CHAPTER

8

Multiple kernel learning

Letizia Squarcina[1], *Umberto Castellani*[2], *Paolo Brambilla*[3, 4]

[1] IRCCS “E. Medea” Scientific Institute, Lecco, Italy; [2] Department of Computer Science, University of Verona, Verona, Italy; [3] Department of Neurosciences and Mental Health, Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy; [4] Department of Psychiatry and Behavioural Neurosciences, UT Houston Medical School, Houston, TX, United States

8.1 Introduction

Kernel-based machine learning algorithms have been extensively used in the last years in the study of brain disorders. Examples of such are Support Vector Machine (SVM) and logistic regression, among the most widely used procedures for automatic classification and regression. A kernel is a function that computes the similarity between two vectors. The simplest example is the linear kernel, which computes the dot product of the input vectors (the dot product between the vectors $a = [a_1, a_2, a_3, \ldots, a_n]$ and $b = [b_1, b_2, b_3, \ldots, b_n]$ is $a \cdot b = \sum_{i=1}^{n} [a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3 \cdots + a_n b_n]\Big)$. The choice of an appropriate kernel is not trivial, especially because it can be a source of bias, which can heavily affect the algorithm’s performances (Zhuang, Wang, Hoi, & Lan, 2011) (for a more in-depth discussion of bias, see Chapter 17). Usually, this choice is made using cross-validation employing a validation set different from the training set, with the aim of choosing the best performing kernel among a set of different functions (Gönen & Alpaydın, 2011). It is to be noted that not only the kernel’s choice requires some expert knowledge but also the optimal solution may be a combination of several kernels. Moreover, although the choice of a particular kernel may take advantage of prior knowledge, it can be difficult to justify (Lanckriet et al., 2004). Multiple kernel learning

https://doi.org/10.1016/B978-0-12-815739-8.00008-0

(MKL) algorithms address this issue, producing a combination of several candidate kernels to be used as classification rule (Wilson, Li, Kuan, & Wang, 2018). Furthermore, MKL algorithms can be used to address another common issue in studies of brain disorders: the heterogeneity of sources of the available data. For example, in the case of brain disorders, possible sources of data include structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), electroencephalography, clinical and behavioral scores, and genomic data. The integration of these modalities within a single statistical analysis may allow a better understanding of the interplay between distinct biological processes underlying a disease. Such integration would be impractical to implement using SVM algorithms, which are not ideally suited for the analysis of different modalities in a single analysis (Pettersson-Yeo et al., 2014). However, it can be achieved with MKL methods, through the combination of different kernels, each one corresponding to an appropriate measure of similarity for a given type of data (Fig. 8.1). Combining multiple kernels has been defined by Noble (2004) as intermediate integration, as opposed to early combination, where features from heterogeneous sources are simply concatenated, and late combination, where different learning algorithms are trained separately and the decisions are then combined (see Chapter 16).

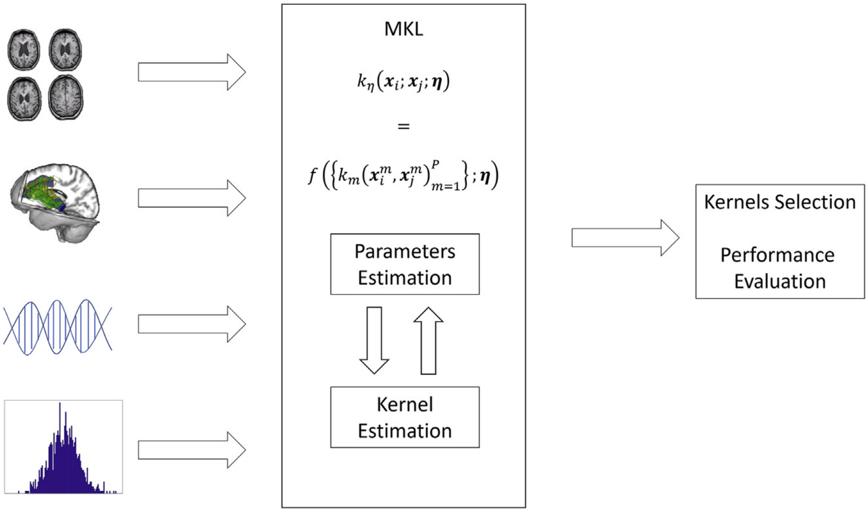
MKL
$k_\eta(\boldsymbol{x}_i; \boldsymbol{x}_j; \boldsymbol{\eta})$
=
$f\left(\left\{k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)_{m=1}^P\right\}; \boldsymbol{\eta}\right)$
Parameters Estimation
Kernel Estimation
Kernels Selection
Performance Evaluation

FIGURE 8.1 Multiple kernel learning methods take into account several features, each one associated to a different kernel. The kernels are then combined, with weights that are computed during an optimization process. The kernel parameters are estimated during the same process. The most relevant features can be selected on the basis of the weight assigned to each of them during the optimization procedure.

A key aspect of MKL algorithms is the function that determines the way the kernels are combined; this can be a simple summation or multiplication with no parameters to be estimated or a more complex function involving the estimation of parameters via optimization or the use of Bayesian approaches. When using more complex functions, MKL algorithms are required to estimate both the parameters that regulate how the kernels are combined and the parameters of the learning algorithm (e.g., *C* parameter in SVM). These two sets of parameters can be estimated at the same time, using a one-stage approach, or in a separate fashion until convergence is achieved, using a two-stage approach (Gönen & Alpaydın, 2011). This characteristic, together with the complexity of the learning algorithm itself, is critical for the performance of MKL algorithms. The vast majority of MKL algorithms involve the minimization of a single objective function that includes parameters that regulate how the kernels are combined as well as the parameters of the learning algorithm, thus employing a single-stage approach (Han, Yang, Li, Liu, & Ma, 2018). Gönen and Alpaydın (2011) demonstrated with a series of experiments that a one-stage SVM-based approach is a reasonable choice in most cases, mainly because it involves efficient minimization procedures. Nevertheless, in recent years, the combination function has been employed in an increasing number of studies (e.g., Han et al., 2018; Lanckriet et al., 2004; Rakotomamonjy, Bach, Canu, & Grandvalet, 2007). It is also to be noted that, although MKL algorithms are for the most part employed in the context of supervised learning, where there is preexisting knowledge of the labels, they can also be utilized in semisupervised learning, where not all labels are known (Li, Bai, Peng, Du, & Ying, 2018), and in unsupervised learning, where labels are not known (e.g., Zhang et al., 2011) (see Chapter 1).

In the context of brain disorders research, MKL has been used to identify the most suitable kernels within a specific dataset (e.g., Surampudi et al., 2018) and to combine different data modalities, with the final aim of understanding the possible interactions between biological processes (Squarcina et al., 2017). In this chapter, we provide an overview of MKL methods, followed by a review of exemplar applications of MKL to brain disorders.

8.2 Method description

In general, kernel functions tend to be considered within the same family of methods as SVM (Vapnik, 1998). The overall aim consists of estimating an optimal hyperplane that is able to separate the samples into two classes (e.g., patients and healthy controls [HCs]). The advantage of using SVM is that such optimal hyperplane is trained in a *maximum* *margin* principle which ensures the best generalization performance in classifying new and unseen subjects. To extend the task to nonlinear classification, the relations between samples are encoded through a well-defined kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that computes a similarity metric between data instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ (Ulas et al., 2010). In particular, with the so-called "kernel trick," a kernel function can be seen as a projection of the original data into a higher dimensional space $k(\boldsymbol{x}_i, \boldsymbol{x}_j) =< \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) >$ where the separation between samples turns out to be linear (Castellani et al., 2012; Ulas et al., 2010).

An interesting and promising approach consists of extending the possibility to exploit different relations between samples and therefore using more than one kernel. This enables the extraction of different features for each sample and the application of a kernel to each of such features. Indeed, a proper measure of similarity can be employed for each different feature (Lanckriet et al., 2004). For example, in the case of different modalities of MRI acquisitions, a different kernel may be assigned to each technique (Peruzzo et al., 2015). Similarly, when several regions of interest (ROIs) are considered, the same feature can be extracted from each region (i.e., cortical thickness) and a single kernel can be associated to each of them. Finally, a proper strategy is exploited to combine all the kernels into a unified one that is used from the SVM for the classification task. We refer to MKL (Fig. 8.1) when the parameters of the kernel combination and those for SVM classification are estimated within the same learning procedure (Gönen & Alpaydın, 2011).

In more details, a kernel that combines P of several kernels k_m can be written as a function f which considers all kernels, using some parameters stored in $\boldsymbol{\eta}$:

$$k_\eta(\boldsymbol{x}_i; \boldsymbol{x}_j; \boldsymbol{\eta}) = f\left(\left\{k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right)_{m=1}^{P}\right\}; \boldsymbol{\eta}\right) \tag{8.1}$$

where $\boldsymbol{x}_i^m$ is the feature m of sample i which is compared with the representation m of sample j (i.e., $\boldsymbol{x}_j^m$) through the kernel k_m. For instance, $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_i^{m\prime}$ can be the thickness and the volume of subject i, respectively. Thus, it is possible to define a similarity measure encoded in the kernel k_m for the comparison of thickness values and another similarity measure encoded in the kernel $k_{m\prime}$ for the comparison of volume values. Finally, parameters η_m and $\eta_{m\prime}$ define how the information from thickness is combined with that from volume. Because a prior value for $\boldsymbol{\eta}$ is not intuitive, the estimation of kernel parameters is delegated to the learning procedure, i.e., parameters $\boldsymbol{\eta}$ are those that maximize the classification accuracy. The way f behaves affects performances and complexity of the resulting MKL algorithm. The definition of the function f is not trivial; it has risen a definite interest in the field of informatics and has been object of many studies in the last years.

8.2.1 Linear multiple kernel learning combination algorithms

The simplest way to combine the different kernels involved in MKL algorithms is a weighted sum or average of the considered kernels. Eq. (8.1) can then be rewritten as

$$k_{\eta}(\boldsymbol{x}_i; \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{p} \eta_m k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right), \eta_m \in \mathbb{R} \tag{8.2}$$

In the simplest scenarios, all η_m can be equally set to 1 to achieve k_η as the unweighted sum of all kernels. Similarly, in the so-called *mean rule*, all weights are set to $\eta_m = 1/P$, so the resulting k_η is an average of all kernels k_m. These are two particular cases of linear MKL algorithms, where there is no need of learning procedures for defining the parameters $\boldsymbol{\eta}$. The only optimization procedures needed is for estimating the learning algorithm parameters, as in the case of single kernel methods (Gönen & Alpaydın, 2011).

In the most interesting scenarios, the weights η_m of Eq. (8.2) can be estimated with a learning approach; thus, an optimization procedure can be used to search for a weighted combination of kernels which maximizes a performance measure. The parameters η_m and those of the SVM are estimated simultaneously in the framework of a single optimization problem (Castellani et al., 2012). Some constraints can be put on the weights: two common limitations on weights are the requirement of positivity ($\eta_{\dot{m}} \in R_+^P$) or the sum of all elements might be set to 1 (convex sum: $\sum_{m=1}^{P} \eta_m = 1$). Restricting the weights to be positive has an important consequence for results interpretation: the relative importance given to each kernel can be measured looking at the weight learnt for each classifier. This is not possible if weights also take negative values. For example, let us suppose that we are classifying a brain disease on the basis of different measures, e.g., cortical thickness, cortical volume, psychological tests, and scales. If each of these features is associated to a different kernel, knowing their weights allows one to infer the relative contribution of each measure to classification; this in turn can provide valuable information on the most relevant biological processes for the disease under consideration.

An alternative approach to involve an MKL strategy is when different basis kernels are applied to the same data features (Gönen & Alpaydın, 2011). Here, the estimated weights of the MKL procedure enable us to select the most effective kernel parameters that usually are computed using a cross-validation method. The idea is to identify the best performance among different kernels, but letting an automated algorithm decide which kernel is best for the problem at hand. Each kernel may use the given features using different parameters (for example, all kernels may be Gaussian kernels which differ for the values of their bandwidth), so that different measures of similarity are applied to the same data.

The estimation of the MKL weights is not a trivial task and many methods have been reported in literature. For example, they can be computed with a heuristic approach, which involves considering the performances of each kernel separately and then combining them. Tanabe and colleagues (Tanabe, Bao Ho, Nguyen, & Kawasaki, 2008) proposed a method where the weight associated with the kernel k_m depends on the accuracy π_m obtained applying only that kernel to classify the data, in relation to the performances obtained by all the other kernels:

$$\eta_m = \frac{\pi_m - \delta}{\sum_{h=1}^{P} \pi_h - \delta} \tag{8.3}$$

where δ is the minimum accuracy obtained with the P learning algorithm. In this way, the weight associated with a specific kernel is directly proportional to its performance, i.e., if the classification with that kernel has a good performance, that kernel will have more importance in the combination than other kernels with a worse discrimination power. However, in this fashion, the classification accuracy is maximized for each kernel (i.e., similarity measure) separately without considering the benefit of using all the information at the same time. To consider all kernels simultaneously, an interesting approach is the so-called Lp-norm MKL (Kloft, Brefeld, Sonnenburg, & Zien, 2011) or group lasso MKL (GL-MKL) (Xu, Jin, Yang, King, & Lyu, 2010). The advantage of GL-MKL over other MKL methods is that it allows an explicit tuning of the sparsity effect on the involved MKL weights. Moreover, a closed form solution for the estimation of the MKL weights can be derived by exploiting the equivalence between group lasso and MKL. Formally, the estimation of MKL weights is given by the following updated equation:

$$\eta_m = \frac{\left|\left|w_m\right|\right|_2^{\frac{2}{p+1}}}{\left(\sum_{h=1}^{P}\left|\left|w_h\right|\right|_2^{\frac{2}{p+1}}\right)} \tag{8.4}$$

where $\left|\left|w_m\right|\right|_2^{\frac{2}{p+1}} = \eta_m^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right)$ is derived from the duality condition (Vapnik, 1998). In this way, an alternating optimization method is employed to learn the parameters: (i) at first, α_i are estimated by fixing η_m and using a standard SVM solver and then (ii) kernel weights are updated using Eq. (8.4). These two phases are repeated until convergence. Note that parameter p regulates the sparsity effect. When p= 1, the method encourages the definition of zero weights and a *competitive* approach is introduced among features by improving the selection effect of MKL. Conversely, when p= 2, a *cooperative* approach is introduced by exploiting complementary information. In particular, when the task is to select the most important features to characterize the disease, a competitive strategy can be employed. Therefore, the MKL approach can be seen as a feature selection method where features associated to weights with very low value (i.e., close to 0) are discarded. In contrast, when the aim is to use all the available information, and therefore all the features, a cooperative approach emphasizing complementary information is more convenient (Peruzzo et al., 2015).

8.2.2 Advanced multiple kernel learning methods

An advanced MKL method is based on a nonlinear kernel combination strategy. A simple example of nonlinear combination is the multiplication:

$$k_\eta(\boldsymbol{x}_i; \boldsymbol{x}_j; \boldsymbol{\eta}) = \prod_{m=1}^{P} \eta_m k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right) \tag{8.5}$$

It has been argued that nonlinear combinations of kernels might yield more information than linear ones. Nevertheless, Gönen and Alpaydın (2011) obtained that linear combinations of complex kernels are more promising than nonlinear combinations. However, learning more informative relations from the data comes at the cost of complexity. This limitation has been addressed in recent work. For example, Gu, Liu, Jia, Benediktsson, and Chanussot (2016) combined several linear kernels obtained by multiplying element by element, and the parameters iteratively computed until convergence.

Another advanced MKL approach is based on the so-called *local kernel* (Gönen & Alpaydin, 2008). The main idea of locality consists of estimating a different weigh for each sample. This leads to the following kernel formulation:

$$k_\eta(\boldsymbol{x}_i; \boldsymbol{x}_j; \boldsymbol{c}_i; \boldsymbol{c}_j) = \sum_{m=1}^{p} \eta_m(\boldsymbol{c}_i) k_m\left(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m\right) \eta_m(\boldsymbol{c}_i) \tag{8.6}$$

where the weight $\eta_m(\boldsymbol{c}_i): R^h \rightarrow R$ is now a function of the local characteristic $\boldsymbol{c}_i$, which depends on the sample i. Possible effective candidates for the weighting function $\eta_m(\boldsymbol{c}_i)$ are the well-known sigmoid function or the softmax function (Gönen & Alpaydin, 2008; Gönen & Alpaydın, 2013; Squarcina et al., 2017). In this approach, rather than estimating the fixed weight values, the learning procedure computes the parameters of the weighting function (e.g., the sigmoid or softmax function). It is interesting to note that when the local characteristic is the original space (i.e., $\boldsymbol{c}_i = \boldsymbol{x}_i$), the estimated weight is guided from the location where the current sample is lying on the input space (Gönen & Alpaydın, 2013). In particular, the classifier is more flexible in performing differently on different regions of the source data. Another interesting and promising effect of the local kernel formulation is proposed in Squarcina et al. (2015; 2017) where the local characteristic $\boldsymbol{c}_i$ is a covariate such as age or gender of the subject. Here, the classifier is able to deal with heterogeneous data that usually introduce a confounding effect. Indeed, the confounding factors are considered as nuisance variable that corrupts the acquired data (Sanderman, Coyne, & Ranchor, 2006). For instance, a common problem in brain data is the integration of datasets that have been acquired with different MRI systems. Rather than analyzing each dataset separately, the MRI system can be encoded as covariate to increase statistical power and improve the classification process (Squarcina et al., 2017). Usually the covariates are treated as confounding variables and their effect is removed from data using statistical methods such as the generalized linear model (McCullagh & Nelder, 1989). Using the local kernel approach, data are effectively adjusted for covariates, leading to improved classification accuracy (Squarcina et al. 2017).

8.3 Applications to brain disorders

MKL is particularly relevant to neuroimaging studies, which often involve the acquisition of multiple types of data using different neuroimaging techniques. Furthermore, it is particularly relevant to studies that combine neuroimaging data with other types of data such as clinical, behavioral, and genomic information. Because MKL is a relatively recent method, its application to brain disorders is still at a preliminary stage. Nevertheless, within the existing literature we can find applications to a range of neuroimaging data, such as sMRI, fMRI, and positron emission tomography (PET), as well as genomic data. For example, Ye et al. (2008) applied MKL to sMRI and genetic data, while Zhang et al. (2011) integrated sMRI, PET, and biological information obtained from the cerebrospinal fluid (CSF). To date, the majority of applications of MKL to brain diseases are related to neurological conditions such as Alzheimer's disease (AD) and cognitive impairment (Raamana, Weiner, Wang, Beg, & Alzheimer's Disease Neuroimaging Initiative, 2015; Schrouff et al., 2018; Ye et al., 2008; Zhang et al., 2011; Zhu, Thung, Adeli, Zhang, & Shen, 2017); these applications were aimed at identifying the brain regions that are most affected in these disorders. Other studies in the field of neurology focused on Parkinson's disease (PD) (Adeli et al., 2017; Castillo-Barnes et al., 2018; Segovia et al., 2017), Tourette syndrome (Wen et al. 2017), lateral amyotrophic sclerosis (Fekete, Zach, Mujica-Parodi, & Turner, 2013), and glioblastoma (Zhang, Li, Peng, & Wang, 2016). The field of psychiatry has also shown interest in MKL, with most applications in psychosis (Castellani et al., 2012; Castro, Martínez-Ramón, Pearlson, Sui, & Calhoun, 2011; Squarcina et al., 2017; Ulas et al., 2010). The next section reviews studies that have used MKL to investigate the biological basis of AD, PD, and psychosis. For each disorder, we consider the processing of the data, the implementation of MKL, and the main findings. Taken collectively, these studies indicate that MKL (i) outperforms single kernel methods when patients with AD, PD, and psychosis are compared against HCs and (ii) allows the identification of potential neuromarkers implicated in these disorders, helping disclose their underlying pathogenesis.

8.3.1 Diagnostic classification of Alzheimer's disease

MKL methods have been widely used for automatic classification in AD. This is partially due to the availability of hundreds of subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (www.loni.ucla.edu/ADNI), a consortium that includes neuroimaging data from HCs, patients with mild cognitive impairment (MCI), and AD, recruited in more than 50 sites. Ye et al. (2008) were the first to apply MKL in AD; the authors used MKL to combine demographic, genetic, and sMRI data from 118 subjects (59 HC and 59 AD) who were part of the ADNI dataset. Kernels from each feature set were combined using a linear combination, where the weights of each kernel and the parameters of the learning algorithm were simultaneously learned via an optimization process. The MKL model achieved a specificity of over 90%, outperforming all single kernel algorithms. Notably, it was also possible to identify the brain regions which were more important for classification. These included the parahippocampus, the hippocampus, the amygdala, and several temporal and occipital regions, all of which are known to be involved in AD. This suggests that MKL could be used to discover and validate potential biomarkers for AD, without making use of previous knowledge about the neuroanatomical basis of the disease that could bias the analysis.

In a subsequent study, Zhang et al. (2011) applied an MKL SVM model to sMRI, PET, and CSF data obtained from 204 subjects (51 AD patients, 91 MCI, and 52 HC), who were also part of the ADNI dataset. The authors used three different kernels, one for each modality, and linearly combined them (using a convex sum approach) into a mixed kernel. The weights associated to each kernel were estimated via grid search and the best performing weights were identified using cross-validation. This allowed the estimation of the SVM parameters using a conventional SVM solver. Results revealed accuracies of over 90% for the classification of AD versus

HC and over 80% for the classifications of MCI versus HC, respectively. Importantly, the combination of multiple kernels for multimodal classification outperformed both the use of a single modality and the simple concatenation of the features extracted from the different modalities.

A common problem in clinical neuroimaging is that the data are often incomplete, meaning that not all modalities may be available for all subjects. Standard strategies for dealing with this problem include (i) discarding subjects with incomplete data, which, however, causes a reduction of the sample size (Zhu et al., 2017) and (ii) imputing missing data. In a recent study, Zhu and colleagues attempted to address this issue using MKL by building a classifier for each modality and then combining the different kernels by summation (Zhu et al., 2017). Block-wise missing data, i.e., large portions of data are missing for one or more blocks of participants, were addressed embedding a nonlinear maximum mean discrepancy (MMD) mapping function, which does not require the same number of samples for each modality, in the MKL approach. The MMD criterion maps the data to a common reproducing kernel Hilbert space, where the different modalities are comparable. Information based on predefined ROIs from sMRI and PET images were used as input features. This approach was able to classify AD versus HC the OASIS database (Open Access Series of Studies, oasis-brains.org). The authors achieved this using a linear combination of kernels, where each kernel corresponded to a brain region as defined by an atlas. This identified several regions that are known to play a key role in memory and cognition.

8.3.2 Diagnostic classification of Parkinson's Disease

The differential diagnosis of parkinsonian syndromes is particularly difficult because of the fact that different pathologies share the same symptoms, especially during the early stages of the illness (Segovia et al., 2017). The gold standard tools for PD diagnosis are Single Photon Emission Computerized Tomography (SPECT) and PET. MKL has been used to combine these two modalities and has been shown to outperform the use of SPECT alone (Segovia et al., 2017). In this work, the authors used MKL to perform a binary classification (idiopathic and nonidiopathic PD) and a multigroup classification (PD, multiple system atrophy, and progressive supranuclear palsy) based on scans from 87 patients with parkinsonian symptoms. They obtained the best results (accuracies around 70%) using SPECT data for the binary classification and PET data for the multigroup classification. The weights of the kernels, specific to each region, were computed by averaging the weights obtained by the voxels within a kernel obtained while training an SVM classifier. This revealed that the striatum, olfactory bulb, thalamus, and supplementary motor area were the most important regions to separate the groups, in line with previous literature. SPECT in conjunction with sMRI was also used to discriminate PD patients from controls using data from 538 subjects who were part of the Parkinson's Progression Markers Initiative (PPMI) database (Adeli et al., 2017). Here, kernels were also linearly combined but, in contrast with previous studies, different types of kernels (linear, radial basis function (RBF), histogram kernel) were used to generate the ensemble kernel. This exploited both the linear and nonlinear properties of the data. The combination of sMRI and SPECT allowed classification accuracies over 90%; however, the analysis of the weights given to kernels suggested that SPECT was more informative than MRI for PD diagnosis. In addition, the putamen and caudate were identified as the regions providing the greatest contribution to classification. In the final application of MKL to PD discussed here, a larger number of potential biomarkers, including SPECT, CSF, plasma, serum, and RNA tests, were used to investigate 388 subjects from the PPMI dataset (Castillo-Barnes et al., 2018). Kernels were linearly combined, using fixed weights computed on the basis of the accuracy of classification based on single linear SVM kernels. The authors report very high accuracies of around 90% using imaging information only, suggesting that the other measures under investigation may not allow detection of PD at the level of the individual patient.

8.3.3 Diagnostics classification of psychosis

The possibility of gaining insight about complex data using MKL has been exploited in several studies of psychosis. For example, Ulas and colleagues applied MKL to sMRI data acquired from 59 patients with schizophrenia and 55 HCs to identify the neuroanatomical regions that are most affected in this disorder (Ulas et al., 2011). The kernels' parameters and weights for each neuroanatomical region were estimated simultaneously using cross-validation on the test set. Following this estimation, the authors performed a linear combination of the different kernels corresponding to the different regions. MKL significantly outperformed SVM in the classification and furthermore provided evidence for the involvement of the thalamus in schizophrenia. The same group (Castellani, Ulas, & Murino, 2011) used MKL to distinguish between 30 patients diagnosed with schizophrenia and 30 HCs using information about the shape of the thalamus, geometrically characterized by the heat diffusion associated with each specific shape. Here, kernels were associated to different scales used to compute the heat diffusion and linearly combined. The weights of the kernels and their parameters were estimated in a single optimization process. The accuracy of the classification with MKL reached 85% using the combination of all features; this was higher than the accuracy achieved using single features separately (78%). Using only the features with the highest four weights, an even higher accuracy of 88% was achieved.

Castro et al. (2011) applied MKL to an fMRI dataset of 54 patients with schizophrenia and 54 HCs. Data were preprocessed using a standard fMRI pipeline and dimensions were reduced using independent component analysis (ICA), from which the authors retained those components related to the temporal lobe and the default mode network. Components were then segmented into 116 regions derived from an atlas. Each MKL kernel corresponded to data coming from a specific region and then kernels were linearly combined. Parameters were tuned recursively for the kernel weights: this involves including all kernels at the start and then removing a kernel at each iteration, with a grid search to tune the kernel's parameters. This approach allowed the authors to obtain a very high classification accuracy (95%) and rank regions on the basis of their importance for discrimination between the two groups. Interestingly, temporal lobe regions and the cingulate gyrus were the most informative, consistent with findings that highlight the importance of the temporal and frontal regions in schizophrenia. The same group later published a follow-up study in which they applied MKL to fMRI data considering both magnitude and phase data, i.e., taking into account complex-valued spatiotemporal data that are part of the acquired fMRI signal (Castro, Gómez-Verdejo, Martínez-Ramón, Kiehl, & Calhoun, 2014). Also in this case, ICA was applied to data, which came from 31 patients with schizophrenia and 21 HCs. Each kernel corresponded to a brain region, and kernels, linear or nonlinear, were combined in a linear fashion. The parameters of the kernels and weights were simultaneously estimated, and an additional sparsity constraint was added, which ensured that nonrelevant regions were discarded from the analysis. The default mode network and the temporal regions, consistently found to be involved with schizophrenia in the literature, were identified as the most prominent for the classification.

More recently, MKL was used to identify relevant regions in recent onset psychosis in a study conducted by Squarcina et al. (2017), using cortical thickness obtained from images acquired from 127 HCs and 127 patients with a first episode of psychosis (FEP). Here, cortical thickness values were divided into 68 regions defined from a morphological atlas, and the 5 best-performing regions in an SVM classifier were used in MKL. Kernels were then linearly combined, with weights depending on subjects' covariates, aimed at reducing their confounding effect. In this way, it was possible to combine data acquired at 1.5T and at 3T in two different MRI machines, and the covariates were taken into account directly during the learning phase and not, as is usually done, in a preprocessing step. Notably, the performance of the different MKL models, which reached an accuracy of 85% when considering data acquired with a 3T scanner, was better than the performance of models that involved a simple concatenation of features and the use of preprocessing to minimize confounding effect. The temporal and frontal areas were detected as the most prominent brain regions for the classification of FEP patients from HCs.

8.4 Conclusion

The use of kernels in machine learning methods allows the identification of an optimal hyperplane for the separation of two classes (e.g., patients with a certain brain disorder of interest and HCs). When different acquisition modalities are available, using a single kernel for all the available data is a disadvantage because data heterogeneity cannot be fully exploited. MKL addresses this issue by, for example, allowing the integration of different kernels derived from different modalities or the automatic selection of the best performing kernel. It has been shown that MKL methods outperform single kernel methods when applied to brain disorders such as AD, PD, and psychosis, while at the same time making it possible to understand which features (e.g., brain region) are the most important for classification. This information can aid the interpretation of the results in the context of multimodal studies, helping identify the biological processes underlying brain disorders.

8.5 Key points

- While kernel-based algorithms are extensively used in brain disorders research, the selection of an appropriate kernel is not trivial.
- Single kernel methods are not suited for multimodal datasets because data heterogeneity cannot be fully exploited.
- In contrast, the heterogeneity of multimodal data can be fully exploited using multiple kernel methods.
- Multiple kernel methods, via optimization, have the added advantage of being able to identify the most prevalent features which distinguish between classes.
- The use of nonlinear combinations models, while capable of yielding more information than linear ones, can be computationally expensive.
- Multiple kernel methods tend to outperform single kernel methods when patients with neurological or psychiatric disorders are compared against HCs.

Acknowledgments

This Chapter was supported by grants from the Italian Ministry of Health to P.B. (RF-2016-02364582).

References

Adeli, E., Wu, G., Saghafi, B., An, L., Shi, F., & Shen, D. (2017). Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson's disease. *Scientific Reports*, *7*, 41069. https://doi.org/10.1038/srep41069.

Castellani, U., Rossato, E., Murino, V., Bellani, M., Rambaldelli, G., Perlini, C., et al. (2012). Classification of schizophrenia using feature-based morphometry. *Journal of Neural Transmission*, *119*(3), 395–404. https://doi.org/10.1007/s00702-011-0693-7.

Castellani, U., Ulas, A., & Murino, V. (2011). *Selecting scales by multiple kernel learning for shape diffusion analysis*. Retrieved from https://hal.inria.fr/inria-00624051.

Castillo-Barnes, D., Ramírez, J., Segovia, F., Martínez-Murcia, F. J., Salas-Gonzalez, D., & Górriz, J. M. (2018). Robust ensemble classification methodology for I123-ioflupane SPECT images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's disease. *Frontiers in Neuroinformatics*, *12*, 53. https://doi.org/10.3389/fninf.2018.00053.

Castro, E., Gómez-Verdejo, V., Martínez-Ramón, M., Kiehl, K. A., & Calhoun, V. D. (2014). A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: Application to schizophrenia. *NeuroImage*, *87*, 1–17. https://doi.org/10.1016/j.neuroimage.2013.10.065.

Castro, E., Martínez-Ramón, M., Pearlson, G., Sui, J., & Calhoun, V. D. (2011). Characterization of groups using composite kernels and multi-source fMRI analysis data: Application to schizophrenia. *NeuroImage*, *58*(2), 526–536. https://doi.org/10.1016/j.neuroimage.2011.06.044.

Fekete, T., Zach, N., Mujica-Parodi, L. R., & Turner, M. R. (2013). Multiple kernel learning captures a systems-level functional connectivity biomarker signature in amyotrophic lateral sclerosis. *PLoS One*, *8*(12), e85190. https://doi.org/10.1371/journal.pone.0085190.

Gönen, M., & Alpaydin, E. (2008). Localized multiple kernel learning. In *Proceedings of the 25th international conference on Machine learning - ICML '08* (pp. 352–359). New York, New York, USA: ACM Press. https://doi.org/10.1145/1390156.1390201.

Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*, 2211–2268. Retrieved from files/1477/JMLR-Gonen-Alpaydin-2011-Multiple_Kernel_Learning_Algorithms.pdf.

Gönen, M., & Alpaydın, E. (2013). Localized algorithms for multiple kernel learning. *Pattern Recognition*, *46*(3), 795–807. https://doi.org/10.1016/J.PATCOG.2012.09.002.

Gu, Y., Liu, T., Jia, X., Benediktsson, J. A., & Chanussot, J. (2016). Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(6), 3235–3247. https://doi.org/10.1109/TGRS.2015.2514161.

Han, Y., Yang, Y., Li, X., Liu, Q., & Ma, Y. (2018). Matrix-regularized multiple kernel learning via (r,p) norms. *IEEE Transactions on Neural Networks and Learning Systems*, 4997–5007. https://doi.org/10.1109/TNNLS.2017.2785329.

Kloft, M., Brefeld, U., Sonnenburg, S., & Zien, A. (2011). Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, *12*(Mar), 953–997. Retrieved from http://www.jmlr.org/papers/v12/kloft11a.html.

Lanckriet, G. R., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M. I., Jordan Lanckriet, M. I., et al. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72. https://doi.org/10.1162/153244304322765649.

Li, X., Bai, Y., Peng, Y., Du, S., & Ying, S. (2018). Nonlinear semi-supervised metric learning via multiple kernels and local topology. *International Journal of Neural Systems, 28*(02), 1750040. https://doi.org/10.1142/S012906571750040X.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall. Retrieved from https://www.crcpress.com/Generalized-Linear-Models-Second-Edition/McCullagh-Nelder/p/book/9780412317606.

Noble, W. S. (2004). Support vector machine applications in computational biology. In *Kernel Methods in computational biology Bernhard Scholkopf, Koji Tsuda, and Jean-Philippe Vert*. The MIT Press.

Peruzzo, D., Castellani, U., Perlini, C., Bellani, M., Marinelli, V., Rambaldelli, G., et al. (2015). Classification of first-episode psychosis: A multi-modal multi-feature approach integrating structural and diffusion imaging. *Journal of Neural Transmission (Vienna, Austria: 1996), 122*(6), 897−905. https://doi.org/10.1007/s00702-014-1324-x.

Pettersson-Yeo, W., Benetti, S., Marquand, A. F., Joules, R., Catani, M., Williams, S. C. R., et al. (2014). An empirical comparison of different approaches for combining multimodal neuroimaging data with support vector machine. *Frontiers in Neuroscience, 8*, 189. https://doi.org/10.3389/fnins.2014.00189.

Raamana, P. R., Weiner, M. W., Wang, L., & Beg, M. F. (2015). & alzheimer's disease neuroimaging initiative, for the A. D. N. *Thickness Network Features for Prognostic Applications in Dementia. Neurobiology of Aging, 36*(Suppl. 1), S91−S102. https://doi.org/10.1016/j.neurobiolaging.2014.05.040.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2007). SimpleMKL. *Journal of Machine Learning Research, 9*, 2491−2521. Retrieved from https://hal.inria.fr/hal-00218338/.

Sanderman, R., Coyne, J. C., & Ranchor, A. V. (2006). Age: Nuisance variable to be eliminated with statistical control or important concern? *Patient Education and Counseling, 61*(3), 315−316. https://doi.org/10.1016/j.pec.2006.04.002.

Schrouff, J., Monteiro, J. M., Portugal, L., Rosa, M. J., Phillips, C., & Mourão-Miranda, J. (2018). Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models. *Neuroinformatics, 16*(1), 117−143. https://doi.org/10.1007/s12021-017-9347-8.

Segovia, F., Górriz, J. M., Ramírez, J., Martínez-Murcia, F. J., Levin, J., Schuberth, M., et al. (2017). Multivariate analysis of 18F-DMFP PET data to assist the diagnosis of parkinsonism. *Frontiers in Neuroinformatics, 11*, 23. https://doi.org/10.3389/fninf.2017.00023.

Squarcina, L., Castellani, U., Bellani, M., Perlini, C., Lasalvia, A., Dusi, N., et al. (2017). Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. *NeuroImage, 145*, 238−245. https://doi.org/10.1016/j.neuroimage.2015.12.007.

Squarcina, L., Perlini, C., Bellani, M., Lasalvia, A., Ruggeri, M., Brambilla, P., & Castellani, U. (2015). Learning with heterogeneous data for longitudinal studies. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351* (pp. 535−542).

Surampudi, S. G., Naik, S., Surampudi, R. B., Jirsa, V. K., Sharma, A., & Roy, D. (2018). Multiple kernel learning model for relating structural and functional connectivity in the brain. *Scientific Reports, 8*(1), 3265. https://doi.org/10.1038/s41598-018-21456-0.

Tanabe, H., Bao Ho, T., Nguyen, C. H., & Kawasaki, S. (2008). Simple but effective methods for combining kernels in computational biology. In *2008 IEEE international conference on research, innovation and Vision for the future in computing and communication technologies* (pp. 71−78). IEEE. https://doi.org/10.1109/RIVF.2008.4586335.

Ulas, A., Duin, R. P. W., Castellani, U., Loog, M., Mirtuono, P., Bicego, M., et al. (2010). Dissimilarity-based detection of schizophrenia. In *2010 first workshop on brain decoding: Pattern recognition challenges in neuroimaging* (pp. 32–35). IEEE. https://doi.org/10.1109/WBD.2010.10.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley. Retrieved from https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034.

Wen, H, Liu, Y, Rekik, I, Wang, S, Zhang, J, Peng, Y, & He, H (2017). Disrupted topological organization of structural networks revealed by probabilistic diffusion tractography in Tourette syndrome children. *Human Brain Mapping*, *38*(8), 3988–4008.

Wilson, C. M., Li, K., Kuan, P., & Wang, X. (2018). *Multiple-kernel learning for genomic data mining and prediction*.

Xu, Z., Jin, R., Yang, H., King, I., & Lyu, M. R. (2010). Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th international conference on International Conference on Machine Learning (ICML'10), Johannes fürnkranz and Thorsten Joachims* (pp. 1175–1182). USA: Omnipress.

Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., et al. (2008). Heterogeneous data fusion for alzheimer's disease study. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining–KDD 08* (p. 1025). New York, New York, USA: ACM Press. https://doi.org/10.1145/1401890.1402012.

Zhang, Y., Li, A., Peng, C., & Wang, M. (2016). Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(5), 825–835. https://doi.org/10.1109/TCBB.2016.2551745.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative, A. D. N. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, *55*(3), 856–867. https://doi.org/10.1016/j.neuroimage.2011.01.008.

Zhuang, J., Wang, J., Hoi, S. C. H., & Lan, X. (2011). Unsupervised multiple kernel learning. *ACML*, *20*, 129–144. Retrieved from http://proceedings.mlr.press/v20/zhuang11/zhuang11.pdf.

Zhu, X., Thung, K.-H., Adeli, E., Zhang, Y., & Shen, D. (2017). Maximum mean discrepancy based multiple kernel learning for incomplete multimodality neuroimaging data. Medical image Computing and computer-assisted Intervention: Miccai. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, *10435*, 72–80. https://doi.org/10.1007/978-3-319-66179-7_9.