A latent discriminative model-based approach for classification of imaginary motor tasks from EEG data

A latent discriminative model-based approach for classification of imaginary motor tasks from EEG data

Jaime F Delgado Saa[1,2] **and Müjdat Çetin**[1]

[1] Signal Processing and Information Systems Laboratory, Sabanci University, Orhanli, Tuzla, 34956 Istanbul, Turkey

[2] Department of Electrical and Electronic Engineering, Universidad del Norte, Barranquilla, Colombia

E-mail: delgado@sabanciuniv.edu and mcetin@sabanciuniv.edu

Received 16 November 2011
Accepted for publication 8 February 2012
Published 14 March 2012
Online at stacks.iop.org/JNE/9/026020

Abstract

We consider the problem of classification of imaginary motor tasks from electroencephalography (EEG) data for brain–computer interfaces (BCIs) and propose a new approach based on hidden conditional random fields (HCRFs). HCRFs are discriminative graphical models that are attractive for this problem because they (1) exploit the temporal structure of EEG; (2) include latent variables that can be used to model different brain states in the signal; and (3) involve learned statistical models matched to the classification task, avoiding some of the limitations of generative models. Our approach involves spatial filtering of the EEG signals and estimation of power spectra based on autoregressive modeling of temporal segments of the EEG signals. Given this time–frequency representation, we select certain frequency bands that are known to be associated with execution of motor tasks. These selected features constitute the data that are fed to the HCRF, parameters of which are learned from training data. Inference algorithms on the HCRFs are used for the classification of motor tasks. We experimentally compare this approach to the best performing methods in BCI competition IV as well as a number of more recent methods and observe that our proposed method yields better classification accuracy.

(Some figures may appear in colour only in the online journal)

1. Introduction

A brain–computer interface (BCI) is a system that provides an alternative communication pathway for patients who have lost their ability to perform motor tasks due to disease or accident [4]. In addition, applications for healthy subjects in the fields of multimedia and gaming have started to incorporate these technologies in recent years as well [15]. BCIs aim to use brain signals to help subjects control external devices and interact with their environment. In the case of execution of (real or imaginary) motor tasks, it is known that electroencephalography (EEG) signals measured over the motor cortex exhibit changes in power related to the movements. These changes primarily involve increase and decrease of power in the alpha (8–13 Hz), sigma (11–15 Hz), beta (18–26 Hz) and low gamma (25–35 Hz) frequency bands [23]. These phenomena are known as event-related synchronization and desynchronization [20]. This information can be used to classify different imaginary motor tasks by comparison of the power levels of the EEG signals recorded in a number of positions on the scalp. In particular, changes of the signal power in different frequency bands with time provide useful information. Based on this observation, methods based on time–frequency analysis of the EEG signals have been proposed [13, 17, 31]. Furthermore, algorithms involving stochastic time series models taking into account changes of the signal power with time, such as hidden Markov models (HMMs) [16, 29, 2, 6], have been used in combination with features describing the temporal behavior of the EEG signals [10, 30]. We share the perspective with this latter body of work that changes in the power of the signals during execution of motor tasks reflect the underlying states in the brain and that the sequence of states provides useful information for discrimination of different imaginary motor tasks. Previous work based on HMMs has shown that this approach provides good results [16, 29, 2, 6]. Nevertheless, if the EEG signal is modeled by an HMM, which is a generative model, the distribution of the data must be estimated and conditional independence assumptions of the data given the underlying states should be incorporated in order to make the inference problem tractable. A remedy for this problem is the use of conditional random fields (CRFs) [12]. Although this is a discriminative model that does not require the estimation of the distribution of the data, there is one more issue for the case of the BCI applications, where, unlike the analysis of sleep EEG signals based on CRFs as proposed in [8], the sequence of states is unknown. A modified CRF method has been proposed for BCI in [3] where the classes are associated with states in the CRF. However, this method does not utilize intermediate states in the EEG signal related to each mental state, which have been proved to increase the performance in HMM-based approaches [16, 29, 2]. This motivates the use of hidden states in CRF. Gunawardana *et al* [9] have proposed a hidden-state CRF with application to phone classification which has been generalized by Sugiura *et al* in [28] to the so-called hierarchical hidden state CRF (HHCRF). Sugiura *et al* have presented an application of HHCRF in EEG signal segmentation in an asynchronous BCI application exhibiting advantages when compared to the generative counterpart, the hierarchical HMM. However, the model proposed in [28] is based on a complicated structure making the parameter estimation and state sequence approximation computationally expensive. Quattoni *et al* [21] have proposed a hidden conditional random field (HCRF) model that uses hidden variables to model the latent structure of the input domain and defines a joint conditional distribution over the class labels and the hidden variables given the observations. Contrary to the work in [9], the HCRF model defined by Quattoni *et al* does not fix the sufficient statistics used in the potential function of the CRF and does not assume Gaussianity of the data, which leads to a more flexible model selection process.

Motivated by the work in [21], we present an HCRF-based approach for classification of imaginary motor tasks in a synchronous BCI scenario, where the labels do not change with time, making it unnecessary to define a top layer with different states as in HHCRF. In our approach, the collected EEG data are first spatially filtered using the common spatial pattern (CSP) technique. We perform feature extraction through time–frequency analysis of the spatially filtered signals based on auto-regressive modeling. Autoregressive models of 1 s intervals of the filtered EEG signals are used to estimate their power spectra, obtaining a representation in time and frequency. Feature selection is performed by selection of frequency bands related to the execution/imagination of motor tasks (alpha, sigma, beta, low gamma). These extracted features constitute the data to be fed to the HCRF. Intermediate brain states are defined and represented by latent variables in the HCRF model. Model parameters are learned from labeled training data, and inference algorithms on HCRFs are used for classification. We present experimental results demonstrating the improvements provided by our HCRF-based approach over the best-performing method in BCI competition IV as well as over the HMM-based method, a CRF-based method, and the recently proposed bispectrum-based approach in [26].

2. HCRFs for BCI

In the task of labeling sequence data, one of the most widely used tools is the HMM [22], a finite automaton which contains discrete-valued states Q emitting a data vector X at each time point; the distribution of the data at each time point depends on the current state. Given that models of this kind are generative, they require computation of the joint probability density function of the observed data samples over multiple time points. In order to make the inference problem tractable, assumptions about independence of the data at each time point conditioned on the states should be made. Such assumptions are violated in many practical scenarios. CRFs are discriminative models that overcome these issues [12], avoiding the need to explicitly model the data distribution as well as the need for the independence assumptions. For *linear-chain* CRFs, Lafferty *et al* [12] define the probability of a particular label sequence $\bar{\mathbf{y}}$ given an observation sequence $\mathbf{x}$ to be of the form

$$P_\theta(\bar{\mathbf{y}}|\mathbf{x}) \propto \exp\left\{\sum_{l\in L_1}\sum_{j=1}^{m} f_{1,l}(\bar{y}_{j-1}, \bar{y}_j, \mathbf{x}, j)\theta_{1,l} + \sum_{l\in L_2}\sum_{j=1}^{m} f_{2,l}(\bar{y}_j, \mathbf{x}, j)\theta_{2,l}\right\}, \quad (1)$$

where j represents the discrete time index, m is the length of the sequence $\mathbf{x}$, $f_{1,l}$ and $f_{2,l}$ are the *CRF features*[3] related to the edges and nodes of the graph, respectively, and are given and fixed. L_1 and L_2 denote the sets of indices for the *CRF features*. One has to estimate the parameters $\theta_{1,l}$ and $\theta_{2,l}$ based on training data. A more detailed description of CRFs is beyond the scope of this paper, for which we refer the reader to [12].

This approach overcomes the problems stated above for HMMs. However, CRFs focus on assigning a label for each observation (e.g. each time point in a sequence), and they neither capture hidden states nor directly provide a way to estimate the conditional probability of a class label for an entire sequence. In the BCI problem, which is of interest in this paper, labels are not available for temporal segments of (training) EEG data recorded during the execution of a motor task, and the central problem of interest is to assign a class label for an entire sequence. As a result, it would be necessary to use a model that facilitates classifying an entire sequence and that involves hidden states. Such a model has been proposed in [21] and is called the HCRF. HCRFs are able to capture intermediate structures through hidden states, combined with the power of discriminative models provided

[3] These are simply called features in the CRF literature. However, to distinguish them from features to be extracted from the EEG signal, we call them CRF features.

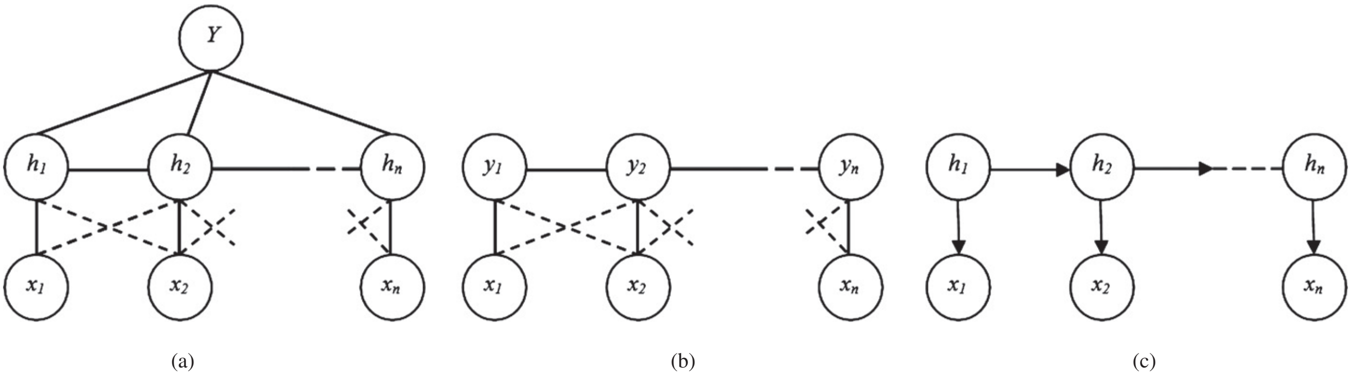

Figure 1. Dynamical statistical models. (a) HCRF model. (b) CRF model. (c) HMM. Dashed lines indicate the possibility of including long range dependences between the input data and the nodes.

by CRFs. Furthermore, unlike CRFs, they also provide a way to estimate the conditional probability of a class label for an entire sequence. An HCRF is constructed as follows. The task is to infer the class y from the data $\mathbf{x}$, where y is an element of the set Y of possible labels for the entire data and $\mathbf{x}$ is the set of vectors of temporal EEG features $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$. The subindex m represents the number temporal observations. The training data consist of a set of labeled samples $(\mathbf{x_i}, y_i)$ for $i = 1, \ldots, n$ where $y_i \in Y$ and $\mathbf{x_i} = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$. For any $\mathbf{x_i}$, a vector of latent variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ is assumed, providing the state sequence of the data. Each possible value for h_j is a member of a finite set H of possible hidden states. The joint probability of the labels and the states given the data is described as

$$P(y, \mathbf{h}|\mathbf{x}, \theta) = \frac{\exp(\Psi(y, \mathbf{h}, \mathbf{x}; \theta))}{\sum_{y',\mathbf{h}} \exp(\Psi(y', \mathbf{h}, \mathbf{x}; \theta))}, \qquad (2)$$

where θ are the parameters of the models and $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is a potential function $\in R$. The conditional probability of the labels given the data can be found by marginalizing out h:

$$P(y|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} \exp(\Psi(y, \mathbf{h}, \mathbf{x}; \theta))}{\sum_{y',\mathbf{h}} \exp(\Psi(y', \mathbf{h}, \mathbf{x}; \theta))}. \qquad (3)$$

Following [21], the estimation of parameter values, using the training data, can be performed by maximizing the following objective function:

$$L(\theta) = \sum_i \log P(y_i|\mathbf{x_i}, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2, \qquad (4)$$

where the first term in (4) is the log-likelihood of the data. The second term is the log of a Gaussian prior with variance σ^2. Given this objective function, various nonlinear optimization algorithms can be used to search for the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$. In our work, we use a quasi-Newton algorithm using Hessian updates based on the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula. Given a new test example $\mathbf{x}$ and parameter values θ^* induced from the training set, the label for the example is taken to be $\arg\max_{y\in Y} P(y|\mathbf{x}, \theta^*)$

HCRFs use undirected graphical structures, with the graph defined by $G = (V, E)$ where V denotes the vertices in the graph and E denotes the edges. Based on this, the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is defined as

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_{j=1}^{m} \sum_{l\in L_1} f_{1,l}(j, y, h_j, \mathbf{x})\theta_{1,l} + \sum_{(j,k)\in E} \sum_{l\in L_2} f_{2,l}(j, k, y, h_j, h_k, \mathbf{x})\theta_{2,l}, \qquad (5)$$

where $f_{1,l}$ and $f_{2,l}$ are the *HCRF features* related to the nodes and edges of the graph, respectively, and are given and fixed. L_1 and L_2 denote the sets of indices for the *HCRF features*. It is important to note that $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ is decomposed into a series of potentially local functions of the hidden variables. This property is the key for efficient inference over such models. If the set of hidden states forms a tree-structured graph, then exact methods for inference and parameter estimation can be used. In particular, the belief propagation algorithm [18] can be used to compute the marginal distributions of hidden states given the data, which can in turn be used in the solution of the classification problem defined above [21]. If the graph G contains cycles, approximate methods such as loopy belief propagation can be used for approximate inference.

Figure 1 shows an HCRF graphical model. The graphical structure of this model encodes which variables are involved in each of the functions defining the *HCRF features* in $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ in equation (5). For example, the chain structure of the hidden variables in the particular graphical model in figure 1 implies that the only hidden variables appearing in the edge *HCRF features* $f_{2,l}$ in equation (5) are those with adjacent indices, i.e. with $|j - k| = 1$. Likewise, in the case in which the possible edges indicated by dashed lines in figure 1 are missing, the node *HCRF features* in equation (5) for the graph in figure 1 would take the form $f_{1,l}(j, y, h_j, x_j)$. Furthermore, since y and x_j are not directly connected, but connected through h_j, $f_{1,l}$ would further decompose into two functions, one expressing the compatibility between y and h_j, and the other between h_j and x_j. Hence, the graphical model contains information directly related to the decomposition of the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$, which in turn specifies how the posterior probability of the labels in equation (3) is expressed in terms of local functions.

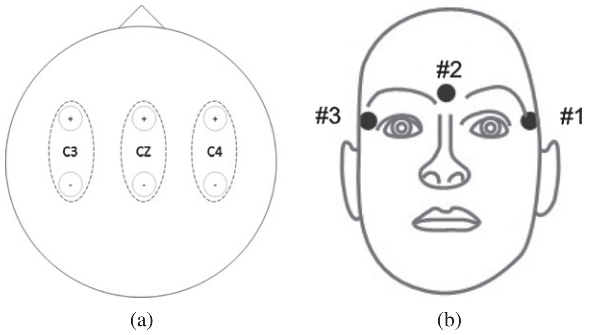

Figure 2. (a) Montage used to extract the signal on C3, C4 and Cz. (b) EOG channels [19].

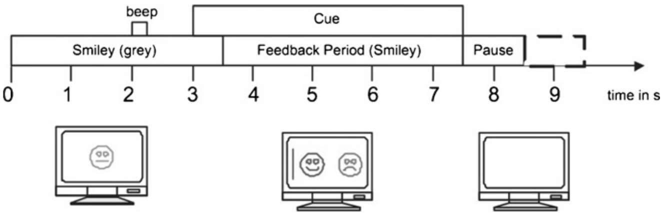

Figure 3. Time scheme for the experimental procedure.

3. Description of the proposed method and experiments

3.1. Problem and data set description

In typical BCI applications based on the imagination of motor activity, the subject is requested to execute imaginary motor tasks following a visual cue. It is known that the imagination of motor activities produces synchronization and/or desynchronization of the electrical signals recorded over the motor cortex and that this process has an asymmetrical spatial distribution during the imagination of the motor task (e.g. imagination of movement of a particular leg produces changes in the power of electrical signals in the contralateral region of the brain). Given a number of training sessions containing data from multiple trials in which the subject has been requested to imagine several motor tasks, the first task is to learn a model. Then, given some new (test) data, the task is to run an inference algorithm to perform classification of the imaginary motor task.

In this work, data set 2b of BCI competition IV [19], which consists of bipolar EEG recordings over scalp positions for electrodes C3, Cz and C4 (see figure 2(a)) in nine subjects, was used. The cue-based BCI paradigm involved two classes, represented by the imagination of the movement of the left hand and the right hand, respectively. The time scheme of the sessions is depicted in figure 3. At the beginning of each trial, a fixation cross and a warning tone are presented. Three seconds later, a cue (indicating left or right movement) is presented and the subject is requested to perform the imaginary movement of the corresponding hand. The data set contains five sessions, three for training and the remaining two for testing. Some of these sessions involved feedback, indicating to the subject how well the imagination of the motor task has been executed, and others did not. In our work, we have used the sessions with feedback. Temporal behavior of the EEG signals could be modified due to the feedback influence [14].

3.2. Artifact reduction

In order to reduce the interference of electrooculographic (EOG) signals in the EEG recordings, linear regression was employed, using the EOG data recorded at $N = 3$ channels using electrode locations shown in figure 2(b). In this approach, the signal recorded by the EEG electrodes is modeled as the summation of the actual underlying EEG signal and the noise, represented by a linear combination of the EOG signals interfering into the EEG electrodes [24]:

$$w(n) = s(n) + u(n).b, \tag{6}$$

where n represents the discrete time index, $w(n)$ and $s(n)$ represent the noisy and the actual EEG signals at M electrodes, and $u(n)$ represents the EOG signal at N electrodes. Representing $w(n)$, $s(n)$ and $u(n)$ at a particular time point as row vectors of appropriate dimensions, b is an *unknown* matrix of size $N \times M$ representing the set of coefficients that explain how the EOG signals have propagated by volume conduction to each of the points on the scalp where the EEG measurements are made. The problem is to recover $s(n)$ from measurements of $w(n)$ and $u(n)$. Given that the EOG signals are large in magnitude compared to the EEG signals, the interference of EEG in the EOG recordings $u(n)$ can be neglected [24]. If we knew b, the original EEG signal could be found by $s(n) = w(n) - u(n).b$. We describe a procedure to estimate b, which can then be used in this equation to estimate $s(n)$. Multiplying the signal $w(n)$ by $u(n)^T$ and taking expectation, we obtain

$$E[u(n)^T w(n)] = E[u(n)^T s(n)] + E[u(n)^T u(n)b]. \tag{7}$$

Under the assumption that there is no correlation between the EEG signal $s(n)$ and the EOG signals $u(n)$, we obtain an expression for estimating the coefficient matrix b:

$$\hat{b} = E[u(n)^T u(n)]^{-1} E[u(n)^T w(n)]. \tag{8}$$

We learn the correlation matrices above and compute $\hat{b}$ using a set of EOG and EEG measurements available in the data set for each one of the subjects as described in [24]. These measurements involve the execution of different ocular movements enabling the estimation of b before the start of the motor task classification sessions. We then use the estimated b in our experiments to estimate $s(n)$ based on data $w(n)$ and $u(n)$ recorded by EEG and EOG electrodes, respectively. The obtained signals are then bandpass filtered in the frequency band of interest for real/imaginary motor activity (8–35 Hz).

3.3. Feature extraction

3.3.1. Spatial filtering. CSPs are spatial filters that are well suited to discriminate mental states characterized by ERS/ERD phenomena [7]. Given the bandpass filtered, labeled EEG signals $s(n)$ ($1 \times M$ row vectors at each time point n) from the training set for each of the two classes C_1 and C_2, we estimate the $M \times M$ sample spatial covariance matrices Σ_{C_1} and Σ_{C_2} of the EEG signals for the two classes.

CSP performs the simultaneous diagonalization of Σ_{C_1} and Σ_{C_2} in such a way that the eigenvalues of the diagonalized matrices sum to 1, that is,

$$V^T \Sigma_{C_1} V = D \qquad \text{and} \qquad V^T(\Sigma_{C_1} + \Sigma_{C_2})V = I, \tag{9}$$

where V is the matrix of generalized eigenvectors, D is a diagonal matrix of eigenvalues and I is the identity matrix. Hence, the EEG signal $s(n)$ at each time point can be transformed from the electrode space to the CSP space through $s(n)V$. We can focus on the jth CSP component by using the filter V_j (jth column of V) and the resulting projected signals $s(n)V_j$. If the signal is from class 1, the variance of the projected signal will be $V_j^T \Sigma_{C_1} V_j = d_j$ (d_j is the corresponding eigenvalue for the eigenvector V_j). Likewise, for signals from class 2, the variance of the projected signal will be $1 - d_j$. Since we are interested in the discrimination of the two classes, it makes sense to use CSP components that emphasize the contrast between the classes. As observed, the filters V_j that provide the best contrast between the two classes are those with large eigenvalues and low eigenvalues, producing large variance for class 1 and low variance for class 2, and vice versa. Then, choosing those particular components corresponding to high and low eigenvalues only, the spatial filtered signal is obtained as follows:

$$c(n) = s(n)W, \tag{10}$$

where W is a matrix whose columns are composed of a subset of the eigenvectors V_j, in particular those with relatively large and small eigenvalues. In our experiments, we linearly transform the EEG signals at $M = 3$ electrodes into two CSP-filtered EEG signals using the largest and the smallest eigenvalues. Once the filters are designed based on the training data in this manner, they are applied on the test data.

3.3.2. Power spectral density estimation using auto-regressive parameters. The power spectrum of the signal is computed by parametric methods involving the calculation of autoregressive (AR) models of the signal. In this paper, we use Burg's method for AR model estimation because it provides better stability than the Yule–Walker method by minimizing the error in the backward and in the forward direction [27]. The power spectrum of the EEG signal is estimated as the frequency response of the auto-regressive model:

$$c_i(n) = \sum_{k=1}^{p} a_k c_i(n-k) + g(n), \tag{11}$$

where the subindex i represents each of the CSP components, n represents the discrete time index, p is the model order, a_k is the kth coefficient of the model and $g(n)$ is the system input or noise function. Then, we can compute the system function in the z-domain:

$$H_i(z) = \frac{C_i(z)}{G(z)} = \left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)^{-1}. \tag{12}$$

The AR spectrum can be obtained by evaluating $H_i(z)$ on the unit circle where $z = \exp(\mathrm{j}\omega)$ [11].

For estimating the AR parameters, we use a 1 s sliding window, over the spatial filtered signals $c(n)$. For each signal segment of 1 s, the model is estimated and the frequency response is obtained. The overlap of the segments was fixed to 90% of the window length. This produces a time–frequency map for each signal. From this time–frequency representation, the features used as input for the HCRF model are selected based on physiological information of the frequency bands related to execution/imagination of motor tasks. Table 1 shows the selected frequency bands used in this work. The features are calculated by taking the average power across frequency and the indicated frequency bands. The frequency resolution used in this work was 1 Hz.

Table 1. Selected frequency bands used as features for the HCRF model.

EEG rhythm	Frequency band (Hz)
Alpha	08–13
Sigma	11–15
Low beta	18–23
High beta	21–26
Low gamma	25–35

3.4. Model selection and classification

EEG feature vectors obtained using the auto-regressive power spectrum as described previously constitute the data $\mathbf{x}$ to be fed to the HCRF-based inference algorithm to be labeled. Since we use five frequency bands and two CSP components, the component x_j of the vector $\mathbf{x}$ at time point j is ten dimensional.

The particular HCRF model used in our work is a special case of the general form appearing in equation (5). In particular, we use a model represented by the graphical structure in figure 1(a), without the presence of the long range dependences indicated by the dashed lines. This leads to decoupling and a number of simplifications in the potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$ of equation (5). First, since y and $\mathbf{x}$ are only connected through $\mathbf{h}$, the node potential function decomposes into two terms, one relating y and $\mathbf{h}$, and the other one relating $\mathbf{h}$ and $\mathbf{x}$. Second, since long range dependences are not present, only x_j (rather than the past and future values present in the input sequence $\mathbf{x}$) is involved in the potential function for h_j. Third, the edge potential function involves cliques formed by consecutive nodes h_j and h_k (where $|j-k| = 1$) and the label y. Putting all of this together, we obtain the following potential function used in our work:

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j f_{1,1}(x_j) \cdot \theta_h[h_j] + \sum_j f_{1,2}(y, h_j) \cdot \theta_l + \sum_{(j,k)\in E} f_{2,1}(y, h_j, h_k) \cdot \theta_e, \tag{13}$$

where we have the *node-data* HCRF feature function $f_{1,1}(x_j) = x_j$. The dot product $f_{1,1}(x_j) \cdot \theta_h[h_j]$ measures the compatibility between the current EEG feature and the state h_j, where $\theta_h[h_j]$ are the weights associated with h_j. The second term, which involves $f_{1,2}(y, h_j) \cdot \theta_l$ measures the compatibility between the current state h_j and the motor task (label) y. Each element of the *node-label weight* vector θ_l contains a weight for a particular pair of values for the label and the hidden state. Hence, θ_l contains weights for all possible values of these variables. The HCRF-feature function $f_{1,2}(y, h_j)$ is an indicator vector, with a value of 1 for the entry corresponding to the particular set of values (y, h_j), and 0 for all the other entries. Hence, the dot product $f_{1,2}(y, h_j) \cdot \theta_l$ simply produces the weight for the particular pair of label and the hidden state (y, h_j). Similarly, the third term, which involves $f_{2,1}(y, h_j, h_k) \cdot \theta_e$ measures the compatibility of the state transition from h_j to h_k and the motor task y. Each element of the *edge weight* vector θ_e contains a weight for a particular triple of hidden state pairs and label. The HCRF-feature function $f_{2,1}(y, h_j, h_k)$ is an indicator vector, with a value of 1 for the entry corresponding to the particular set of values (y, h_j, h_k), and 0 for all the other entries. As the potential function in (13) can be written in the same form as (5) and the graphical structure modeling the hidden state transitions is a chain, algorithms such as belief propagation can be used for inference [21, 5].

One important issue in the BCI problem treated here is that the number of different brain states encountered during the execution or imagination of motor tasks is not obvious. In order to find the number of states that explain the signal well, a fourfold cross validation is performed over the training data, with possible values of 2, 3 and 4 for the number of distinct states[4]. From this set of models, with different numbers of hidden states, the model which provides the best classification accuracy after the cross-validation process, over the training data, is selected.

Once the model is selected, classification is performed by assigning the label y for a test sequence $\mathbf{x}$ as follows:

$$\hat{y} = \arg\max_{y \in Y} P(y/\mathbf{x}; \theta^*). \tag{14}$$

4. Results

We evaluate the performance of the HCRF-based approach presented above on BCI Competition IV data set 2b. The number of hidden states in the HCRF model was selected using a fourfold cross-validation on the training data. Table 2 shows the final selection of the number of hidden states in the HCRF model for each subject. The selected model for each subject was used to classify the data in the test sessions identified in the data set as B0X04E and B0X05E, with X indicating the respective subject.

We compare the results of our approach to the top three results in the competition for this data set. In addition, we also present a comparison with an HMM-based approach with Gaussian outputs and with a CRF-based method (using the same features employed for the HCRF model).

For the case of HMM, the number of hidden states and the number of Gaussian mixtures were selected by cross-validation. The graph for the HMM is depicted in figure 1(c). For the CRF model, there are no hidden states and the number of states is equal to the number of labels, as in the previous work on CRFs for BCI applications [3]. The CRF graphical model used here is depicted in figure 1(b). Given this graph structure, two class of feature functions are used, node features and edge features, as shown in equation (1). The parameters in the CRF model are learned through a quasi-Newton algorithm using Hessian updates based on the BFGS formula, which is the same procedure we use for HCRFs. As the BCI competition rules require, the HMM, CRF and HCRF models all produce an output (predicted class) for each time point. Table 2 shows the cross-validation accuracy for the HMM and the CRF model, as well as the number of states selected and used.

Table 2. Cross-validation accuracy (correct classification percentage) on training data and the number of states used for HCRFs, CRFs and HMMs. Note that in the HCRFs and HMMs the number of states makes reference to hidden states and is selected by cross-validation. For the CRF model, there are no hidden states, and the number of states is equal to the number of labels, in this case two (left- and right-hand motor imagery).

	HCRF		CRF		HMM	
Subject	CV-Acc	states	CV-Acc	states	CV-Acc	states
B01	83	2	80	2	79	2
B02	68	3	57	2	61	2
B03	50	3	45	2	54	2
B04	99	2	99	2	100	2
B05	95	2	89	2	96	2
B06	85	2	84	2	82	2
B07	90	2	91	2	90	2
B08	87	3	87	2	89	2
B09	87	2	89	2	90	3

Table 3. Comparison of the proposed HCRF-based approach with the top three methods in BCI competition IV as well as with HMM- and CRF-based techniques in terms of classification accuracy (kappa values). Bold is used to indicate the highest average rate obtained.

Subject	Chin.	Gan	Coyle	HMM	CRF	HCRF
B01	0.40	0.42	0.19	0.43	0.49	0.60
B02	0.21	0.21	0.12	0.16	0.23	0.32
B03	0.22	0.14	0.12	0.08	−0.03	0.06
B04	0.95	0.94	0.77	0.94	0.94	0.97
B05	0.86	0.71	0.57	0.86	0.73	0.87
B06	0.61	0.62	0.49	0.66	0.73	0.78
B07	0.56	0.61	0.38	0.63	0.46	0.63
B08	0.85	0.84	0.85	0.80	0.70	0.88
B09	0.74	0.78	0.61	0.71	0.48	0.81
Average	0.60	0.58	0.46	0.59	0.53	**0.66**

All methods used for comparison in table 3 use spatial filters (CSP) in the pre-processing stage or as for the case of the winner of the competition, an enhanced version of it, the filter bank CSP (FBCSP) [1]. Furthermore, a comparison with a recently published method based on the bispectrum of the EEG signal [26] is presented in table 4.

Following the methodology used in the competition, we use the kappa values [25] as the metric for comparing different methods:

$$\kappa = \frac{C \times P_{cc} - 1}{C - 1}, \tag{15}$$

where C is the number of classes and P_{cc} is the probability of correct classification[5]. Relatively larger kappa values indicate

[4] The value of 1 was not considered because it is physically inconsistent with phenomena involving changes (synchronization and desynchronization) in the EEG signal.

[5] Equation (15) takes this simple form given that the same number of samples for each class is available for each subject in each session.

Table 4. Comparison between the bispectrum + LDA approach and the proposed HCRF-based approach. 04E and 05E denote two distinct sessions in the test data. Max kappa refers to picking the best kappa value for each subject across the two sessions (following the analysis in [26]). Bold is used to indicate the highest average rate obtained.

	Shahid *et al* [26]			HCRF		
Subject	04E	05E	Max kappa	04E	05E	Max kappa
B01	0.64	0.44	0.64	0.70	0.51	0.70
B02	0.33	0.25	0.33	0.33	0.38	0.38
B03	0.29	0.15	0.29	0.11	0.00	0.11
B04	0.96	0.89	0.96	1.00	0.94	1.00
B05	0.60	0.68	0.68	0.88	0.86	0.88
B06	0.64	0.73	0.73	0.74	0.85	0.85
B07	0.43	0.57	0.57	0.56	0.75	0.75
B08	0.69	0.94	0.94	0.79	0.96	0.96
B09	0.81	0.68	0.81	0.84	0.78	0.84
Average	0.60	0.59	0.66	**0.66**	**0.67**	**0.72**

Table 5. Student's *t*-test results (*p*-values) evaluating the statistical significance of the difference between the performance of the proposed HCRF-based method and the other methods based on the results in table 3.

Subject	*p*-Value
HCRF versus Chin	0.0683
HCRF versus Gan	0.0178
HCRF versus Coyle	0.0013
HCRF versus HMM	0.0076
HCRF versus CRF	0.0011

better performance. According to the competition rules, the time course of the kappa value is calculated and the maximum kappa value is selected in reporting the results for each method.

The results of our experiments are shown in table 3. We observe that the method proposed in this paper provides higher kappa values than the top algorithms in the BCI competition, and the dynamic classifiers based on HMMs and CRFs. The proposed method outperforms all three algorithms from the BCI competition in eight out of nine subjects and produces an average kappa value of 0.66 compared to 0.60 for the winner of the competition. Table 5 presents the results of Student's *t*-test to evaluate the statistical significance of the difference between the performance of our approach and the methods of the BCI competition as well as the CRF- and HMM-based methods.

The methods from BCI competition IV we have compared against do not use the EOG data for artifact removal. Chin *et al* and Gan *et al* filter the EEG data for EOG artifact removal without using the EOG data, and Coyle *et al* do not perform EOG artifact removal at all. In order to ensure fairness in our comparisons with these methods, we have repeated our experiments without any EOG artifact removal. In this case, our HCRF-based approach has produced an average kappa value of 0.65. Note that our average kappa value with EOG artifact reduction was 0.66. Thus, our approach without EOG artifact reduction still performs better than the top three methods from the BCI competition which we compare against, two of which perform EOG artifact reduction using the EEG data.

The time course of the kappa value produced by our approach for each subject in each evaluation session is shown in figure 4. Given the structure of the model as depicted in figure 1(a), the HCRF model does not provide output for each sample point. Then, the plots in figure 4 are obtained by simulating an online experiment where data from the beginning of a trial to the current time point are used. In this way, the model calculates the likelihood of the sequence for each class and provides an output for each sample point. (Note that the evaluation requirements in the competition require that the algorithms provide outputs for each sample point.) The discussion on the time course of the kappa values also helps us contrast HCRFs with CRFs for synchronous BCI problems. If we plotted similar time courses for the CRF-based method whose results were presented in table 3 in comparison with our HCRF-based approach, we would observe that the time course is constant. CRFs are sequential labeling models able to model the extrinsic dynamics of the labels given the data. However, there is no label dynamics in a synchronous BCI paradigm, that is, during a trial no transitions among class labels occur. This will be learned by the CRF model generating a strong bias to remain in the same label during the trial. Then, an error in the label based on information at the beginning of a trial will propagate in time to the end of the trial. This explains why CRFs are not well suited for synchronous BCI applications and is also the reason for their rather poor performance, presented in table 3. A solution for this is proposed by [3] where the transitions are not modeled directly. However, as the signals (or EEG features) obtained during each trial are assumed to belong to the same state (which is also the label in this case), temporal intrinsic dynamics of data for each class are not exploited, contrary to what is actually achieved by the HCRF-based approach proposed in this paper. In this paper, we have compared the HCRF- and CRF-based models using one particular type of feature and one particular classification methodology. While we have not chosen these pieces to favor one model versus the other, we acknowledge that other choices (e.g. as in [3]) might lead to different performance results.

We also compare our HCRF-based approach to the recent work in [26] where a high order statistic method involving the bispectrum of the EEG signal, together with the linear discriminant analysis (LDA) was used for classification of motor imaginary tasks. The results are presented in table 4 following the methodology employed in [26]. These results demonstrate that our proposed HCRF-based approach outperforms the method in [26] on the BCI competition data set.

It is important to note that for dynamic models presented here (HCRF, CRF and HMM), and contrary to what is observed in LDA classifiers, the higher accuracy is obtained toward the end of the trial, which means that the time point of good performance is known *a priori* and the algorithm does not have to be optimized for a specific trial length.

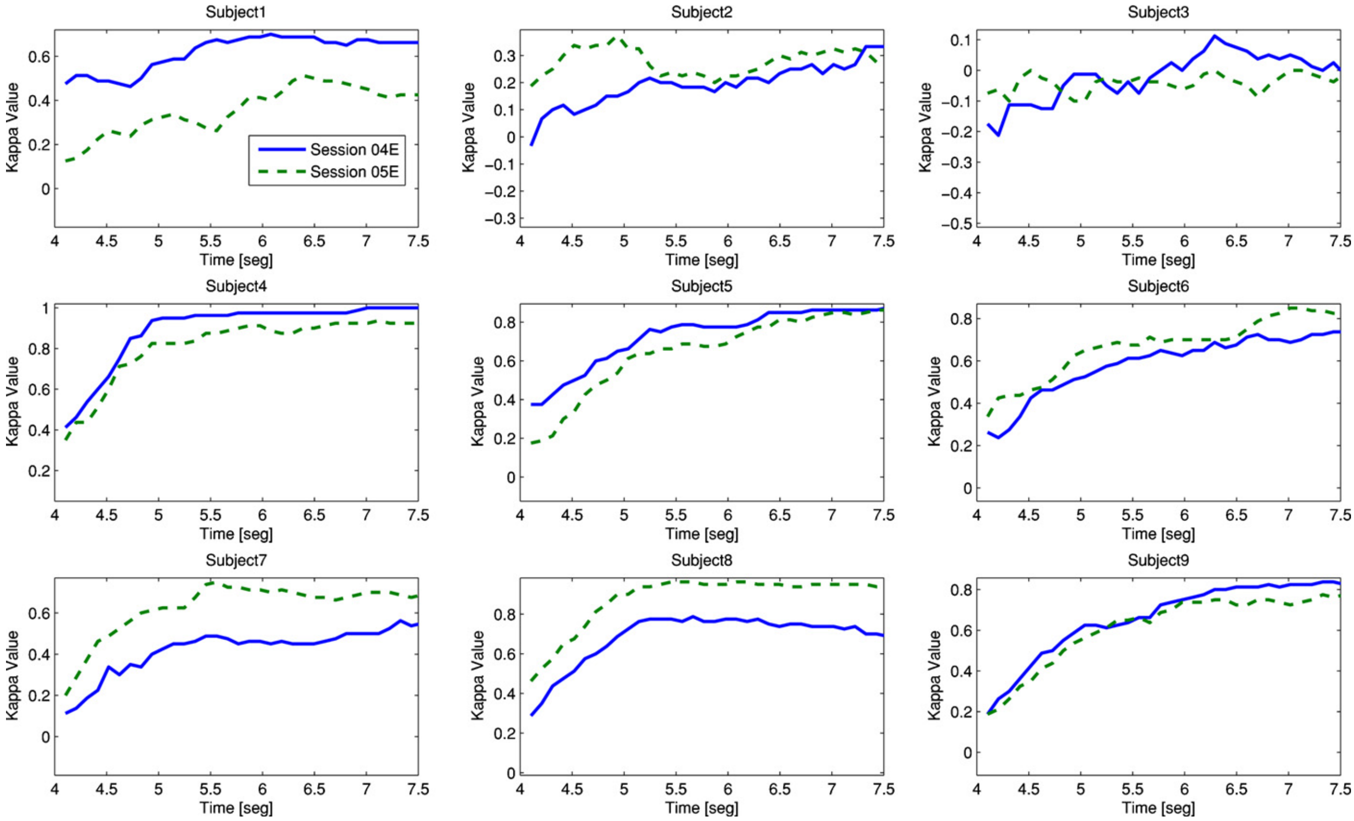

Figure 4. Time course of the kappa values for the proposed method in evaluation sessions 04E and 05E.

5. Conclusion

We have proposed a new method for classification of imaginary motor tasks, based on HCRFs. The autoregressive modeling of the CSP components, followed by the computation of the power spectrum and the selection of the frequency bands according to neurophysiological information, produces the feature vector that is fed to the HCRF-based classifier. Although subject-dependent selection of the frequency bands could lead to higher accuracy, we have opted here for common frequency bands for all subjects making the approach more general, which, given the performance obtained, shows the robustness of this method. Furthermore, the discriminative nature of the model proposed makes it unnecessary to model the distribution of the data or make assumptions about independence. Experimental results demonstrate the improvements in the classification accuracy provided by this approach over other methods. In addition, this method is based on modeling the temporal changes of the EEG signal and the analysis of the state sequences could provide insights into the physical phenomena underlying the execution of the imaginary motor tasks. This last point raises an interesting question about the physiological meaning of the states, which is the focus of our future work. One potential disadvantage of the proposed HCRF-based method (as well as the CRF- and HMM-based methods considered in our experiments) is the higher computational load as compared to simple classifiers used in BCI, such as the LDA. This is because the calculation of the likelihoods of sequences is computationally costly as compared to simple linear classification. Algorithmic and computational improvements are needed for applicability of these methods in real-time BCI applications.

Acknowledgments

This work was partially supported by the Scientific and Technological Research Council of Turkey under grant 111E056, Sabanci University Internal grant IACF-11-00889 and by a Turkish Academy of Sciences Distinguished Young Scientist Award.

References

[1] Ang K K, Chin Z Y, Zhang H and Guan C 2008 Filter bank common spatial pattern (FBCSP) in brain–computer interface *IEEE Joint Conf. Neural Networks (IEEE World Congress on Computational Intelligence)* pp 2390–7

[2] Argunsah A O and Cetin M 2010 AR-PCA-HMM approach for sensorimotor task classification in EEG-based brain–computer interfaces *20th Int. Conf. on Pattern Recognition* pp 113–6

[3] Awaad Shiekh Hasan B and Gan J Q 2011 Conditional random fields as classifiers for three-class motor-imagery brain–computer interfaces *J. Neural Eng.* **8** 025013

[4] Birbaumer N and Cohen L G 2007 Brain–computer interfaces: communication and restoration of movement in paralysis *J. Physiol.* **579** 621–36

[5] Bor W S, Quattoni A, Morency L-P, Demirdjian D and Darrell T 2006 Hidden conditional random fields for gesture recognition *IEEE Computer Society Conf.* vol 2 pp 1521–7

[6] Delgado Saa J F and Cetin M 2011 Modeling differences in the time–frequency presentation of EEG signals through

HMMs for classification of imaginary motor tasks *Technical Report, Sabanci University ID SU-FENS-2011/0003* available at http://research.sabanciuniv.edu/16498

[7] Dornhege G *et al* 2007 *Toward Brain–Computer Interfacing* (Cambridge, MA: MIT Press) chapter 13

[8] Gang L and Wanli M 2007 Subject-adaptive real-time sleep stage classification based on conditional random field *AMIA Annu. Symp. Proc.* pp 488–92

[9] Gunawardana A, Milind M, Acero A and Platt J C 2005 Hidden conditional random fields for phone classification *Interspeech* pp 1117–20

[10] Hjorth B 1970 EEG analysis based on time domain properties *Electroencephalogr. Clin. Neurophysiol.* **29** 306–10

[11] Jansen B, Bourn J and Ward J 1981 Autoregressive estimation of short segment spectra for computerized EEG analysis *IEEE Trans. Biomed. Eng.* **BME-28** 630–7

[12] Lafferty J D, McCallum A and Pereira F 2001 Conditional random fields: probabilistic models for segmenting and labeling sequence data *Proc. 18th Int. Conf. on Machine Learning* (San Francisco, CA: Morgan Kaufmann) pp 282–9

[13] Magjarevic R, Yonas A, Prihatmanto A S and Mengko T L 2010 Time–frequency features combination to improve single-trial EEG classification *World Congress on Medical Physics and Biomedical Engineering* vol 25/4 ed O Dössel and W C Schlegel (Berlin: Springer) pp 805–8

[14] Neuper C, Schlögl A and Pfurstcheller G 1999 Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery *Clin. Neurophysiol.* **16** 373–82

[15] Nijholt A, Reuderink B and Bos D O 2009 Turning shortcomings into challenges: brain–computer interfaces for games *Intelligent Technologies for Interactive Entertainment* vol 9 ed O Akan, P Bellavista and J Cao (Berlin: Springer) pp 153–68

[16] Obermaier B, Guger C, Neuper C and Pfurtscheller G 2001 Hidden Markov models for online classification of single trial EEG data *Pattern Recognit. Lett.* **22** 1299–309

[17] Palaniappan R 2005 Brain–computer interface design using band powers extracted during mental tasks *Proc. 2nd Int. IEEE EMBS Conf. on Neural Engineering* pp 321–4

[18] Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Francisco, CA: Morgan Kaufmann)

[19] Pfurstcheller G *et al* 2008 BCI competition IV, Data Set 2b available at http://www.bbci.de/competition/iv/

[20] Pfurtscheller G and Silva F L da 1999 Event-related EEG/MEG synchronization and desynchronization: basic principles *Clin. Neurophysiol.* **110** 1842–57

[21] Quattoni A, Bor W S, Morency L-P, Collins M and Darrell T 2007 Hidden conditional random fields *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 1848–52

[22] Rabiner L R 1989 A tutorial on hidden Markov models and selected applications in speech recognition *Proc. IEEE* **77** 257–86

[23] Sanei S and Chambers J A 2007 *EEG Signal Processing* (Chichester: Wiley) chapter 1

[24] Schlögl A, Keinrath C, Zimmermann D, Scherer R, Leeb R and Gert P 2007 A fully automated correction method of EOG artifacts in EEG recordings *Clin. Neurophysiol.* **118** 98–104

[25] Schögl A and Kronegg J 2007 *Toward Brain–Computer Interfacing* (Cambridge, MA: MIT Press) chapter 19

[26] Shahjahan S and Girijesh P 2011 Bispectrum-based feature extraction technique for devising a practical brain–computer interface *J. Neural Eng.* **8** 025014

[27] Stoica P and Moses R L 1997 *Introduction to Spectral Analysis* (New York: Prentice-Hall)

[28] Sugiura T, Goto N and Hayashi A 2007 A discriminative model corresponding to hierarchical HMMs *Proc. 8th Int. Conf. on Intelligent Data Engineering and Automated Learning* (Berlin: Springer) pp 375–84

[29] Suk H-I and Lee S-W 2010 Two-layer hidden Markov models for multi-class motor imagery classification *1st Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging* pp 5–8

[30] Vidaurre C, Krämer N, Blankertz B and Schlögl A 2009 Time domain parameters as a feature for EEG-based brain–computer interfaces *Neural Netw.* **22** 1313–9

[31] Zhendong M, Dan X and Hu J 2009 Classification of motor imagery EEG signals based on time–frequency analysis *Int. J. Digital Content Technol. Appl.* **3** 116–9