A Localized MKL Method for Brain Classification with Known Intra-class Variability

Aydın Ulaş[1,⋆], Mehmet Gönen[2], Umberto Castellani[1], Vittorio Murino[1,3], Marcella Bellani[4], Michele Tansella[4], and Paolo Brambilla[5]

[1] Department of Computer Science, University of Verona (UNIVR), Verona, Italy
mehmetaydin.ulas@univr.it
[2] Aalto University, Department of Information and Computer Science, Finland
[3] Istituto Italiano di Tecnologia (IIT), Genova, Italy
[4] Dpt Public Health & Community Medicine, Psychiatry, ICBN, UNIVR, Italy
[5] IRCCS "E. Medea" Scientific Institute, Udine, Italy

Abstract. Automatic decisional systems based on pattern classification methods are becoming very important to support medical diagnosis. In general, the overall objective is to classify between healthy subjects and patients affected by a certain disease. To reach this aim, significant efforts have been spent in finding reliable biomarkers which are able to robustly discriminate between the two populations (i.e., patients and controls). However, in real medical scenarios there are many factors, like the gender or the age, which make the source data very heterogeneous. This introduces a large intra-class variation by affecting the performance of the classification procedure. In this paper we exploit how to use the knowledge on heterogeneity factors to improve the classification accuracy. We propose a *Clustered Localized* Multiple Kernel Learning (CLMKL) algorithm by encoding in the classication model the information on the clusters of apriory known stratifications.

Experiments are carried out for brain classification in Schizophrenia. We show that our algorithm performs clearly better than single kernel Support Vector Machines (SVMs), linear MKL algorithms and canonical Localized MKL algorithms when the gender information is considered as apriori knowledge.

Keywords: brain imaging, magnetic resonance imaging, computer-aided diagnosis, localized multiple kernel learning, schizophrenia.

1 Introduction

Advanced pattern recognition methods have demonstrated their growing importance in the medical domain for the definition of new decisional systems able to support medical diagnosis. In particular, in neuroscience the use of brain classification methods represents a recent and relevant trend aiming at discriminating healthy subjects from patients having a certain mental disorder [13,17]. This leads to a two-class classification problem that is addressed by using for instance

⋆ Strada Le Grazie 15, 37134, Verona, VR, Italy.

F. Wang et al. (Eds.): MLMI 2012, LNCS 7588, pp. 152–159, 2012.

discriminative learning methods like Support Vector Machine (SVM). However, in practical situations the performance of classifiers are highly affected by intra-class variations. For instance in brain classification there is a general diversity of the brain properties between male and female (in both patients and controls). In this paper, we propose a new brain classification method which encodes explicitly the known intra-class variability into the classification model. To this aim, we benefit from recently proposed *localized Multiple Kernel Learning* (LMKL) [8] approaches. In LMKL, the idea is to define a decision function whose parameters depend on the input data, i.e., *localized* information. In practice, similar to classifier selection [19,12], the localized estimates are used to select or combine kernels [8]. In our work, we specialize this approach to design a new model which embeds the information on the clusters of a pre-defined data stratification (i.e., male and female). Instead of adopting a *sample-specific* localization like in [8], we introduce a *cluster localized* approach to set up the combination scheme. The difference of our method from LMKL is that instead of letting the algorithm choose the partitioning, we use apriori partitioning based on expert knowledge. We call our method *Clustered Localized Multiple Kernel Learning* (`CLMKL`). In this fashion, according to the spirit of Multiple Kernel Learning (MKL) methods [9], we learn a separate combination of input kernels for each cluster.

MKL methods have been recently proposed on the medical domain to detect Alzheimer's disease [10,6]. In Castro et al. [4], a recursive composite kernel method is applied for schizophrenia. In these works, MKL approach was employed to integrate/select different factors of the disease. Note that also our MKL formulation can deal naturally with different sources of information as shown in the experiments. We evaluate our method on brain classification for Schizophrenia detection. Several experimental configurations are evaluated as well as a comparison with other MKL classification methods by showing a clear improvement of our method. Our method allows to train all data in different clusters together, thus avoiding the reduction of training examples and over-training. This can clearly be seen when we compare our method with separating the clusters and training/testing a single model for each cluster.

The paper is organized as follows: in Section 2, we introduce the MKL framework and our methodology, we show our experiments and results in Section 3 and we conclude in Section 4.

2 Methodology

2.1 Multiple Kernel Learning

The assumption behind kernel methods is to transform linearly unseparable data into a higher dimensional (possibly with infinite dimension) space where it is possible to separate the classes linearly [18]. The support vector machine (SVM) in this sense is a discriminative classifier which is based on the theory of structural risk minimization proposed for binary classification problems. Given a sample of N training instances $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i$ is the D-dimensional input vector and $y_i \in \{-1, +1\}$ is its class label, SVM finds the linear discriminant

with the maximum margin in the feature space induced by a mapping function $\Phi\colon \mathbb{R}^D \rightarrow \mathbb{R}^S$. Considering the dual formulation with the "kernel trick", the discriminant function can be rewritten as

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

where $k\colon \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ is called the *kernel function* (similarity between instances of data) and $\boldsymbol{\alpha}$ denotes the dual variables corresponding to each training sample.

There are several kernel functions successfully used in the literature, such as the linear kernel, the polynomial kernel, and the Gaussian kernel. Selecting the kernel function and its parameters is an important issue in training. Generally, a separate validation set is used to choose the best performing kernel among a set of kernels. Recently, multiple kernel learning (MKL) methods have been proposed [2,15], which learn a combination k_η instead of selecting a specific kernel and its corresponding parameters. The simplest way is to combine the kernels as a weighted sum which corresponds to the linear MKL:

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

with $\eta_m \in \mathbb{R}$. Different versions of this approach differ in the way they put restrictions on the kernel weights: [2,15,16]. This is similar to classifier combination [14] in the sense that instead of choosing a single classifier, we select a set of classifiers and let the algorithm do the picking. MKL can be used for selecting/combining a set of different kernels which correspond to different notions of similarity or can be used to combine different sources of information probably with different dimensions which in our case correspond to different parts of the brain. In this work, we compare our method with `RBMKL` and `SMKL`, where `RBMKL` denotes the rule-based MKL algorithm that trains an SVM with the mean of the combined kernels [5], `SMKL` is the iterative algorithm of [16] that uses projected gradient updates and trains single-kernel SVMs at each iteration.

2.2 Our Method

Given a set of base classifiers, the idea behind classifier combination [14] is to find a function to accurately combine the decisions of individual base classifiers. Classifier selection [19,12] is different than classifier combination in the sense that the combination is also based on the input data point through a gating function [11]. In a similar setting, Gönen and Alpaydın [8] propose a data-dependent formulation called localized multiple kernel learning (`LMKL`) that combines kernels using weights calculated from a gating model where the gating model $\eta_m(\cdot\,|\,\cdot)$, parameterized by $\mathbf{V}$, assigns a weight to the feature space obtained with $\Phi_m(\cdot)$. Then the combined kernel matrix is represented as

$$k_\eta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{P} \eta_m(\boldsymbol{x}_i|\mathbf{V}) k_m(\boldsymbol{x}_i, \boldsymbol{x}_j) \eta_m(\boldsymbol{x}_j|\mathbf{V}).$$

This gating function can be formulated to be learned from the data so that the similarity is computed using multiple kernels where the kernel weights not only depend on kernel functions but on the input data. This can be done in an unsupervised way using the stratifications in the training data but in some applications the stratification of input data can be known apriori. For instance in medical applications the population can be subdivided into males and females. The crucial step is to formulate a good gating function to incorporate this apriori information and in this work we propose a gating function in order to take into account the knowledge of *intra*-class variability. In a medical application, the overall aim is the classification between healthy subjects and patients affected by a certain disease (i.e., two-class classification). In particular, we want the gating function to behave differently w.r.t. the gender (i.e., two apriori known subject stratifications). With this idea in mind, we embed the apriori clustering information and we formulate the following gating function based on softmax:

$$\eta_m(\boldsymbol{x}|\mathbf{V}) = \sum_{c=1}^{K} \delta_c(c_{\boldsymbol{x}}) \frac{\exp(v_c^m)}{\sum_{h=1}^{P} \exp(v_c^h)} \qquad \forall m \tag{1}$$

where $\mathbf{V} = \{v_1^m, v_2^m, \ldots, v_K^m\}_{m=1}^P$ are the weights (v_i^m is the weight if ith cluster and mth kernel), K is the number of clusters, $c_{\boldsymbol{x}}$ denotes the cluster of $\boldsymbol{x}$, and $\delta_c(c_{\boldsymbol{x}})$ is the Kronecker delta where $\delta_c(c_{\boldsymbol{x}}) = 1$ if $c_{\boldsymbol{x}} = c$, and 0 otherwise. We will refer to our method as `CLMKL` throughout the text. With this formulation, we get a constant set of weights for each cluster (gender). When the similarity between a data point and another one within the same cluster is computed, the same weights are used. But, this effect is reduced when the similarity is computed between two data points belonging to different clusters. For example, if the weight of a kernel is 0, when we compute the similarity between two data points belonging to two different clusters, this kernel is ignored. Only the kernels with nonzero weights contribute to the computation of similarities between inter-cluster data points. The gating model parameters are computed using alternating optimization: first, the kernel weights are fixed and the SVM parameters are estimated by standard solvers (i.e., libSVM), second, the SVM parameters are fixed and kernel weights are estimated by a gradient descent procedure. The two steps are iterated until convergence (starting from a random initialization of the weights). The gating function is chosen in order to enforce the weights to be in the interval between 0 and 1.

3 Experiments and Results

3.1 Data Set

The study population used in this work consists of 42 patients (21 male, 21 female) who were being treated for schizophrenia and 40 controls (19 male, 21 female) with no DSM-IV axis I disorders and had no psychiatric disorders

among first-degree relatives. Diagnoses for schizophrenia were corroborated by the clinical consensus of two psychiatrists. T1 weighted structural MRI scans were acquired with a 1.5 Tesla machine and to minimize biases and head motion, restraining foam pads were used. The original image size is 384x512x144; these images are then rotated and realigned to a resolution of 256x256x192. After this alignment, they were segmented into specific brain regions called Regions of Interest (ROIs) manually by experts following a specific protocol for each ROI [3]. In this work, we use three ROIs from the two hemispheres of the brain summing upto a total of six different brain regions: Dorsolateral prefrontal cortex (*ldlpfc* and *rdlpfc*), Entorhinal Cortex (*lec* and *rec*), and Thalamus (*lthal* and *rthal*) which are found to be impaired in schizophrenic patients.

Preprocessing. After the alignment and ROI tracing, DARTEL [1] tools within SPM software [7] was used to pre-process the data. Initially, images are segmented into grey and white matter in *Native* and *DARTEL imported* spaces. The DARTEL imported images have lower resolution than the original images but are used to spatially align to standard MNI atlas. In the second step, DARTEL template generation is applied which creates an average template from the input data while simultaneously aligning white and grey matter. In this step, the flowfields of the registration are also computed which will be used to segment the MNI space normalized images into ROIs. In the final step, the DARTEL template is used to spatially normalize all images into standard MNI space. In this way, smoothed (12 mm Gaussian), and Jacobian scaled grey matter images are constructed which is general practice in neuroimaging applications.

Feature extraction. The images at the end of the preprocessing pipeline are the intensity probability maps which are then used to construct the features for our classification experiments. Since we already have ROI segmented source images, using the flow fields computed in the second step of preprocessing; we create the intensity maps for every subject and ROI instead of extracting a single set of features for the whole brain. Since the ROIs have different bounding boxes, the sizes of these images are not the same for all subjects. By applying thresholding at 0.2 level, we compute histograms of probability maps for every subject and ROI. Number of bins in each histogram is chosen to be 40 which showed the best performance in our experiments. As a result, we have a data set of six different ROIs, 82 subjects with a feature vector of size 40 which we apply our classification pipeline.

3.2 Experiments

In our first set of experiments, we show how our algorithm behaves when presented with only one data source. We compare `CLMKL` with `SVM` which is the single SVM on the feature set, `CONCAT` which is the concatenation of the feature and the gender information and `LMKL` mentioned in Section 2. We used linear kernels as base kernels in all our experiments because the number of parameters to optimize is fewer. We use a Leave-One-Out (LOO) validation scheme by training

Table 1. Accuracies on schizophrenia detection data set using one data source only

ROI	SVM	CLMKL	CONCAT	LMKL
ldlpfc	54.88	**73.17**	54.88	65.85
rdlpfc	70.73	**76.83**	70.73	70.73
lec	71.95	**81.71**	73.17	74.39
rec	67.07	**74.39**	69.51	69.51
lthal	70.73	**79.27**	78.05	71.95
rthal	71.95	**74.39**	69.51	68.29

all the methods using all but one data point ($\boldsymbol{x_i}$) and testing if we can get the correct classification on $\boldsymbol{x_i}$. We do this for all $\boldsymbol{x_i}$ and the percentage of correct classifications over all subjects is the accuracy which we report in all our tables. We can see the accuracies for single data source in Table 1. Our method is always the most accurate method and better than single SVMs, the concatenation and the canonical `LMKL`.

In our second set of experiments, we combine all the ROIs to see the effect of our algorithm compared to other MKL algorithms, the concatenation of features, and the results of separately training the male and female subjects. We can see the results in Table 2. `CONCAT` shows the accuracy of a linear SVM when the features of all ROIs are concatenated. This time we can clearly see the advantage of our method. We obtain the best results when we use `CLMKL`, which uses the apriori clustering information without depending on the input data point. This makes sense because as the number of parameters increase, it gets harder to optimize and `LMKL` may be stuck in a local minima. We can deduce two results from this table. We can see that our method includes the gender information in terms of apriori knowledge and has better accuracy than the linear MKL methods and single SVMs. Second, when we divide the data into male/female subsets and train accordingly, we increase accuracy (male/female separated accuracies are better than training male/female together because the anatomic similarities make it easier to classify the same gender) but not as much as `CLMKL`. This is what we expect because the number of subjects in the training set gets smaller, but this is also why our method is superior to other MKL methods and canonical localized algorithms.

Table 2. Accuracies on schizophrenia detection data set using all ROIs

Method	Together	Male	Female
`SVM`	71.95	75.00	76.19
`RBMKL`	71.95	82.50	71.43
`SMKL`	71.95	85.00	71.43
`CONCAT`	69.51		
`LMKL`	73.17		
`CLMKL`	**90.24**		

Table 3. Accuracies on schizophrenia detection data set by adding gender information.

Method	w/o gender	w/ gender
CONCAT	69.51	70.73
RBMKL	71.95	69.51
SMKL	71.95	69.51
CLMKL	**90.24**	

To be fair, we also include the gender information as another data source (kernel) and compare the accuracy of CLMKL also with this case. We can see from Table 3 that also in this case CLMKL is superior to other methods. What we observe from this table is that when we add the gender as another data source to the classification system, it becomes important and can change the model significantly. In the SMKL and RBMKL cases, we see that this creates a problem and the accuracies actually decrease.

4 Conclusions

In this paper we benefit from the knowledge of heterogenity factors in order to deal with intra-class variability and therefore improve the performance of automatic medical decisional systems.

We propose a new *localized* Multiple Kernel Learning algorithm which takes into account the information on the clusters of a known subject stratification. We evaluate our CLMKL method on a dataset of Schizophrenic patients and healthy controls by showing a substantial improvement of our approach in comparison to several other methods. In particular, the gender factor was considered as prior information in order to properly encode the variability between male and female. Even when a single data source is considered, our CLMKL shows improvements over classical SVM algorithms. Moreover, we observe a further improvement of our method when data from multiple sources is considered (in our case different ROIs of the brain). The strength of our method also shows when we compare our method with the models trained/tested on male/female separated data. Our method allows to train all data together, thus avoiding small number of subjects and overfitting.

Our future work will address the exploitation of other known heterogenity factors like age, education level, and other meta information coming from the subject's interview. Moreover, the method can be evaluated on other features coming from other image modalities.

References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage 38(1), 95–113 (2007)
2. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML 2004, pp. 41–48 (2004)

3. Baiano, M., Perlini, C., Rambaldelli, G., Cerini, R., Dusi, N., Bellani, M., Spezzapria, G., Versace, A., Balestrieri, M., Mucelli, R.P., Tansella, M., Brambilla, P.: Decreased entorhinal cortex volumes in schizophrenia. Schizophr. Res. 102(1-3), 171–180 (2008)
4. Castro, E., Martínez-Ramon, M., Pearlson, G., Sui, J., Calhoun, V.: Characterization of groups using composite kernels and multi-source fMRI analysis data: Application to schizophrenia. NeuroImage 58(2), 526–536 (2011)
5. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press (2000)
6. Filipovych, R., Resnick, S.M., Davatzikos, C.: Multi-Kernel Classification for Integration of Clinical and Imaging Data: Application to Prediction of Cognitive Decline in Older Adults. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (eds.) MLMI 2011. LNCS, vol. 7009, pp. 26–34. Springer, Heidelberg (2011)
7. Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (eds.): Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press (2007)
8. Gönen, M., Alpaydın, E.: Localized multiple kernel learning. In: ICML 2008, pp. 352–359 (2008)
9. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. JMLR 12, 2181–2238 (2011)
10. Hinrichs, C., Singh, V., Xu, G., Johnson, S.: Predictive markers for AD in a multimodality framework: An analysis of MCI progression in the ADNI population. NeuroImage 55(2), 574–589 (2011)
11. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation 3, 79–87 (1991)
12. Kang, H.J., Doermann, D.: Selection of classifiers for the construction of multiple classifier systems. In: ICDAR 2005, vol. 2, pp. 1194–1198 (2005)
13. Kawasaki, Y., Suzuki, M., Kherif, F., Takahashi, T., Zhou, S.Y., Nakamura, K., Matsui, M., Sumiyoshi, T., Seto, H., Kurachi, M.: Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage 34(1), 235–242 (2007)
14. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. Wiley-Interscience (2004)
15. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.: Learning the kernel matrix with semidefinite programming. JMLR 5, 27–72 (2004)
16. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. JMLR 9, 2491–2521 (2008)
17. Ulaş, A., Duin, R., Castellani, U., Loog, M., Mirtuono, P., Bicego, M., Murino, V., Bellani, M., Cerruti, S., Tansella, M., Brambilla, P.: Dissimilarity-based detection of schizophrenia. International Journal of Imaging Systems and Technology 21(2), 179–192 (2011)
18. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons (1998)
19. Woods, K., Philip Kegelmeyer Jr., W., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. IEEE TPAMI 19(4), 405–410 (1997)