Computational and Comparative Study on Multiple Kernel Learning Approaches for the Classification Problem of Alzheimer's Disease

Ahlam Mallak[1,2], Jeonghwan Gwak[1,3], Jong-In Song[1,3], and Sang-Woong Lee[1,2(✉)]
for the Alzheimer's Disease Neuroimaging initiative

[1] National Research Center for Dementia, Gwangju 61452, Republic of Korea
`ahlam.mallak@ymail.com, james.han.gwak@gmail.com,`
`jisong@gist.ac.kr, swlee@chosun.ac.kr`
[2] Department of Computer Engineering, Chosun University, Gwangju 61996, South Korea
[3] School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Abstract. Several classification methods have been proposed for assisting computer-aided diagnosis of Alzheimer's disease (AD). Among them, classification methods including (i) support vector machines (SVM), and (ii) generalized multiple kernel learning (GMKL) are getting increasing attention in recent studies. Nevertheless, there is little research on the comparison among these methods to find a better classification framework and further analysis of brain imaging features in the study of AD. To deal with this issue, we carry out exhaustive comparative study in this work to evaluate efficiency of these different classification methods. For the experiments, we used FreeSurfer mean cortical thickness dataset downloaded from the ADNI database (adni.loni.usc.edu) baseline data. The classification accuracy (in classifying the three classes CN, LMCI, AD) of comparative methods has been evaluated using 3-fold cross validation. From the comparative study, we could observe that GMKL is the most promising framework if the sufficient training data can be provided.

Keywords: Alzheimer's disease · Support vector machines · Multiple kernel learning · Generalized multiple kernel learning

1   Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that causes degenerative changes in the neurons, which results in progressive loss of memory and several other cognitive functions. Alzheimer's disease is the most common form of dementia [1]. Currently, AD is generally detected at a late stage at which treatment can only slow the progression of cognitive decline. This is especially important for individuals with late mild cognitive impairment (LMCI), who are at high risk to develop AD in the near future comparing to cognitively normal (CN) or the other mild cognitive impairment (MCI) groups.

The main goal of this study is to elucidate different classification approaches based on kernel machines in differentiating the groups of AD, LMCI, and CN subjects given mean FreeSurfer cortical thickness data for 365 subjects of age range between 70 and 80 years old, which include 107 CN subjects, 76 AD patients and 182 LMCI subjects.

Data used in the preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

The organization of the paper is as follows: Sect. 2 introduces the kernel machines including SVM, multiple kernel learning (MKL), and generalized MKL (GMKL). In Sect. 3, we briefly mention the data used in this work. Section 4 describes the experimental results and comparative study. Finally, the conclusions and future work are discussed in Sect. 5.

2   Kernel Machine Algorithms

2.1   Support Vector Machines (SVM)

In this section we briefly sketch the SVM algorithm and its motivation. A more detailed description on SVM can be found in [2, 3].

Let us start things off with a fairly simple case of two linearly separable classes. Given a two-class, separable data set $D = \{(x_i, y_i)\}_{i=1}^{l}$ of labeled examples, where $y_i \in \{-1, 1\}$ we want to determine which one, among infinitely many linear classifiers separating the data, will have the smallest generalization error. One good choice is the hyperplane (i.e., the decision boundary) that makes the maximum margin between the two classes in which margin is defined as the sum of the distances of the hyperplane from the closest point (i.e., the support vector) of the two classes.

For non-separable two-class problems, we can still consider the hyperplane that maximizes the margin and minimizes misclassification error. The tradeoff between the margin and misclassification error is controlled by a positive constant chosen in advance. It can be shown that the solution to this problem is a linear classifier given as

$$f(x) = sign\left(\sum_{i+1}^{i} \alpha_i y_i x^T x_i + b\right) \tag{1}$$

When the data is not linearly separable, slack variables $\xi_1, \ldots, \xi_N$ with $\xi_i \geq 0$ are introduced as

$$y_i(w.x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N \tag{2}$$

The purpose of the variables ξ_i is to allow misclassified points, which have their corresponding $\xi_i > 1$ Therefore, $\sum \xi_i$ has an upper bound on the number of training errors. While the first term is minimized to control learning capacity as in the separable case, the second term is added to control the number of misclassified points.

The input data is mapped into a high-dimensional feature space through some nonlinear mapping chosen a priori [4].

Given a kernel $K(x_i, y_i) = \varphi(x_i).\varphi(x_j)$ where $\varphi(x)$ is the mapping function of x in the feature space, only K is needed in the training algorithm and φ is never used explicitly. Conversely, given a symmetric positive kernel $K(x, y)$, Mercer's theorem [2] indicates that there exists a mapping φ such that $K(x, y) = \varphi(x) \cdot \varphi(y)$. Then, the decision function becomes

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \tag{3}$$

SVMs are originally designed for binary classification. Since three classes are available in the classification process, we need an appropriate multiclass-based classification method such as multiclass SVMs [4]. There are two possible ways for such purpose by combining several binary classifiers (i.e., SVMs): (1) "one against one" [5] that applies pairwise comparisons between classes by combining two binary classifiers, and (2) "one against the others" [6] that compares a given class with all the others putting together. According to the comparison study [7], it is known that the accuracies of these methods are almost the same. As a consequence, most researches have been chosen the one with the lowest complexity and thus "one against the others." is the commonly adopted approach in many practical applications. In this work, we also used the approach.

2.2 Multiple Kernel SVM (MK-SVM)

As discussed in Sect. 2.1, SVMs are one of the kernel approaches that can be used efficiently to solve classification or regression problems [12]. For solving non-linear separable problems using SVMs, we need to create a function called the kernel function $K(x, x')$. Let us assume that $\{x_i, y_i\}_{i=1}^{l}$ is the used learning set, where x_i is a subject with some feature(s) belongs to some feature space X and y_i is the intended label for some pattern of subjects x_i. Then, the learning problem that the kernel learning needs to solve is formulated as

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K(x_i, x) + b^*, \tag{4}$$

where α_i^* and b^* are the coefficients to be learned from the training process, while $K(\cdot, \cdot)$ is a given positive definite kernel.

Many recent studies have proven that using multiple kernels can enhance the performance of the learning process, when compared to single kernel solutions [8, 9]. Thus, a new approach has been developed as a convex combination of basis kernels $K(x, x')$

$$K(x, x') = \sum_{m=1}^{M} d_m K_m(x, x'), \text{ with } d_m \geq 0, \sum_{m=1}^{M} d_m = 1, \tag{5}$$

where M shows the number of kernels used, while K_m is the basis kernel which can be any classical kernel (such as Gaussian kernels) with different parameter values. Thus, the main idea for multiple kernel learning is finding the suitable kernel parameters (i.e., weights) d_m that gives the optimal solution of the problem as well as finding the optimal classification coefficients α_i^*.

2.3 Generalized Multiple Kernel Learning (GMKL)

GMKL is an optimized algorithm of MKL aims to learn the optimal parameters of the SVM. Specifically, it estimates the kernel parameters d as the MKL (in Sect. 2.2) does, which enables it finds more optimal solutions for this function $(x) = w^t \boldsymbol{\varphi}_d(\boldsymbol{x}) + b$.

The GMKL process is summarized as follows [10]:

(1) Choose a non-convex formulation that is because the kernel combinations $w_k^t w_k$ and the weights for each kernel should not approach to zero.
(2) The regularizer $r(d)$ should be placed in the objective and given a scalar parameter within it.
(3) We can relax the constraint $d \geq 0$ by making it more generalized. As a result, the learned kernel parameters are required to be positive and definite values.
(4) We need to check whether the gradient descent algorithm whether it is still applicable or not, by checking the gradient of the regularizer $\nabla_d r$ if it exists or not.

The optimization problem $f(x) = \sum_{i=1}^{n} \alpha_i \sum_{m=1}^{M} d_m k_m(x, x_i) + b$ can be divided into two nested loops; inner and outer. The outer loop $\sum_{i=1}^{n} \alpha_i k(x, x_i) + b$ is used to learn SVM parameters. While, the inner loop $k(x, x_i) = \sum_{m=1}^{M} d_m k_m(x, x_i)$ is used to learn the kernel parameters (weights) d.

From the above operations, compared to MKL, GMKL can learn general kernel combinations which is subject to general regularizations on the kernel operations.

3   Classification Study Data

All source imaging data used in this paper consist of 1.5 T T1-weighted MRI volumes in the NIfTI format downloaded from the ADNI1: Complete 1Yr 1.5T Data Collection. All images were processed using three neuroimaging software pipelines: (1) Free-Surfer, (2) Advanced Normalization Tools (ANTs), and (3) Mindboggle the result of this process are tables consist of many features from many different regions of the brain.

FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) is an open source software suite that is used for human brain MRI processing and analysis. Many processes can be done using FreeSurfer, such as skull stripping, image registration, subcortical segmentation, cortical surface reconstruction, fMRI analysis and much more.

In this work, FreeSurfer mean cortical thickness data were used for AD, LMCI and CN subjects of age range between 70 and 80 years old. Table 1 lists the results of (clinical) diagnosis and demographics of the dataset.

4   Experimental Results

We tackle the multiclass classification problem of the baseline FreeSurfer mean thickness data to find the best framework to study Alzheimer's disease. The classification framework adopts the different kernel methods: (1) multi-class SVM using one vs all multiclass classification and the Gaussian radial basis function (rbf) kernel with a quadratic programming method to separate the hyperplane, (2) the standard MKL algorithm, and (3) the GMKL algorithm. Both MKL and GMKL used 15 rbf kernels to measure the efficiency in terms of accuracy and computational time for the comparison purpose. Tables 2 and 3 show the experimental results for each group of different number of training samples. Note that for each method, the highest accuracy in Table 2 and the lowest execution time in Table 3 are highlighted in boldface.

Table 2 lists the classification accuracy results. From the experimental reults, we could see that multi-class SVM tended to perform the worst with the accuracy of only 60 % when almost all the data is used for training and the rest for testing. The relatively poor performance was due to the fact that multi-class SVM uses only one kernel comparing to the other techniques which combine many kernels in different scenarios. In contrast, the highest accuracy was obtained for GMKL when only ten training samples were used and the rest fo testing with the accuracy up to 80 %. In addition, there is a very tight coupling between the two in MKL and GMKL in terms of the accuracy.

The comparison between the previous methods was also done for computational time (in terms of elapsed time) as well. Although GMKL achieved the highest

Table 1. Statistics of the dataset used for training and testing.

	No. of subjects	Diagnosis results (CN/LMCI/AD)	Gender	
			Male	Female
ADNI1 data	365	107/182/76	222	143

Table 2. Comparison results of accuracy

No. of training samples	Multi-class SVM (%)	MKL (%)	GMKL (%)
10	49.57	49.80	**80**
20	49.56	49.85	**80**
30	49.85	**53.01**	**80**
50	**50.47**	51.79	68
100	50.18	49.33	65
200	50.90	–	66.5
360	60	–	–

Table 3. Comparison results of execution time for both the training and testing phases

No of training samples	Multi-class SVM (sec.)	MKL (sec.)	GMKL (sec.)
10	1.308	**0.349**	540.60
20	0.557	0.358	490.20
30	0.552	0.446	453.60
50	0.510	0.771	431.46
100	0.464	1.711	290.90
200	0.361	–	**174.32**
360	**0.300**		
		–	–

classification accuracy, it consumed the largest time of roughly 540 s. However, both multi-class SVM and the standard MKL achieved the accuracy of almost 60 % and 54 % in only 0.3 and 0.446 s, respectively. As shown in Table 3.

Finally, the performance of the three methods multi-class SVM, MKL and GMKL in terms of classification accuracy was also validated using the 3-fold cross-validation technique. In 3-fold cross-validation, the overall dataset is divided randomly into 3 equal sized partitions. Each time a single partition is used for testing and the remaining two partitions (i.e., 2/3 from the total dataset) are used for the training process. This process is repeated three times (i.e., same as the number of partitions or folds), so that each time one different partition is used for testing. Finally, the accuracy is measured in each time and the overall accuracy is the averaged accuracy from the three trials. As a result, as shown in Table 4, the accuracy of each method after the validation using 3-fold cross validation was 49.58 % for multi-class SVM, 52.61 % for MKL and 75.75 % for GMKL. Similar to the previous results, we observed that GMKL has the highest accuracy among all these kernel methods when we validate using 3-fold cross validation method.

Table 4. Comparison results of accuracy using 3 cross-validation

No of training samples	Multi-class SVM (%)	MKL (%)	GMKL (%)
	49.58	52.61	75.75

5   Conclusions and Future Work

Several classification methods have been studied for the computer-aided diagnosis of Alzheimer's disease. Classification methods such SVM, MKL and GMKL have been widely used in recent studies. However, there is lack of comparisons for these methods to find a better framework for classification and analysis of brain imaging features in the study of Alzheimer's disease. In this paper, the efficiency of the classification methods including SVM, the standard MKL, and GMKL were compared in terms of accuracy and execution time. From the experimental results, we could verify that GMKL achieved the highest accuracy among all the others although it requires more computational time. To extend this work, we can include other methods such as generalized and/or adaptive multiple kernel learning approaches. Also, other different multimodal features (such as PET imaging and other clinical facts) can be incorporated in the framework to build more robust framework, which is also one of our research agendas.

References

1. Hebert, L.E., Beckett, L.A., Scherr, P.A., Evans, D.A.: Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050. Alzheimer Dis. Assoc. Disord. **15**(4), 169–173 (2001)
2. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
3. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 1–25 (1995)
4. Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Trans. Neural Netw. **10**(5), 1055–1064 (1999)
5. Boser, B. E., Guyon, I. M., Vapnik, V. N.: A training algorithm for optimal margin classifier. In: Proceeding of 5th ACM Workshop on Computational Learning Theory, Pittsburgh, PA, pp. 144–152 (1992)
6. Pontil, M., Verri, A.: Support vector machines for 3-D object recognition. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 637–646 (1998)
7. Blanz, V., Schölkopf, B., Bülthoff, H., Burges, C., Vapnik, V., Vetter, T.: Comparison of view-based object recognition algorithms using realistic 3D models. In: Malsburg, C., Seelen, W., Vorbrüggen, J.C., Sendhoff, B. (eds.) ICANN 1996. LNCS, vol. 1112, pp. 251–256. Springer, Heidelberg (1996). doi:10.1007/3-540-61510-5_45
8. Lanckriet, G.R.G., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S.: Kernel-based data fusion and its application to protein function prediction in yeast. In: Pacific symposium on Biocomputing, p. 300 (2004)
9. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. J. Mach. Learn. Res. **9**, 2491–2521 (2008)
10. Varma, M., Babu, B.R.: More generality in efficient multiple kernel learning. In: Proceedings of the International Conference on Machine Learning, Montreal, Canada, pp. 1065–1072 (2009)