Classifying EEG for Brain Computer Interfaces Using Gaussian Process

Mingjun Zhong[†1], **Fabien Lotte**[†], **Mark Girolami**[‡], **Anatole Lécuyer**[†]
zmingjun@irisa.fr, fabien.lotte@irisa.fr, girolami@dcs.gla.ac.uk, anatole.lecuyer@irisa.fr
[†]IRISA, Campus de Beaulieu, F-35042 Rennes Cedex, France
[‡]Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland UK

Abstract

Classifying electroencephalography (EEG) signals is an important step for proceeding EEG-based brain computer interfaces (BCI). Currently, kernel based methods such as support vector machine (SVM) are the state-of-the-art methods for this problem. In this paper, we apply Gaussian process (GP) classification to binary classification problems of motor imagery EEG data. Comparing with SVM, GP based methods naturally provide probability outputs for identifying a trusted prediction which can be used for post-processing in a BCI. Experimental results show that the classification methods based on Gaussian process perform similar with kernel logistic regression and probabilistic SVM in terms of predictive likelihood, but outperform SVM and K-Nearest Neighbor (KNN) in terms of 0-1 loss class prediction error.

Keywords: Gaussian process; brain computer interfaces; support vector machine; EEG

1 Introduction

Brain computer interface (BCI) is a new technique that translates specific electrophysiological signals from mere reflections of central nervous system (CNS) into specific commands, aiming at accomplishing the intent of the people who lost their voluntary muscle control [21]. A variety of methods, such as electroencephalography (EEG), magnetoencephalography (MEG), electrocorticography (ECoG), positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and optical imaging, could be used for monitoring those electrophysiological signals related to brain activities. However, at present it is likely that EEG and related methods are the most popular methods for offering a practical BCI. Classifying EEG is a main task in the translation algorithm step of an EEG-based BCI. Recent reviews have shown that most common classification methods which are largely used in BCI are non-probabilistic methods, and among which support vector machine (SVM) is likely to be an efficient one and has been popularly employed for classifying EEG in BCI [12, 8, 1]. It has been known that the class predictive probability outputs of a new feature vector are of importance in practical recognition circumstances [3, 15]. Unfortunately, SVM absents this quantity, and an additional technique for translating the SVM outputs into probabilities has been proposed [15]. However, this method may not really give a good approximate of the probability output [17].

Both kernel logistic regression (KLR) [17] and Gaussian process (GP) [13, 20, 10, 16] methods can naturally give probability outputs for classification problems. Both methods require adapting to the covariate functions by tuning some hyper-parameters, which could be done by using

[1]Corresponding author. Tel: 33 (0)2.99.84.74.83; Fax: 33 (0)2.99.84.71.71

k-fold cross validations. For KLR, the non-linear functions building the relationships between targets and feature vectors associate with some weights, which are required to learn and adapt to the training data sets. So obviously KLR is a parametric method. In Bayesian view, those weights are assumed to follow some prior distributions such as a gaussian prior with one hyper-parameter variance requiring to be tuned. In contrast, GP is a non-parametric method and thus no weights are required to be estimated. For GP based methods, the non-linear functions are assumed to follow the only required Gaussian process priors, which associate with some covariate functions. So the hyper-parameters in the covariate functions are the only parameters to be tuned. Considering these hard benefits, GP is suggested to be employed for EEG classification problems in this paper. Exact methods are impossible for tackling the GP classification, and various approximation methods have recently been developed with high agreements [13, 20, 5, 4]. Several approximation methods for Gaussian process classification are employed in this paper and the experimental results show that across all the data sets employed, all GP based approximation methods consistently give similar results and outperform SVM and K-Nearest Neighbor (KNN) in terms of 0-1 loss class prediction error. It is known that KLR, the probabilistic SVM (pSVM) of Platt, and GP based methods, which can give probability outputs, are essentially kernel based methods. Hence for comparison purposes, both KLR and pSVM were also applied to the same EEG data sets. Despite the advantages of GP described, no significant differences in terms of predictive likelihood were shown between these probabilistic methods, applying to the three EEG data sets.

2 Gaussian Process for Binary Classification

Gaussian process model for binary classification problem is described in this section. Suppose a feature vector $\mathbf{x} \in \mathbf{R}^{D\times 1}$ is corresponding to a binary variable $t \in \{-1, 1\}$. We have N such observations denoting $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}_{i=1}^N$, and conveniently denote $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$, $\mathbf{t} = (t_1, \cdots, t_N)^T$. The aim is to infer a classifier by using the observations which is able to assign a new feature vector $\mathbf{x}^*$ to one of the two classes with a certain agreement. In order to make prediction, we make a function transformation for the feature vectors such that $\mathbf{f} : \mathbf{X} \longrightarrow \mathbf{f}(\mathbf{X})$. Note that $\mathbf{f}(\mathbf{X}) = (f_1(\mathbf{x}_1), \cdots, f_N(\mathbf{x}_N))^T$ and for simplicity $f_i(\mathbf{x}_i)$ is denoted as f_i. Rather than specifying an explicit form for each of the function f_i we assume that this nonlinear transformation corresponds to a Gaussian process (GP) prior such that $\mathbf{f}(\mathbf{X}) \sim \mathcal{N}_{\mathbf{f}}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X}))$ where $\mathbf{C}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X})$ is the covariance matrix defined by kernel functions which are related to a set of hyper-parameters $\boldsymbol{\theta}$. It should be noted that the ijth element of $\mathbf{C}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{X})$ can be defined by some kernel functions, e.g. the Gaussian kernel $c(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\phi \sum_{d=1}^D (x_{id} - x_{jd})^2 + \lambda\}$ where we denote the hyper-parameters $\boldsymbol{\theta} = \{\phi, \lambda\}$. Other kernels could also be used and found in [10]. We employ an auxiliary variable vector $\mathbf{y} = (y_1, \cdots, y_N)^T$ for the noise model such that $y_n = f_n(\mathbf{x}_n) + \mathcal{N}(0, 1)$ which defines a non-linear regression between $\mathbf{y}$ and $\mathbf{X}$. The relationship between y_n and t_n is as follows:

$$t_n = -1 \quad \text{if} \quad y_n < 0; \qquad t_n = 1 \quad \text{if} \quad y_n \geq 0. \tag{1}$$

The posterior over the hidden variables can be represented as follows using Bayes' rule

$$p(\mathbf{f}, \mathbf{y}|\mathcal{D}, \boldsymbol{\theta}) = \frac{P(\mathbf{t}|\mathbf{y}) p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{\int\int P(\mathbf{t}|\mathbf{y}) p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{y} d\mathbf{f}} \tag{2}$$

Gibbs sampler and variational Bayes have been developed for approximating this joint posterior by using an approximating ensemble of factored posteriors such that $p(\mathbf{f}, \mathbf{y}|\mathcal{D}, \boldsymbol{\theta}) \approx Q(\mathbf{f})Q(\mathbf{y})$,

details of which can be found in [4]. It has been shown that given a new feature vector $\mathbf{x}^*$, the predictive probability of it belonging to class 1 can be represented as

$$P(t^* = 1|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}^*) = \Phi\left(\frac{\widetilde{f}^*}{\sqrt{1+\widetilde{\sigma}_*^2}}\right) \tag{3}$$

with $\widetilde{f}^* = \widetilde{\mathbf{y}}^T \left(\mathbf{I} + \mathbf{C}_{\boldsymbol{\theta}}\right)^{-1} \mathbf{c}^*$ where $\mathbf{c}^* = (c(\mathbf{x}_1, \mathbf{x}^*), \cdots, c(\mathbf{x}_N, \mathbf{x}^*))^T$ and $\widetilde{\mathbf{y}}^T$ is the expectation of $Q(\mathbf{y})$, and $\widetilde{\sigma}_*^2 = c^* - (\mathbf{c}^*)^T \left(\mathbf{I} + \mathbf{C}_{\boldsymbol{\theta}}\right)^{-1} \mathbf{c}^*$ where $c^* = c(\mathbf{x}^*, \mathbf{x}^*)$. Note that $\Phi(\cdot)$ denotes the cumulative function of the standard normal distribution, i.e. the *probit* function.

As an alternative approach, the hidden vector $\mathbf{y}$ in (2) can be integrated out such that,

$$\int P(\mathbf{t}|\mathbf{y}) p(\mathbf{y}|\mathbf{f}, \mathbf{X}) d\mathbf{y} = \prod_{n=1}^{N} \int P(t_n|y_n) p(y_n|f_n) dy_n = \prod_{n=1}^{N} \Phi(t_n f_n) \tag{4}$$

Note that this *probit* function could be directly replaced by a *logistic* function. However, we only shortly describe the main idea of the GP method using probit function which enables the predictive posterior to be analytically tractable. Therefore, the posterior distribution function over $\mathbf{f}$ given hyper-parameters can then be represented as follows

$$p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}) = \frac{P(\mathbf{t}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{\int P(\mathbf{t}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{f}} = \frac{\mathcal{N}_{\mathbf{f}}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}}) \prod_{n=1}^{N} \Phi(t_n f_n)}{p(\mathbf{t}|\boldsymbol{\theta})}. \tag{5}$$

For a new feature vector $\mathbf{x}^*$, the predictive likelihood of it belonging to class 1 can be represented as follows

$$P(t^* = 1|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}^*) = \int P(t^* = 1|f^*) p(f^*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}^*) df^* \tag{6}$$

where

$$p(f^*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}^*) = \int p(f^*|\mathbf{f}, \mathbf{X}, \boldsymbol{\theta}, \mathbf{x}^*) p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}) d\mathbf{f} \tag{7}$$

Note that the posterior distribution $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta})$ is non-Gaussian which makes the predictive distribution described to be analytically intractable. Various approximation methods are required to be employed to represent it as a Gaussian form such that $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\theta}) \approx \mathcal{N}_{\mathbf{f}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. These include Laplace approximation [20] and expectation propagation [11, 5] approximations. The predictive distribution then has a closed form of (3) with $p(f^*|\mathcal{D}, \boldsymbol{\theta}, \mathbf{x}^*) \approx \mathcal{N}(\widetilde{f}^*, \widetilde{\sigma}_*^2)$ where $\widetilde{f}^* = (\mathbf{c}^*)^T \mathbf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}$ and $\widetilde{\sigma}_*^2 = c^* - (\mathbf{c}^*)^T \left(\mathbf{C}_{\boldsymbol{\theta}}^{-1} - \mathbf{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma} \mathbf{C}_{\boldsymbol{\theta}}^{-1}\right) \mathbf{c}^*$ where c^* and $\mathbf{c}^*$ have the same meanings with those in equation (3). These approximations will be used to infer classifiers for the EEG data sets described below.

3 Data Sets

The data used for this study correspond to the EEG data set IIIb of the BCI competition III [2]. This data set gathers the EEG signals recorded for three subjects who had to perform motor imagery i.e. to imagine left or right hand movements. Hence, the two classes to be identified were "Left" and "Right".

The EEG were recorded by the Graz team, using bipolar electrodes C3 and C4 (that are located over the motor cortex area), and were filtered between 0.5 and 30 Hz. Subject 1 took part in a virtual reality experiment [6] where the detection of left or right imagined hand movements triggered a camera rotation towards the left or right, respectively, in a virtual room. Subjects 2 and 3 took part in a "basket" experiment where the detection of left or right hand movement made a falling ball, displayed on the screen, move towards the left or the right. The aim was to reach one of the two baskets located on the bottom left and bottom right of the screen [19]. For subject 1, 320 trials were available in the training set, whereas the test set was composed of 159 trials. For subject 2 and 3, both the training and the test set were composed of 540 trials. More details about this data set can be found in [2].

4 Feature Extraction

For further classification, it is first necessary to extract features from these EEG signals. In order to do so, we chose to used Band Power (BP) features. Such features correspond to the power of the signal, in specific frequency bands. They are simply obtained by band-pass filtering the signal, squaring it and averaging it over a given time window [14]. Such features are very popular and efficient for motor imagery as imagination of hand movements is known to trigger amplitude changes in the α ($\approx$ 8-13 Hz) and β ($\approx$ 16-24 Hz) rhythms, over the motor cortex area [14].

The main drawback of such features is that subject-specific frequency bands, in which computing the BP, must be identified before using them. Actually, the optimal frequencies for discriminating between left and right hand movements vary from subject to subject [14]. Moreover, and independenlty from the features used, it is necessary to identify, for each subject, the optimal time window in which extracting the features, in order to achieve maximal discrimination. To achieve these two goals, we used a method based on statistical analysis which was successfully used in previous BCI studies [7, 9]. It should be noted that these calibration steps were performed before entering the classification procedures with the aim of identify the frequency bands and the time window to be used. Once identified, these frequency bands and time window will be used without any modification in the classification procedures.

To identify the subject-specific frequency bands, we used a paired t-test which compared the BP means between both classes, for every 2 Hz wide frequency bands between 1 Hz and 30 Hz, with a step of 1 Hz. As expected from the litterature [14], the frequencies for which the BP achieved the highest discrimination were found in the α and β bands, which support the use of such features (see Fig. 1).

Adjacent significant frequencies (with probability of type I error below $\alpha = 0.01$) were gathered into a single frequency band. Then, for every frequency band, a *shrinking* step was performed which consisted in reducing the frequency band (making it 1 Hz shorter) and computing a new statistic for this band. If the new statistic was higher than the previous one, the shrunk frequency band was selected. The shrinking process was repeated until the statistics could not be increased any further.

To identify the optimal time window in which BP features will be extracted, we performed the statistical analysis mentionned above for several time window, and selected the one with the highest mean value of significant statistics. The parameters used for BP feature extraction are summed up in Table 1. In this table, the window start value is given in seconds after the feedback presentation.

Thus, this BP feature extraction method represents each trial by a four dimensionnal feature

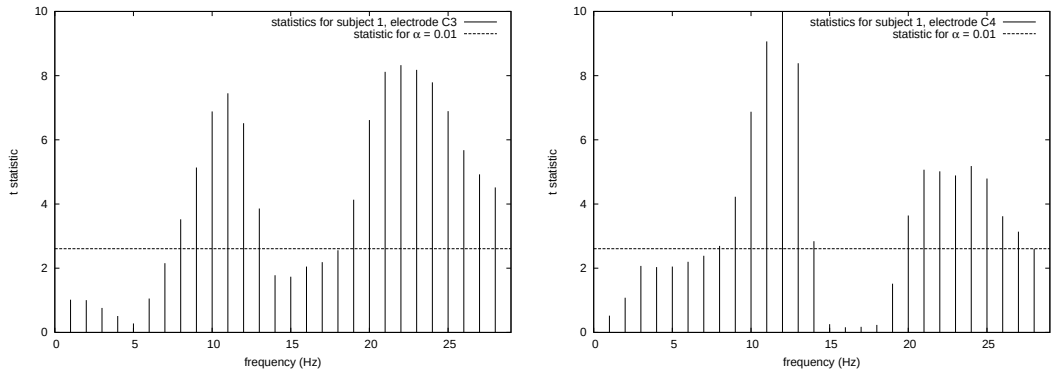

Figure 1: T statistics obtained with the BP features extracted for each frequency, for electrodes C3 (on the left) and C4 (on the right) for Subject 1, in the optimal time window (see below for the determination of this time window). The dashed line represents the significance threshold for $\alpha = 0.01$.

Table 1: Parameters of band power feature extraction for each subject.

Subject	C3 α band (Hz)	C3 β band (Hz)	C4 α band (Hz)	C4 β band (Hz)	window start (s)	window length (s)
1	11	21-29	11-13	21-27	0.4	2.5
2	8-13	20-24	11-14	20-29	1.4	1.5
3	9-12	21-22	11-12	18-25	1.4	1.5

vector: $[C3_\alpha, C3_\beta, C4_\alpha, C4_\beta]$ in which Cp_y is the BP value for eletrode Cp in the y band. These feature vectors will be used as input data for the following classification step.

5 Results

Five approaches for Gaussian process classification were applied to the three data sets described in section 3, using the Band Power features presented in section 4. The approximation methods employed are expectation propagation (EP) [11, 5], variational Bayes (VB) [4], Gibbs sampler [4] and Laplace approximation [20]. For Laplace approximation we consider both the probit and logistic functions in the noise model. For comparison purposes, SVM, KNN, pSVM and KLR were also employed for tackling this problem. For KLR, each weight was assumed to follow a Gaussian distribution with zero mean and variance σ^2. Note that it would be interesting to consider a sparse prior such as a Laplacian or a student-t distribution, which induces sparsity of the weights. After the prior was fixed, the weights were then learned by using the Laplace approximation. For all the methods used here, we employ the kernel function $c(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\phi \sum_{d=1}^{D}(x_{id} - x_{jd})^2 + 2\}$ where one hyper-parameter ϕ is required to be tuned. In order to obtain the relatively optimal hyper-parameters, ten-fold cross-validation (CV) was employed for tuning them. Note that one more hyper-parameter C, i.e. the box constraint parameter, was also optimized using ten-fold CV for SVM, and also the parameter k of KNN,

	S1			S2			S3		
	PL	PE	PT	PL	PE	PT	PL	PE	PT
EP	-0.340	**10.691**	0.329	-0.556	28.333	2.187	-0.483	24.814	2.172
VB	-0.340	**10.691**	0.047	-0.542	27.592	0.5	-0.491	24.814	0.437
GIBBS	-0.375	**10.691**	53.125	-0.540	27.037	123.891	-0.482	25.740	125.75
PL	-0.342	**10.691**	0.203	-0.561	27.962	1.938	-0.487	24.814	1.5
LL	-0.341	11.320	0.047	-0.542	27.222	0.47	-0.484	24.814	0.37
PSVM	-0.384	15.094	0.015	-0.542	**25.556**	0.078	-0.540	25.370	0.078
KLR	-0.359	12.578	0.016	-0.558	27.407	0.109	-0.483	25.370	0.094
SVM	-	13.836	0.016	-	29.074	0.078	-	25.555	0.062
KNN	-	14.465	0.078	-	37.777	0.172	-	**24.629**	0.171

Table 2: The prediction error of SVM and KNN, and the log-predictive likelihood (PL) and prediction error (PE) of Gibbs sampler, EP, VB, probit Laplace (PL) and logistic Laplace (LL) approximations, the probabilistic SVM of Platt [15] (pSVM) and kernel logistic regression (KLR), when applied to the data sets obtained from Subject 1 (S1), Subject 2 (S2) and Subject 3 (S3). The total prediction time (PT) in seconds of each learned classifiers when applied to the test data sets are also shown. Best results are highlighted in bold.

i.e. the number of neighbors. After the hyper-parameters were selected using the training data sets, the classifier was obtained by learning the training data set and then applied to the test data set. The results of the log-predictive likelihood (PL) and the 0-1 loss class prediction error (PE) of those methods employed are shown in Table 2. Note that the PL and PE are respectively defined as $\frac{1}{M}\sum_{i=1}^{M}\log\{P(t_i^*=t_{true}^i|\mathcal{D},\boldsymbol{\theta},\mathbf{x}_i^*)\}$ and $\frac{100}{M}\sum_{i=1}^{M}\mathcal{I}(t_{pred}^i\neq t_{true}^i)$, where M is the number of test samples, t_{true}^i and t_{pred}^i denote the true and predicted labels of the i^{th} test sample respectively, and $\mathcal{I}(\cdot)$ denotes the indicator function. The results show that there are no obvious differences in terms of predictive likelihood between those probabilistic methods employed for the current binary classification problems. However, the results show that GP outperforms SVM in terms of prediction error across all the data sets. On the other hand, except to the third data set, GP is superior to KNN in terms of the prediction error. As a by-product we collected the total prediction time of each learned classifier when applied to the test data sets. The results suggest that except to the Gibbs sampler all the classification methods are likely to be efficient enough for some real time applications such as a BCI system. Note that the experiments presented here were done using Matlab-6.5 under Windows XP, running on an Intel Pentium 4 CPU 3.4GHz, with 1GB RAM.

6 Discussions and Conclusions

Binary classification based on Gaussian process has been applied to EEG. Experimental results have shown that all the GP methods employed in this paper give similar performance on the used EEG data, in terms of predictive likelihood. It has been shown that GP outperforms SVM and KNN in terms of prediction error on the EEG data sets employed. Experimental results indicate that no evidence was shown that KLR or pSVM are better than GP based methods in terms of predictive likelihood. Therefore, following the hard advantages of GP described in the introduction section, we suggest using GP based methods in BCI. Furthermore, when classifying a new test sample to one of the k classes, the classifier which can produce predictive probabilities of the test sample is of great usefulness in practical recognition circumstances. This posterior probability, which can facilitate the separation of inference and decision, essentially

represents the uncertainty in the prediction in a principal manner [3]. The SVM only produces a threshold value which is not a probability for making a decision for a new test sample. This has been a shortcoming for SVM. By contrast, as we have seen, the GP-based classification method can naturally produce posterior probabilities. Importantly, these probability outputs can be used for further processing for a BCI system. For example, this quantity can isolate the test feature vector which has great uncertainty with similar class posterior probability values of binary classification problems. In this case, the subject might not attend the designed tasks and the data sample is not suitable for further use in a BCI system. In our observation, there is a case that the predictive probabilities of misclassification for some data samples are very high. The reason might be that the subject was actually doing an opposite task with respect to the expected one. Imagine a motor task of imaging left (class -1) and right (class $+1$) hand movements in an experiment, the subject was asked to imagine left hand movement for instance. Unfortunately, the subject actually imagined right hand movement which is opposite to the task. The predictive probability of classifying the sample to class $+1$ in this case should be very high, though the data sample is labeled as class -1. Besides, GP-based classification could be used for an asynchronous BCI, in which no cue stimulus is used and the subject can intend a specific mental activity as he wishes. The posterior probability can then be used as a quantity for detecting the mental events and discriminating them from noise and nonevents [18]. These have shown that GP provides a suitable quantity for further processing for a BCI.

7 Acknowledgements

This work was supported by grant ANR05RNTL01601 of the French National Research Agency and the National Network for Software Technologies within the Open-ViBE project.

References

[1] Bashashati, A., Fatourechi, M., Ward R.K., Birch, G.E., 2007. A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. Jouranl of neural Engineering, 4, R32-R57.

[2] Blankertz, B., Muller, K. R., Krusienski, D. J., Schalk, G., Wolpaw, J. R., Schlogl, A., Pfurtscheller, G., Millan, J. D. R., Schroder, M., Birbaumer, N., 2006. The BCI Competition III: Validating Alternative Approaches to Actual BCI Problems. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14(2), 153-159.

[3] Duda, R. O., Hart, P. E., 1973. Pattern Classification and Scene Analysis, John Wiley & Sons.

[4] Girolami, M., Rogers, S., 2006. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Computation, 18(8), 1790-1817.

[5] Kuss, M., Rasmussen, C. E., 2005. Assessing approximate inference for binary gaussian process classification. Journal of Machine Learning Research, 6, 1679-1704.

[6] Leeb, R., Scherer, R., Lee, F., Bischof H., Pfurtscheller, G., 2004. Navigation in Virtual Environments through Motor Imagery. 9th Computer Vision Winter Workshop, CVWW'04, 99-108.

[7] Lotte, F., 2006. The use of Fuzzy Inference Systems for classification in EEG-based Brain-Computer Interfaces. Proceedings of the third international Brain-Computer Interface workshop and training course, 12-13.

[8] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces. Journal of Neural Engineering, 4, R1-R13.

[9] Lotte, F., Lécuyer, A., Lamarche, F., and Arnaldi, B., 2007. Studying the use of fuzzy inference systems for motor imagery classification. IEEE Transactions on Neural System and Rehabilitation Engineering, 15(2), 322-324.

[10] MacKay, D. J. C., 2003. Information Theory, Inference, and Learning Algorithms, Cambridge Press.

[11] Minka, T., 2001. A family of algorithm for approximate Bayessian inference, MIT.

[12] M*ü*ller, K. R., Krauledat, M., Dornhege, G., Curio, G., Blankertz, B., 2004. Machine learning techniques for brain-computer interfaces. Biomedical Engineering, 49(1), 11-22.

[13] Neal, R., 1998. Regression and classification using gaussian process priors, in: Dawid, A.P., Bernardo, M., Berger, J.O., Smith, A.F.M. (Eds.), Bayessian Statistics 6, Oxford University Press, 475-501.

[14] Pfurtscheller, G., Neuper, C., 2001. Motor Imagery and Direct Brain-Computer Communication. proceedings of the IEEE, 89(7), 1123-1134.

[15] Platt, J. C., 1999. Probabilities for Support Vector Machines, in: Smola, A., Bartlett, P., Sch*ö*lkopf, B., Schuurmans, D. (Eds.), Advances in Large Margin Classifiers, MIT Press, 61-74.

[16] Rasmusen,C.E., Williams,C.K.I., 2006. Gaussian Processes for Machine Learning, the MIT Press.

[17] Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1, 211-244.

[18] Townsend, G., Graimann, B., Pfurtscheller, G., 2004. Continuous EEG classification during motor imagery-simulation of an asynchronous BCI. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 12(2), 258-265.

[19] Vidaurre, C., Schlogl, A., Cabeza, R., Pfurtscheller, G., 2004. A fully on-line adaptive Brain Computer Interface. Biomed. Tech. Band, Special issue, 49, 760-761.

[20] Williams, C. K. I., Barber, D., 1998. Bayessian classification with gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), 1342-1352.

[21] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., Vaughan, T. M., 2002. Brain-computer interfaces for communication and control. Clinical Neurophysiology, 113(6), 767-791.