A MULTIMODAL MULTIPLE KERNEL LEARNING APPROACH TO ALZHEIMER'S DISEASE DETECTION

Michele Donini[1,2,3], João M. Monteiro[1,2], Massimiliano Pontil[2,3], John Shawe-Taylor[2] and Janaina Mourao-Miranda[1,2] for the Alzheimer's Disease Neuroimaging Initiative*

(1) Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

(2) Department of Computer Science, University College London, London, UK

(3) Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa, Italy

Abstract

In neuroimaging-based diagnostic problems, the combination of different sources of information as MR images and clinical data is a challenging task. Their simple combination usually does not provides an improvement if compared with using the best source alone. In this paper, we deal with the well known Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset tackling the *AD versus Control* task. We use a recently proposed multiple kernel learning approach, called EasyMKL, to combine a huge amount of basic kernels in synergy with a feature selection methodology, pursuing an optimal and sparse solution to facilitate interpretability. Our new approach, called EasyMKLFS, outperforms baselines (e.g. SVM) and state-of-the-art methods as recursive feature elimination and SimpleMKL.

1 Introduction

We study the problem of combining information from different sources in a high dimensional space using only a small set of examples for training our model. In this context Multiple Kernel Learning (MKL) provides an effective approach to

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

identify which information is discriminative for a specific task considering each source of information as a kernel [1, 2].

Our case study is the problem of classifying patients with possible Alzheimer's disease combining MRI images and other clinical/demographic information. Alzheimer's disease (AD) is a neurodegenerative disorder that accounts for most cases of dementia. We use a subset of the AD Neuroimaging Initiative (ADNI*) dataset combining MR images with a list of clinically relevant information and reaching 150k different features for 227 individuals. Specifically, we collected 168130 features from the voxels of the images and 50 clinical information.

A number of studies have tried to establish the conditions under which feature selection can improve classification accuracy for diagnosis from neuroimaging data and most of them did not show improvement in performance, unless it is guided by domain knowledge [3, 4]. In our case the ratio among the number of features and the number of examples is high (around 660) highlighting the difficulty of our task. In the present study we investigate the use of MKL to combine different sources of information and select the most informative ones. The proposed approach has the potential to improve the model performance and facilitate the interpretability of the models.

We start from EasyMKL [5], a recent MKL algorithm, that is able to manage a large amount of kernels and we combine it in synergy with a new Feature Selection (FS) approach. Our aim is to evaluate and select a set of specific features for our task. We compare our approach with SVM [6]) as the baseline approach, as well as a state-of-the-art MKL approach (SimpleMKL [7]) and recursive feature elimination (RFE) [8]. Our idea is to combine a huge amount of basic (i.e. not very informative) kernels in order to generate a better representation for our neuroimaging-based diagnostic problem pursuing the creation of an optimal kernel.

Summarizing, the main contributions of this paper are two-fold. Firstly, we introduce a new methodology to combine a MKL approach using a huge number of basic kernels and a FS approach in order to improve the prediction performance. This new procedure, called EasyMKLFS, automatically selects the relevant information obtaining sparse models. Secondly, exploiting our EasyMKLFS, we tackle a challenging real-world problem, outperforming the previous state-of-the-art methods and providing a solution with a high level of interpretability.

The paper is organized as follows. In Section 2 we briefly review the notation, MKL methods and EasyMKL. The main part of the paper is Section 3 where we present our main contribution, EasyMKLFS (Section 3.1). Experimental results with the ADNI data are detailed in Section 4. Finally, In Section 4.2 we discuss the results and draw our conclusions.

2 Background

In the next sections, we will introduce the classical MKL framework and a recent MKL algorithm called EasyMKL. Firstly, we introduce the notation used in this paper.

Considering the classification task, we define the training examples as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ in a set $\mathcal{X}$, y_i with values $+1$ or -1. For our case, it is possible to consider the generic set $\mathcal{X}$ equal to $\mathbb{R}^m$, with a very large m. Then, $\mathbf{X} \in \mathbb{R}^{\ell \times m}$ denotes the matrix where examples are arranged in rows. The i^{th} example is represented by the i^{th} row of $\mathbf{X}$, namely $\mathbf{X}[i,:]$ and the r^{th} features by the r^{th} column of $\mathbf{X}$, namely $\mathbf{X}[:,r]$.

2.1 Multiple Kernel Learning (MKL)

MKL [9, 1] is one of the most popular paradigms used to learn kernels in real world applications [10, 11]. The kernels generated by these techniques are combinations of a prescribed set of basic kernels $\mathbf{K}_1, ..., \mathbf{K}_R$ with a constraint in the form: $H_R^q = \{\mathbf{x} \mapsto \mathbf{w} \cdot \boldsymbol{\phi}_{\mathbf{K}}(\mathbf{x}) : \mathbf{K} = \sum_{r=1}^{R} \eta_r \mathbf{K}_r, \boldsymbol{\mu} \in \Psi_q, \|\mathbf{w}\|_2 \leqslant 1\}$ with $\Psi_q = \{\boldsymbol{\mu} : \boldsymbol{\mu} \succcurlyeq 0, \|\boldsymbol{\mu}\|_q = 1\}$ and considering the function $\boldsymbol{\phi}_{\mathbf{K}}$ as the feature mapping from the input space to the feature space. The value q being the kind of mean used, is typically fixed to 1 or 2.

Using this formulation, we are studying the family of sums of kernels in the feature space. It is well know that the sum of two kernels can be seen as the concatenation of the features contained in both the RKHS [12]. Extending the same idea, the weighted sum of a list of basic kernels can be seen as a weighted concatenation of all the features contained in all the RKHS (where the weights are the square roots of the learned weights η_k).

The problem of searching for a combinations of basic kernels can be rephrased as a regularization problem in which the prediction function is the sum of functions f_r in the RKHS of kernel $\mathbf{K}_r$ and the regularization term is the combination of the norms of the f_r in the associated RKHS [13]. For example if $q = 1$ this is a parametric version of the group Lasso [14], see e.g. [15].

These algorithms are supported by several theoretical results that bound the *estimation error* (i.e. the difference between the true error and the empirical margin error). These bounds exploit the *Rademacher complexity* applied to the combination of kernels [15, 16, 17, 18, 19].

Existing MKL approaches can be divided in two main categories. In the first category, *Fixed or Heuristic*, some fixed rule is applied to obtain the combination. They usually get results scalable with respect to the number of kernels combined but their effectiveness critically depends on the domain at hand. They use a parameterized combination function and find the parameters of this function generally by looking at some measure obtained from each kernel separately, often giving a suboptimal solution (since no information sharing among the kernels is exploited).

On the other hand, the *Optimization based* approaches learn the combination parameters by solving a single optimization problem directly integrated in the learning machine (e.g. structural risk based target function) or formulated as a different model (e.g. alignment, or other kernel similarity maximization) [7, 9, 20].

2.2 EasyMKL

EasyMKL [5] is a recent MKL algorithm able to combine sets of basic kernels by solving a simple quadratic optimization problem. Besides its proved empirical effectiveness, a clear advantage of EasyMKL compared to other MKL methods is its high scalability with respect to the number of kernels to be combined. Specifically, its computational complexity is constant in memory and linear in time.

EasyMKL finds the coefficients $\boldsymbol{\eta}$ that maximize the margin on the training set, where the margin is computed as the distance between the convex hulls of positive and negative examples. In particular, the general problem EasyMKL tries to optimize is the following:

$$\max_{\boldsymbol{\eta}:||\boldsymbol{\eta}||_2=1} \min_{\boldsymbol{\gamma}\in\Gamma} \ \boldsymbol{\gamma}^\top \mathbf{Y}(\sum_{r=0}^{R} \eta_r \mathbf{K}_r)\mathbf{Y}\boldsymbol{\gamma} + \lambda||\boldsymbol{\gamma}||^2. \tag{1}$$

where $\mathbf{Y}$ is a diagonal matrix with training labels on the diagonal, and λ is a regularization hyper-parameter. The domain Γ represents two probability distributions over the set of positive and negative examples of the training set, that is $\Gamma = \{\boldsymbol{\gamma} \in \mathbb{R}_+^\ell \ \ | \sum_{y_i=+1} \boldsymbol{\gamma}_i = 1, \sum_{y_i=-1} \boldsymbol{\gamma}_i = 1\}$. Note that any element $\boldsymbol{\gamma} \in \Gamma$ corresponds to a pair of points, the first in the convex hull of positive training examples and the second in the convex hull of negative training examples. At the solution, the first term of the objective function represents the obtained margin, that is the (squared) distance between a point in the convex hull of positive examples and a point in the convex hull of negative examples, in the compounded feature space.

The objective function in Eq. 1 can be interpreted as the dual problem of a regularized empirical objective function using the kernel $\sum_{r=1}^{R} \eta_r \mathbf{K}_r$. This equation is a minimax problem that can be reduced to a simple quadratic problem with a technical derivation described in [5]. The solution of the quadratic problem is an optimal $\boldsymbol{\gamma}^*$ for the original min-max formulation. Due to the particular structure of EasyMKL, it is sufficient to provide the average kernel of all the trace-normalized basic kernels ($\mathbf{K}^A = \frac{1}{R}\sum_{r=1}^{R} \frac{\mathbf{K}_r}{Tr(\mathbf{K}_r)}$). From $\boldsymbol{\gamma}^*$, it is easy to obtain the optimal weights for the single basic kernels $\mathbf{K}_r$ by using the following formula

$$\eta_r = \gamma^{*T}\mathbf{Y} \ (\mathbf{K}_r/Tr(\mathbf{K}_r)) \ \mathbf{Y}\gamma^*, \ \ \forall r = 1, \ldots, R. \tag{2}$$

In the following sections, we will refer to this algorithm as EasyMKL [1].

3 Feature learning using MKL

In the last years, the importance of combining a large amount of kernels to learn an optimal representation became clear [5]. As presented in the previous section, new methods can combine thousands of kernels with acceptable computational complexity contrasting the previous idea that kernel learning is shallow in general. In fact, having MKL algorithms which are scalable opens a new

[1] EasyMKL implementation: github.com/jmikko/EasyMKL

scenario for MKL. While standard MKL algorithms typically cope with a small number of strong kernels and try to combine them (each kernel representing a different view of the same task). In this case, the kernels are individually well designed by experts and their optimal combination hardly leads to a significant improvement of the performance with respect to, for example, a simple averaging combination. In the new scenario, the MKL paradigm can be exploited to combine a very large amount of basic kernels, aiming at boosting their combined accuracy in a way similar to feature weighting [2].

Theoretical results prove that the combination of a large number of kernels using the MKL paradigms is able to add only a small penalty in the *generalization error*, as presented in [15, 17, 18, 19]. In fact, if we consider the class of linear function in the feature space of a kernel $\mathbf{K}$ as $\mathcal{F}_{\mathbf{K}} = \{\mathbf{x} \rightarrow \mathbf{w} \cdot \phi_{\mathbf{K}}(\mathbf{x}) : \|\mathbf{w}\|_2 = 1\}$, the bound on the generalization error of MKL gives only a logarithmic additive dependence with respect to the number of kernels if we learn a function in the family: $\mathcal{F} = \{\sum_{r=1}^{R} a_r f_r : f_r \in \mathcal{F}_{\mathbf{K}_i}, a_r \in \mathbb{R} \geq 0, \sum_{r=1}^{R} a_r \leq B\}$.

In this way, tacking a set of linear basic kernels that are evaluated over a single feature, the connection between MKL and feature learning is clear. The single kernel weight is, in fact, the weight of the feature. Using this framework, we can weight the information contained into bunch of features, evaluated in different ways (i.e. using different kernels that can consider different subsets of features).

A we noted earlier, MKL can be interpreted as a nonparametric version of the group Lasso. Therefore, it admits a Bayesian interpretation, which is similar to that for the Lasso, see [21]. Specifically the prior is of the form $c \exp(-\mu \sum_{r=1}^{R} \|w_r\|)$, where w_r is the weight vector associated to the r^{th} feature map/kernel, c a positive constant and $\mu \geq 0$.

3.1 EasyMKL and Feature Selection

In this section we present our approach to combine MKL and feature selection (FS). We start from EasyMKL with a large family of linear single-feature kernels as basic kernels (i.e. $\mathbf{K}_r = \mathbf{X}[:, r]\mathbf{X}[:, r]^T \ \ \forall r = 1, \ldots, R = m$). Due to the particular definition of this algorithm, we are able to combine efficiently millions of kernels. Given the kernel generated by the average of the trace normalized basic kernels $\mathbf{K}^A = \sum_{r=1}^{R} \frac{\mathbf{K}_r}{Tr(\mathbf{K}_r)}$, EasyMKL produces a list of weights $\boldsymbol{\eta} \in \mathbb{R}^R$, one for each kernel. Fixing a threshold $\rho > 0$, it is possible to remove all the kernels with a weight less or equal to ρ, considering them not sufficiently informative for our task. In this way we are able to inject sparsity in our final model. All the single-feature kernels $\mathbf{K}_r$ with a weight $\eta_r > \rho$ are weighted and summed obtaining a new kernel $\mathbf{K}^* = \sum_{r:\eta_r > \rho}^{R} \eta_r \frac{\mathbf{K}_r}{Tr(\mathbf{K}_r)}$. Algorithm 1 summarizes our approach, called EasyMKLFS. It is important to note that if $\rho = 0$ we are performing the standard MKL approach over R basic kernels.

The same procedure can not be easily exploited with the standard MKL algorithms because of the high amount of memory that they require to combine a large family of kernels. In this sense, EasyMKL becomes fundamental in order to efficiently achieve our goal. In line 7 of Algorithm 1, the amount of memory

Algorithm 1 - EasyMKLFS: feature selection and weighting by using EasyMKL. $\mathbb{O}_{\ell,\ell}$ is the zero-matrix in $\mathbb{R}^{\ell,\ell}$.

Require: $\mathbf{X} \in \mathbb{R}^{\ell,m}, \mathbf{y} \in \{-1,1\}^{\ell}, \lambda \geq 0, \rho > 0$
Ensure: A kernel matrix $\mathbf{K}^* \in \mathbb{R}^{\ell,\ell}$
$\mathbf{K}^A = \mathbb{O}_{\ell,\ell} \quad \mathbf{K}^* = \mathbb{O}_{\ell,\ell}$
$R = m$
for $r = 1$ **to** R **do**
  $\mathbf{K} = \frac{\mathbf{X}[:,r]\mathbf{X}[:,r]^T}{Tr(\mathbf{X}[:,r]\mathbf{X}[:,r]^T)}$
  $\mathbf{K}^A = \mathbf{K}^A + \frac{1}{R}\mathbf{K}$
end for
$\boldsymbol{\eta}$ =EasyMKL($\mathbf{K}^A, \mathbf{X}, \mathbf{y}, \lambda$)
for $r = 1$ **to** R **do**
  if $\eta_r > \rho$ **then**
    $\mathbf{K} = \frac{\mathbf{X}[:,r]\mathbf{X}[:,r]^T}{Tr(\mathbf{X}[:,r]\mathbf{X}[:,r]^T)}$
    $\mathbf{K}^* = \mathbf{K}^* + \eta_r\mathbf{K}$
  end if
end for

required by the storage of the kernels is independent with respect to the number of combined kernels R.

4 Experiments using Real Data

In this section we will present our results using EasyMKLFS on the ADNI dataset for the task of classifying patients with possible Alzheimer's disease versus healthy controls, combining structural magnetic resonance (MR) images and clinical and demographic information.

4.1 Experimental settings

The T1 weighted MRI scans were segmented into grey matter probability maps using SPM12, normalised using Dartel, converted to MNI space with voxel size of 2mm×2mm×2mm and smoothed with a Gaussian filter with 2 mm FWHM. A mask was then generated, this selected voxels which had an average probability of being grey matter equal or higher than 10% for the whole dataset. This resulted in 168130 voxels per subject being used.

We combine features derived from the MR images (each voxel was considered as a single feature) with 50 selected clinical and demographic features. In the following we will refer to (linear single-feature) basic kernels or directly to features without distinction.

We performed a balanced accuracy comparison (i.e., the average between sensitivity and specificity) among five different approaches. The first is the vanilla SVM [6], using a single kernel that is the average of all the considered features (i.e. linear SVM). It is used as baseline to understand the difficulty of the task. The second approach is SVM RFE [8], that is the standard recursive

feature elimination approach. RFE considers the importance of individual features in the context of all the other features and it has the ability to eliminate redundancy and improves the generalization accuracy [4]. The third comparison is with SimpleMKL [7], a well known MKL iterative algorithm by Rakotomamonjy that implements a linear approach with kernel weights in a simplex. Basically SimpleMKL works by repeating two main steps: a) SVM optimization problem defined on current weights; b) Updating of the kernel weights using a gradient function. The last comparison is with EasyMKL, a recent MKL algorithm presented in Section 2.2. Finally, our EasyMKLFS is the combination of MKL and FS approach, presented in section 3.1.

In our experiments, we consider different subsets and different fragmentations of the whole information contained in the ADNI dataset. The considered linear kernels (or features) are divided in 6 different sets. **I** considers the average of all the voxel features and represents the whole image in one single linear kernel. It represents the naive way to combine the information of a single MR image. **C** is the linear kernel with the average of all the clinical features in one single kernel and is the most simple way to exploit the clinical information. $\mathbf{I}+\mathbf{C}$ is the kernel with the average of all the voxels and all the clinical features and represents the average of all the considered information. $\mathbf{I}$ & $\mathcal{C}$ is the family of basic kernels that contains a single linear kernel for the whole image plus one linear kernel for each clinical feature. In this case we are able to tune the importance of the single clinical information and made the correct trade-off between clinical information and MR image. $\mathcal{V}$ is the family of basic kernels (or basic features) that contains one linear kernel for each voxel (i.e. 168130 different basic kernels). Each single voxels can be weighted or selected highlighting the relevant voxels of the MR image. Finally, $\mathcal{V}$ & $\mathcal{C}$ is the family of basic kernels (or basic features) that contains one kernel for each voxel plus one linear kernel for each clinical information.

All the experiments are performed using an average of 5 repetition of a nested 10-fold cross-validation. We fixed the same distribution of the age of the patients among all the subsets. The validation of the hyper-parameters has been performed in the family of $C \in \{0.1, 1, 5, 25\}$ for the SVM parameter, $\lambda \in \{\frac{v}{1-v} : v = 0.0, 0.1, \ldots, 0.9, 1.0\}$ for the EasyMKL parameter, $\rho \in \{\frac{i}{m} : i = 0, 1, \ldots, 20\}$ (where m is the number of the features) for the EasyMKLFS parameter. We fixed the percentage of dropped features at each step of the RFE approach equal to the 5% (using higher percentages deteriorates the results).

We performed experiments in two different experimental settings. In the first setting, we maintained all the ADNI clinical information as features. In our second setting, we remove the clinical information that have a high direct correlation with respect to the labels. We evaluate a controlled false discovery rate (FDR) [22] using an individual p-value test for each feature, with a confidence of 0.01. FDR is a powerful method for correcting for multiple comparisons.The remaining clinical information after this selection are 33. The idea is to prove that the improvement of the results is not specifically related to the clinical information that are directly used by experts to generate the labels.

4.2 Experimental results and discussion

The results for the two settings are depicted in Table 1 (using all the 50 clinical information) and Table 2 (using the subset of 33 clinical information).

In both the settings, SVM reaches is maximal balanced accuracy using all the information (image and clinical). On the other hand, the increase of the performance adding the clinical information is not significant. Using only the clinical information, SVM obtains a balanced accuracy of 68% exploiting all the 50 clinical features. Conversely, it is not able to generate a valid model using only the subset of 33 clinical features selected by using the FDR method.

The FS baseline, i.e. RFE, outperforms SVM and the standard MKL approach (EasyMKL and SimpleMKL with only 51 or 34 basic kernels). As the SVM, RFE is not able to have a significative improvement adding the clinical information to the data. This experiment highlights how the combination of few kernels does not seem to be the correct way to exploit a MKL settings.

Considering one single kernel for each voxel, SimpleMKL is not able to handle the optimization problem (due to the required memory). EasyMKL, using only the voxels information, obtains a balanced accuracy of 86%, that is comparable to the RFE baseline. This performance increases when we apply the FS phase (using our algorithm EasyMKLFS), obtaining a balanced accuracy of 87%.

Finally, considering the $\mathcal{V}$ & $\mathcal{C}$ family of basic kernels, using EasyMKL we obtain a significative improvement of the performance due to the correct weighting of the clinical information. The balanced accuracy obtains another significative increase applying EasyMKLFS to the $\mathcal{V}$ & $\mathcal{C}$ family. In fact, using all the features, we start with an 89% of balanced accuracy for EasyMKL to a 96% for EasyMKLFS. In the second settings, with only 33 clinical variables, EasyMKL obtains 88% and EasyMKLFS 92%.

From these results it is clear that combining MR images and clinical information improves the prediction performance of the model if we are able to select the correct trade-off among different voxels and clinical features, as in the $\mathcal{V}$ & $\mathcal{C}$ set of basic kernels.

In Table 3 the required memory of the different MKL methods is presented. As already noted, SimpleMKL requires a huge amount of memory to handle large family of basic kernels. For example, generating one linear kernel for each voxel, we have to provide more than 50 Gb of memory to store all the required information. EasyMKL and our EasyMKLFS use a fixed amount of memory independently with respect to the number of kernels, due to the particular definition of the optimization problem (see Sections 2.2 and 3.1).

An important characteristic of neuroimaging-based diagnostic algorithms is providing model interpretability, i.e. in a clinical context is important to identify which features are driving the predictions. Figure 1 shows the selection frequency (for the RFE) or the average of the weights $\boldsymbol{\eta}$ (for EasyMKL and EasyMKLFS) of the tested approaches overlayed onto an anatomical brain template, which can be used as surrogate of consistency. These maps show that all approaches find brain areas previously identified as important for neuroimaging-based diagnosis of dementia (e.g. bilateral hippocampus and amygdala), however the SVM RFE also selects features across the whole brain potentially related to noise, while the EasyMKLFS selects almost exclusively voxels within the hippocampus and amygdala.

Algorithm	Kernels	R	Bal. Acc. %
SVM	$\mathbf{I}$	1	84.08 ± 6.94
SVM	$\mathbf{C}$	1	68.73 ± 9.68
SVM	$\mathbf{I} + \mathbf{C}$	1	84.80 ± 6.87
SVM RFE	$\mathcal{V}$	–	86.34 ± 6.93
SVM RFE	$\mathcal{V}$ & $\mathcal{C}$	–	86.93 ± 4.76
SimpleMKL	$\mathbf{I}$ & $\mathcal{C}$	51	84.44 ± 6.68
EasyMKL	$\mathbf{I}$ & $\mathcal{C}$	51	84.78 ± 6.76
SimpleMKL	$\mathcal{V}$	168130	Out of memory
EasyMKL	$\mathcal{V}$	168130	86.12 ± 4.54
EasyMKL	$\mathcal{V}$ & $\mathcal{C}$	168180	88.80 ± 7.02
EasyMKLFS	$\mathcal{V}$	168130	86.91 ± 5.12
EasyMKLFS	$\mathcal{V}$ & $\mathcal{C}$	168180	96.14 ± 3.55

Table 1: Comparisons of 5 repetitions of a nested 10-fold cross-validation balanced accuracy using all the clinical information contained in ADNI. R is the number of generated kernels as input of the algorithm.

Algorithm	Kernels	R	Bal. Acc. %
SVM	$\mathbf{C}$	1	50.00 ± 0.00
SVM	$\mathbf{I} + \mathbf{C}$	1	84.10 ± 7.92
SVM RFE	$\mathcal{V}$ & $\mathcal{C}$	–	86.53 ± 5.99
SimpleMKL	$\mathbf{I}$ & $\mathcal{C}$	34	84.29 ± 11.78
EasyMKL	$\mathbf{I}$ & $\mathcal{C}$	34	84.47 ± 7.28
EasyMKL	$\mathcal{V}$ & $\mathcal{C}$	168163	87.97 ± 6.59
EasyMKLFS	$\mathcal{V}$ & $\mathcal{C}$	168163	92.38 ± 7.27

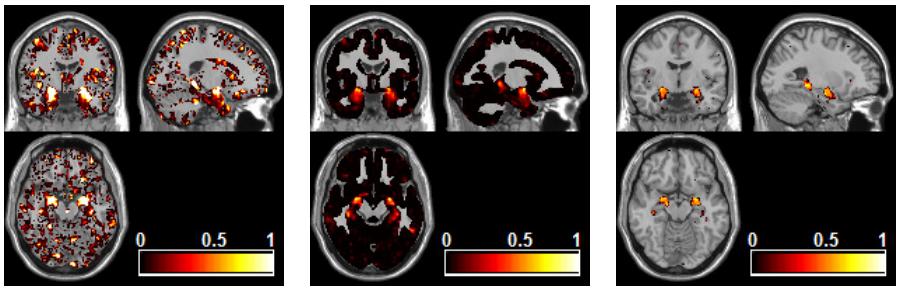
Table 2: Comparisons of 5 repetitions of a nested 10-fold cross-validation balanced accuracy using the clinical information selected by a FDR procedure.

Algorithm	Kernels	Memory	Memory (real)
SimpleMKL	$\mathbf{I}$ & $\mathcal{C}$	$\mathcal{O}(R\ell^2)$	$\sim$ 15 Mb
EasyMKL	$\mathbf{I}$ & $\mathcal{C}$	$\mathcal{O}(\ell^2)$	293 Kb
SimpleMKL	$\mathcal{V}$	$\mathcal{O}(R\ell^2)$	$\sim$ 50 Gb
EasyMKL	$\mathcal{V}$ & $\mathcal{C}$	$\mathcal{O}(\ell^2)$	293 Kb
EasyMKLFS	$\mathcal{V}$ & $\mathcal{C}$	$\mathcal{O}(\ell^2)$	293 Kb

Table 3: Required memory for different methods to handle different families of basic kernels.

5 Conclusion

In this paper, we presented EasyMKLFS, an extension of a MKL approach called EasyMKL. Exploiting this new algorithm, we shown how the selection of the relevant feature, in synergy with the correct trade-off between MR images and clinical information is able to improve the prediction performance for the challenging task Alzheimer's disease versus healthy controls of the ADNI



(a) SVM RFE with $\mathcal{V}$ & $\mathcal{C}$. (b) EasyMKL with $\mathcal{V}$ & $\mathcal{C}$. (c) EasyMKLFS with $\mathcal{V}$ & $\mathcal{C}$.

Figure 1: Comparison of voxels selection frequency (RFE) and weights (EasyMKL and EasyMKLFS), overlayed onto an anatomical template.

dataset.

References

[1] Mehmet Gonen and Ethem Alpaydin, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[2] Michele Donini, Fabio Aiolli, and Via Trieste, "Feature and kernel learning," in *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015, number April, pp. 22–24.

[3] Carlton Chu, Ai Ling Hsu, Kun Hsien Chou, Peter Bandettini, and ChingPo Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, 2012.

[4] Benson Mwangi, Tian Siva Tian, and Jair C. Soares, "A review of feature reduction techniques in Neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[5] Fabio Aiolli and Michele Donini, "EasyMKL: a scalable multiple kernel learning algorithm," *Neurocomputing*, pp. 1–10, 2015.

[6] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[8] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[9] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," *Twentyfirst international conference on Machine learning ICML 04*, vol. 69, pp. 6, 2004.

[10] Serhat Bucak, Rong Jin, and Anil Jain, "Multiple Kernel Learning for Visual Object Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1354–1369, 2014.

[11] Eduardo Castro, Vanessa Gómez-Verdejo, Manel Martínez-Ramón, Kent a. Kiehl, and Vince D. Calhoun, "A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: Application to schizophrenia," *NeuroImage*, vol. 87, pp. 1–17, 2014.

[12] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[13] Charles A. Micchelli and Massimiliano Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[14] Lukas Meier, Sara Van De Geer, and Peter Bühlmann, "High-dimensional additive modeling," *Annals of Statistics*, vol. 37, no. 6 B, pp. 3779–3821, 2009.

[15] Andreas Maurer and Massimiliano Pontil, "Structured sparsity and generalization," *Journal of Machine Learning Research*, vol. 13, pp. 671–690, 2012.

[16] Nathan Srebro and Shai Ben-David, "Learning Bounds for Support Vector Machines with Learned Kernels," *19th Annual Conference on Learning Theory, COLT 2006*, pp. 169–183, 2006.

[17] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh, "Generalization bounds for learning kernels," *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 247—-254, 2010.

[18] Zakria Hussain and John Shawe-Taylor, "Improved Loss Bounds for Multiple Kernel Learning," vol. 15, no. 2004, pp. 370–377, 2011.

[19] Zakria Hussain and John Shawe-Taylor, "A Note on Improved Loss Bounds for Multiple Kernel Learning," *arXiv preprint arXiv:1106.6258*, vol. 15, no. 2004, pp. 1–11, 2011.

[20] Manik Varma and Bodla Rakesh Babu, "More generality in efficient multiple kernel learning," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8, 2009.

[21] Trevor Park and George Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[22] Yoav Benjamini and Yosef Hochberg, "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 2016.