ℓ_p Norm Multiple Kernel Fisher Discriminant Analysis for Object and Image Categorisation

Fei Yan Krystian Mikolajczyk Mark Barnard Hongping Cai Josef Kittler
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, Surrey, GU2 7XH, UK
{f.yan, k.mikolajczyk, mark.barnard, hongping.cai, j.kittler}@surrey.ac.uk

Abstract

In this paper, we generalise multiple kernel Fisher discriminant analysis (MK-FDA) such that the kernel weights can be regularised with an ℓ_p norm for any $p \geq 1$, in contrast to existing MK-FDA that uses either l1 or l2 norm. We present formulations for both binary and multiclass cases and solve the associated optimisation problems efficiently with semi-infinite programming. We show on three object and image categorisation benchmarks that by learning the intrinsic sparsity of a given set of base kernels using a validation set, the proposed ℓ_p MK-FDA outperforms its fixed-norm counterparts, and is capable of producing state-of-the-art performance. Moreover, we show that our ℓ_p MK-FDA outperforms the ℓ_p multiple kernel support vector machine (ℓ_p MK-SVM) which has been recently proposed. Based on this observation and our experience with single kernel FDA and SVM, we argue that the almost century-old FDA is still a strong competitor of the popular SVM.

1. Introduction

Object and image categorisation is one of the most active fields in computer vision. Recently, this area of research has seen rapid progress due to advances in both feature design [18, 20, 25] and machine learning. In the latter, continuing the success of kernel methods [26, 27], multiple kernel learning (MKL) [16] has been reported to produce state-of-the-art performance on several benchmarks [30, 9, 32]. These MKL techniques are essentially multiple kernel support vector machines (MK-SVMs) in the sense that they maximise the SVM [26, 27] type of objective, i.e., the margin between two classes.

In contrast to SVM, Fisher discriminant analysis (FDA) [8] maximises the ratio of projected between and within class scatters. Since its introduction in the 1930s, FDA has stood the test of time. Equipped recently with kernelisation [19, 2] and efficient implementation [3], FDA has established itself as a strong competitor of SVM. In many comparative studies, FDA is reported to offer comparable or even better performance than SVM [19, 3, 33].

In [13, 33], a multiple kernel FDA (MK-FDA) is introduced, where an ℓ_1 norm is used to regularise the kernel weights. ℓ_1 regularisation tends to produce sparse selection results, which may lead to a loss of information. [31] recently proposed to replace the ℓ_1 regularisation with an ℓ_2 version. Experiments in [31] suggest that the regularisation norm can have a significant impact on the classifier performance, and one should choose ℓ_1 or ℓ_2 regularisation based on the intrinsic sparsity of the given set of base kernels. The ℓ_2 MK-FDA formulation in [31] is only for binary problems.

In this paper, we extend [13, 33, 31] to a general ℓ_p norm regularisation. This is achieved by approximating the ℓ_p norm constraint in the optimisation problem using Taylor expansion. We carry out experiments on three object and image categorisation benchmarks and show that by selecting the regularisation norm p on an independent validation set, the intrinsic sparsity of the given set of base kernels can be learnt. As a result, using the learnt optimal norm p in the proposed ℓ_p MK-FDA offers better performance than ℓ_1, ℓ_2, or ℓ_∞ MK-FDAs. In particular, we show that when applied to carefully designed kernels, such a scheme is capable of producing state-of-the-art performance. Moreover, we show that our ℓ_p MK-FDA outperforms the ℓ_p MK-SVM [15] which has been recently proposed.

The rest of this paper is organised as follows. In Section 2, we introduce previous work that is related to this paper. We then present ℓ_p MK-FDA, first for binary case then for multiclass case, in Section 3. Experimental evidence showing the advantage of ℓ_p MK-FDA over other MKL techniques is provided in Section 4. Finally conclusions are given in Section 5.

2. Related Work: Multiple Kernel Learning

Let us for now consider a binary classification problem. Suppose one is given n $m \times m$ training kernel matrices $K_j, j = 1, \cdots, n$ and m class labels $y_i \in \{1, -1\}, i = 1, \cdots, m$, where m is the number of training samples. The original formulation of MKL [16] considers a linear convex combination of these n base kernels: $K = \sum_{j=1}^{n} \beta_j K_j, \quad \beta_j \geq 0, \quad ||\boldsymbol{\beta}||_1 = 1$. Geometrically, taking the sum of two kernel matrices can be interpreted as taking the Cartesian product of the two associated feature spaces. The goal of MKL is then to learn the "optimal" scaling of the feature spaces, such that the "separation" of the two classes in the augmented feature space is maximised.

[16] proposes to use the margin as a measure of separation and formulates the resulting ℓ_1 MK-SVM optimisation problem as a semi-definite program (SDP). The efficiency of ℓ_1 MK-SVM was improved significantly in later works [1, 28, 23]. ℓ_1 regularisation is known to produce sparse solutions [24], which may not always be desirable since the information carried in the zero-weighted kernels is lost. To overcome this problem, non-sparse MK-SVMs based on ℓ_2 regularisation and the general case of ℓ_p ($p \geq 1$) regularisation have been proposed in [14, 15]. Other works on the regularisation norm in MK-SVM include composite kernel learning [29] and mixed norm kernel learning [21].

In parallel to MK-SVMs, another line of research focuses on multiple kernel learning for Fisher discriminant analysis [13, 33, 31]. In MK-FDA, the FDA type of class separation criterion, i.e., the ratio of the projected between and within class scatters, is considered instead of the margin criterion in SVM. The ℓ_1 MK-FDA in [33] is derived for both binary and multiclass cases. However, similar to ℓ_1 MK-SVM, it suffers from the "over-selectiveness" problem. This is overcome by its ℓ_2 counterpart in [31], but the formulation in [31] is for binary problems only. It is thus the goal of this paper to extend existing MK-FDA methods to a general ℓ_p regularisation for both binary and multiclass problems.

3. ℓ_p Norm Multiple Kernel FDA

In this section we first present our ℓ_p MK-FDA for binary problems and then for multiclass problems. In both cases, we first give problem formulation, then solve the associated optimisation problem using semi-infinite programming.

3.1. Binary Case

Problem formulation Our goal is to learn optimal kernel weights $\boldsymbol{\beta} \in \mathbb{R}^n$ for the linear combination of n kernels under the ℓ_p constraint: $K = \sum_{j=1}^{n} \beta_j K_j, \beta_j \geq 0, ||\boldsymbol{\beta}||_p^p \leq 1$ for any $p \geq 1$, such that the ratio criterion of FDA is maximised. The $p \geq 1$ requirement is to ensure that the triangle inequality is satisfied and that $||\cdot||_p$ defines a norm.

Let m^+ be the number of positive training samples, and $m^- = m - m^+$ the number of negative training samples. For a given kernel K, we assume it has been centred in its feature space [26]. Let μ^+ and μ^- be the centroids of the positive and negative samples in the feature space, respectively, and C^+ and C^- be the covariance matrices of the two classes, respectively. The between class scatter S_B and within class scatter S_w are defined as:

$$S_B = \frac{m^+ m^-}{m}(\mu^+ - \mu^-)(\mu^+ - \mu^-)^T \quad (1)$$

$$S_W = m^+ C^+ + m^- C^- \quad (2)$$

The objective of single kernel FDA is to find the projection direction $\mathbf{w}$ in the feature space that maximises $\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$, or equivalently, $\frac{\mathbf{w}^T \frac{m}{m^+ m^-} S_B \mathbf{w}}{\mathbf{w}^T S_T \mathbf{w}}$, where $S_T = S_B + S_W$ is the total scatter matrix. In practice a regularised version,

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^T \frac{m}{m^+ m^-} S_B \mathbf{w}}{\mathbf{w}^T (S_T + \lambda I) \mathbf{w}} \quad (3)$$

is maximised to improve generalisation and numerical stability [19], where λ is a small positive number.

Exploring the link between FDA and regularised least squares (RLS) and using the duality theory of optimisation, it is proved in [33] that the maximal value of (3) is given by (up to an additive constant determined by the labels):

$$J_1^* \sim \min_{\boldsymbol{\alpha}} (\frac{1}{4}\boldsymbol{\alpha}^T (I + \frac{1}{\lambda} K)\boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{a}) \quad (4)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\mathbf{a} = (\frac{1}{m^+}, \cdots, \frac{1}{m^+}, \frac{-1}{m^-}, \cdots, \frac{-1}{m^-})^T \in \mathbb{R}^m$ contains the centred labels. Now consider the case where the kernel K can be chosen from linear combinations of a set of base kernels. The kernel weights must be regularised somehow to make sure (4) remains meaningful and does not become arbitrarily large. In this paper, we propose to impose an ℓ_p regularisation on the kernel weights for any $p \geq 1$: $\mathcal{K} = \{K = \sum_{j=1}^{n} \beta_j K_j : \boldsymbol{\beta} \geq \mathbf{0}, ||\boldsymbol{\beta}||_p^p \leq 1\}$. It directly follows that the optimal K maximising (4) is found by solving:

$$\max_{\boldsymbol{\beta}} \min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad \text{s.t. } \boldsymbol{\beta} \geq \mathbf{0}, \ ||\boldsymbol{\beta}||_p^p \leq 1 \quad (5)$$

where

$$S(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{4\lambda}\boldsymbol{\alpha}^T \sum_{j=1}^{n} \beta_j K_j \boldsymbol{\alpha} + \frac{1}{4}\boldsymbol{\alpha}^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{a} \quad (6)$$

Note that putting an ℓ_p constraint on $\boldsymbol{\beta}$ or penalizing $\mathbf{w}$ by an ℓ_q norm are equivalents with $p = q/(2-q)$ [15, 29]. When $p = 1$ we have the ℓ_1 MK-FDA; while $p = \infty$ leads to $q = 2$, and MK-FDA reduces to the regular kernel FDA with concatenation of feature spaces. In this paper, however, we are interested in the general case of any $p \geq 1$.

(5) is an optimisation problem with a quadratic objective and a general p^{th} order constraint. We exploit the idea from ℓ_p MK-SVM [15] and use second order Taylor expansion to approximate the norm constraint:

$$||\boldsymbol{\beta}||_p^p \approx \frac{p(p-1)}{2}\sum_{j=1}^{n}\tilde{\beta}_j^{p-2}\beta_j^2 - \sum_{j=1}^{n}p(p-2)\tilde{\beta}_j^{p-1}\beta_j + \frac{p(p-3)}{2} + 1 := \nu(\boldsymbol{\beta}) \quad (7)$$

where $\tilde{\beta}_j$ is the current estimate of β_j in an iterative process, which will be explained in more detail in the next section. Substituting (7) into (5) we arrive at the binary ℓ_p MK-FDA optimisation problem:

$$\max_{\boldsymbol{\beta}} \min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad \text{s.t. } \boldsymbol{\beta} \geq \mathbf{0}, \ \nu(\boldsymbol{\beta}) \leq 1 \quad (8)$$

Solving the optimisation problem with SIP A semi-infinite program (SIP) is an optimisation problem with finite number of variables $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ on a feasible set described by infinitely many constraints [12]. Following the same arguments as in [28, 33], it is straightforward to show that (8) is equivalent to a SIP:

$$\max_{\theta,\boldsymbol{\beta}} \quad \theta \quad (9)$$
$$\text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \quad \nu(\boldsymbol{\beta}) \leq 1, \quad S(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \theta \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^m$$

We adapt the wrapper algorithm proposed in [28] to solve (9). This algorithm is based on a technique called column generation, where the basic idea is to divide a SIP into an inner sub-problem and an outer sub-problem. The algorithm alternates between solving the two sub-problems until convergence. At step t, assuming the current optimal $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$ have been obtained in the outer sub-problem, the inner sub-problem identifies the constraint that maximises the constraint violation for $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$:

$$\boldsymbol{\alpha}^{(t)} := \arg\min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(t)}) \quad (10)$$

Observing that (10) is an unconstrained quadratic program, $\boldsymbol{\alpha}^{(t)}$ is obtained by solving the following linear system [33]:

$$(\frac{1}{2}I + \frac{1}{2\lambda}\sum_{j=1}^{n}\beta_j^{(t)}K_j)\boldsymbol{\alpha}^{(t)} = \mathbf{a} \quad (11)$$

If $\boldsymbol{\alpha}^{(t)}$ satisfies constraint $S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}) \geq \theta^{(t)}$ then solution $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$ is optimal. Otherwise, the constraint is added to the set of constraints and the algorithm proceeds to the outer sub-problem of step $t+1$.

The outer sub-problem is also called the restricted master problem. At step t, it computes the optimal $(\theta^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$ in (9) for a restricted subset of constraints:

$$(\theta^{(t+1)}, \boldsymbol{\beta}^{(t+1)}) = \arg\max_{\theta,\boldsymbol{\beta}} \theta \quad (12)$$
$$\text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \ \nu(\boldsymbol{\beta}) \leq 1, \ S(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}) \geq \theta \ \forall r = 1, \cdots, t$$

Table 1. An iterative algorithm for solving the SIP problem (9)

- **Initialisation**: $S^{(0)} = 1$, $\theta^{(1)} = -\infty$, $\beta_j^{(1)} = n^{-1/p}$ for $j = 1, \cdots, n$
- **for** $t = 1, 2, \cdots$ **do**
  - Compute $\boldsymbol{\alpha}^{(t)} = \arg\min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(t)})$ using (11)
  - Compute $S^{(t)} := S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$
  - **if** $|1 - \frac{S^{(t)}}{\theta^{(t)}}| \leq \epsilon$ **break**
  - Compute $(\theta^{(t+1)}, \boldsymbol{\beta}^{(t+1)}) = \arg\max_{\theta,\boldsymbol{\beta}} \theta$ in (12), where $\nu(\boldsymbol{\beta})$ is defined as in (7) with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$.
- **end for**

When $p = 1$, $\nu(\boldsymbol{\beta}) \leq 1$ reduces to a linear constraint. As a result, (12) becomes a linear program (LP) and ℓ_p MK-FDA reduces to the ℓ_1 MK-FDA in [33]. When $p > 1$, (12) is a quadratically constrained linear program (QCLP) with one quadratic constraint $\nu(\boldsymbol{\beta}) \leq 1$ and $t + n$ linear constraints. This can be solved by off-the-shelf optimisation tools such as Mosek [1]. Note that at time $t+1$, $\nu(\boldsymbol{\beta})$ is defined as in (7) with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$, i.e., the current estimate of $\boldsymbol{\beta}$.

Normalised maximal constraint violation is used as a convergence criterion. The algorithm stops when $|1 - \frac{S^{(t)}}{\theta^{(t)}}| \leq \epsilon$, where $S^{(t)} := S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$ and ϵ is a pre-defined accuracy parameter. This iterative algorithm for solving the ℓ_p binary MK-FDA SIP problem is summarised in Table 1. It is a special case of a set of SIP algorithms known as exchange methods, which are guaranteed to converge [12].

3.2. Multiclass Case

In this section we consider the multiclass case. Let c be the number of classes, and m_k be the number of training samples in the k^{th} class. In multiclass FDA, the following objective is commonly maximised [33]:

$$J_2(W) = \text{trace}(\frac{W^T S_B W}{W^T(S_T + \lambda I)W}) \quad (13)$$

where W is the projection matrix, and $S_T = S_B + S_W$. More specifically, S_W is defined in a similar way as in (2) but with c classes, and $S_B = \phi(X)HH^T\phi(X)^T$, where $\phi(X) = (\phi(x_1), \phi(x_2), \cdots, \phi(x_m))$ is the set of m training samples in the feature space, and $H = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_c)$ is an $m \times c$ matrix with $\mathbf{h}_k$ defined as:

$$\mathbf{h}_k(i) = \begin{cases} \sqrt{\frac{m}{m_k}} - \sqrt{\frac{m_k}{m}} & \text{if } y_i = k \\ -\sqrt{\frac{m_k}{m}} & \text{if } y_i \neq k \end{cases}$$

[1] http://www.mosek.com

Similar to the binary case, using duality theory and the connection between FDA and RLS, [33] shows that the maximal value of (13) is given by (up to an additive constant determined by the labels):

$$J_2^* \sim \min_{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_c} \sum_{k=1}^{c} (\frac{1}{4}\boldsymbol{\alpha}_k^T (I + \frac{1}{\lambda} K)\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^T \mathbf{h}_k) \quad (14)$$

where $\boldsymbol{\alpha}_k \in \mathbb{R}^m$ for $k = 1, \cdots, c$. When choosing from linear combinations of a set of base kernels with kernel weights regularised with an ℓ_p norm, we use again second order Taylor expansion (7) to approximate the norm constraint and arrive at the multiclass ℓ_p MK-FDA optimisation problem:

$$\max_{\boldsymbol{\beta}} \min_{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_c} S(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c, \boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta} \geq \mathbf{0},\ \nu(\boldsymbol{\beta}) \leq 1 \quad (15)$$

where $\nu(\boldsymbol{\beta})$ is defined as in (7) and

$$\begin{aligned} & S(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c, \boldsymbol{\beta}) \quad (16) \\ = & \sum_{k=1}^{c} (\frac{1}{4\lambda}\boldsymbol{\alpha}_k^T \sum_{j=1}^{n} \beta_j K_j \boldsymbol{\alpha}_k + \frac{1}{4}\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^T \mathbf{h}_k) \end{aligned}$$

Again similar to the binary case, (15) can be formulated as a SIP:

$$\begin{gathered} \max_{\theta,\boldsymbol{\beta}} \quad \theta \quad (17) \\ \text{s.t. } \boldsymbol{\beta} \geq \mathbf{0},\ \ \nu(\boldsymbol{\beta}) \leq 1,\ \ S(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c, \boldsymbol{\beta}) \geq \theta \\ \forall \boldsymbol{\alpha}_k \in \mathbb{R}^m, k = 1, \cdots, c \end{gathered}$$

and the SIP (17) can be solved with the same column generation algorithm in Table 1. In the inner sub-problem, the only difference is that here c linear systems need to be solved, one for each $\mathbf{h}_k$:

$$(\frac{1}{2} I + \frac{1}{2\lambda} \sum_{j=1}^{n} \beta_j^{(t)} K_j) \boldsymbol{\alpha}_k^{(t)} = \mathbf{h}_k \quad (18)$$

When $p = 1$, the outer sub-problem reduces to an LP and our formulation reduces to that in [33]. For $p > 1$, the outer sub-problem is a QCLP with one quadratic constraint and $t + n$ linear constraints, as in the binary case. This is easily seen by rearranging (16):

$$\begin{aligned} & S(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_c, \boldsymbol{\beta}) \quad (19) \\ = & \sum_{j=1}^{n} \beta_j (\sum_{k=1}^{c} \frac{1}{4\lambda} \boldsymbol{\alpha}_k^T K_j \boldsymbol{\alpha}_k) + \sum_{k=1}^{c} (\frac{1}{4}\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^T \mathbf{h}_k) \end{aligned}$$

4. Experiments

In this section we first show experimental results of the binary formulation on PASCAL VOC2007 [6], and then the results of the multiclass formulation on Caltech101 [7] and Oxford Flower17 [22]. For both formulations, kernels are centred in its feature space, and normalised to have a unit trace.

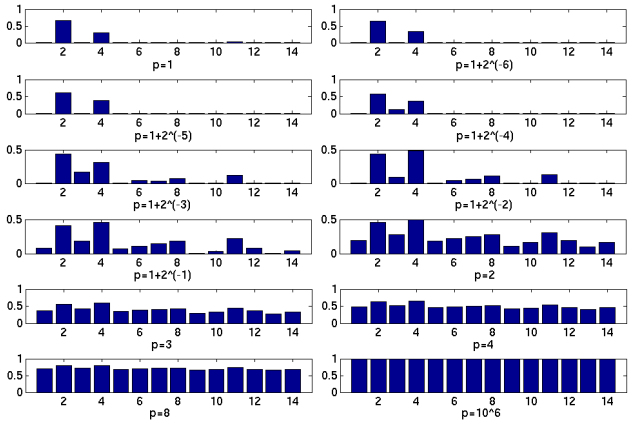

Figure 1. VOC2007: kernel weights learnt on the training set in ℓ_p MK-FDA with various p values. "cat" class.

4.1. Pascal VOC2007

Pascal VOC2007 is a multilabel dataset consisting of 20 object categories. It is divided into training, validation and test sets, with 2501, 2510 and 4952 images respectively. To tackle this multilabel problem, the classification of the 20 object classes is treated as 20 independent binary problems, and average precision (AP) is used to measure the performance of each binary classifier.

We generate 14 base kernels by combining 7 colour variants of local descriptors in [25] and two distance functions, namely, spatial pyramid match kernel (SPMK) [17, 11] and radial basis function (RBF) kernel with χ^2 distance [34]. We first perform supervised dimensionality reduction to improve its discriminability following [4]. The descriptors with reduced dimensionality are clustered with k-means to learn codewords [5]. The soft assignment scheme in [10] is then exploited to generate a histogram for each image as its representation. Finally, the two distance functions are applied to the histograms to build kernels.

We study the effect of the regularisation norm, and compare the performance of ℓ_p MK-SVM [15] and the proposed ℓ_p MK-LDA. We used the ℓ_p MK-SVM implementation in the Shogun machine learning toolbox [28]. For ℓ_p MK-FDA, we implemented it in Matlab, and the associated optimisation problem was solved using the Mosek software. For both methods, we learn the parameter p on the validation set from 12 values: $\{1, 1+2^{-6}, 1+2^{-5}, 1+2^{-4}, 1+2^{-3}, 1+2^{-2}, 1+2^{-1}, 2, 3, 4, 8, 10^6\}$. In ℓ_p MK-SVM, the trade-off parameter C is learnt jointly with p from 10 values: $\{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$. Similarly, in ℓ_p MK-FDA, the parameter λ, which essentially serves the same function as C, is learnt jointly with p from 10 values that are logarithmically evenly spaced over 10^{-8} to 10^{+1}.

Plotted in Fig. 1 are the weights learnt on the training set in ℓ_p MK-FDA with various p values for the "cat" class, where for each p value, the weights learnt with the opti-

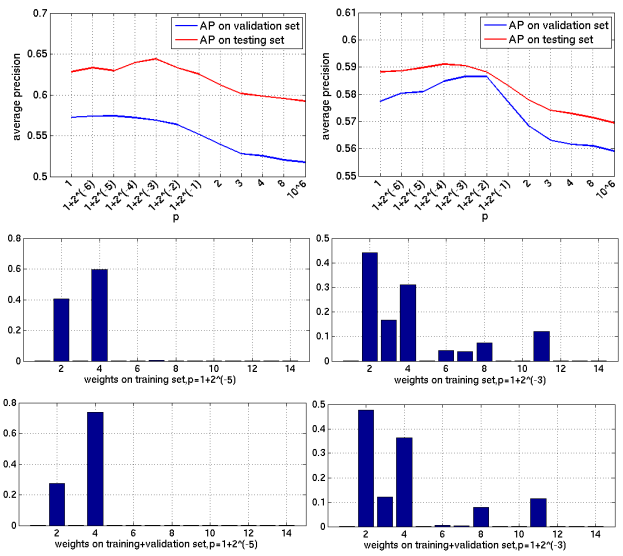

Figure 2. VOC2007: learning the norm p on a validation set. Left column: "dinningtable" class. Right column: "cat" class. Top row: APs on the validation set and test set with various p values; middle row: kernel weights on the training set with the optimal $\{p, \lambda\}$ combination; bottom row: kernel weights on the training+validation set with the same $\{p, \lambda\}$ combination.

mal λ value are plotted. It is clear that as p increases, the sparsity of the learnt weights decreases. As expected, when $p = 10^6$ (practically infinity), the kernels weights become ones, i.e., ℓ_p MK-FDA becomes ℓ_∞ MK-FDA. Note also that $p = 1$ and $p = 2$ are equivalent to ℓ_1 and ℓ_2 MK-FDAs respectively.

Next, we plot in Fig. 2 top-left the APs on the validation set and test set for the "dinningtable" class with various p values, where again for each p value, the APs with the λ value that gives the best AP on the validation set are plotted. It is clear that the two curves match well, which implies that learning p in addition to λ should help. Shown in the middle and bottom rows of the same column are the learnt kernel weights with the optimal $\{p, \lambda\}$ combination on the training set and on the training + validation set, respectively. Since for the "dinningtable" class the optimal p found on the validation set is $1+2^{-5}$, both sets of weights are sparse. This means for this particular class, the intrinsic sparsity of the set of base kernels is high.

Similarly, the right column of Fig. 2 shows the results for the "cat" class. We observe again that the AP on the validation set and that on the test set show similar patterns. However, for the "cat" class, the optimal p on the validation set is found to be $1 + 2^{-3}$, which implies that the intrinsic sparsity of the kernels is lower.

When keeping the norm p fixed at 1, 2, 10^6 and learning only the C/λ parameter, the ℓ_p MK-SVM/MK-FDA reduces to ℓ_1, ℓ_2 and ℓ_∞ MK-SVM/MK-FDA, respectively.

Table 2. VOC2007: Average precisions of 8 MKL methods

	MK-SVM				MK-FDA			
	ℓ_1	ℓ_2	ℓ_∞	ℓ_p	ℓ_1	ℓ_2	ℓ_∞	ℓ_p
aeroplane	78.8	79.7	79.6	79.6	**80.4**	80.1	79.5	**80.4**
bicycle	63.4	64.7	65.0	64.7	**69.9**	68.5	67.6	**69.9**
bird	57.3	60.6	61.0	61.0	61.6	63.6	61.9	**64.8**
boat	71.1	70.2	70.1	71.1	**72.4**	71.2	70.0	**72.4**
bottle	29.1	29.7	29.9	29.7	29.1	**30.4**	29.7	29.9
bus	62.9	64.2	64.9	65.5	66.2	**67.5**	66.1	66.7
car	77.9	78.6	78.8	78.8	81.4	80.8	79.5	**81.9**
cat	56.7	56.4	56.4	57.1	57.8	57.8	56.9	**58.8**
chair	52.3	52.8	53.0	53.0	**53.5**	53.3	52.5	**53.5**
cow	38.7	40.3	41.4	41.4	**46.4**	43.6	41.5	**46.4**
din. table	52.4	56.1	57.3	56.6	62.5	61.2	59.2	**62.8**
dog	42.8	43.9	45.8	44.6	46.0	46.0	**46.1**	45.9
horse	78.9	80.2	80.6	80.6	81.0	81.6	81.1	**82.2**
moterbike	66.3	66.6	66.8	66.8	67.7	68.6	67.7	**69.5**
person	86.7	87.8	88.0	88.0	89.1	88.8	88.1	**89.3**
pot. plant	31.8	39.7	41.0	40.5	41.2	**43.1**	42.6	39.5
sheep	40.2	44.8	46.0	46.0	47.0	46.4	44.4	**49.5**
sofa	44.0	43.2	43.8	44.0	43.9	45.4	43.7	**46.8**
train	81.3	82.2	82.4	82.4	**85.2**	85.0	84.2	85.1
tvmonitor	53.3	53.2	53.7	53.7	55.2	55.8	54.1	**56.6**
MAP	58.3	59.8	60.3	60.3	61.9	61.9	60.8	**62.6**

The APs and mean APs (MAPs) of the 8 MKL methods are shown in Table 2. The results in Table 2 demonstrate that learning p indeed improves the performance of MK-FDA. It is also worth noting that the performance of ℓ_p MK-FDA is comparable to that in [32], which is the best reported on this benchmark to the best of our knowledge. Another interesting observation is that for any norm MK-FDA consistently outperforms MK-SVM, which is widely considered the state-of-the-art classification method.

The algorithm described in Table 1 for solving the SIP in ℓ_p MK-FDA is known to be efficient [28, 33]. In our experiments, we observe that the speed of our implementation of ℓ_p MK-FDA is comparable to that of the ℓ_p MK-SVM in the Shogun toolbox. For each binary problem in the VOC2007 dataset, both methods take approximately 500 seconds to learn the weights of the 14 kernels on a single core of an AMD Opteron Processor. For both methods, the stopping threshold is set to 10^{-4}.

4.2. Caltech101

Caltech101 is a multiclass object recognition benchmark with 101 object categories. For multiclass problems, only the ℓ_1 and ℓ_∞ MK-SVMs have been implemented in the Shogun toolbox. As a result, we only compare our multiclass ℓ_p MK-FDA with ℓ_1, ℓ_2, ℓ_∞ MK-FDAs and ℓ_1, ℓ_∞ MK-SVMs. We follow the popular practice of using 15 randomly selected images per class for training, up to 50 randomly selected images per class for testing, and computing the average accuracy over all classes. This process

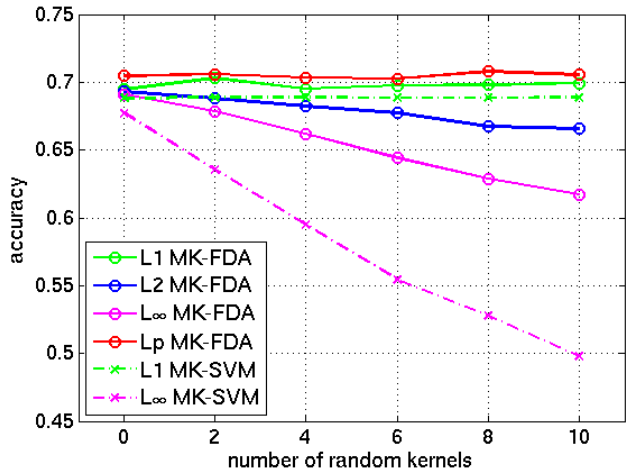

Figure 3. Caltech101: Accuracy of ℓ_1, ℓ_2, ℓ_∞, ℓ_p MK-FDAs and ℓ_1, ℓ_∞ MK-SVMs with various noise levels.

is repeated 3 times, and we report the mean of the average accuracies. In order to learn the parameters (p and λ for ℓ_p MK-FDA, λ for ℓ_1, ℓ_2, ℓ_∞ MK-FDAs, and C for ℓ_1, ℓ_∞ MK-SVMs), in each of the 3 runs, we randomly split the 15 training images from each class into a training set of 10 images and a validation set of 5 images. This process is repeated 3 times, and the mean of the average accuracies on the validation set is used for choosing the parameters.

We generate kernels in a similar way as in the VOC2007 experiments. In addition to these "informative" kernels, we also construct 10 RBF kernels from 10 sets of random vectors. To test the robustness of the ℓ_p MK-FDA, we repeat experiments 6 times. We start with only the informative kernels, and add two more random kernels in each subsequent run. p, λ and C are learnt from the same sets of values as in the VOC2007 experiments.

The performance of the four MK-FDAs and the two MK-SVMs with various numbers of random kernels is shown in Fig. 3. As expected, ℓ_1 MK-FDA and ℓ_1 MK-SVM are very robust to noise, while the performance of ℓ_2 MK-FDA, ℓ_∞ MK-FDA and ℓ_∞ MK-SVM drops significantly as the noise level increases. On the other hand, by tuning the regularisation norm p the intrinsic sparsity of a kernel set can be learnt. As a result, ℓ_p MK-FDA outperforms all the fixed-norm MKL methods regardless of the level of noise in the kernel set.

We can also observe that both ℓ_1 MK-FDA and ℓ_∞ MK-FDA outperform their MK-SVM counterparts. This is consistent with the VOC07 results. We believe this observation is important given that SVM and SVM based MKL are widely accepted as the state-of-the-art classifier in almost all object categorisation systems, and it highlights the significance of our contribution with the proposed ℓ_p MK-FDA.

Table 3. Flower17: Comparison of 8 MKL methods

method	accuracy	parameters
product	85.5 ± 1.2	C
averaging	84.9 ± 1.9	C
MKL (SILP)	85.2 ± 1.5	C
MKL (Simple)	85.2 ± 1.5	C
CG-Boost	84.8 ± 2.2	C
LP-β	85.5 ± 3.0	$C_j, j = 1, \cdots, n$ and $\delta \in (0, 1)$
LP-B	85.4 ± 2.4	$C_j, j = 1, \cdots, n$ and $\delta \in (0, 1)$
ℓ_p MK-FDA	$\mathbf{86.7 \pm 1.2}$	p and λ jointly

4.3. Oxford Flower17

Oxford Flower17 dataset consists of flower images from 17 categories of flowers with 80 images per category. This dataset comes with three predefined splits into train (17×40 images), validation (17×20 images) and test (17×20 images) sets. Moreover the authors of [22] precomputed 7 distance matrices using various features, and put the matrices online [2]. We downloaded these distance matrices and followed the same procedure as in [9] to compute 7 kernel matrices: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-D(\mathbf{x}_i, \mathbf{x}_j)/\gamma)$, where γ is the mean of the pairwise distances.

Table 3 compares the ℓ_p MK-FDA proposed in this paper and 7 kernel combination techniques discussed in [9]. Note that these methods are directly comparable since they share the same kernel matrices and the same splits. In Table 3, the performance of ℓ_p MK-FDA is achieved by learning p and λ from the same sets of values as in the VOC2007 and Caltech101 experiments. For the other 7 methods, the corresponding entries are taken directly from [9].

In Table 3, "product" and "sum" refer to the two simplest kernel combination methods, namely, taking the element-wise geometric mean and arithmetic mean of the kernels, respectively; "MKL (SILP)" and "MKL (Simple)" are essentially ℓ_1 MK-SVM; while "CG-Boost", "LP-β" and "LP-B" are three boosting based kernel combination methods. We can see that these boosting based methods, although performing well on other datasets according to [9], fail to outperform the baseline methods "product" and "averaging". The ℓ_p MK-FDA on the other hand shows a relatively significant improvement over all the methods discussed in [9]. The optimal p value is found to be 1.5. The number of parameters that need to be learnt in these methods is also compared in Table 3.

5. Conclusions

In this paper, we have generalised MK-FDA such that the kernel weights can be regularised with an ℓ_p norm for any $p \geq 1$. We have presented formulations for both binary and multiclass cases and solved the associated optimisation

[2] http://www.robots.ox.ac.uk/ vgg/research/flowers/index.html

problems efficiently with semi-infinite programming. We have demonstrated on three object and image categorisation benchmarks that by learning the intrinsic sparsity of a given set of base kernels using a validation set, the proposed ℓ_p MK-FDA outperforms its fixed-norm counterparts, and is capable of producing state-of-the-art performance. Moreover, we have shown that our ℓ_p MK-FDA outperforms the ℓ_p MK-SVM from [15]. Based on this observation and our experiments with single kernel FDA and SVM, we argue that the almost century-old FDA is still a strong competitor of the popular SVM. Code for the proposed ℓ_p MK-FDA and kernels for the datasets used in this paper are available online at http://www.featurespace.org.

Acknowledgement

This work has been supported by EU IST-2-045547 VIDI-Video Project.

References

[1] F. Bach and G. Lanckriet. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

[2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

[3] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. In *International Conference on Data Mining*, 2007.

[4] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *PAMI*, 2010.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV workshop on Statistical Learning in Computer Vision*, 2004.

[6] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[7] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[8] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[9] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[10] J. Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[11] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[12] R. Hettich and K. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 35(3):380–429, 1993.

[13] S. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel fisher discriminant analysis. In *ICML*, 2006.

[14] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg. Non-sparse multiple kernel learning. In *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

[15] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Efficient and accurate lp-norm mkl. In *NIPS*, 2009.

[16] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] D. Lowe. Distincetive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[19] S. Mika. Kernel fisher discriminants. PhD Thesis, University of Technology, Berlin, Germany, 2002.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[21] J. Nath, G. Dinesh, S. Raman, C. Bhattacharyya, A. Ben-Tal, and K. Ramakrishnan. On the algorithmics and applications of a mixed-norm based kernel learning formulation. In *NIPS*, 2009.

[22] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

[23] A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu. Simplemkl. *JMLR*, 9:2491–2521, 2008.

[24] G. Ratsch. Robust boosting via convex optimization. PhD Thesis, University of Potsdam, Potsdam, Germany, 2001.

[25] K. Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *CVPR*, 2008.

[26] B. Scholkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[27] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[28] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[29] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. In *ICML*, pages 1040–1047, 2008.

[30] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV*, 2007.

[31] F. Yan, J. Kittler, K. Mikolajczyk, and A. Tahir. Non-sparse multiple kernel learning for fisher discriminant analysis. In *International Conference on Data Mining*, 2009.

[32] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *ICCV*, 2009.

[33] J. Ye, S. Ji, and J. Chen. Multi-class discriminant kernel learning via convex programming. *JMLR*, 9:719–758, 2008.

[34] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007.