A Review of Adaptive Feature Extraction and Classification Methods for EEG-Based Brain-Computer Interfaces

Shiliang Sun and Jin Zhou

***Abstract*— A brain-computer interface (BCI) is a system that allows its users to control external devices which are independent of peripheral nerves and muscles with brain activities. Electroencephalogram (EEG) signals are electrical signals collected from the scalp. They are frequently used in brain-computer interaction. However, EEG signals which change over time are highly non-stationary. One major challenge in current BCI research is how to extract features of time-varying EEG signals and classify the signals as accurately as possible. An effective BCI should be robust against and adaptive to the dynamic variations of brain activities. Adaptive learning in a BCI system, a rapidly developing application of machine learning, would be an effective approach to conquer the challenge. This paper reviews representative adaptive feature extraction and classification methods for EEG-based BCIs and further discusses some important open problems which can hopefully be useful to promote the research of the BCIs.**

***Index Terms*— Brain-Computer Interface, Electroencephalogram, Adaptive Feature Extraction, Adaptive Classification, Machine Learning**

I. Introduction

A brain-computer interface (BCI) is a communication and control system in which messages or commands do not depend on the brain's normal output pathways of peripheral nerves and muscles [63]. It is very helpful to assist patients with damaged motor functions, such as completely paralyzed patients with amyotrophic lateral sclerosis. BCIs can also be widely used in the area of medicine and biometric identification. Research on BCIs gains more and more interest in recent years. It is an interesting, vibrant and highly interdisciplinary research topic at the interface among medicine, psychology, neurology, signal processing and machine learning [11], [28], [57], [63].

As Fig. 1 shows, BCIs can be seen as a pattern recognition system [13]. Its aim is to translate brain activities into commands for a robot or other devices. In order to achieve this goal, firstly signals from the brain are acquired by electrodes mounted on the scalp or in the head and subsequently the specific features of these signals will be extracted (e.g, amplitudes of evoked potentials, band powers or power spectral density values). Then these features are classified and translated into commands to control a device. In this paper, we focus on one kind of neurophysiological signals, namely electroencephalogram (EEG) signals that are electrical brain activities recorded from electrodes placed on the scalp. EEG signals are noninvasive, inexpensive, and relatively convenient to acquire compared with other signals like magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) [63].

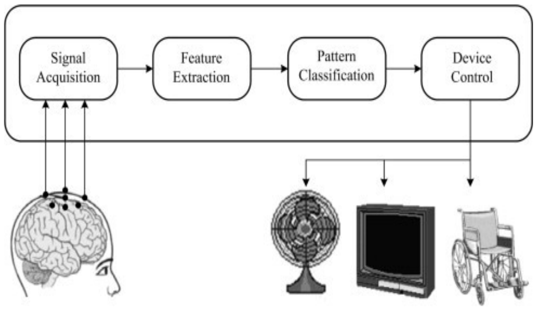

Fig. 1. A general EEG-based BCI [43]

However, for an EEG-based BCI, the recorded EEG signals are highly non-stationary [59]. They usually change over time due to both biological and technical causes such as the anatomical differences between subjects, the variability between different sessions, subject attention, mental state, amplifier noise and ambient noise [57]. Moreover, from a neuroscience perspective, the oscillatory components of EEG have distinct and non-stationary characteristics. The high variability of EEG signals makes it difficult to classify different EEG signals accurately. Therefore, to improve the performance of existing BCIs, adaptive learning is a very important issue in EEG-based BCI research. With respect to adaptive learning for EEG signals, one can choose to update feature extractors or alternatively classifiers. However, up to now, there is not so much work addressing this problem. This paper aims to promote this kind of research by briefly summarizing current adaptive feature extraction and classification methods used in EEG-based BCIs. Moreover, we give several open problems which can be useful for further progress in the area of BCI research.

The remainder of this paper proceeds as follows. In Section II, we review some representative adaptive feature extraction approaches used in EEG-based BCIs. Analogically, in Section III, adaptive classification methods are introduced. Then, in Section IV, we attempt to provide some open problems which may be helpful for further research of BCIs. Finally, we provide concluding remarks in Section V.

Shiliang Sun and Jin Zhou are with the Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, P. R. China (email: slsun@cs.ecnu.edu.cn, jinjin.zhou12@gmail.com).

This work is supported by the National Natural Science Foundation of China under Projects 61370175 and 61075005, and Shanghai Knowledge Service Platform Project (No. ZF1213).

II. Adaptive feature extraction methods

In a BCI system, the extracted features are non-stationary since EEG signals may rapidly vary over time. Thereby adaptive feature extraction methods are needed. Here we introduce some representative methods.

A. Adaptive common spatial patterns

Among various feature extraction methods of EEG signals, the common spatial pattern (CSP) method is one of the most effective in discriminating different classes of motor imaginary tasks such as left and right hand movements [12], [42]. For the two-class discrimination problems, CSP can be seen as a linear spatial filter which finds a projection direction that maximizes the variance of one class and simultaneously minimizes the variance of the other class. This method is based on the simultaneous diagonalization of two covariance matrices and has been successfully applied to the classification of movement-related EEG signals [27].

Let us briefly review the CSP method [27]. Suppose the raw EEG data of a single trial contains two different types of brain activities such as left and right hand movements. C_L and C_R are the matrices of data coming from trials of two classes (left and right). The normalized spatial covariance of the EEG signals can be given as

$$C = \frac{EE^\top}{trace(EE^\top)}. \tag{1}$$

E is the signal matrix sized $N \times T$, where N is the number of channels and T is the number of samples per channel. The covariance matrix of each class can be computed as the average of all single covariances belonging to one class, which are given as

$$S_L = \frac{C_L C_L^\top}{N_L}, \qquad S_R = \frac{C_R C_R^\top}{N_R}, \tag{2}$$

where N_L (N_R) is the number of channels belonging to the class of left (right). CSP tries to find a spatial filter which is calculated by the projection matrix Q. Then the data X can be transformed as $X_{CSP} = Q^\top X$. CSP considers the following optimization problem [42]:

$$\max J(\mathbf{w}) = \frac{\mathbf{w}^\top S_L \mathbf{w}}{\mathbf{w}^\top S_R \mathbf{w}}. \tag{3}$$

The solution of Eq. (3) is obtained by calculating a simultaneous diagonalization of S_L and S_R and the sum of the two diagonalized matrices is the identity matrix. Then we construct a matrix Q, composed of the first q and last q of $\mathbf{w}$, which correspond to the first q and last q ordered eigenvalues. According to $X_{CSP} = Q^\top X$, we can maximally separated the two classes by their variance.

As mentioned above, CSP can be seen as a linear approach. That means the short-term variation in EEG signals is ignored. Therefore, CSP may be likely to fail to discover the underlying change of the EEG data [47]. Moreover, due to the inherent variability of EEG patterns, the discriminative directions for classification usually shift over time. In order to resolve this problem, adaptive techniques that could take into account the non-stationary property of EEG signals are necessary in signal processing for BCIs [24], [32].

Sun et al. [48] proposed an adaptive method named as ACSP to update the corresponding covariance matrix by adding a variability coefficient θ. ACSP reflects the idea of weighted average. Simply speaking, given K EEG trials $x(k)$ from a certain class, the covariance of the original non-adaptive CSP is $C(k) = \frac{1}{K}\Sigma_{k=1}^{K} x(k)x^\top(k)$ and the CSP feature extractors generally adopt fixed covariances. In [48], Sun et al. updated the covariance matrix as $C(k) = \theta C(k-1) + (1-\theta)x(k)x^\top(k)$. This method contains not only historical information $C(k-1)$ but also the newly added information $x(k)x^\top(k)$, which can reflect the time-varying characteristics of EEG signals. Chen et al. [9] proposed a similar adaptive method to rectify the projection matrix Q, which makes the projection matrix automatically change with the new samples. This strategy provides a guarantee that Q can always reflect the optimal projection direction. Additionally, Zhao et al. [68] considered that CSP has poor adaptability since it is a batch-type algorithm. To resolve this problem, they proposed a new algorithm called the incremental common spatial pattern (ICSP) algorithm which trains the CSP on-line and proposed a novel formula for adapting a common spatial pattern trained on a block of recording data. The authors showed that using ICSP to update the spatial components on-line makes it more suitable for the non-stationary EEG signals.

B. Semi-supervised feature extraction

As we all know, EEG signals are time-varying, for example, the discriminant features of EEG signals recorded in the training and test sessions are often different. The non-stationary characteristic of EEG signals often degrades the performance of traditional feature extractors. To alleviate this limitation, the semi-supervised feature extraction methods are proposed for EEG classification.

Semi-supervised feature extraction is suitable for extracting the features of time-varying EEG signals, because it matches the requirement of small training sets owing to the need of short calibration sessions and alleviates the time-variances between the training and test sessions. Semi-supervised learning can learn with few labeled data and a large number of unlabeled data jointly where the test sessions are regarded as unlabeled data.

Sun [42] proposed extreme energy ratio (EER), which is a feature extractor to learn spatial filters for EEG signals classification. EER tries to discover source signals whose average energies from two conditions are the most different. The discriminative EER criterion is given as

$$\max / \min \quad \frac{\phi^\top S_L \phi}{\phi^\top S_R \phi}, \tag{4}$$

where S_L and S_R are the covariances of each class. For EEG data X, a spatial filter is donated by $\phi_{N\times 1}$ and the spatially filtered signal will be $\phi^\top X$. EER shows the theoretical equivalence and gives the computational savings in comparison with CSP. However, it is totally supervised.

By improving previous EER, Tu and Sun [54] further proposed two semi-supervised feature extraction methods. They are named as semi-supervised temporally smooth EER (STSEER) and semi-supervised importance weighted EER (SIWEER), respectively.

STSEER constructs a regularization term R on the preservation of the temporal manifold of test samples and adds it as a constraint to the learning of spatial filters. The STSEER criterion is given by adding the regularization term as follows:

$$\max \frac{\phi^\top S_L \phi}{(1-\alpha)\phi^\top S_R \phi + \alpha R}, \tag{5}$$

$$\max \frac{\phi^\top S_R \phi}{(1-\alpha)\phi^\top S_L \phi + \alpha R}, \tag{6}$$

where α is a constant defined by users to adjust the desired level of temporal locality to be preserved. SIWEER is obtained by introducing two *importance* weights, which are defined as *inter-trial importance* and *intra-trial importance*, respectively. By exploiting the distribution information of test samples, SIWEER assigns the two kinds of weights to training data points and trials to improve the estimation of covariance matrices. For the outliers or the "noisy" samples which strongly diverge from the distribution of test samples, it is expected that the two weights will be comparatively small so that SIWEER can effectively reduce the negative influence of the outliers and noisy samples. Thereby, the spatial filters learned by SIWEER can be not only robust to noisy samples but also adaptive to the test samples.

Besides, there are other semi-supervised feature extraction methods. For example, Lee et al. [20] proposed semi-supervised nonnegative matrix factorization (SSNMF) which is superior to standard NMF in extracting discriminative features from EEG data by combining the data matrix and the class label matrix into NMF.

C. Adaptive autoregressive parameters

Autoregressive (AR) parameters have been used successfully in EEG analysis and the AR model can well describe the stochastic nature of EEG [17]. However, the model parameters of AR are assumed to be unchanged over time. This makes the AR model unsuitable for analyzing the time-varying EEG signals. To address the problem of non-stationary EEG signals analysis, an approach referred to as the adaptive autoregressive (AAR) model is proposed and has been used in EEG-based BCIs [34], [37], [38].

A pth-order AAR model describes the signal in the following form:

$$y_t = a_{1,t}y_{t-1} + a_{2,t}y_{t-2} + \cdots + a_{p,t}y_{t-p} + x_t, \tag{7}$$

where $a_{1,t} \cdots a_{p,t}$ are the AAR model parameters and t describes discrete time points. In the ideal case, x_t is a purely random or white noise process with zero mean and variance σ_x^2. The difference to an AR model is that the parameters $a_{1,t} \cdots a_{p,t}$ vary with time.

In [39], Schlogl et al. proposed several available algorithms to estimate AAR parameters and described their update equations, including the least-mean-square (LMS) algorithm [38], different Kalman filtering (KF) algorithms [3], the recursive-least-squares (RLS) algorithm (special form of KF) [34], and the recursive AR (RAR) algorithm. Moreover, the authors used a relative error variance (REV) as a criterion for comparing these algorithms and drew a conclusion that given a model order of p=10, the lowest error rate was reached by KF. For the detailed parameters updating process, see [39].

Adaptive autoregressive parameter estimation is a suitable method for an EEG-based BCI. The AAR method can completely describe EEG signals, but depends on a number of parameters such as the model order and update coefficient. Their correct estimation is difficult and needs experience [37]. Moreover, note that the previous algorithms of AAR model parameters are mostly extracted from each single channel features separately. However, for multi-channel EEG, the single channel can not provide enough information. To resolve this problem, Wang et al. [62] proposed an improved multivariate adaptive autoregressive (MVAAR) models to extract features of multi-channel EEG signals.

D. Wavelet packet transform

Due to the non-stationary property of EEG signals, the wavelet packet transform (WPT) can better describe the signals than the fast Fourier transform (FFT) or autoregressive (AR) parameter model as it depicts the information in various time windows and frequency bands [64], [65]. As a result of the various time windows, the WPT is able to capture non-stationary information such as frequency variation but the FFT or AR cannot.

The WPT is implemented by means of a filter bank whose structure is shown in Fig. 2, where H is a high pass filter and L is a low pass filter, $S_0(0)$ denotes the initial signal space, $S_j(k)$ is the decomposed subspace, j is the decomposition level, and k is the index of the subspace. The WPT offers many alternative signal decompositions which need obey two rules: first, from the top view, there is no overlap among the decomposed signals. Second, the decomposed signals can be added together to recover the original signal length. For example, $S_1(0)$ and $S_1(1)$, or the shadowed subspaces. How to select the packet best basis which provides the most appropriate subbands for signal representation is the key of adaptive EEG signals feature extraction.

The non-adaptive method usually uses subband energies contained in the last decomposition level as features. The dimensionality of the feature vector would be quite high. To overcome this disadvantage, Yang et al. [66] introduced an adaptive method using wavelet packet best basis decomposition (WPBBD). There are two kinds of adaptive methods: subject-based adaptation which constructs a wavelet packet best basis fitted for each subject and non-subject-based adaptation which constructs a uniform wavelet packet best basis for all subjects.

The steps for the subject-based adaptation are as follows: Firstly, original EEG signals are decomposed to a given

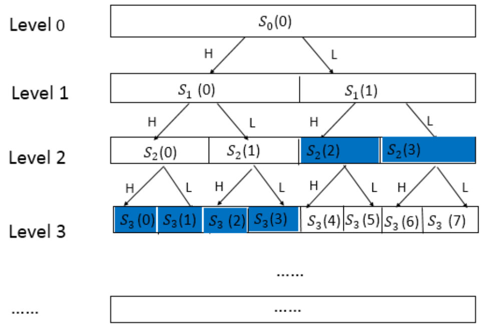

Fig. 2. The structure of WPT

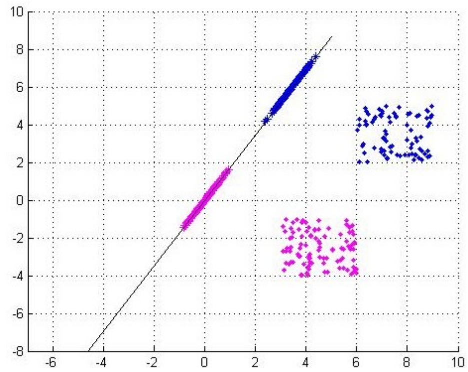

Fig. 3. The optimal projection derection

level by the WPT. Secondly, for each subject, the best basis algorithm is used to find the best-adapted basis. There are different algorithms for acquiring the best basis (e.g., [10] proposed an entropy-based algorithm to select the best basis based on the WPT). Finally, subband energies contained in the best basis are used as effective features.

Analogically, for the non-subject-based adaptation, the steps are almost the same as subject-based adaptation. The difference is that its best basis is selected using the training samples from all subjects rather than each subject. That is to say, for the non-subject-based adaptation, all subjects use the subband energies contained in the common basis as features which cannot be fit for each subject. Experiment results also showed that the subject-based adaptation can provide better performance compared with non-subject-based adaptation and no adaptation.

III. Adaptive Classifiers

The non-stationary nature of EEG signals means that a classification model built earlier using the previous data is not able to well reflect the changes that have already taken place to the signals. So adaptive updates to the classification model are needed. This section briefly summarizes some adaptive classifiers used to design a BCI system.

A. Adaptive linear discriminant analysis

The aim of linear discriminant analysis (LDA) is to use a hyperplane to separate the data representing the different classes [13], [23]. As Fig. 3 shows, for a two-class problem, LDA can find the optimal projection which can maximize the distance between the two classes means and minimize the interclass variances. The separating hyperplane is perpendicular to the projection direction. The hyperplane of two different classes can be written as

$$y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0, \tag{8}$$

$$\mathbf{w} = (\Sigma_1 + \Sigma_2)^{-1} \times (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \tag{9}$$

$$w_0 = -\frac{1}{2} \times \mathbf{w}^\top \times (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2), \tag{10}$$

where Σ_i is the covariance matrix for each class, $\boldsymbol{\mu}_i$ is the mean for each class $(i = 1, 2)$ and $\mathbf{x}$ is the feature vector. If $y(\mathbf{x}) > 0$, the input is classified as class 2, and otherwise, as class 1.

To classify the time-varying EEG signals better, the adaptive LDA classifier is needed. Kalman adaptive LDA (KALDA) is an adaptive version of LDA based on Kalman filtering in which the Kalman gain changes the update coefficient and varies the adaptation speed according to the property of the data [61]. KALDA is a supervised classifier. Analogously, Vidaurre et al. [58] also proposed a supervised adaptive LDA and described how to update the class means $\boldsymbol{\mu}$ and the global covariance matrix Σ to classify the time-varying data.

In a supervised scenario, all the trials before the current time are given a true label. This is relatively easier for the adaptation. However, class information is usually not available in real BCI tasks since different unrelated task factors or electrode montages might affect the signals. In this case, class information is not available for the adaptation of the systems. This motivated unsupervised adaptation without label information in practical BCI scenarios.

In [58], [60], the authors proposed three different unsupervised adaptive LDA. The first one just updates the means $\boldsymbol{\mu}_i(t)$ according to $\boldsymbol{\mu}_i(t) = (1-\eta)\boldsymbol{\mu}_i(t-1)+\eta\mathbf{x}(t)$ but does not change the weight vector $\mathbf{w}$. The second and third one update $\boldsymbol{\mu}$ and $\mathbf{w}$ in different ways. Further, the authors compared each of the unsupervised LDA versus no adaptation. In addition, Blumberg et al. [7] proposed an unsupervised adaptive LDA method for simulated online clustering of EEG patterns, in which the expectation maximization (EM) method was used to update the mean values as well as covariances of the class distributions continuously in time. Liu et al. [22] proposed different unsupervised adaptations of LDA involving incremental adaptation, GMM-based adaptation, and improved GMM (iGMM)-based adaptation to update the LDA classifier for a simulated online BCI scenario.

This technique is simple and has a very low computational requirement. The main drawback is its linearity, and thus it may provide poor results on complex nonlinear EEG data. This can be resolved by using a kernel function [5].

B. Adaptive support vector machines

A support vector machine (SVM) also discriminates classes by constructing a linear optimal hyperplane, which is induced from the maximum margin principle between two classes [8], [30].

Given the training set $S = \{(x_1, y_1), \cdots (x_m, y_m)\}$, points $x_i \in R^d$ and the corresponding labels $y_i \in \{-1, 1\}$, the linear hyperplane between two classes in a feature space mapped by $\varphi(\mathbf{x})$ can be written as

$$\mathbf{w}^\top \varphi(\mathbf{x}) + b = 0 \quad \mathbf{w} \in R^d, \quad b \in R. \tag{11}$$

The optimization problem for SVM classification is formulated as

$$\min_{\mathbf{w}, b, \boldsymbol{\varepsilon}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \varepsilon_i, \tag{12}$$

$$s.t. \quad y_i(\mathbf{w}^\top \varphi(\mathbf{x}) + b) \geq 1 - \varepsilon_i,$$
$$\varepsilon_i \geq 0, \quad i = 1 \cdots m,$$

where the constant C controls the balance between the margin and empirical loss. The large margin principle is minimizing $\frac{1}{2} \|\mathbf{w}\|^2$ with $2/ \|\mathbf{w}\|$ being the margin between the two separating hyperplanes $\mathbf{w}^\top \varphi(\mathbf{x}) + b = 1$ and $\mathbf{w}^\top \varphi(\mathbf{x}) + b = -1$. The SVM classifier would be

$$c = sign(\mathbf{w}^\top \varphi(\mathbf{x}) + b). \tag{13}$$

If $c > 0$, the input is classified as class 2, and otherwise, as class 1. Such an SVM enables classification using linear decision boundaries, and is known as a linear SVM [23]. In addition, it can also resolve nonlinear decision boundaries by using a kernel function that recovers nonlinear boundaries between classes. The kernel generally used in BCI research is the radial basis function (RBF) kernel.

Because of the high variability of recorded brain signals in a BCI system, the SVM classification accuracy will degrade with time. In order to maintain the classification accuracy and overall performance of the system, online classification and adaptive schemes which modify BCI classification parameters in real time systems are particularly important [1]. Oskoei et al. [29] updated Training Data Set (TDS) by inserting fresh samples into TDS using supervised or unsupervised methods. The proposed adaptive schemes based on the online SVM can effectively improve the classification performance on real BCI data.

To reduce the time-consuming training sessions, there are also semi-supervised SVM learning algorithms by using large amounts of unlabeled data. For example, Bennett Demiriz [6] proposed a semi-supervised SVM which can be implemented using mixed integer programming. But the computation of mixed integer programming will become more and more complex as the number of unlabeled data increases. Therefore, Qin and Li [35] introduced a batch-mode incremental training method which divides the original unlabeled data set into several subsets for the semi-supervised SVM. The parameters of the SVM will adjust gradually with the addition of these subsets, which can reduce the computational complexity. Additionally, Li and Guan [21] proposed a new semi-supervised SVM in which the feature extraction and classification are jointly performed with iterations.

C. Adaptive Bayesian classifiers

The aim of a Bayesian classifier is to assign a feature vector to the class with the highest probability [13]. The Bayes rule is used to compute the posteriori probability that the feature vector belongs to a given class.

For the Bayesian classifiers with Gaussian mixture modes (GMM), the posteriori probability can be given as

$$\begin{aligned} p(c_k|x) &= \frac{p(c_k)p(x|c_k)}{p(x)} \\ &= \frac{p(c_k) \sum_{i=1}^{N_k} a_k^i G(x|\gamma_k^i, \sigma_k^i)}{\sum_{j=1}^{K} p(c_j) \sum_{j=1}^{N_k} a_k^j G(x|\gamma_k^j, \sigma_k^j)}. \end{aligned} \tag{14}$$

Here $p(x|c_k) = \sum_{i=1}^{N_k} a_k^i G(x|\gamma_k^i, \sigma_k^i)$, $s.t.$, $\sum_{i=1}^{N_k} a_k^i = 1$, which is assumed to be the weighted combination of N_k Gaussian probability density functions. Then we classify an unknown pattern x to the class with the highest posterior probability. For the current GMM formulated in Eq. (14), the parameters are unchanged. However, for on-line application of BCI systems, the recorded data are varying dynamically. This necessitates adaptive Bayesian classifiers which can update parameters using new added samples and then classify forthcoming samples.

The adaptive update of GMM-based Bayesian statistical classifiers has been recently studied in several papers [24]–[26] which are among the earliest ones discussing the problem of online EEG signals classification based on Bayesian classifiers and stochastic gradient methods (SGMs). Moreover, Sun et al. [45] used the stochastic approximation method (SAM) instead of SGM for learning adaptive Bayesian classifiers. SAM is a batch processing algorithm, which adopts a pool of samples to calculate gradients and update parameters. With SAM, the parameters of the mean values and covariance matrices of the Bayesian classifiers can be simultaneously updated in a batch mode. Both SGM and SAM are applicable to adaptively update parameters of Bayesian classifiers. The difference is the number of samples used to calculate gradients. SGM only adopts one sample to update subsequent parameters while SAM integrates more than one sample to compute gradients. Additionally, to accelerate convergence, the decorrelated gradient is adopted for updating the parameters of the classifier adaptively [47].

There are other dynamic Bayesian models for non-stationary adaptive classification. For example, Yoon et al. [67] described a Bayesian model based on a sequential Monto Carlo filter, which can sequentially predict the decision boundary in BCI time series. Sykacek et al. [52] proposed a method which used the variational Kalman filtering as an inference technique for adaptive Bayesian classifiers.

D. Adaptive neural networks

Neural networks (NN) are mostly used classifiers in BCIs (see, e.g., [2], [16]), because they can provide a well-established framework for pattern-recognition problems. Generally speaking, neural networks consist of an input layer, the hidden layer (which can have one or several layers) and an output layer. Multi-layer perceptron (MLP) are very popular NN used in classification problems. However, MLP are universal approximations, which make these classifiers sensitive to overtraining especially for the non-stationary data as EEG [23]. Therefore, to classify the time-varying signals, classifiers which can process temporal data are necessary. Here, we briefly present some adaptive neural networks applied to BCIs.

Probabilistic neural networks (PNN), introduced by Specht [40], [41] in the early 1990s, are derived from Bayes decision networks and kernel based estimators of probability density function (PDF). Briefly speaking, if the PDF of each class is known, then an unknown x will be assigned to the class j as

$$p_j f_j(x) > p_i f_i(x), \quad all \quad j \neq i \tag{15}$$

where p_k, $f_k(x)$ are separately the prior probability and the PDF of the class k. PNN estimate the probability density function using the nonparametric, or parametric methods for each class based on the training samples.

In classical PNN, the estimating methods can be applied where probability distributions do not change with time. However, for designing an online BCI system, an important issue is that the brain signals are characterized by significant subject-to-subject variations and time-varying probability distributions. Due to these variabilities, adaptive probabilistic neural networks (APNN) are generated, which can work in a non-stationary environment for the classification of EEG signals. Rutkowski [36] proposed a recursive version of the discriminate function and formulated the problem of pattern classification in a time-varying environment as a prediction problem due to the fact that on the basis of a learning sequence of length N, a pattern in the moment $N+k, k \geq 1$, should be classified (for details, see [15], [36]).

Besides APNN, there are other adaptive neural networks which can be usually used to classify the EEG signals in the online experiments, such as adaptive logic networks (ALN) [18], the finite impulse response neural networks (FIRNN) [14], Gamma dynamic neural networks (GDNN) [4]. They are not described in detail here.

E. Dynamic combination of classifiers

The main advantage of ensemble techniques is that the effect of the combination of classifiers is very likely to outperform a single one. Ensemble learning can effectively improve weak classifiers and thereinto bagging, boosting and random subspace are three powerful and popular representatives [51]. It is widely acknowledged that an effective ensemble learning system should consist of individuals that are not only accurate, but are diverse as well, that is, a good balance should hold between diversity and individual performance [49], [50].

Sun [44] proposed a new combination rule named as W-LWCA (weighting by local within-class accuracies) to assign weights to individual classifiers in a multiple classifier system by exploiting local within-class accuracies. Distance metric learning is adopted to determine the within-class nearest neighbors. Compared with W-LA (weighting by local accuracies), W-LWCA considers the category relationship (within-class or between-class) between the test example and its neighbors and tends to make up for the blind spot of W-LA. Experimental results also showed that W-LWCA can provide better performance than majority voting and W-LA. Moreover, in [55], Tu and Sun employed a two-lever ensemble strategy to dynamically and locally combine the outcomes of a robust classifier and an adaptive classifier to reach a single decision output. Owing to the distribution differences between training and testing data, the weights in the final model are sensitive to target examples. To alleviate this problem, Tu and Sun [56] further proposed a method to dynamically assign weights to different test examples by making use of additional classifiers called model-friendly classifiers. Through this, we can judge which base models predict well on a specific test example and simultaneously give the most suitable weights to different examples.

IV. Open problems

In this section we present several important open problems which can be very useful for further progress in the area of BCI research.

A. Transfer learning for BCIs

Improving classification performance of EEG-based BCI systems is an urgent need today [33]. For new subjects, a long training phase in which they concentrate on prescribed mental tasks is needed to construct subject-specific feature extractors and classifiers. However, the training session is a very boring and time-consuming process especially for many disabled users due to their cognitive impairments and concentration problems. Therefore, reducing the training time is very necessary. Transfer learning is an effective approach to reduce the training time.

The essence of transfer learning is the ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks [31]. In an EEG-based BCI system, we can utilize the samples collected from other subjects to aid the subject whose brain signals would be classified in the test session. Through this, transfer learning can help the subjects reduce the training time effectively. However, how to properly transfer the source data to target data is the key point. There have been some related work, such as the session transfer strategy [19], subject transfer strategy [55] and multi-source part-based transfer learning [46]. Therefore, a natural question to ask is whether we can find more strategies to effectively realize the transfer learning of a BCI system.

B. Semi-supervised learning for BCIs

As EEG signals are time-varying, the variance of signals between the training and test sessions may be large. To alleviate this problem, an approach called semi-supervised learning method is appropriate. Semi-supervised learning method can effectively alleviate the time-variance between the training and test sessions and combine the limited labeled data together with a large number of unlabeled data coming from the test sessions for effective function learning. Through semi-supervised learning, the user training time is reduced and the classifiers can be adaptive to test samples. For example, Tu and Sun have shown different semi-supervised feature extraction methods to classify the non-stationary EEG signals [53], [54]. It would be interesting to explore more semi-supervised learning methods tailored for BCIs.

C. Active learning for BCIs

Active learning is an effective learning method which can actively query the user for labels of the most informative examples. Due to this, the number of examples needed to learn can often be much lower than the corresponding supervised learning case.

Active learning to BCIs can actually be seen as experimental design. It will actively collect signals which are the most valuable under a given criterion to label. For example, in a two-class trail, if the classification accuracy of one class is lower than the other class, then in subsequent process we will actively select more signals of this class to improve its accuracy. Through this, active learning can effectively reduce the user training time and improve the BCI performance. However, as far as we know, there is not much such work reported for BCIs.

V. Conclusion

This paper has surveyed representative adaptive feature extraction and classification methods used to design a BCI. As EEG signals change over time, the adaptation of feature extractors and classifiers is a very important and necessary issue in EEG-based BCI research. In this paper, we summarize four main adaptive feature extraction methods, including adaptive common spatial patterns, semi-supervised feature extraction, adaptive autoregressive parameters, and wavelet packet transform. Empirically the performance of adaptive feature extraction is better than non-adaptive because the adaptive ones can catch hold of the variability of EEG signals. Furthermore, we review some representative adaptive classification methods: adaptive linear discriminant analysis, adaptive support vector machines, adaptive Bayesian classifiers, adaptive neural networks and dynamic combination of classifiers. Several open problems have also been provided, which we think are important for the development of future BCIs.

In a word, adaptive learning can effectively boost up the performance of existing BCIs. Moreover, exploring other machine learning methods which are suitable for the characteristics of EEG signals will also be an interesting direction for BCIs in the future.

References

[1] O. AlZoubi, R. A. Calvo, and R. H. Stevens, "Classification of EEG for affect recognition: an adaptive approach," *Lecture Notes in Computer Science*, vol. 5866, pp. 52-61, 2009.

[2] C. W. Anderson, and Z. Sijercic, "Classification of EEG signals from four subjects during five mental tasks," *Proceedings of the Conference on Engineering Applications in Neural Networks*, pp. 407-414, 1996.

[3] M. Arnold, X. H. R. Milner, H. Witte, R. Bauer, and C. Braun, "Adaptive AR modeling of nonstationary time series by means of Kalman filtering," *IEEE Transactions on Biomedical Engineering*, vol. 45, pp. 553-562, 1998.

[4] A. B. Barreto, A. M. Taberner, and L. M. Vicente, "Classification of spatio-temporal EEG readiness potentials towards the development of a brain-computer interface," *Proceedings of the IEEE SouthEastcon Conference*, pp. 99-102, 1996.

[5] G. Baudat, and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural computation*, vol. 12, pp. 2385-2404, 2000.

[6] K. P. Bennett, and A. Demiriz, "Semi-supervised support vector machines," *Proceedings of Neural Information Processing Systems*, pp. 368-374, 1998.

[7] J. Blumberg, J. Rickert, S. Waldert, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Adaptive classification for brain computer interfaces," *Proceedings of the 29th annual international conference of the IEEE EMBS*, pp. 2536-2539, 2007.

[8] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[9] C. Chen, W. Song, J. Zhang, Z. Hu, and H. Xu, "An Adaptive Feature Extraction Method for Motor-Imagery BCI Systems," *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 275-279, 2010.

[10] R. Coifman, and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, pp. 713-718, 1992.

[11] E. A. Curran, and M. J. Stokes, "Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems," *Brain and Cognition*, vol. 51, pp. 326-336, 2003.

[12] G. Dornhege, B. Blankertz, and G. Curio, "Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials," *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering*, pp. 595-598, 2003.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, second edition, John Wiley & Sons, 2000.

[14] E. Haselsteiner, and G. Pfurtscheller, "Using time-dependent neural networks for EEG classification," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, pp. 457-463, 2000.

[15] M. K. Hazrati, and A. Erfanian, "An online EEG-based brainCcomputer interface for controlling hand grasp using an adaptive probabilistic neural network," *Medical Engineering & Physics*, vol. 32, pp. 730-739, 2010.

[16] A. Hiraiwa, K. Shimohara, and Y. Tokunaga, "EEG topography recognition by neural networks," *Engineering in Medicine and Biology Magazine*, vol. 9, pp. 39-42, 1990.

[17] B. H. Jansen, J. R. Bourne, and J. W. Ward, "Autoregressive estimation of short segment spectra for computerized EEG analysis," *IEEE Transactions on Biomedical Engineering*, vol. 9, pp. 630-638, 1981.

[18] A. Kostov, and M. Polak, "Parallel man-machine training in development of EEG-based cursor control," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, pp. 203-205, 2000.

[19] M. Krauledat, M. Tangermann, B. Blankertz, and K. R. Müller, "Towards zero training for brain-computer interfacing," *PLoS One*, vol.3, pp. 2967, 2008.

[20] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *Signal Processing Letters*, vol. 17, pp. 4-7, 2010.

[21] Y. Li, and C. Guan, "Joint feature re-extraction and classification using an iterative semi-supervised support vector machine algorithm," *Machine Learning*, vol. 71, pp. 33-53, 2008.

[22] G. Liu, G.Huang, J. Meng, D. Zhang, and X. Zhu, "Improved GMM with parameter initialization for unsupervised adaptation of brain-computer interface," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 26, pp. 681-691, 2010.

[23] F. Lotte, M. Congedo, A. Lcuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, pp. 1-13, 2007.

[24] J. Millan, "On the need for on-line learning in brain-computer interfaces," *Proceedings of the International Joint Conference on Neural Networks*, pp. 2877-2882, 2004.

[25] J. R. Millan, F. Renkens, J. Mouriño, and W. Gerstner, "Noninvasive brain-actuated control of a mobile robot by human EEG," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1026-1033, 2004.

[26] J. R. Millan, F. Renkens, J. Mouriño, and W. Gerstner, "Brain-actuated interaction", *Artificial Intelligence*, vol. 159, pp. 241-259, 2004.

[27] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clinical Neurophysiology*, vol. 110, pp. 787-798, 1999.

[28] M. A. L. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, pp. 403-407, 2001.

[29] M. A. Oskoei, J. Q. Gan, and H. Hu, "Adaptive schemes applied to online SVM for BCI data classification," *Proceedings of 31st Annual International Conference of the IEEE EMBS*, pp. 2600-2603, 2009.

[30] M. A. Oskoei, and H. Hu, "Support vector machine-based classification scheme for myoelectric control applied to upper limb," *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 1956-1965, 2008.

[31] S. J. Pan, and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345-1359, 2010.

[32] L. Parra, C. Alvino, A. C. Tang, B. A. Pearlmutter, N. Yeung, A. Osman, and P. Sajda, "Linear spatial integration for single trial detection in encephalography," *NeuroImage*, vol. 7, pp. 223-230, 2002.

[33] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, "Current trends in Graz brain-computer interface (BCI) research," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, pp. 216-219, 2000.

[34] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger, "Sparability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *IEEE Transactions on Rehabilitation Engineering*, vol. 6, pp. 316-325, 1998.

[35] J. Qin, and Y. Li, "An Improved Semi-Supervised Support Vector Machine Based Translation Algorithm for BCI Systems," *Proceedings of 18th International Conference on Pattern Recognition*, pp. 1240-1243, 2006.

[36] L. Rutkowski, "Adaptive probabilistic neural networks for pattern classification in time-varying environment," *IEEE Transactions on Neural Networks*, vol. 15, pp. 811-827, 2004.

[37] A. Schlögl, D. Flotzinger, and G. Pfurtscheller, "Adaptive Autoregressive Modeling used for Single-trial EEG Classification-Verwendung eines Adaptiven Autoregressiven Modells für die Klassifikation von Einzeltrial-EEG-Daten," *Biomedizinische Technik*, vol. 42, pp. 162-167, 1997.

[38] A. Schlögl, K. Lugger, and G. Pfurtscheller, "Using adaptive autoregressive parameters for a brain-computer-interface experiment," *Proceedings of the 19th Annual International Conference on Medicine and Biology Society*, pp. 1533-1535, 1997.

[39] A. Schlogl, S. J. Roberts, and G. Pfurtscheller, " A criterion for adaptive autoregressive models," *Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society*, pp. 1581-1582, 2000.

[40] D. F. Specht, "Probabilistic neural networks," *Neural networks*, vol. 3, pp. 109-118, 1990.

[41] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, pp. 568-576, 1991.

[42] S. Sun, "The extreme energy ratio criterion for EEG feature extraction," *Lecture Notes in Computer Science*, vol. 5164, pp. 919-928, 2008.

[43] S. Sun, "Extreme energy difference for feature extraction of EEG signals," *Expert Systems with Applications*, vol. 37, pp. 4350-4357, 2010.

[44] S. Sun, "Local within-class accuracies for weighting individual outputs in multiple classifier systems," *Pattern Recognition Letters*, vol. 31, pp. 119-124, 2010.

[45] S. Sun, Y. Lu, Y. Chen, "The stochastic approximation method for adaptive Bayesian classifiers: Towards online brain-computer interfaces," *Neural Computing and Applications*, vol. 20, pp. 31-40, 2011.

[46] S. Sun, Z. Xu, and M. Yang, "Transfer learning with part-based ensembles," *Lecture Notes in Computer Science*, vol. 7872, pp. 271-282, 2013.

[47] S. Sun, and C. Zhang, "Learning on-line classification via decorrelated LMS algorithm: application to brain-computer interfaces," *Proceedings of 8th International Conference on Discovery Science*, pp. 215-226, 2005.

[48] S. Sun, and C. Zhang, "Adaptive feature extraction for EEG signal classification," *Medical and Biological Engineering and Computing*, vol. 44, pp. 931-935, 2006.

[49] S. Sun, and C. Zhang, "Subspace ensembles for classification," *Physica A: Statistical Mechanics and Its Applications*, vol. 385, pp. 199-207, 2007.

[50] S. Sun, and Q. Zhang, "Multiple-view multiple-learner semi-supervised learning," *Neural Processing Letters*, vol. 34, pp. 229-240, 2011.

[51] S. Sun, C. Zhang, and D. Zhang, "An experimental evaluation of ensemble methods for EEG signal classification," *Pattern Recognition Letters*, vol. 28, pp. 2157-2163, 2007.

[52] P. Sykacek, S. J. Roberts, and M. Stokes, "Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 719-727, 2004.

[53] W. Tu, and S. Sun, "Semi-supervised feature extraction with local temporal regularization for EEG Classification," *Proceedings of the International Joint Conference on Neural Networks*, pp. 75-80, 2011.

[54] W. Tu, and S. Sun, "Semi-supervised feature extraction for EEG classification," *Pattern Analysis and Applications*, vol. 16, pp. 213-222, 2012.

[55] W. Tu, and S. Sun, "A subject transfer framework for EEG classification," *Neurocomputing*, vol. 82, pp. 109-116, 2012.

[56] W. Tu, S. Sun, "Dynamical ensemble learning with model friendly classifiers for domain adaptation", *Proceedings of the 21st International Conference on Pattern Recognition*, pp. 1181-1184, 2012.

[57] T. M. Vaughan, W. J. Heetderks, L. J. Trejo, W. Z. Rymer, M. Weinrich, M. M. Moore, A. Kübler, B. H. Dobkin, N. Birbaumer, E. Donchin, E. W. Wolpaw, and J. R. Wolpaw, "Guest editorial brain-computer interface technology: brain-computer interface technology: a review of the second international meeting," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 94-109, 2003.

[58] C. Vidaurre, M. Kawanabe, P. von Bunau, B. Blankertz, and K. R. Muller, "Toward unsupervised adaptation of LDA for brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 587-597, 2011.

[59] C. Vidaurre, R. Scherer, R. Cabeza, A. Schlögl, A., and G. Pfurtscheller, "Study of discriminant analysis applied to motor imagery bipolar data," *Medical & Biological Engineering & Computing*, vol. 45, pp. 61-68, 2007.

[60] C. Vidaurre, A. Schlögl, B. Blankertz, M. Kawanabe, and K. R. Müller, "Unsupervised adaptation of the LDA classifier for brain-computer interfaces," *Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course*, pp. 122-127, 2008.

[61] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 550-556, 2007.

[62] J. Wang, G. Xu, L. Wang, and H. Zhang, "Feature extraction of brain-computer interface based on improved multivariate adaptive autoregressive models", *Proceedings of the IEEE International Conference on Biomedical Engineering and Informatics*, pp. 895-898), 2010.

[63] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767-791, 2002.

[64] Y. Wu, and R. Du, "Feature extraction and assessment using wavelet packets for monitoring of machining processes," *Mechanical Systems and Signal Processing*, vol. 10, pp. 29-53, 1996.

[65] Q. Xu, H. Zhou, Y. Wang, and J. Huang, "Fuzzy support vector machine for classification of EEG signals using wavelet-based features," *Medical Engineering & Physics*, vol.37, pp. 858-865, 2009.

[66] B. Yang, G. Yan, R. Yan, and T. Wu, "Adaptive subject-based feature extraction in brain-computer interfaces using wavelet packet best basis decomposition," *Medical Engineering & Physics*, vol. 29, pp. 48-53, 2007.

[67] J. W. Yoon, S. J. Roberts, M. Dyson, and J. Q. Gan, "Adaptive classification for brain computer interface systems using sequential Monte Carlo sampling," *Neural Networks*, vol. 22, pp. 1286-1294, 2009.

[68] Q. Zhao, L. Zhang, A. Cichocki, and J. Li, "Incremental common spatial pattern algorithm for BCI," *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 2656-2659, 2008.