Composite Kernels for Automatic Relevance Determination in Computerized Diagnosis of Alzheimer's Disease

Murat Seckin Ayhan, Ryan G. Benton,
Vijay V. Raghavan, and Suresh Choubey

Center for Advanced Computer Studies
University of Louisiana at Lafayette
Lafayette, LA, USA 70503
{msa4307,rbenton,vijay}@cacs.louisiana.edu
Quality Operations, GE Healthcare
3000 N. Grandview Blvd., Waukesha, WI 53118
suresh.choubey@med.ge.com

Abstract. Voxel-based analysis of neuroimagery provides a promising source of information for early diagnosis of Alzheimer's disease. However, neuroimaging procedures usually generate high-dimensional data. This complicates statistical analysis and modeling, resulting in high computational complexity and typically more complicated models. This study uses the features extracted from Positron Emission Tomography imagery by 3D Stereotactic Surface Projection. Using a taxonomy of features that complies with Talairach-Tourneau atlas, we investigate composite kernel functions for predictive modeling of Alzheimer's disease. The composite kernels, compared with standard kernel functions (i.e. a simple Gaussian-shaped function), better capture the characteristic patterns of the disease. As a result, we can automatically determine the anatomical regions of relevance for diagnosis. This improves the interpretability of models in terms of known neural correlates of the disease. Furthermore, the composite kernels significantly improve the discrimination of MCI from Normal, which is encouraging for early diagnosis.

Keywords: Statistical learning, Classification, Bayesian methods, Gaussian processes, Positron emission tomography.

1 Introduction

Alzheimer's disease (AD) is one major cause of dementia. It is progressive, degenerative and fatal. Various fairly accurate diagnostic tests are available; however, a conclusive diagnosis is only possible through an autopsy. Mild Cognitive Impairment (MCI) is a transitional state between normal aging and AD. MCI shares features with AD and it is likely to progress to AD at an accelerated

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 126–137, 2013.

rate [1]. However, an MCI case may lead to other disorders, as well. Thus, MCI patients form a heterogeneous group with subcategories [1].

One promising source of information for the early diagnosis of AD is Positron Emission Tomography (PET) scans. In [2], the utility of 3D Stereotactic Surface Projection (3D-SSP) in AD diagnosis was demonstrated. The metabolic activity scores based on the PET-scans were shown to enable the localization of cortical regions with abnormalities. 3D-SSP provides both statistical analysis and standardization of PET imagery so that an objective, data-driven analysis is accomplished [3].

In [4], the accuracy of dementia diagnosis provided by radiologists has been compared to that of computer-based diagnostic methods. Utilizing Support Vector Machines (SVMs), they concluded that the accuracy of computerized diagnosis is equal to or better than that of radiologists. A general adoption of computerized methods for visual image interpretation for dementia diagnosis is recommended by [4,5].

In [6], two well-known classification algorithms, Naïve Bayes (NB) and SVMs, have been benchmarked for automated diagnosis of AD. An analysis of features extracted from PET imagery via 3D-SSP revealed strong dependencies between the predictiveness of features and their corresponding cortical regions' cognitive and physiological characteristics. For instance, the posterior cingulate cortex is greatly involved in memory and is deemed to characterize *early-to-moderate AD* [5]. The features obtained from this region, which constitutes a very small portion of the brain, are highly predictive of the disease [6]. On the other hand, visual cortex is usually spared until very late stages of AD [7]. As a result, features from this region are not as predictive [6]. In addition, the most of features obtained via 3D-SSP are highly correlated due to their spatial properties. In [8], to cope with feature correlations, certain regions of the brain containing characteristic patterns of AD were handpicked based on the domain-knowledge.

SVMs and Gaussian Processes (GPs) are two examples of kernel machines. Given the characteristic patterns of AD, simple kernel functions, such as a Gaussian-shaped one (eq.3), may fail to capture the input structure. To remedy this situation, in this paper, we propose a composite kernel strategy to automatically determine the anatomical regions of relevance for diagnosis.

In this study, we mine the brain imaging data supplied by the Alzheimer's Disease Neuroimaging Initiative (ADNI) [1]. The data collection is composed of 3D-PET scans of human brains. However, such neuroimaging procedures usually end up generating high-dimensional data. This complicates statistical analysis and modeling, resulting in high computational complexity and typically more complicated models. Furthermore, the cost of labeled data is high since the data gathering process involves expensive imaging procedures and domain-experts. As a result, sample sizes are small and this is a well-recognized problem in statistical machine-learning. By using composite kernel functions, we aim to discover relevant subspaces given the high-dimensional data.

[1] `http://adni.loni.ucla.edu/`

2 Gaussian Processes for Regression

For regression problems, we aim to predict the output of a real-valued function $y = f(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, ..., x_D)$ and D is the number of dimensions. Thus, we seek to learn an appropriate function that maps inputs to outputs, and GPs enable us to do inference in the function-space (eq.1).

> "A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution." [9, p.13].

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \text{ where} \\ m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \end{aligned} \tag{1}$$

Accordingly, $f(\mathbf{x})$ and $f(\mathbf{x}')$ are jointly Gaussian. Thus, given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ where $i = 1...N$, we obtain an N-dimensional random vector $\mathbf{f}$.

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K) \tag{2}$$

A *GP-prior* (eq.2) specifies the prior distribution over the latent variables. Once combined with the likelihood of data, it gives rise to a *GP-posterior* in function space. This Bayesian treatment promotes the smoothness of predictive functions [12] and the prior has an effect analogous to the quadratic penalty term used in maximum-likelihood procedures [9].

In GPs terminology, a kernel is a covariance function that estimates the covariance of two latent variables $f(\mathbf{x})$ and $f(\mathbf{x}')$ in terms of input vectors $\mathbf{x}$ and $\mathbf{x}'$. The choice of the covariance function $k(\mathbf{x}, \mathbf{x}')$ in eq.1 is important because it dictates the covariance matrix K in eq.2 and eq.6. A typical covariance function, known as *squared-exponential* (SE) covariance function, is

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \text{exp} \left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2} \right), \tag{3}$$

where ℓ and σ_f are the bandwidth (length-scale) and scale parameters, respectively. Furthermore, the idea of length-scale parameter ℓ can be specialized for individual dimensions (eq.4) so that irrelevant features are effectively turned off by large length-scales during model selection:

$$k_{ARD}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \text{exp} \left(-\sum_{i=1}^{D} \frac{(x_i - {x'}_i)^2}{2\ell_i^2} \right). \tag{4}$$

This process is known as *Automatic Relevance Determination* (ARD) [10,11], which determines good features while training. However, ARD is computationally-expensive for high-dimensional data; the cost is $O(N^2)$ per hyperparameter [9].

Neural network (NN) covariance function is another interesting example:

$$k_{NN}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \text{sin}^{-1} \left(\frac{2\tilde{\mathbf{x}}^T \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1 + 2\tilde{\mathbf{x}}^T \Sigma \tilde{\mathbf{x}})(1 + 2\tilde{\mathbf{x}}'^T \Sigma \tilde{\mathbf{x}}')}} \right), \tag{5}$$

where $\tilde{\mathbf{x}} = (1, \mathbf{x})^T$ is an augmented input vector and Σ is a covariance matrix[2] for *input-to-hidden* weights $\mathbf{w}$ [9,12]. A GP with NN covariance function (eq.5) can be viewed as emulating a NN with a single hidden layer.

GPs framework supports many covariance functions. Moreover, one can build up a covariance function as the sum of several covariance functions, each of which processes certain parts of inputs [12]. Clearly, information processing capabilities of GPs are mostly determined by the choice of covariance function. The impact of the covariance function is larger for small to medium-sized datasets [13].

2.1 Learning of Hyperparameters

Many covariance functions have adjustable parameters, such as ℓ and σ_f in eq.3. In this regard, learning in GPs is equivalent to finding suitable parameters for the covariance function. Given the target vector $\mathbf{y}$ and the matrix X that consists of training instances, this is accomplished by maximizing the log marginal likelihood function:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{N}{2}\log 2\pi, \qquad (6)$$

where σ_n is due to the Gaussian noise model, $y_i = f_i + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

2.2 Predictions

GP regression yields a predictive Gaussian distribution (eq.7):

$$f_*|X, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\bar{f}_*, \mathbb{V}[f_*]), \text{ where} \qquad (7)$$

$$\bar{f}_* = \mathbf{k}_*^T(K + \sigma_n^2 I)^{-1}\mathbf{y} \qquad (8)$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(K + \sigma_n^2 I)^{-1}\mathbf{y} \qquad (9)$$

and $\mathbf{k}_*$ is a vector of covariances between the test input $\mathbf{x}_*$ and the training instances. Eq.8 gives the mean prediction $\bar{f}_*$, which is the *empirical risk minimizer* for any symmetric loss function [9]. Eq.9 yields the predictive variance.

3 Gaussian Processes for Classification

GP classification is a generalization of *logistic regression.* For binary (0/1) classification, a sigmoid function (eq.10) assigns the class probability:

$$p(y_* = 1|f_*) = \lambda(f_*) = \frac{1}{1 + \exp(-f_*)}. \qquad (10)$$

Compared to the regression case, GP models for classification require a more sophisticated treatment due to discrete target variables, such that $y_* \sim$ *Bernoulli*$(\lambda(f_*))$. Thus, we resort to approximation methods. Expectation Propagation (EP) [14] is heavily used for GP learning. It delivers accurate marginals, reliable class probabilities and faithful model selection [15].

[2] $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma)$

4 GPs versus SVMs

Both GPs and SVMs exploit kernels. However, their objectives are quite different. SVMs are large margin classifiers and their goal is to maximize distances from decision boundaries. On the other hand, GPs are Bayesian and they are designed for likelihood maximization.

Training a typical SVM with a *Radial Basis Function* (RBF)[3] [16] involves a grid search for model parameters, such as C (penalty parameter) and γ. However, for a large number of parameters, the grid search becomes prohibitively expensive. Furthermore, SVMs require a validation set for the search, which results in a smaller training set.

Thanks to Bayesian model selection for GPs, a large number of hyperparameters can be approximated by maximizing marginal likelihood (eq.6). Also note that GP models do not require a validation set to be used for the optimization of model parameters. As a result, more of data can be used for training, which is desirable when the sample size is small.

5 Data and Processing

Table 1 describes the demographics of the patients in our data collection, which is composed of 391 PET scans and is broken into three groups: Normal, MCI and AD. The images covered a period between October 25, 2005 and August 16, 2007. The metabolic activity of the cerebral cortex is extracted with respect to the 3D-SSP using a GE proprietary application known as Cortex ID. As a result, an ordered list of 15964 predefined points is obtained (Fig. 1, Fig. 2 and Table 2). Each *voxel* is assigned a *z-score*, which measures how many standard deviations the metabolic activity departs from its expected mean. The mean is estimated from a healthy control group [2]. Voxels are also grouped according to Talairach-Tourneau atlas (Fig. 2 and Table 2).

Table 1. Demographic data on ADNI scans (extended from [8])

		Gender		*Ethnicity*			Race		
	Avg. Age	M	F	Hispanic or Latino	Not Hispanic	Unknown	African	Asian	Caucasian
Normal	76.1	64	37	0	97	4	1	0	100
MCI	75.6	163	67	6	219	5	4	0	226
AD	77.4	35	25	0	56	4	0	1	59

[3] $k_{RBF}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right)$

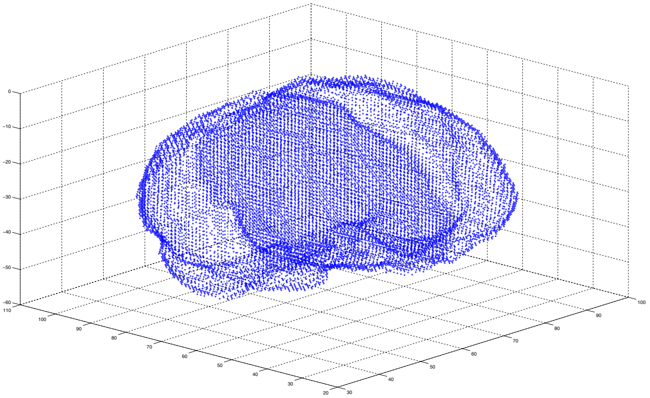

Fig. 1. Cortex extracted via 3D Stereotactic Surface Projection (reprint from [8])

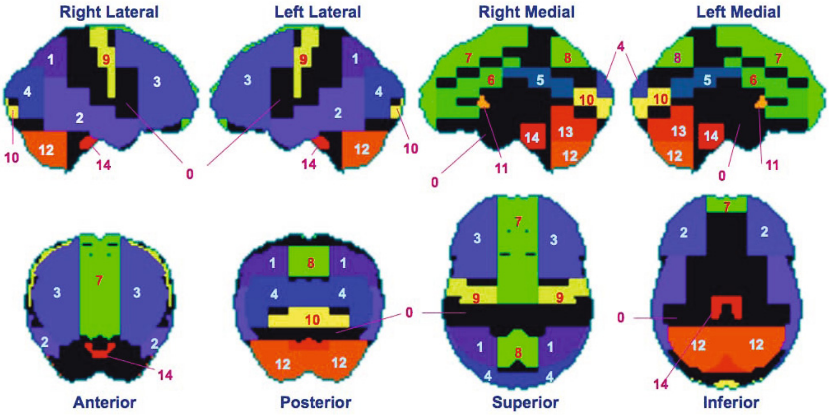

Fig. 2. Taxonomy of cortical regions (reprint from [8])

6 Composite Kernels

A composite kernel consists of many kernels. We introduce two composite kernels: i) SE (eq.11) and ii) NN composite kernels.

$$\begin{aligned} k_{SEcomposite}(\mathbf{x}, \mathbf{x}') &= \sigma_{f_0}^2 \exp\left(-\frac{\|\mathbf{x}_0 - \mathbf{x}'_0\|^2}{2{\ell_0}^2}\right) + ... + \\ & \quad \sigma_{f_{14}}^2 \exp\left(-\frac{\|\mathbf{x}_{14} - \mathbf{x}'_{14}\|^2}{2{\ell_{14}}^2}\right), \end{aligned} \tag{11}$$

where each region (denoted by a subvector $\mathbf{x}_i, i \in \{0, 1, 2, ..., 14\}$) is assigned a local kernel function. The scale parameters (σ_{f_i}) indicate the relevance of regions. This can also be seen as *L2-regularization* by which irrelevant regions

Table 2. Region mapping table (reprint from [8])

Region ID	*Anatomical Region*	*Size (# of voxels)*	*Region Ratio*
0	Other	5456	0.3418
1	Parietal Association Cortex	572	0.0358
2	Temporal Association Cortex	1296	0.0812
3	Frontal Association Cortex	2148	0.1346
4	Occipital Association Cortex	810	0.0507
5	Posterior Cingulate Cortex	368	0.0231
6	Anterior Cingulate Cortex	626	0.0392
7	Medial Frontal Cortex	1636	0.1025
8	Medial Parietal Cortex	412	0.0258
9	Primary Sensorimotor Cortex	390	0.0244
10	Visual Cortex	410	0.0257
11	Caudate Nucleus	34	0.0021
12	Cerebellum	1064	0.0666
13	Vermis	442	0.0277
14	Pons	300	0.0188

are turned off entirely, instead of dealing with individual voxels. This achieves an efficient ARD at the region level. By replacing the simple SE covariance functions with the simple NN covariance functions (eq.5), we derive the NN composite kernel. On the other hand, a grid search with such parameter-rich kernels would be prohibitive. For composite kernels, we utilize GPML toolbox[4], since GP learning has computational advantages in this respect.

7 Experiments

In order to estimate generalization performances of the specified algorithms, we applied 10-fold cross-validation (CV). For SVMs [16], we used a single RBF kernel and a grid search. For GPs, we used BFGS[5] for 100 iterations. Our performance metrics are classification accuracy, precision (eq.12) and recall (eq.13). Table 3, Table 4 and Table 5 present averages of 10 classification tasks. Fig. 3, Fig. 4 and Fig. 5 show the average (mean) accuracies and comparison intervals. The confidence level is 95% and according to Tukey–Kramer method, two means are significantly different if their comparison intervals do not overlap.

$$\text{Precision} = \frac{\text{\# of True positives}}{\text{\# of (True positives + False positives)}} \tag{12}$$

$$\text{Recall} = \frac{\text{\# of True positives}}{\text{\# of (True positives + False negatives)}} \tag{13}$$

[4] `http://www.gaussianprocess.org/gpml/code/matlab/doc/index.html`

[5] A quasi-Newton method for solving unconstrained optimization problems.

Table 3. Normal vs. AD: Classification performance

	SVM RBF	*GP SE*	*GP NN*	*GP SE Composite*	*GP NN Composite*
Accuracy	92.50	62.50	94.38	92.50	94.38
Precision	1.00	0.00	0.98	0.92	0.97
Recall	0.80	0.00	0.87	0.88	0.88

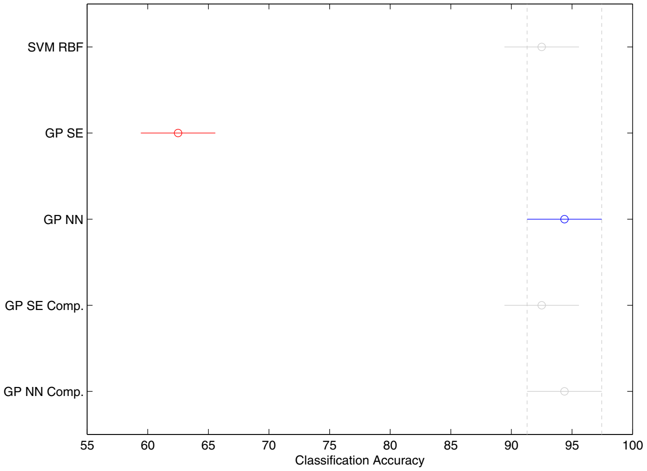

Fig. 3. Normal vs. AD: Performance comparison

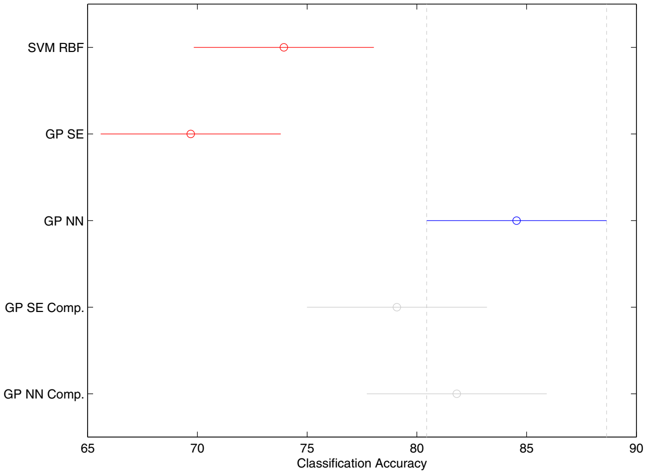

Fig. 4. Normal vs. MCI: Performance comparison

Table 4. Normal vs. MCI: Classification performance

	SVM RBF	*GP SE*	*GP NN*	*GP SE Composite*	*GP NN Composite*
Accuracy	73.94	69.70	84.55	79.09	81.82
Precision	0.73	0.70	0.87	0.84	0.86
Recall	0.98	1.00	0.93	0.87	0.89

Table 5. MCI vs. AD: Classification performance

	SVM RBF	*GP SE*	*GP NN*	*GP SE Composite*	*GP NN Composite*
Accuracy	79.31	79.31	84.14	81.38	82.76
Precision	0.00	0.00	0.72	0.59	0.69
Recall	0.00	0.00	0.40	0.25	0.35

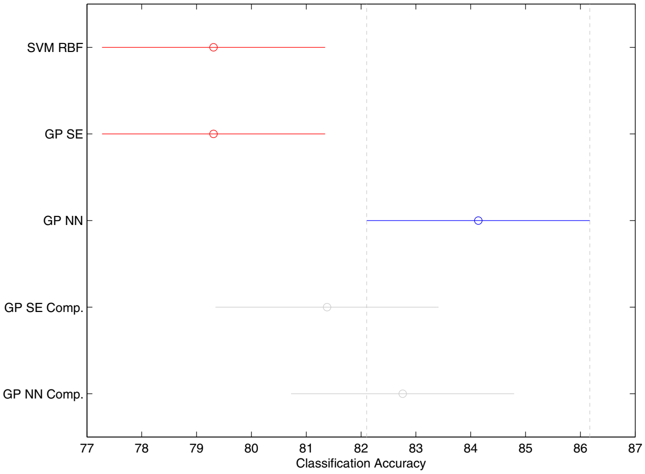

Fig. 5. MCI vs. AD: Performance comparison

Table 3 shows that an SVM with a standard configuration can be farily accurate. However, its recall measure indicates that it has failed to identify some AD cases. Table 4 shows that it is highly biased towards MCI class when utilized to separate MCI from Normal. This leads to high recall, but low precision. For GPs, the use of a simple SE covariance function leads to majority predictors. For instance, despite the classification accuracy of 62.50% in Table 3, precision and recall measures indicate that the diagnosis attempts have always failed[6], which may be attributed to

[6] Number of true positives (AD predictions) is zero.

the presence of a large number of correlated features. Due to the quadratic form in the exponent of the covariance function, even the slightest change in feature values easily causes the covariance between f_i and f_j to tend to zero, which is undesirable. In Table 4 and Table 5, GPs with SE covariance function always predict MCI, which is not the case. In Table 5, SVM also induces a majority predictor. Based on these results, we, therefore, conclude that a simple Gaussian-shaped (SE or RBF) kernel is inappropriate for our problem.

A single NN kernel gives rise to the most accurate classifier in each task (Table 3, Table 4, Table 5). However, composite kernels are competitive with the NN kernel (Fig. 3, Fig. 4, Fig. 5) and when utilized for GP learning, they significantly outperform the simple Gaussian-shaped kernel in the discrimination of AD and MCI from Normal (Fig. 3 and Fig. 4).

Table 5 shows that all the classifiers have difficulties in discriminating AD from MCI (Fig. 5). Recall that MCI is a transitional state and it shares features with AD. As a result, a good separation is difficult. Nevertheless, a GP-classifier with NN covariance function significantly outperforms the SVM and GPs with SE covariance function.

Fig. 6 and Fig. 7 show the *normalized* mean scale parameters (σ_f) assigned to anatomical regions in cases of SE and NN composite kernels, respectively. Posterior cingulate cortex is shown to be the most crucial region for the discrimination of Normal and AD cases. It is also important for the discrimination of MCI and AD cases. This is quite sensible because the posterior cingulate cortex is deemed to characterize early-to-moderate AD [5]. Primary sensorimotor cortex was utilized as a reference region for calculating z-scores in [2]. It plays a major role for MCI-AD separation here, as well. In regards to the discrimination of MCI from Normal, ARD resorts to more regions in order to account for the heterogeneity of MCI group. In short, all anatomical regions are weighted with respect to their relevance to the classification task.

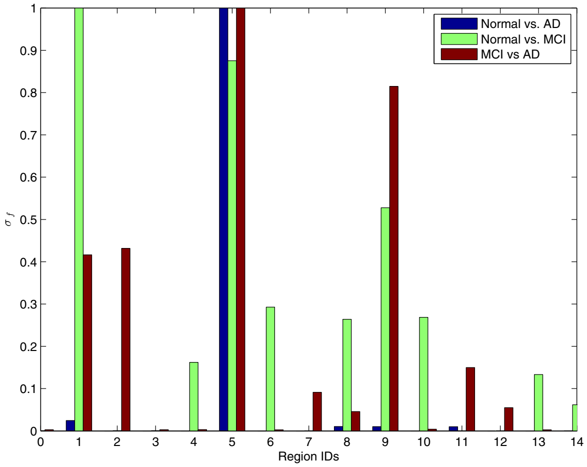

Fig. 6. Normalized relevance scores via SE composite kernel

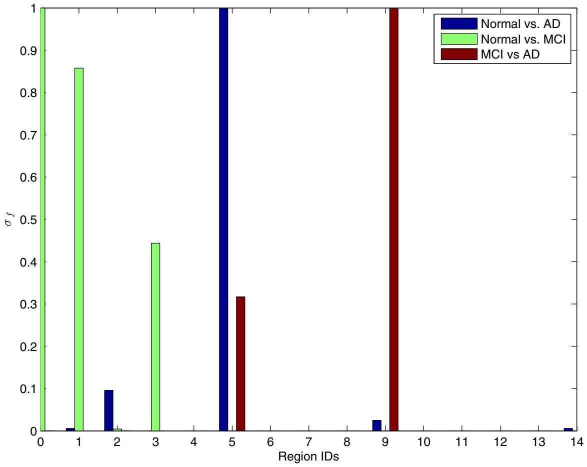

Fig. 7. Normalized relevance scores via NN composite kernel

8 Conclusion

Voxel-based analysis of neuroimagery provides an objective and reliable examination of cortical abnormalities. However, from a machine learning perspective, we need to confront major challenges when modeling neural correlates of dementia. One is the high-dimensionality of data resulting from neuroimaging. Also, sample sizes are small, which aggravates the situation.

In this study, we utilized GPs for predictive modeling of AD via composite kernels. The composite kernels respond to characteristic patterns of the disease. As a result, we automatically determine the anatomical regions of relevance for diagnosis. This improves the interpretability of models in terms of neural correlates of the disease. In terms of classification accuracy, the composite kernels are competitive with or better than simple kernels. Moreover, composite kernels significantly improve the discrimination of MCI from Normal, which is encouraging for early diagnosis of AD. Last but not the least, we shift the ARD from voxel level to region level. This allows us to significantly reduce the computational burden.

Acknowledgments. In preparation of this manuscript, we have used portions of text, tables and graphics from [8] that appeared in IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) 2012. Section 1, Section 2, Section 3 and Section 5 were derived from [8].

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`www.loni.ucla.edu/ADNI`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

References

1. Petersen, R.C., Doody, R., Kurz, A., Mohs, R.C., Morris, J.C., Rabins, P.V., Ritchie, K., Rossor, M., Thal, L., Wingblad, B.: Current Concepts in Mild Cognitive Impairment. Arch. Neurol. 58(12), 1985–1992 (2001)
2. Minoshima, S., Frey, K.A., Koeppe, R.A., Foster, N.L., Kuhl, D.E.: A Diagnostic Approach in Alzheimer's Disease Using Three-dimensional Stereotactic Surface Projections of Fluorine-18-FDG PET. Journal of Nuclear Medicine 36(7), 1238–1248 (1995)
3. Matsuda, H.: Role of Neuroimaging in Alzheimer's Disease, with Emphasis on Brain Perfusion SPECT. Journal of Nuclear Medicine 48(8), 1289–1300 (2007)
4. Kloppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S.J.: Accuracy of Dementia Diagnosis - A Direct Comparison Between Radiologists and A Computerized Method. Brain: A Journal of Neurology 131(11), 2969–2974 (2008)
5. Imabayashi, E., Matsuda, H., Asada, T., Ohnishi, T., Sakamoto, S., Nakano, S., Inoue, T.: Superiority of 3-dimensional Stereotactic Surface Projection Analysis Over Visual Inspection in Discrimination of Patients With Very Early Alzheimer's Disease From Controls Using Brain Perfusion SPECT. Journal of Nuclear Medicine 45(9), 1450–1457 (2004)
6. Ayhan, M.S., Benton, R.G., Raghavan, V.V., Choubey, S.: Exploitation of 3D Stereotactic Surface Projection for Predictive Modelling of Alzheimer's Disease. Int. J. Data Mining and Bioinformatics 7(2), 146–165 (2013)
7. Herholz, K., Adams, R., Kessler, J., Szelies, B., Grond, M., Heiss, W.D.: Criteria for the diagnosis of Alzheimer's disease with positron emission tomography. Dementia and Geriatric Cognitive Disorders 1(3), 156–164 (1990)
8. Ayhan, M.S., Benton, R.G., Raghavan, V.V., Choubey, S.: Utilization of domain-knowledge for simplicity and comprehensibility in predictive modeling of Alzheimer's disease. In: Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 265–272. IEEE Computer Society, Washington, DC (2012)
9. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Second printing. MIT Press, Cambridge (2006)
10. MacKay, D.J.C.: Bayesian Methods for Backpropagation Networks. In: Models of Neural Networks II. Springer (1993)
11. Neal, R.M.: Bayesian Learning for Neural Networks. Lecture Notes in Statistics. Springer (1996)
12. Williams, C.K.I., Barber, D.: Bayesian Classification with Gaussian Processes. IEEE Trans. Pattern Anal. Mach. Intell. 20(12), 1342–1351 (1998)
13. Duvenaud, D., Nickisch, H., Rasmussen, C.E.: Additive Gaussian Processes. In: Proceedings of the 25th Annual Conference on Neural Information Processing Systems 2011, pp. 226–234. Curran Associates, Inc., Red Hook (2011)
14. Minka, T.P.: Expectation Propagation for Approximate Bayesian Inference. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, pp. 362–369. Morgan Kaufmann Publishers Inc., San Francisco (2001)
15. Nickisch, H., Rasmussen, C.E.: Approximations for binary Gaussian process classification. Journal of Machine Learning Research 9, 2035–2078 (2008)
16. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`