2) \\ \text{s.t.} \quad & K\alpha + \vec{1}\, b = y + \xi & (3) \\ & \vec{1}_i^T \xi = 0 \ \text{ for } i = 1,2 & (4) \end{aligned}$$

1. Working out ξ in (3):

$$\xi(\alpha, b) = K\alpha + \vec{1}\, b - y$$

2. Optimization problem (2):

$$\begin{aligned} \min_{\alpha,b} \quad & \{\alpha^T KK\alpha - lb^2 - 2y^T K\alpha + \\ & + y^T y + \mu_N \|\alpha\|^2\} \\ \text{s.t.} \quad & \vec{1}_i^T \xi(\alpha, b) = 0 \text{ for } i = 1,2 \end{aligned}$$

3. b depends on α as (4):

$$b(\alpha) = -\frac{1}{l} 1_l K\alpha + 1_l y$$

Proof(II)

- It is straightforward to show that the following proof is also valid for other regularization functions.

$$\min_{\alpha,b,\xi} \quad \|\xi\|^2 + \mu_N \|\alpha\|^2 \tag{5}$$

$$\text{s.t.} \quad K\alpha + \vec{1}\, b = y + \xi \tag{6}$$

$$\vec{1}_i^T \xi = 0 \ \text{ for } i = 1, 2 \tag{7}$$

4. Substituting $b(\alpha)$ we have an optimization problem with no constraints:

$$\begin{aligned} \min_\alpha \quad & \{\alpha^T K \left(I - 1_l\right) K\alpha \\ & -2y^T \left(I - 1_l\right) K\alpha + \\ & + \frac{\mu_N}{2} \|\alpha\|^2 + D\} \end{aligned} \tag{8}$$

Proof(II)

- It is straightforward to show that the following proof is also valid for other regularization functions.

$$\min_{\alpha,b,\xi} \quad \|\xi\|^2 + \mu_N \|\alpha\|^2 \tag{5}$$

$$\text{s.t.} \quad K\alpha + \vec{1} b = y + \xi \tag{6}$$

$$\vec{1}_i^T \xi = 0 \;\; \text{for } i = 1, 2 \tag{7}$$

4. Substituting $b(\alpha)$ we have an optimization problem with no constraints:

$$\begin{aligned} \min_\alpha \quad & \{\alpha^T K (I - 1_l) K\alpha \\ & -2y^T (I - 1_l) K\alpha + \\ & + \frac{\mu_N}{2} \|\alpha\|^2 + D\} \end{aligned} \tag{8}$$

5. Minimum value of Equation (8) is the same as those of the regularized KQPFS when $\mu_N = \mu_Q$.

$$G_\mu(\alpha) = \frac{1}{2}\alpha^T (Q_K + \mu_Q I)\alpha - F_K^T \alpha$$

Outline

Complexity

Standard KFDA

$$O(l^3) + 2l(l_1^2 + l_2^2) + 5l^2 + l_1^2 + l_2^2 + 7l$$

- Depends on the prior distributions of classes.

KQPFS

$$O(l^3) + 2l^3 + 4l^2$$

- Independent of the prior distributions of classes.

When is KQPFS faster than KFDA?

$$(l_1^2 + l_2^2)(2l + 1) + 5l^2 + 7l \gg 2l^3 + 4l^2$$

- Prior distributions of the class labels are highly unbalanced.

Outline

Empirical Equivalence

- Thirteen artificial and real world datasets were considered from the Rätsch benchmark repository.
- *Optimal* parameter values are known (Rätsch benchmark repository).
  - Width of the Gaussian kernel σ: $K(x, y) = e^{\frac{\|x-y\|^2}{\sigma}}$.
  - Regularization parameter μ_N.
  - $\mu_Q = \mu_N$.

For every training set

$$\cos\left(\alpha^*_{\text{KFD}}\ ,\ \alpha^*_{\text{KQPFS}}\right) = 1 \Longrightarrow \cos\left(w^*_{\text{KFD}}\ ,\ w^*_{\text{KQPFS}}\right) = 1$$

Computational Cost

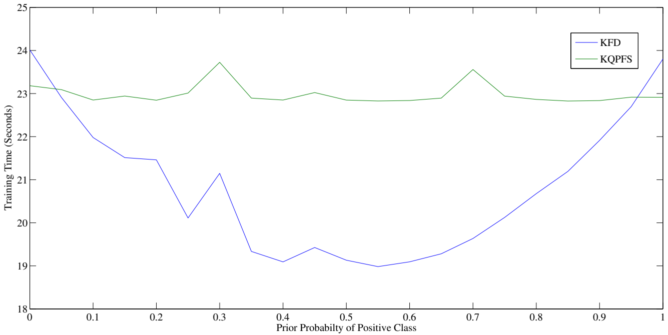


Figure: Abalone. Training time in seconds for the KFD and KQPFS algorithms.

Outline

Conclusions

- **Reformulation** of the Quadratic Programming Feature Selection (QPFS) method in a kernel space (KQPFS).
- **Proof** of the equivalence between KQPFS direction and KFD direction.
  - **New interpretation** of the KFD vector: direction which minimizes the covariance among features and maximizes the covariance of each feature with the target class in the kernel space.
  - **New solution** for KFD disregarding the explicitly dependence on the kernelized between and within scatter matrices.
  - **More efficient computation** of the Kernel Fisher direction when the classes are highly unbalanced.

Thank you!