Learning with Heterogeneous Data for Longitudinal Studies

Letizia Squarcina[1], Cinzia Perlini[1], Marcella Bellani[1], Antonio Lasalvia[1], Mirella Ruggeri[1], Paolo Brambilla[3], and Umberto Castellani[2,⋆]

[1] University of Verona, Department of Psychiatry, Verona, Italy
[2] University of Verona, Department of Computer Science, Verona, Italy
[3] Department of Neurosciences and Mental Health, Psychiatric Clinic, Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy

Abstract. Longitudinal studies are very important to understand cerebral structural changes especially during the course of pathologies. For instance, in the context of mental health research, it is interesting to evaluate how a certain disease degenerates over time in order to discriminate between pathological and normal time dependent brain deformations. However longitudinal data are not easily available, and very often they are characterized by a large variability in both the age of subjects and time between acquisitions (follow up time). This leads to heterogeneous data that may affect the overall study. In this paper we propose a learning method to deal with this kind of heterogeneous data by exploiting covariate measures in a Multiple Kernel Learning (MKL) framework. Cortical thickness and white matter volume of the left middle temporal region are collected from each subject. Then, a subject-dependent kernel weighting procedure is introduced in order to obtain the correction of covariate effect simultaneously with classification. Experiments are reported for First Episode Psychosis detection by showing very promising results.

Keywords: Support Vector Machines, Multiple Kernel Learning, Longitudinal study, First Episode Psychosis.

1 Introduction

The analysis of anatomical variability over time is a relevant topic in neuroimaging research to characterize the progression of a certain disease [8]. The overall aim is to detect structural changes in spatial and time domains that are able to differentiate between patients and healthy controls. To this end, longitudinal studies are considered where the same subject is observed repeatedly at several time points [4]. In the simplest scenarios only two time points are defined, namely *baseline* and *follow up*. In this fashion a classification method, like Support Vector Machine (SVM) [18], can be designed for each time point and, more interestingly, for the differential degeneracy measurements to evaluate how

⋆ Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 535–542, 2015.
DOI: 10.1007/978-3-319-24574-4_64

the dynamics of the observed features may change across the populations (i.e., patients and controls). Unfortunately, a relevant issue in longitudinal studies is the lack of reliable datasets. In many cases collected data are heterogeneous due to the large variability of the considered population. In particular, there are different factors that may contribute to brain deformations such as gender, age, follow up time, and so on. In general, these factors are considered as *nuisance* variables that cause a *dispersion* of the observed values (i.e., the main variable) by possible affecting the final analysis result [15]. For instance, in brain classification if the volume reduction is dependent on both the analysed disease and the age of subject, it may be difficult to discriminate between a young patient and an elderly control. A common approach to deal with this problem consists of introducing some correction procedure based on statistical regression. As an example the *Generalized Linear Model* (GLM) [13] can be employed to predict and remove the effect of the confounding covariates. In general this procedure is carried out as a pre-processing before classification. In this paper our intent is to integrate the data correction into the learning phase by defining a classification model that explicitly takes into account of the nuisance covariates. We exploit Multiple Kernel Learning (MKL) as a powerful and natural approach to combine different sources of information [7] that has already been successfully applied in the neuro-imaging field [2,8]. The main idea behind MKL methods [1,7] is to learn an optimal weighted combination of several kernels while simultaneously training the classifier ([7]). In general, each kernel is associated to a specific feature and their combination is carried out aiming at exploiting interdependent information from the given features. In this work, we propose to adapt the standard MKL formulation to enable the correction of the covariate effect in the kernel space by overcoming the hypotesis of linear progression. We show that this approach leads to a subject-specific kernel weighting scheme where each weight is locally dependent on the subject covariate. A similar strategy is known in literature as *localized* kernel learning [6,7] where the idea consists of assigning different weights to a kernel in different regions of the input space. We evaluate our method to improve the classification performance of longitudinal study [4] to automatically detect First Episode Psychosis (FEP) [10,17]. Research on the first stages of psychosis is very important to understand the causes of the disease, excluding the consequence of chronicity and treatment on the study outcomes. In our study we evaluate the brain variation by analysing several cortical features, namely the *volume* and the surface *thickness* [3,5]. In our paper we focus on the *middle temporal* region since it is already known its relation with psychosis [10,17]. We considered as confounding covariate the age of subjects since in our population the span of age is quite large and heterogeneous. For each subject two MRI scans are available, i.e., baseline and follow up. Classification is carried out for the two time points, and for the differential values of volume and thickness. In the latter experiment the follow up time is considered as covariate together with the age. Experiments show promising results by reporting a clear improvement when covariates are integrated in the learning phase. Moreover our study evidences that also brain deformation over the time can be used to discriminate

between patients and controls by confirming the idea that speed of structural changes is faster in mental disorders.

2 CO-MKL: Multiple Kernel Learning with Covariate

2.1 Background

The overall idea of kernel methods is to project non-linearly separable data into a higher dimensional space where the linear class separation becomes feasible [18]. Support Vector Machine (SVM) is an example of discriminative classifier belonging to this class of methods for binary classification problems. Given a training set of N samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i \in R^d$ is a vector of the input space and $y_i \in \{-1, +1\}$ is the sample class label, SVM finds the best discriminative hyperplane in the feature space according to the maximum margin principle [18]. Therefore, after training, the discriminative function is defined as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) \rangle + b, \tag{1}$$

where Φ is the mapping function, and α_i are the dual variables. Considering the so called "kernel trick" the dot product in Equation 1 is replaced by a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) \rangle$. Several kernel functions can be defined such as the linear kernel, the polynomial kernel, and the Gaussian Radial Basis Function (RBF) kernel. An interesting extension of SVMs is represented by Multiple Kernel Learning (MKL) methods[7], recently proposed in [1,11]. MKL aims at selecting or combining a set of different kernels that usually describes different data sources. The simplest combination strategy is given by the linear MKL model:

$$k_w(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{P} w_m k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m), \tag{2}$$

where $w_m \in \mathbb{R}$, $\boldsymbol{x}_i = \{\boldsymbol{x}_i^m\}_{m=1}^{P}$, $\boldsymbol{x}_i^m \in R^{d_m}$, and d_m is the size of the mth feature. In particular, MKL methods enable the estimation of the kernel weights and the coefficients of the associated predictor in a single learning framework.

2.2 Proposed Approach

In our method we address the problem of integrating the confounding covariate in a supervised fashion by including the data correction in the learning phase. We assume the knowledge of covariate for each subject. Therefore, each sample is now represented by $\{(\boldsymbol{x}_i, \boldsymbol{c}_i, y_i)\}$ where $\boldsymbol{c}_i \in R^h$ encodes the covariates associated to sample i. Rather than working on the original space, our main idea consists of correcting the effect of the covariate in the kernel space. More specifically, the decision function reported in Equation 1 is redefined as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i \langle w(\boldsymbol{c}_i)\Phi(\boldsymbol{x}_i), w(\boldsymbol{c})\Phi(\boldsymbol{x}) \rangle + b, \tag{3}$$

where a weighting function $w(\boldsymbol{c})\colon R^h \rightarrow R$ is introduced to take into account of the covariates. Note that in this fashion all the covariate are simultaneously considered in the correction procedure. In order to deal with multiple features the proposed correction method is extended to a MKL framework by combining Equations 3 and 2:

$$k_w^c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{P} \langle w_m(\boldsymbol{c}_i)\Phi(\boldsymbol{x}_i^m), w_m(\boldsymbol{c}_j)\Phi(\boldsymbol{x}_j^m)\rangle = \sum_{m=1}^{P} w_m(\boldsymbol{c}_i)k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)w_m(\boldsymbol{c}_j). \quad (4)$$

As an important assessment we observe that kernel definition of Equation 4 has the same form of the so called *localized kernel* [6]. The difference regards the meaning of the weighting function $w_m(.)$ that in [6] is defined on the original space rather than on the covariates. This leads to a different interpretation of the weighting effect. In [6] the weighting aims at estimating different classification criteria at different region of the input space [16]. In our approach the main contribution of the weighting function is to attenuate the conditioning of nuisance variables. Moreover, we can consider the covariate as a sort of *side* information that are included in the classification model. The weighting function $w_m(.)$ is defined as a softmax model[7]. The model parameters of the proposed MKL with covariate (CO-MKL) are computed using standard alternating optimization[7]: first, the kernel weights are fixed and standard solvers (i.e., libSVM) are used to estimate the SVM parameters, second, the SVM parameters are fixed and a gradient descent strategy is employed to compute kernels weights. The two steps are iterated until convergence.

3 Materials and Methods

The pipeline we adopted can be schematically visualized in Figure 1. After data collection, extracted features and covariates are fed to the classifier. This

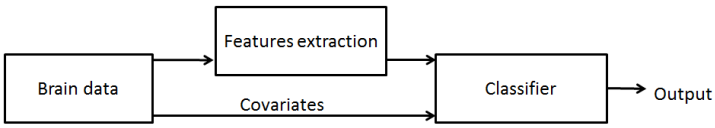

Fig. 1. Features, i.e. thickness and volume of temporal white and gray matter, were extracted from brain data, and features were used for classification together with covariates (age or months between different acquisitions).

study was conducted in the frame of Psychosis Incident Cohort Outcome Study (PICOS), a multi-site naturalistic research study, which aimed at analyzing clinical, social, cognitive, genetic and imaging features of FEP patients [12]. 22 First

Episode of Psychosis patients (FEP, 12 males, mean age 37.5 ± 7.9 y.o. at baseline, 39.4 ± 7.6 y.o. at follow up) and 22 healthy controls (HC, 12 males, 39.0 ± 11.7 y.o. at baseline, 42.0 ± 12.2 y.o. at follow up) underwent two Magnetic Resonance Imaging sessions (months between acquisitions (Δ_m) 21.1 ± 8.4 for FEP, 36.7 ± 7.8 for HC) with a 1.5 Siemens Symphony scanner. The two populations were matched for age and sex both at baseline and at follow up (t-test, $p>0.05$). The difference in months between the two scans (follow up time) was significantly higher for controls (t-test, $p<0.05$). To avoid bias in the classification we normalized the values independently in the two populations in order to obtain two vectors ranging from 0 to 1, with no statistical difference (t-test, $p>0.05$). T1-weighted images(256x256x18 voxels, 1x1x5 mm^3, TR 2160 ms, TE 47 ms, flip angle 90^o) were acquired for all subjects. Data were analyzed using FreeSurfer [3] and volume and cortical thickness were computed for the left medial temporal gray and white matter. With this methodology meshes of the boundaries between white and gray matter and between gray matter and cerebrospinal fluid are built. Thickness is defined as the distance between the meshes. Volume is computed assigning a label corresponding to a brain structure to each voxel of the image based on probabilistic information [5]. Thickness and volume are computed for several regions of interest (ROIs). We chose to restrict our analyses to the medial temporal white and gray matter, because of the involvement of this area of the brain with early psychosis. Mean region volume at baseline (μ_A^V) and at follow up (μ_B^V) and mean region thickness at baseline (μ_A^T) and follow up (μ_B^T) were computed for each subject, and used as features for the classifier. We also computed the difference in mean thickness and volume for each subject and used it as features $\Delta\mu^X = \frac{\mu^X{}_A - \mu^X{}_B}{\mu^X{}_A}$, where X is T or V in the case of thickness or volume, respectively.

All experiments were conducted with radial basis functions (RBF) kernels. The optimal values for free parameters were chosen with a search in a grid as shown in [9]. Cross validation was done with a leave-one-out procedure for all classification methods. For each subject, the complete set of features is $\boldsymbol{x} = [\mu_A^T, \mu_B^T, \mu_A^V, \mu_B^V, \Delta\mu^T, \Delta\mu^V]$ and that of covariates is $\boldsymbol{c} = [\text{age}_A, \text{age}_B, \Delta_m]$ where age_A and age_B are age at baseline and at follow up.

4 Results

We evaluated classification methods using baseline, follow-up, and difference values, considering as covariate baseline age, follow-up age and baseline age in conjunction with the normalized follow up time in months (Δ_m), respectively. For comparison, we employed standard SVM on single features and simple MKL [14] on both thickness and volume, to evaluate if the introduction of covariates in the CO-MKL method overtakes the effect of considering more than one feature simultaneously. We also used the covariates as features in SVM and MKL classification of baseline and follow-up data, to ensure that these values do not discriminate between the classes when used as features. More in details, analyses were carried out as follows:

- SVM classification using either middle temporal thickness or volume as features, adjusted for age using GLM. When considering the difference in thickness of volume ($\Delta\mu^T$ and $\Delta\mu^V$ respectively) the adjustment with GLM was done also for differences in follow up months.
- MKL where features are both thickness and volume adjusted for differences using GLM.
- CO-MKL with a single feature that can be the mean volume (at baseline, or at follow up) or thickness (at baseline, or at follow up), or the difference $\Delta\mu^X$ (thus $\boldsymbol{x} = \mu^T_{AorB}$ or $\boldsymbol{x} = \mu^V_{AorB}$ or $\boldsymbol{x} = \Delta\mu^X$). The covariate is age of subjects (at baseline or at follow up) or both age at baseline and follow up time in the case of the difference ($\boldsymbol{c} = \text{age}_A$ or age_B, or $\boldsymbol{c} = [\text{age}_A, \Delta_m]$).
- CO-MKL with multiple features at baseline, or at follow up, or for differences ($\boldsymbol{x} = [\mu^T_A, \mu^V_A]$ or $\boldsymbol{x} = [\mu^T_B, \mu^V_B]$ or $\boldsymbol{x} = [\Delta\mu^T, \Delta\mu^V]$). The covariate $\boldsymbol{c}$ is age of subjects for baseline and follow up, or both age at baseline and follow up time in the case of differences ($\boldsymbol{c} = \text{age}_A$ or $\boldsymbol{c} = \text{age}_B$, or $\boldsymbol{c} = [\text{age}_A, \Delta_m]$).
- MKL where both main variables and covariates are used as features. Therefore at baseline $\boldsymbol{x} = [\mu^T_A, \mu^V_A, age_A]$, at follow up $\boldsymbol{x} = [\mu^T_B, \mu^V_B, \text{age}_B]$, and for the difference $\boldsymbol{x} = [\Delta\mu^T, \Delta\mu^V, \text{age}_A, \Delta_m]$.
- SVM with the concatenation of both main variables and covariates.

Table 1. Results from classification from data of thickness and volume of middle temporal gray and white matter.

	Baseline			**Follow up**			Δ		
	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec
CO-MKL									
Thickness	70.4	54.6	86.4	**81.8**	77.3	86.4	70.5	59.1	81.8
Volume	**72.7**	54.6	89.0	68.2	59.0	77.3	**72.7**	63.6	81.8
SVM									
Thickness	65.9	50.0	81.8	61.3	63.6	59.0	56.8	59.0	54.6
Volume	72.7	50.0	95.5	65.9	54.5	77.3	56.9	72.7	42.0

All results are reported in Table 1 and 2. We observe that CO-MKL reached higher values of accuracy than SVM where data are adjusted for age or between scans (82% vs 66%), demonstrating that correcting for nuisance variables in a MKL framework is more efficient than adjusting the values with a linear model. Moreover, CO-MKL with both thickness and volume values as features did not improve the accuracy in respect with employing CO-MKL using only the thickness as feature, but reaches better results than standard simple MKL used with both features μ^V and μ^T, demonstrating that out methods improves the classification performances in merging different information. As expected, MKL using covariates as features did not reach good values of accuracy, with a maximum of 66% in the case of follow up data. Also using SVM concatenating all variables of interest did not improve classification results, demonstrating once again that age and follow up time are useful when used to correct the classification but not if considered as features.

Table 2. Results from classification from data of thickness and volume of middle temporal gray and white matter considered together.

	Baseline			**Follow up**			Δ		
	Acc	Sens	Spec	Acc	Sens	Spec	Acc	Sens	Spec
CO-MKL	**70.5**	68.2	72.7	**81.8**	86.4	77.3	**70.45**	81.8	59.1
MKL	43.2	61.4	27.3	61.3	50.0	72.7	59.1	77.3	31.8
MKL + cov.	59.1	45.5	72.7	47.7	54.5	40.9	54.5	31.8	77.2
CON - SVM	72.7	50.0	95.4	65.9	54.6	77.3	65.9	59.1	72.7

It is to be noted that the use of both thickness and volume simultaneously did not improve performances in this specific case, but could allow a more comprehensive evaluation of degeneracy processes in respect to the analysis of a single brain feature. Moreover, even if classification using $\Delta\mu^X$ did not reach very high values, an accuracy of around 70% can be considered promising, since changes in brain structure are not easily detectable in psychosis patients, and points towards possible modifications in the rate of brain changes in patients in respect to healthy.

5 Conclusions

In this work, we propose CO-MKL to exploit supervised correction of the effects of confounding variables, with the aim of compensating possible unbalanced dataset. In fact, the presence of heterogeneity in longitudinal datasets is very common in neuropsychiatry research. We showed how automatic classification improves when the spurious differences between populations are taken into account into a process of supervised correction instead of ignored or integrated in simple models, as GLM. We focused our analysis to the middle temporal region of the brain, because it is known that it is involved in psychosis. We obtained high accuracies, up to more than 80%, demonstrating that brain structure and in particular thickness and volume of definite ROIs are markers of psychosis also at early stages. Moreover, we showed that the progression of the disease by itself, which we measured computing the difference in thickness or volume over time, could be employed to distinguish healthy from patients, suggesting that there could be a change in the speed of brain deformation in psychiatric diseases. The number of subjects we considered is limited: we plan on recruiting more patients and controls, as well as extending the analysis to other confounds to use as covariates, as for example medications or clinic scales both at baseline and at follow-up.

References

1. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML 2004, pp. 41–48 (2004)
2. Castro, E., Martínez-Ramón, M., Pearlson, G., Sui, J., Calhoun, V.D.: Characterization of groups using composite kernels and multi-source fmri analysis data: Application to schizophrenia. Neuroimage 58, 526–536 (2011)

3. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9(2), 179–194 (1999)
4. Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N.: Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. International Journal of Computer Vision 103(1), 22–59 (2013)
5. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C.E.A.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33(3), 341–355 (2002)
6. Gönen, M., Alpaydın, E.: Localized multiple kernel learning. In: ICML 2008, pp. 352–359 (2008)
7. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2181–2238 (2011)
8. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., Initiative, T.A.D.N.: Predictive markers for ad in a multi-modality framework: An analysis of mci progression in the adni population. Neuroimage 55, 574–589 (2011)
9. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2010). http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
10. Kuroki, N., Shenton, M.E., Salisbury, D.F., Hirayasu, Y., Onitsuka, T., Ersner-Hershfield, H., Yurgelun-Todd, D., Kikinis, R., Jolesz, F.A., McCarley, R.W.: Middle and inferior temporal gyrus gray matter volume abnormalities in first-episode schizophrenia: an MRI study. Am. J. Psychiatry 163(12), 2103–2110 (2006)
11. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.: Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5, 27–72 (2004)
12. Lasalvia, A., Tosato, S., Brambilla, P., Bertani, M., Bonetto, C., Cristofalo, D.E.A.: Psychosis Incident Cohort Outcome Study (PICOS). A multisite study of clinical, social and biological characteristics, patterns of care and predictors of outcome in first-episode psychosis. Background, methodology and overview of the patient sample. Epidemiol. Psychiatr. Sci. 21(3), 281–303 (2012)
13. Peter, P.M., Nelder, J.: Generalized Linear Models. Chapman and Hall (1989)
14. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. Journal of Machine Learning Research 9, 2491–2521 (2008)
15. Sederman, R., Coyne, J.C., Ranchor, A.V.: Age: Nuisance variable to be eliminated with statistical control or important concern? Patient Education and Counseling 61(37), 315–316 (2006)
16. Segate, N., Blanzieri, E.: Fast and scalable local kernel machines. Journal of Machine Learning Research 11, 1883–1926 (2010)
17. Tang, J., Liao, Y., Zhou, B., Tan, C., Liu, W., Wang, D., Liu, T., Hao, W., Tan, L., Chen, X.: Decrease in temporal gyrus gray matter volume in first-episode, early onset schizophrenia: an MRI study. PLoS One 7(7), e40247 (2012)
18. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons (1998)