Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Absent Multiple Kernel Learning

Xinwang Liu
School of Computer
National University
of Defense Technology
Changsha, China, 410073

Lei Wang
School of Computer Science
and Software Engineering
University of Wollongong
NSW, Australia, 2522

Jianping Yin, Yong Dou
School of Computer
National University
of Defense Technology
Changsha, China, 410073

Jian Zhang
Faculty of Engineering
and Information Technology
University of Technology Sydney
NSW, Australia, 2007

Abstract

Multiple kernel learning (MKL) optimally combines the multiple channels of each sample to improve classification performance. However, existing MKL algorithms cannot effectively handle the situation where some channels are missing, which is common in practical applications. This paper proposes an absent MKL (AMKL) algorithm to address this issue. Different from existing approaches where missing channels are firstly imputed and then a standard MKL algorithm is deployed on the imputed data, our algorithm directly classifies each sample with its observed channels. In specific, we define a margin for each sample in its own relevant space, which corresponds to the observed channels of that sample. The proposed AMKL algorithm then maximizes the minimum of all sample-based margins, and this leads to a difficult optimization problem. We show that this problem can be reformulated as a convex one by applying the representer theorem. This makes it readily be solved via existing convex optimization packages. Extensive experiments are conducted on five MKL benchmark data sets to compare the proposed algorithm with existing imputation-based methods. As observed, our algorithm achieves superior performance and the improvement is more significant with the increasing missing ratio.

Introduction

Multiple kernel learning (MKL) has been an active topic in machine learning community during the last decade (Lanckriet et al. 2004; Rakotomamonjy et al. 2008; Cortes, Mohri, and Rostamizadeh 2012; Liu et al. 2012; Xu et al. 2010; Gönen 2012; Liu et al. 2014). Current research on MKL mainly focuses on improving the efficiency of MKL algorithms (Rakotomamonjy et al. 2008; Xu et al. 2010; Orabona and Luo 2011; Gönen and Alpaydın 2011), designing non-sparse and nonlinear MKL algorithms (Xu et al. 2010; Yan et al. 2012; Kloft et al. 2011), developing two-stage MKL algorithms (Cortes, Mohri, and Rostamizadeh 2012; Kumar et al. 2012) and integrating radius information into traditional margin-based MKL algorithms (Do et al. 2009; Gai, Chen, and Zhang 2010; Liu et al. 2013). These research works on MKL usually take the following assumption: all channels (or base kernels) of each sample are observed. However, this assumption may not hold anymore when some channels of samples become missing, which is common in neuroimaging (Yuan et al. 2012), computational biology (Chechik et al. 2008), medical analysis (Marlin 2008), etc. This issue is known as missing value (Smola, Vishwanathan, and Hofmann 2005; Marlin 2008) or absent data learning (Chechik et al. 2008), and it has been studied in the literature (Smola, Vishwanathan, and Hofmann 2005; Chechik et al. 2008; Ghahramani and Jordan 1993; Yuan et al. 2012). Nevertheless, the research on designing novel MKL algorithms to effectively handle absent channels has not been explored in existing MKL literature.

The violation to this assumption makes existing MKL algorithms unable to handle the above issue. A straightforward remedy may firstly impute the absent channels with zero (known as zero-filling in the literature) or mean value (mean-filling) or by more sophisticated expectation-maximization (EM-filling) (Ghahramani and Jordan 1993) approach. A standard MKL algorithm is then deployed on the imputed data. Such imputation methods usually work well when missing ratio is small. However, they may produce inaccurate imputation with the increase of missing ratio, and this could deteriorate the performance of the subsequent MKL.

Different from the imputation approach, this paper proposes to directly classify samples with absent channels without imputation. It is inspired by the concept of sample-based margin developed in (Chechik et al. 2008). In this paper, we first immigrate the concept of sample-based margin into multiple kernel-induced feature spaces, and maximize the minimum of all sample-based margins. On one hand, this approach is able to effectively handle the issue of absent channels. On the other hand, it yields a difficult optimization problem due to the minimization over all training samples. In this paper, we reformulate it as a convex optimization problem and make it readily solvable via off-the-shelf optimization packages. We highlight the main contributions of this paper as follows:

- Our work extends existing MKL algorithms by enabling them to handle samples with absent channels, which broadens the application scope of MKL.
- By applying the representer theorem (Schölkopf, Herbrich, and Smola 2001), we reformulate the optimization problem as a convex one in its primal space. It can be readily solved by existing convex optimization packages

such as CVX (CVX Research 2012).

- We conduct comprehensive experiments to compare the proposed algorithm with existing imputation-based methods on multiple MKL benchmark data sets. The results verify the superiority of the proposed algorithm, especially in the presence of intensive absence of channels.

We end up this section by discussing the differences between our work and the work in (Chechik et al. 2008) which inspires us. Both papers classify samples with missing observations by maximizing the sample-based margin. However, they have the following important differences: (1) Our algorithms directly work on kernel matrices and handle the case where some entries of kernel matrices are missing. However, the algorithm in (Chechik et al. 2008) only considers the absence of input features; (2) Different from developing a two-step iterative algorithm to solve the resulting optimization problem as in (Chechik et al. 2008), we design a novel algorithm that directly solves the problem in the primal space by employing the representer theorem (Schölkopf, Herbrich, and Smola 2001). More importantly, the latter consistently achieves significant improvements over the former, as validated by our experimental results; and (3) In addition, our work assumes that some *channels*[1] of samples are missing, while the work in (Chechik et al. 2008) studies the problem that some *individual features* of samples are missing. Form this perspective, the work in (Chechik et al. 2008) can be treated as a special case when a linear kernel is applied to all channels and each single feature is viewed as a channel.

Related Work

The Sample-Based Margin

The sample-based margin is firstly proposed in the seminal work (Chechik et al. 2008) and applied to absent data learning where some features of a sample are missing. An important assumption for the sample-based margin is that the learnt classifier should have consistent parameters across samples, even if those samples do not reside in the same space, i.e., having different sets of observed features. Based on this assumption, the margin $\rho_i(\boldsymbol{\omega})$ for the i-th $(1 \leq i \leq n)$ sample is defined as $\rho_i(\boldsymbol{\omega}) = y_i(\boldsymbol{\omega}^{i\top}\mathbf{x}_i)/\|\boldsymbol{\omega}^i\|$, where $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is a training data set, $\mathbf{x}_i$ is characterized by a subset of features from a full set $\mathcal{F}$, $y_i \in \{+1, -1\}$, $\boldsymbol{\omega}$ is the normal vector of SVMs on the full feature set $\mathcal{F}$, $\boldsymbol{\omega}^i$ is a vector obtained by taking the entries of $\boldsymbol{\omega}$ that are relevant to $\mathbf{x}_i$, namely those for which the sample $\mathbf{x}_i$ has observed features, and $\mathbf{f}(\mathbf{x}_i) = \boldsymbol{\omega}^{i\top}\mathbf{x}_i$ is the decision score of $\mathbf{x}_i$.

The above equation defines the margin for each sample in its own relevant space. As can be seen, it will reduce to a traditional margin as in SVMs when all samples have a full set $\mathcal{F}$. From this perspective, the work in (Chechik et al. 2008) provides an elegant approach to handling samples with absent features. Though bearing such advantages, the maximization over the sample-based margins makes the resultant optimization problem more difficult to solve than the one in traditional SVMs. In (Chechik et al. 2008), a two-step iterative optimization procedure is proposed to solve the problem. However, the optimization in (Chechik et al. 2008) is non-convex and the global optimum is not guaranteed to be obtained.

Multiple Kernel Learning

MKL provides an elegant framework which not only learns a data-dependent optimal kernel for a specific application but also integrates multiple heterogeneous channels (Rakotomamonjy et al. 2008; Xu et al. 2010; Cortes, Mohri, and Rostamizadeh 2010). In MKL, each sample can be treated as a concatenation of multiple base kernel mappings. Specifically, each sample in MKL takes the form of

$$\phi(\cdot) = [\phi_1^\top(\cdot), \phi_2^\top(\cdot), \cdots, \phi_m^\top(\cdot)]^\top, \tag{1}$$

where $\{\phi_p(\cdot)\}_{p=1}^m$ are the feature mappings corresponding to m pre-defined base kernels $\{\kappa_p(\cdot,\cdot)\}_{p=1}^m$, respectively. Based on this definition, the seminal work in MKL (Lanckriet et al. 2004) proposes to optimize the problem in Eq. (2),

$$\begin{aligned} &\min_{\boldsymbol{\omega},b,\boldsymbol{\xi}} \frac{1}{2}\left(\sum\nolimits_{p=1}^m \|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}\right)^2 + C\sum_{i=1}^n \xi_i, \\ &s.t.\ y_i\left(\sum\nolimits_{p=1}^m \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) + b\right) \geq 1-\xi_i,\ \xi_i \geq 0,\ \forall i, \end{aligned} \tag{2}$$

where $\boldsymbol{\omega}_p$ is the normal vector corresponding to the p-th base kernel, b is the bias, $\boldsymbol{\xi}$ consists of the slack variables and $\mathcal{H}_p$ is a Hilbert space corresponding to base kernel κ_p. Note that the MKL formulation in Eq. (2) can be equivalently rewritten as those in (Rakotomamonjy et al. 2008; Xu et al. 2010) according to (Xu et al. 2010).

As can be seen in Eq. (1), each sample in current MKL is assumed to be $\phi(\mathbf{x}_i) = [\phi_1^\top(\mathbf{x}_i), \phi_2^\top(\mathbf{x}_i), \cdots, \phi_m^\top(\mathbf{x}_i)]^\top$. A question naturally raised is that how to perform MKL when some channels, i.e., part of $\{\phi_p(\mathbf{x}_i)\}_{p=1}^m$, are absent in a sample[2]? In the following parts, we extend the sample-based margin to address this situation and call it absent multiple kernel learning (AMKL).

Absent Multiple Kernel Learning

The Sample-based Margin in AMKL

We are given n training samples $\{(\phi(\mathbf{x}_i), y_i)\}_{i=1}^n$ and a missing matrix $\mathbf{s} \in \{0,1\}^{n\times m}$, where $\phi(\mathbf{x}_i)$ is defined as in Eq. (1) and $s(i,p) \in \{0,1\}$ $(1 \leq i \leq n,\ 1 \leq p \leq m)$ indicates whether the p-th channel of the i-th sample is absent or not. Specifically, $s(i,p) = 0$ implies absent and $s(i,p) = 1$ otherwise.

Similar to (Chechik et al. 2008), we assume that a normal vector should be consistent across samples, no matter whether they have the same observed channels or not. Under this assumption, we define the margin for the i-th $(1 \leq$

[1] Each channel can be composed of a group of features.

[2] Note that the absence of channels essentially leads to missing entries in base kernel matrices in practice. We present this problem from the perspective of channel absence for the sake of clarity and convenience of analysis.

$i \leq n$) sample as,

$$\rho_i(\boldsymbol{\omega}) \quad = \quad \frac{y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) \right)}{\sum_{p=1}^{m} s(i,p) \|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}}, \tag{3}$$

where $\boldsymbol{\omega} = [\boldsymbol{\omega}_1^\top, \cdots, \boldsymbol{\omega}_m^\top]^\top$ and $\boldsymbol{\omega}_p \, (1 \leq p \leq m)$ are the normal vectors corresponding to the whole channels and the p-th channel, respectively. $\|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}$ is the norm in a Hilbert space induced by the p-th base kernel.

As can be seen, Eq. (3) defines the margin in multiple kernel-induced feature spaces for the samples with absent channels, i.e., some of $\{\phi_p(\mathbf{x}_i)\}_{p=1}^m \, (1 \leq i \leq n)$ are absent, as indicated by $s(i,p)$. At this point, we have extended the sample-based margin in (Chechik et al. 2008), where some features of the samples are missing, to MKL where some channels of the samples are missing. In AMKL, we propose to maximize the minimum of all sample-based margins so that the resultant classifier separates the two classes as far as possible. This objective is fulfilled as in Eq. (4),

$$\max_{\boldsymbol{\omega}} \left(\min_{1 \leq i \leq n} \frac{y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) \right)}{\sum_{p=1}^{m} s(i,p) \|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}} \right). \tag{4}$$

This problem is more difficult to solve than a traditional MKL one because the denominator varies with samples.

The proposed C-AMKL

We now show how to reformulate the optimization problem in Eq. (4) as a convex one such that it could be solved via existing convex optimization packages such as CVX (CVX Research 2012). We first express the optimization problem in Eq. (4) as a constrained one,

$$\max_{\boldsymbol{\omega}} \min_{1 \leq i \leq n} \frac{1}{\sum_{p=1}^{m} s(i,p) \|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}}, \; s.t. \; y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) \right) \geq 1, \tag{5}$$

which can be equivalently reformulated (Rakotomamonjy et al. 2008) as in Eq. (6),

$$\begin{aligned} \min_{\boldsymbol{\omega}, \boldsymbol{\gamma} \in \mathcal{Q}} \; & \max_i \; \frac{1}{2} \sum_{p=1}^{m} s(i,p) \frac{\|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}^2}{\gamma_p}, \\ & s.t. \; y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) \right) \geq 1, \forall i, \end{aligned} \tag{6}$$

where $\mathcal{Q} = \{\boldsymbol{\gamma} \in \mathbb{R}^m | \sum_{p=1}^m \gamma_p = 1, \, 0 \leq \gamma_p \leq 1, \, \forall p\}$.

The problem in Eq. (6) can be further rewritten as

$$\begin{aligned} \min_{\boldsymbol{\omega}, \boldsymbol{\gamma} \in \mathcal{Q}} \; & u \\ & s.t. \; y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) \right) \geq 1, \, \forall i, \\ & \frac{1}{2} \sum_{p=1}^{m} s(i,p) \frac{\|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}^2}{\gamma_p} \leq u, \, \forall i. \end{aligned} \tag{7}$$

After adding the bias term b and slack variables $\boldsymbol{\xi}$ to deal with non-separable cases, we obtain

$$\begin{aligned} \min_{\boldsymbol{\omega}, \boldsymbol{\gamma} \in \mathcal{Q}, b, \boldsymbol{\xi}, u} \; & u + C \sum_{i=1}^{n} \xi_i \\ & s.t. \; y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\omega}_p^\top \phi_p(\mathbf{x}_i) + b \right) \geq 1 - \xi_i, \, \xi_i \geq 0, \, \forall i, \\ & \frac{1}{2} \sum_{p=1}^{m} s(i,p) \frac{\|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}^2}{\gamma_p} \leq u, \forall i. \end{aligned} \tag{8}$$

As can be seen in Eq. (8), the objective function as well as the first and third constraints are all linear in variables. Besides, the second constraint, which is a quadratic cone, is also convex. As a consequence, the optimization problem in Eq. (8) is convex. However, solving such a problem directly is still difficult since the feature mapping functions $\{\phi_p(\cdot)\}_{p=1}^m$ are usually not explicitly known. A commonly used trick to handle this problem is to derive its dual problem where $\{\phi_p(\cdot)\}_{p=1}^m$ appears in the form of inner product in the kernel-induced feature space. Nevertheless, it is hard to solve its dual problem due to the complicated ratio optimization caused by the second constraint of Eq. (8). Here we show that these issues can be removed via applying the well-known representer theorem (Schölkopf, Herbrich, and Smola 2001). According to this theorem, a solution of $\boldsymbol{\omega}_p$ in Eq. (8) should take the form of

$$\boldsymbol{\omega}_p = \sum_{i=1}^{n} \alpha_i \kappa_p(\mathbf{x}_i, \cdot), \forall p \tag{9}$$

where $\kappa_p(\cdot, \cdot)$ is the p-th base kernel. According to the kernel reproducing property (Aronszajn 1950), we have

$$\mathbf{f}_p(\mathbf{x}) = \langle \boldsymbol{\omega}_p, \kappa_p(\mathbf{x}, \cdot) \rangle_{\mathcal{H}_p} = \sum_{i=1}^{n} \alpha_i \kappa_p(\mathbf{x}_i, \mathbf{x}), \tag{10}$$

and

$$\|\boldsymbol{\omega}_p\|_{\mathcal{H}_p}^2 = \langle \boldsymbol{\omega}_p, \boldsymbol{\omega}_p \rangle_{\mathcal{H}_p} = \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa_p(\mathbf{x}_i, \mathbf{x}_j). \tag{11}$$

With Eq. (10) and (11), the optimization problem in Eq. (8) can be equivalently rewritten as

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathcal{Q}, b, \boldsymbol{\xi}, u} \; & u + C \sum_{i=1}^{n} \xi_i \\ & s.t. \; y_i \left(\sum_{p=1}^{m} s(i,p) \boldsymbol{\alpha}^\top \mathbf{K}_p(:, \mathbf{x}_i) + b \right) \geq 1 - \xi_i, \xi_i \geq 0, \forall i, \\ & \frac{1}{2} \sum_{p=1}^{m} s(i,p) \frac{\boldsymbol{\alpha}^\top \mathbf{K}_p \boldsymbol{\alpha}}{\gamma_p} \leq u, \, \forall i, \end{aligned} \tag{12}$$

where $\mathbf{K}_p(:, \mathbf{x}_i) = [\kappa_p(\mathbf{x}_1, \mathbf{x}_i), \, \cdots, \, \kappa_p(\mathbf{x}_n, \mathbf{x}_i)]^\top$.

The optimization problem in Eq. (12) is called convex AMKL (C-AMKL) in this paper, which has some intuitive explanation. Specifically, the first constraint implies that the resultant classifier classifies samples with the observed channels, i.e., those for which $s(i,p) \neq 0$, without imputation. The second constraint takes the maximum of the reciprocal of each sample-based margin. After that, the objective function minimizes this maximum to control the generalization performance of the learnt classifier. We apply the CVX package to solve the problem in Eq. (12). After obtaining $\boldsymbol{\alpha}$ and b, the decision score of a test sample $\mathbf{x}_t$ is

$$\hat{y}(\mathbf{x}_t) = \sum_{p=1}^{m} t(p) \boldsymbol{\alpha}^\top \mathbf{K}_p(:, \mathbf{x}_t) + b, \tag{13}$$

where $\mathbf{t} \in \{0,1\}^m$ is a vector indicating the absence of the channels for $\mathbf{x}_t$.

Discussion

The difference between AMKL and zero-filling MKL (ZF-MKL) is the way in calculating the margin. ZF-MKL first imputes the absent channels with zeros and maximizes the margin defined on the imputed samples. Differently, in AMKL the margin of each sample is calculated in its own

relevant space and the minimum of these sample-based margins is maximized. Note that calculating the margin in a relevant space does not imply that the unobserved channels are set to zero. This difference can be clearly seen from Eq.(3). For ZF-MKL, all the term $s(i,p)$ in the denominator will simply become one and this will result in a denominator different from the one optimized by AMKL.

Besides the difference in calculating the margin, mean-filling MKL (MF-MKL) fills the absent channels with the value averaged on the samples for which the channels are observed. Note that both zero-filling and mean-filling may not be accurate in the presence of intensive channel absence, which would deteriorate the performance of the learnt classifier. The above drawback of ZF-MKL and MF-MKL will be shown in our experiments.

Experimental Results

In this section, we conduct experiments to compare the proposed C-AMKL, with the commonly used imputation methods, including ZF-MKL, MF-MKL and EM-MKL. The mean margin of all sample-based margins MKL (MM-MKL) is also included into comparison. A variant of C-AMKL, termed mean margins MKL (MM-MKL), which maximizes the mean (instead of the minimum) of all sample-based margins is also included for comparison. In addition, an algorithm called "classifiers fusion (CF)" is also compared as a baseline. In this algorithm, multiple classifiers are trained on each channel separately using available data only. In the test phase, the decision score of a test sample is the mean of scores calculated by the multiple classifiers.

We show how to construct the absent matrix $\mathbf{s} \in \{0,1\}^{n\times m}$ on the training data, where n and m are the number of training samples and channels. In specific, we randomly generate a row of $\mathbf{s}$ and set its first $round(\varepsilon_0 * m)$[3] smallest values as zeros and the rest as ones, respectively. We repeat this process for each row of $\mathbf{s}$. The absent matrix on test data is generated in the same way. The parameter ε_0, termed missing ratio in this paper, will affect the performance of the above algorithms. Intuitively, the larger the value of ε_0 is, the worse the performance that the algorithms can achieve. In order to show this point in depth, we compare these algorithms with respect to ε_0. In specific, ε_0 on all the five data sets is set to be $[0:0.1:0.9]$, where $\varepsilon_0 = 0$ implies no channel missing.

The aggregated performance is used to evaluate the goodness of the above algorithms. Taking the aggregated F1score for example, it is obtained by averaging the averaged F1 score achieved by an algorithm with different ε_0. We repeat this procedure ten times and report the averaged aggregated results and standard deviation. Furthermore, to conduct a rigorous comparison, the *paired student's t-test* is performed. The regularization parameter C for each algorithm is chosen from an appropriately large range $[10^{-1}, 1, \cdots, 10^4]$ by 5-fold cross-validation on the training data.

The above algorithms are evaluated on five benchmark MKL data sets, including psortPos, psortNeg, plant data sets, the protein fold prediction data set and Caltech101. In the first part, we report these algorithms on binary classification tasks, where the one-versus-one strategy is adopted to convert the psortPos, psortNeg and plant data sets to six, ten, and six binary classification problems. After that, we compare their performance on two multi-class classification tasks, i.e., the protein fold prediction and Caltech101 data sets. By following the same settings in (Zien and Ong 2007), the F1score is used to measure the performance of each algorithm on the PsortPos and PsortNeg data sets, while the matthew correlation coefficient (MCC) is used for the plant data set. The classification accuracy is taken as a measurement on the last two data sets.

Table 1: Aggregated F1score comparison with statistical test on psortPos. Each cell represents mean aggregated F1score $\pm$ standard deviation. Boldface means no statistical difference from the best one (p-value $\geq$ 0.05).

	C-AMKL	MM-MKL	ZF-MKL	MF-MKL	CF
C1 v.s. C2	**93.82 ±0.25**	93.17 ±0.32	92.80 ±0.31	92.55 ±0.43	91.04 ±0.27
C1 v.s. C3	**97.93 ±0.14**	96.80 ±0.22	96.52 ±0.22	95.74 ±0.48	94.23 ±0.17
C1 v.s. C4	**92.52 ±0.48**	91.86 ±0.43	91.63 ±0.28	91.18 ±0.56	89.80 ±0.29
C2 v.s. C3	**98.66 ±0.31**	96.05 ±0.65	95.67 ±0.63	94.44 ±1.95	94.66 ±0.39
C2 v.s. C4	**93.50 ±0.70**	90.35 ±1.10	88.82 ±1.21	89.07 ±0.71	85.51 ±0.84
C3 v.s. C4	**83.61 ±0.87**	73.35 ±1.96	71.82 ±0.97	75.50 ±2.19	66.17 ±1.07

Results on the protein subcellular localization[4]

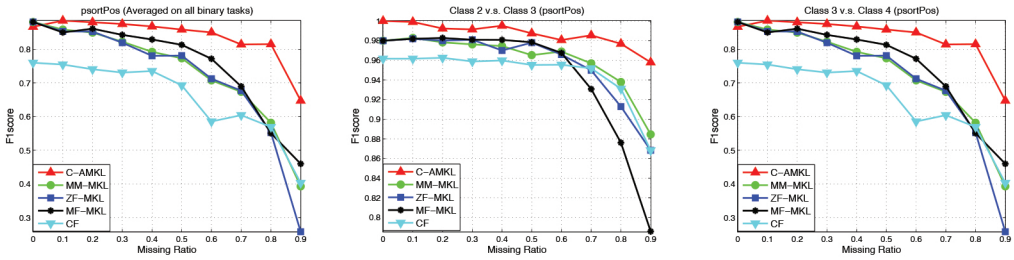
Results on psortPos: Binary Classification Figure 1(a) is the F1score of these algorithms averaged on all pairwise binary tasks and Figures 1(b)-1(c) are the results on two binary classification tasks shown as examples. As observed: (1) The proposed C-AMKL algorithm (in red) are on the top in all sub-figures, indicating its best performance; (2) The improvement of C-AMKL is more significant with the increase of missing ratio. When the missing ratio is 0.9, C-AMKL improves the second best algorithm (MF-MKL) by over 18% on class 3 vs. class 4 task (see Figure 1(c)); and (3) The variation of C-AMKL is much smaller with the increase of the missing ratio when compared with other algorithms. It implies that C-AMKL is more stable, which is a desired characteristic for a good classifier.

We attribute the superiority of C-AMKL to the sample-based margin maximization. In contrast, ZF-MKL, MF-MKL and EM-MKL algorithms maximize the margin on the imputed samples. As can be seen, such imputation is not reliable when channel absence is intensive, leading to poor performance.

The aggregated F1score, standard deviation and the p-value of statistical test are reported in Table 1, where the

[3] $round(\cdot)$ denotes a rounding function.

[4] http://raetschlab.org//suppl/protsubloc/



(a) Averaged (b) Class2 vs. Class3 (c) Class3 vs. Class4

Figure 1: F1score comparison of the above algorithms with variation of missing ratio on psortPos.

Table 2: Aggregated F1score comparison with statistical test on psortNeg.

	C-AMKL	MM-MKL	ZF-MKL	MF-MKL	CF
C1 v.s. C2	**96.27 ±0.16**	94.46 ±0.58	94.40 ±0.33	91.70 ±0.52	93.84 ±0.33
C1 v.s. C3	**90.53 ±0.26**	89.29 ±0.51	89.52 ±0.57	88.98 ±0.89	88.90 ±0.35
C1 v.s. C4	**95.25 ±0.19**	94.26 ±0.35	93.31 ±0.50	90.95 ±0.41	92.91 ±0.19
C1 v.s. C5	**95.87 ±0.24**	94.95 ±0.35	94.83 ±0.31	93.05 ±0.36	94.58 ±0.16
C2 v.s. C3	**95.40 ±0.23**	94.67 ±0.19	94.66 ±0.19	92.92 ±0.88	93.50 ±0.34
C2 v.s. C4	**95.22 ±0.19**	95.40 ±0.25	**94.91 ±1.03**	93.75 ±0.28	93.63 ±0.21
C2 v.s. C5	**97.34 ±0.19**	96.61 ±0.21	96.33 ±0.16	93.87 ±0.73	96.08 ±0.31
C3 v.s. C4	**89.48 ±0.56**	88.52 ±0.39	86.16 ±0.58	88.75 ±0.45	77.16 ±0.57
C3 v.s. C5	**89.46 ±0.47**	88.73 ±0.43	88.50 ±0.55	87.10 ±0.52	86.88 ±0.28
C4 v.s. C5	**90.96 ±0.27**	**90.92 ±0.34**	**91.03 ±0.35**	89.73 ±0.41	87.26 ±0.21

highest accuracy and those whose differences from the highest one are not statistically significant are shown in bold. As observed, C-AMKL usually significantly outperforms MM-MKL, ZF-MKL, MF-MKL and CF, which is consistent with our observations in Figure 1.

Results on psortNeg: Binary Classification The F1score achieved by the above algorithms with different missing ratio is plotted in Figure 2. As can be seen, C-AMKL is still on the top in most of the cases. The improvement of C-AMKL over the others becomes more significant with the increase of missing ratio. It outperforms the second best one (ZF-MKL) by seven percent on the class1 vs. class2 task (Figure 2(b) when the missing ratio is 0.9. We also report the mean aggregated F1score, standard deviation and the p-value of statistical test in Table 2. Again, C-AMKL demonstrates statistically significant improvement over the others.

Results on plant: Binary Classification The MCC of the above algorithms with different missing ratio on the plant data set is plotted in Figure 3. Again, C-AMKL obtains superior performance to the others. Also, when the missing ratio is high, C-AMKL demonstrates more improvement. The corresponding mean aggregated MCC, standard deviation and the statistical results are reported in Table 3. We can see that C-AMKL is consistently superior to other ones.

Table 3: Aggregated MCC comparison with statistical test on plant.

	C-AMKL	MM-MKL	ZF-MKL	MF-MKL	CF
C1 v.s. C3	**75.81 ±1.04**	73.53 ±2.21	70.46 ±1.45	69.34 ±1.90	57.62 ±0.81
C1 v.s. C3	**93.20 ±0.71**	88.33 ±1.28	86.70 ±1.00	80.43 ±1.35	90.77 ±0.54
C1 v.s. C4	**91.21 ±0.55**	86.59 ±1.14	85.92 ±0.62	85.17 ±0.37	85.32 ±0.49
C2 v.s. C3	**92.05 ±0.44**	89.78 ±0.77	88.73 ±0.72	85.06 ±1.15	87.29 ±0.58
C2 v.s. C4	**84.91 ±0.49**	79.08 ±1.27	77.42 ±1.72	72.16 ±1.28	69.87 ±1.01
C3 v.s. C4	**87.67 ±0.61**	84.13 ±0.87	82.18 ±0.63	78.24 ±1.11	83.44 ±0.71

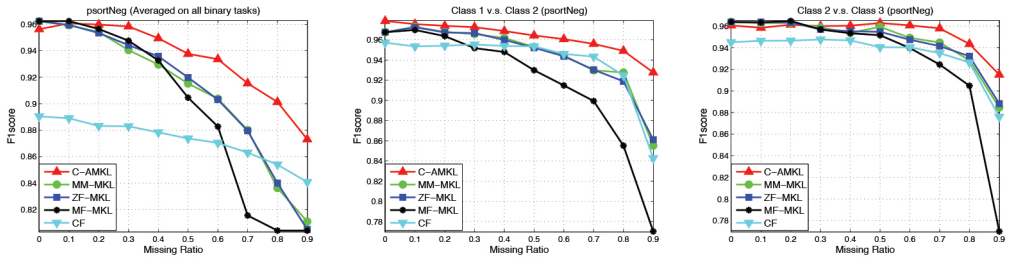
The above experimental results demonstrate the advantages of the proposed C-AMKL. At the same time, we notice that C-AMKL may become slightly inferior to the others when the missing ratio is relatively small, for example, Figure 2(c) and 3(c). We conjecture that this subtle difference is caused by the approach to solving SVMs. Specifically, C-AMKL solves the SVMs in the primal space, while the other algorithms solve it in the dual space. Though the two approaches are conceptually equivalent, they may produce slightly different solutions in practice and this in turn leads to different classification results.

Results on protein fold prediction[5]

Besides the above five algorithms, we also compare the EM-filling approach (EM-MKL) and max-margin absent features (MMAF) algorithm (Chechik et al. 2008) on the protein fold prediction data set since its input features are available[6]. It is a 27-class classification task. To make these two algorithms applicable, we first concatenate the features of

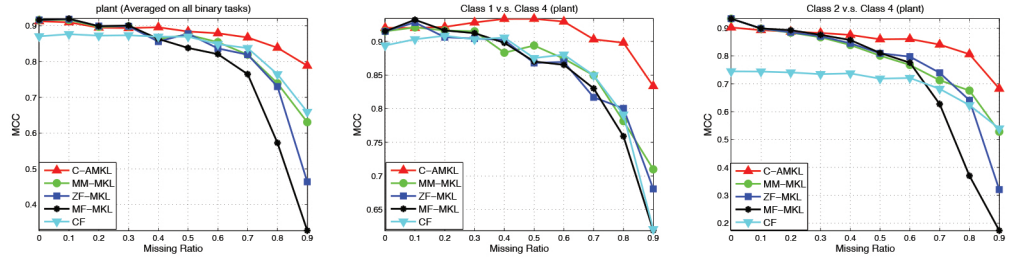
[5]http://mkl.ucsd.edu/dataset/protein-fold-prediction/

[6]Note that EM-filling and max-margin absent features need to access the raw input features, which are not publicly available from the internet for other data sets.



(a) Averaged (b) Class1 vs. Class2 (c) Class2 vs. Class3

Figure 2: F1score comparison of the above algorithms with variation of missing ratio on psortNeg.



(a) Averaged (b) Class1 vs. Class4 (c) Class2 vs. Class4

Figure 3: MCC comparison of the above algorithms with the variation of missing ratio on plant.

Table 4: Aggregated accuracy comparison with statistical test on protein fold prediction.

C-AMKL	MM-MKL	ZF-MKL	MF-MKL	CF	MMAF	EM-MKL
$\mathbf{51.99 \pm 0.62}$	48.13 ± 0.75	48.04 ± 0.86	42.25 ± 0.79	44.50 ± 0.56	48.64 ± 0.46	44.24 ± 1.39

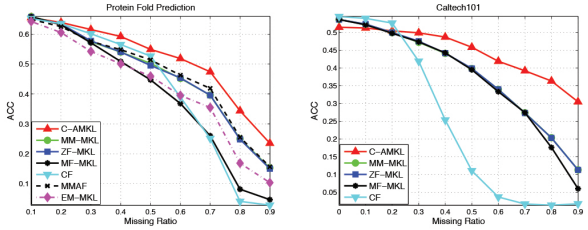
each sample from all channels into a long vector. Afterwards, EM-MKL firstly imputes the missing features with the EM algorithm (Ghahramani and Jordan 1993) and then performs SimpleMKL (Rakotomamonjy et al. 2008) on the imputed data. Differently, MMAF classifies each sample directly in the concatenated space without imputation.

As plotted in Figure 4(a), the proposed C-AMKL shows significant improvement after the missing ratio reaches 0.4. When the missing ratio equals to 0.7, C-AMKL gains nearly eight percent improvement over the second best one (MMAF). The mean aggregated classification accuracy, standard deviation and the statistical test results are reported in Table 4. Again, C-AMKL achieves significantly better classification accuracy than the rest ones.

Results on Caltech101[7]

The Caltech101 used in our experiments has a 101 object categories, with 15 randomly selected samples per class for training and up to 50 randomly selected samples per class for testing. The classification accuracy of the above algorithms on Caltech101 is plotted in Figure4(b). The proposed C-AMKL is significantly better than the others after the missing ratio is larger than 0.2. When the missing ratio reaches 0.9, C-AMKL achieves nearly 19% improvement over the second best one (MM-MKL). We also report the mean aggregated classification accuracy, standard deviation and the statistical test results in Table 5. Again, C-AMKL achieves the highest classification accuracy.

[7]http://kahlan.eps.surrey.ac.uk/featurespace/web/mkl/



(a) protein fold prediction (b) Caltech101

Figure 4: Classification accuracy comparison of the above algorithms with variation of missing ratio.

Table 5: Aggregated accuracy comparison with statistical test on Caltech101.

C-AMKL	MM-MKL	ZF-MKL	MF-MKL	CF
$\mathbf{44.58}$	38.04	38.07	37.13	24.81
$\mathbf{\pm 0.14}$	± 0.26	± 0.24	± 0.19	± 0.82

From the above experiments, we conclude that the proposed C-AMKL: (1) effectively addresses the issue of channel absence in MKL, and (2) achieves superior performance over the comparable ones, especially in the presence of intensive absence.

Conclusion

While MKL algorithms have been used in various applications, they are not able to effectively handle the scenario where there are some absent channels in samples. To address this issue, this paper proposes to maximize the minimum of all sample-based margins in the multiple kernel-induced feature spaces. After that, we propose C-AMKL to solve the optimization problem. Extensive experiments have demonstrated the effectiveness of our proposed algorithms, especially when the missing ratio is relatively high. In the future, we plan to improve the computational efficiency of C-AMKL by solving it via more advanced optimization techniques. Moreover, considering the radius of minimum enclosing ball (MEB) may vary due to the channel absence of samples, it is worth trying to integrate the sample-based radius information to further improve our approach.

Acknowledgements

This work was supported by the National Basic Research Program of China (973) under Grant No. 2014CB340303, the National Natural Science Foundation of China (project No. 61125201, 61403405, 60970034, 61170287 and 61232016).

References

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society* 68(3):337–404.

Chechik, G.; Heitz, G.; Elidan, G.; Abbeel, P.; and Koller, D. 2008. Max-margin classification of data with absent features. *Journal of Machine Learning Research* 9:1–21.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2010. Two-stage learning kernel algorithms. In *ICML*, 239–246.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2012. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research* 13:795–828.

CVX Research, I. 2012. CVX: Matlab software for disciplined convex programming, version 2.0. http://cvxr.com/cvx.

Do, H.; Kalousis, A.; Woznica, A.; and Hilario, M. 2009. Margin and radius based multiple kernel learning. In *ECML/PKDD (1)*, 330–343.

Gai, K.; Chen, G.; and Zhang, C. 2010. Learning kernels with radiuses of minimum enclosing balls. In *NIPS*, 649–657.

Ghahramani, Z., and Jordan, M. I. 1993. Supervised learning from incomplete data via an em approach. In *NIPS*, 120–127.

Gönen, M., and Alpaydın, E. 2011. Multiple kernel learning algorithms. *Journal of Machine Learning Research* 12:2211–2268.

Gönen, M. 2012. Bayesian efficient multiple kernel learning. In *ICML*.

Kloft, M.; Brefeld, U.; Sonnenburg, S.; and Zien, A. 2011. l_p-norm multiple kernel learning. *Journal of Machine Learning Research* 12:953–997.

Kumar, A.; Niculescu-Mizil, A.; Kavukcuoglu, K.; and III, H. D. 2012. A binary classification framework for two-stage multiple kernel learning. In *ICML*.

Lanckriet, G. R. G.; Cristianini, N.; Bartlett, P. L.; Ghaoui, L. E.; and Jordan, M. I. 2004. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* 5:27–72.

Liu, X.; Wang, L.; Yin, J.; and Liu, L. 2012. Incorporation of radius-info can be simple with simplemkl. *Neurocomputing* 89:30–38.

Liu, X.; Wang, L.; Yin, J.; Zhu, E.; and Zhang, J. 2013. An efficient approach to integrating radius information into multiple kernel learning. *IEEE Transactions on Cybernetics* 43(2):557–569.

Liu, X.; Wang, L.; Zhang, J.; and Yin, J. 2014. Sample-adaptive multiple kernel learning. In *AAAI*, 1975–1981.

Marlin, B. M. 2008. *Missing Data Problems in Machine Learning*. Ph.D. Dissertation, Department of Computer Science, University of Toronto.

Orabona, F., and Luo, J. 2011. Ultra-fast optimization algorithm for sparse multi kernel learning. In *ICML*, 249–256.

Rakotomamonjy, A.; Bach, F. R.; Canu, S.; and Grandvalet, Y. 2008. Simplemkl. *Journal of Machine Learning Research* 9:2491–2521.

Schölkopf, B.; Herbrich, R.; and Smola, A. J. 2001. A generalized representer theorem. In *COLT/EuroCOLT*, 416–426.

Smola, A. J.; Vishwanathan, S. V. N.; and Hofmann, T. 2005. Kernel methods for missing variables. In Cowell, R. G., and Ghahramani, Z., eds., *AISTATS05*, 325–332.

Xu, Z.; Jin, R.; Yang, H.; King, I.; and Lyu, M. R. 2010. Simple and efficient multiple kernel learning by group lasso. In *ICML*, 1175–1182.

Yan, F.; Kittler, J.; Mikolajczyk, K.; and Tahir, M. A. 2012. Non-sparse multiple kernel fisher discriminant analysis. *Journal of Machine Learning Research* 13:607–642.

Yuan, L.; Wang, Y.; Thompson, P. M.; Narayan, V. A.; and Ye, J. 2012. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *KDD*, 1149–1157.

Zien, A., and Ong, C. S. 2007. Multiclass multiple kernel learning. In *ICML*, 1191–1198.