Probabilistic Methods in Multi-Class Brain-Computer Interface

Ping Yang, Xu Lei, Tie-Jun Liu, Peng Xu, and De-Zhong Yao

Abstract—Two probabilistic methods are extended to research multi-class motor imagery of brain-computer interface (BCI): support vector machine (SVM) with posteriori probability (PSVM) and Bayesian linear discriminant analysis with probabilistic output (PBLDA). A comparative evaluation of these two methods is conducted. The results shows that: 1) probabilistic information can improve the performance of BCI for subjects with high kappa coefficient, and 2) PSVM usually results in a stable kappa coefficient whereas PBLDA is more efficient in estimating the model parameters.

Index Terms—Bayesian linear discriminant analysis, brain-computer interface, kappa coefficient, support vector machine.

1. Introduction

Brain-computer interface (BCI) is a new communication channel that directly translates brain activities into control commands or messages for peripheral equipments. Such a device may give disabled people direct control over a computer application or a neuro-prosthesis[1][2]. In laboratory studies and practical applications, accuracy and information transfer rates (ITR) are the two important factors for measuring the performance of BCI, especially the ITR which is directly relevant to the real-time communication between human and the environment. At present, however, BCI applicability is severely limited by its low ITR and low accuracy. A feasible way to increase the ITR of a BCI system is to change the usual binary parameters.

To have a real high ITR, a multi-class BCI must have a considerably high accuracy. However, when the number of brain patterns increases, great difficulties in both signal processing (feature extraction) and machine learning (pattern classification) stand out. In current practice, some classifiers, such as Fisher discriminant analysis (FDA), multilayer perception, nearest neighbor classifiers, and some combined algorithms[6] have been introduced for multi-class BCIs. However, most of them only produce uncalibrated values and unconfident results.

In recent years, some probabilistic methods, such as Gaussian processes[7] and Bayesian learning[8] have been introduced to improve the robustness and generalization of BCIs. Furthermore, a probabilistic method may provide confidence level of the output that is meaningful for further post-processing such as classifiers combination[9][10]. Specifically, some studies showed that linear discriminant analysis (LDA), a very simple method[11], may give a good result. Motivated by the success of LDA in BCI, Hoffman et al. developed an evidence framework based Bayesian LDA (BLDA) and certified its usefulness in a P300-based BCI[10]. In addition, support vector machine (SVM)[12][13] was also certified to give high quality results in BCI[9][10], but standard SVM does not provide probabilistic output, it only produces uncalibrated value. In 1999, SVM with posteriori probabilistic output (PSVM) was first introduced[14], which can produce a useful calibrated posterior probability to enable post-processing and it has been successfully applied in face recognition. However, these algorithms have not been applied to solve multi-class problem in BCI.

In this work, firstly, one-versus-one common spatial patterns (CSP) method[5] is suggested to estimate the different activities of four motor imagery tasks including imaging movements of left hand, right hand, foot, and tongue. Then, PSVM and PBLDA are used to classify the four motor imagery tasks with probabilistic outputs.

2. Datasets and Feature Extraction

Two datasets were used in this study, dataset 1 was provided by the BCI-Lab in BCI Competition III 2005 (Data set III a), and dataset 2 was from our own experiments. The main task was to perform imaging left hand, right hand, foot or tongue movements according to a cue. The experiment design for dataset 2 was similar to that of dataset 1. The experiment consisted of several runs (6 runs or 9 runs) and each run contained 40 trials about 9 minutes. For each subject, all trails were split into a training set (120 trials for subject L1 and P1, and 180 trials for subject K1, K6, and Y1) and unlabeled test set (the some size of training sets), and more detail can be found in [15]. In offline analysis, the data was down-sampled to 100 Hz and re-referenced to common average reference.
2.1 Neurophysiological Facts and Algorithm Procedure

Event related synchronization and event related desynchronization (ERS/ERD) phenomenon\cite{6} can be observed over sensorimotor cortex during motor imagery tasks, and the experimental observations show that particular mental tasks have related effects on the spatial distribution of electroencephalograph (EEG) at \( \mu \) (8 Hz to 13 Hz) and \( \beta \) (18 Hz to 26 Hz) rhythms. For further classification, it is necessary to extract features from these EEG signals based on this phenomenon. In this work, in order to solve the multi-class motor imagery problems with binary classifiers, one-versus-one method\cite{5} is used to change the multi-class problems to a few binary problems. Then, binary classifiers are adopted for the different pairs of classes and the final result is obtained by majoring voting. The procedure is shown in Fig. 1.

Based on ERD/ERS, CSP has proved to be a highly effective method to extract feature vectors from two motor imaginary tasks multi-channel EEG data\cite{17,18}. In this paper, one-versus-one CSP is employed to estimate the different activities of the four motor imagery tasks. After CSP feature extraction, the features of single-trials are chosen as the log-variance of the CSP projected signals.

3. Classification Algorithms

In this section, two classification algorithms with probabilistic outputs are introduced and the strategies of major voting for multi-class problem are illuminated.

3.1 SVM with Posteriori Probability (PSVM)

SVM has been adopted for classification in various domains\cite{6,12,19}. In general, the decision function of SVM does not have meaning of probability and it only produces an uncalibrated value. Let

\[
S_{XY} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}
\]

be the training sets composed of feature vector \( x \in \mathbb{R}^p \) and the corresponding binary states \( y \in \{1, 2\} \). Then, the non-linear discriminate function is

\[
f(x) = \sum_{i=1}^{N} y_i a_i K(x, x_i) + b
\]

where \( N \) is the number of the support vectors, \( a_i \) is the positive Lagrangian multiplier, \( K(x, x_i) \) is kernel function, and \( b \) is the bias term. In our study, \( K \) is a Gaussian kernel. The binary classification rule \( d: X \rightarrow Y = \{1, 2\} \) is defined as

\[
d(x) = \begin{cases} 
1 & \text{for } f(x) > 0 \\
2 & \text{for } f(x) < 0
\end{cases}
\]

From this classification rule, SVM only produces an arbitrary decision value without meaning of probability. However, the probabilistic output of the classifier is useful in post-processing, for example, combination of a few classifiers to obtain the final outputs for a small part of test samples. Recently, some probabilistic models were proposed to translate the uncalibrated value of discriminant function to a posteriori probabilistic distribution\cite{14}. Specifically, when given the value of the discriminant function \( f(x) \) for a two class problem, the main task of the translation is to estimate parameters of a posteriori distribution \( P_{\gamma \alpha}(y_0 \mid f(x), \theta) \). The posterior probabilistic distribution is modeled by a sigmoid function.

\[
P_{\gamma \alpha} (1 \mid f(x), \theta) = \frac{1}{1 + \exp(a_1 f(x) + a_2)}.
\]

This function is determined by parameters \( \theta = [a_1, a_2]^T \). The log-likelihood function is defined as

\[
L(\theta \mid S_{XY}) = \sum_{i=1}^{M} \ln P_{\gamma \alpha}(y_i \mid f(x_i), \theta).
\]

The parameters \( \theta = [a_1, a_2]^T \) can be estimated by the maximum-likelihood method:

\[
\theta = \arg \max_{\theta} L(\theta \mid S_{XY}) = \arg \max_{\theta} \sum_{i=1}^{M} \log P_{\gamma \alpha}(y_i \mid f(x_i), \theta).
\]

After the parameters \( a_1 \) and \( a_2 \) estimated, we can translate the uncalibrated value of discriminant function \( f(x) \) to a posteriori probabilistic output.

3.2 BLDA with Probabilistic Output (PBLDA)

PBLDA is based on the evidence framework for Bayesian regression and has been certified very useful to a P300-based BCI\cite{19}. It is assumed that the target \( y \) and feature vectors \( x \) are linearly related with white Gaussian noise \( n \):

\[
y = \alpha^T x + n.
\]

From (6) we can obtain the likelihood function for the weights \( \alpha \):

\[
p(S_{XY} \mid \beta, \alpha) = \left( \frac{\beta}{2 \pi} \right)^{\frac{M}{2}} \exp \left( -\frac{\beta}{2} \| \hat{X}^T \alpha - Y \| \right)
\]

where \( Y \) denotes a vector contained all the train targets, \( \hat{X} \) denotes the matrix obtained from the horizontal stacking of the training feature vectors, \( \beta \) denotes the inverse variance of the noise, and \( M \) denotes the number of training samples. In Bayesian analysis, the prior distribution of the weights \( \alpha \) is given by
\[ p(w | \alpha) = \left( \frac{\alpha}{2\pi} \right)^{n/2} \left( \frac{e}{2\pi} \right)^{1/2} \exp \left( -\frac{1}{2} \frac{w^T \Gamma(\alpha) w}{\alpha} \right) \]  

where \( \Gamma(\alpha) \) is a \( n+1 \) dimensional, diagonal matrix and the prior distribution for the weights is an zero mean Gaussian distribution with variance \( 1/\alpha \), \( \alpha \) is a very small bias values for overcoming the danger of overfitting, \( n \) is the dimension of feature vectors. From the likelihood function and prior distribution, the posterior distribution can be computed by using Bayes rule:

\[ p(w | \beta, x, S_{xy}) = \frac{p(S_{xy} | \beta, w) p(w | \alpha)}{\int p(S_{xy} | \beta, w) p(w | \alpha) dw} \]  

As the prior distribution and likelihood distribution are Gaussian, the posterior distribution is also Gaussian and the distribution can be determined by the mean and covariance:

\[ m = \beta (\beta \hat{x} \hat{x}^T + \Gamma(\alpha))^{-1} XY \]  
\[ C = \left( \beta \hat{x} \hat{x}^T + \Gamma(\alpha) \right)^{-1}. \]

From the posterior distribution and likelihood function, the predictive distribution can be obtained by inputting a new test feature vector \( \hat{x} \):

\[ p(\hat{y} | \beta, \alpha, \hat{x}, S_{xy}) = \int p(\hat{y} | \beta, \hat{x}, w)p(w | \beta, \alpha, S_{xy}) dw \]  

where \( p(\hat{y} | \beta, \hat{x}, w) \) is one dimension. So the predictive distribution can be characterized by its mean \( u \) and variance \( \sigma^2 \):

\[ u = m^T \hat{x} \]  
\[ \sigma^2 = 1/\beta + \hat{x}^T C \hat{x}. \]

Then we can calculate the probability that feature \( \hat{x} \) belongs to class label \( y=1 \) (similar to class label \( y=-1 \)):

\[ p(\hat{y} \geq 0 | \beta, \alpha, \hat{x}, S_{xy}) = \Phi \left( \frac{u}{\sigma} \right) \]

where \( \Phi \) denotes the standard normal cumulative distribution function.

Obviously, the probabilistic output depends on the mean \( u \) and variance \( \sigma^2 \) calculated from (13) and (14). Therefore, both the posterior and predictive distributions of \( w \) depend on the parameters \( \alpha \) and \( \beta \). In PBLDA, the problem of parameter selection can efficiently be solved by maximum likelihood methods:

\[ \alpha = \frac{D}{\sum_{i=1}^{N} C_i + m_i^2} \]  
\[ \beta = \frac{M}{\text{tr}(\hat{x} \hat{x}^T C) + \| \hat{x}^T m - Y \|^2} \]

where \( D \) denotes the dimension of feature vectors and \( M \) denotes the number of training number. The partial derivatives for \( \alpha \) and \( \beta \) depend on the posterior mean \( m \). Thus, an iterative scheme is used: 1) Compute \( \alpha \) and \( \beta \) for an initial value of \( \alpha \) and \( \beta \), 2) the hyperparameters are updated according to (16) and (17), 3) Equations (10), (11), (13), and (14) are used to obtain the predictive distribution, 4) the probability that feature \( \hat{x} \) belongs to class 1 by standard normal cumulative distribution function is calculated. For the EEG datasets we tested, the converged solution of hyperparameter optimization can be achieved typically after eight to fifty iterations.

### 3.3 Majority Voting

In previous, we introduced two probabilistic methods for binary classification problems. The one versus one decomposition transforms the \( N \)-class problems into \( N(N-1)/2 \) binary classification problems. Now the final decisions can be made by the voting of all the binary classifiers. Let \( p_i(i | j; x) \) denote the probability of feature \( x \) belonging to class \( i \) when the classification is made between class \( i \) and class \( j \). Then, we can use voting method to obtain the probability of feature \( x \) belonging to class \( i \) with all binary classifiers (named as \( p(i | x) \)).

\[ p(i | x) = \frac{\sum_{j=1}^{N} p_i(i | j; x)}{\sum_{k=1}^{N} \sum_{j=1}^{N} p_k(k | j; x)} \]

The majority votes based on multi-class classifier assign the input \( x \) into such class \( y \in \{1,2,3,4\} \) having the majority of votes:

\[ i = \arg \max_i p(i | x), \quad i \in \{1,2,3,4\}. \]

### 3.4 Performance Measure—the Kappa Coefficient

The kappa coefficient\(^{[20]}\) is an evaluation criterion for unifying different number classification problems. In the \( N \) class problem, the proper performance measure of the classifier is described by its confusion matrix\(^{[20]}\).

If the frequency of all \( N \) classes occur equally, kappa coefficient \( k \) and accuracy \( acc \) can be related to

\[ k = \frac{acc - 1/2}{1/2}. \]

In our research, there were an equal number of trails of each class in every session. So, we used this simplified equation to obtain the performance of BCI.

### 4. Results

Table 1 shows the results of three different algorithms for five subjects in four motor imaginary BCI experiments.
All algorithms were tested in 20-fold cross-validation loops. All trails were split into training sets (180 trails for subject K1 and Y1; and 120 trails for K6, L1, and PI) and unlabeled test sets (half of the rest trails for the expansion samples procedure and half for the final test process). As a model selection procedure, the values of the SVM parameters (the regularization constant and Gaussian kernel argument) were estimated by a 2×5-folds cross validated with the whole training set for different subjects. The STPRTTool software was used to execute the one-versus-one SVM procedure.

As we can see in Table 1, PSVM and PBLDA (probabilistic methods) classify trails better than standard SVM (arbitrary algorithm). It is illuminated that probabilities or posterior probabilities is required when a classifier is making a small part of an overall decision and the classification outputs should be combined for the overall decisions. So, this result also confirms that probability is necessary for the classification of some uncertain samples and probabilistic information can improve the performance of classification.

The result of PSVM is more stable than the result of PBLDA for subjects with low kappa coefficient (such as P1 and Y1). The reason is that some experiential assumptions were used in the model-estimated of PBLDA, so the obtained probability may be inaccurate in some situation. But this problem does not exist in posteriori probability methods. So posteriori probability methods can obtain more stable results and probability methods are more direct for representing uncertainty.

It is clear that when we applied PBLDA algorithm, the average kappa coefficients are nearly equivalent to the results of PSVM. However, the PBLDA algorithm can obtain probabilistic output directly and quickly, so PBLDA algorithm is more suited for practice application than PSVM algorithm.

5. Conclusions

In our work, we applied the PSVM and PBLDA for the classification of four motor imaginary patterns. These are efficient methods for reducing the training efforts in a BCI system. Classification results with probabilistic outputs can conquer unclassifiable problem when expanding one-versus-one classification methods to multi-class. The stabilization and validity of these methods for classification was testified by an application to two data sets. With these methods, we obtained a better result than traditional classification algorithm (such as SVM). Especially, to solve the problem of model-building with experiential assumptions, we can explore more prior information and refine the Bayesian model in PBLDA. Furthermore, with these probabilities, some post-processing procedures can be adopted for increasing robustness and generalization of classification. Also, we can refuse giving the last result according to probability in practical BCI (like control the wheelchair for handicapped), and consequently can improve the control accuracy. Clearly, these methods have a great potential in reducing training effort and encourage us to perform a valuable online test in the future.

References

Xu Lei was born in Chongqing, China, in 1982. He received B.S. degree in information and computational science in 2005, from UESTC. He is now pursuing the Ph.D. degree in biomedical engineering with UESTC. His research interests include EEG classification, EEG inverse problem, and EEG/IMRI fusion.

Tie-Jun Liu was born in Liaoning, China, in 1976. He received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, in 1999 and 2002, both in electrical engineering. He received the Ph.D. degree in medical science and engineering from UESTC in 2008. He is currently working with UESTC. His research interest includes brain computer interface.

Peng Xu received the B.S and Ph.D. degrees in biomedical engineering from UESTC in 2000 and 2006, respectively. He was a postdoctoral at Neural Systems and Dynamics Laboratory, University of California, Los Angeles from April 2007 to February 2009. Currently, he works with school of life science and technology, UESTC. His research interest includes biomedical signal processing, biomedical models, and medical image analysis.

De-Zhong Yao was born in Chongqing, China, 1965. He received the Ph.D. degree in applied geophysics from the Chengdu University of Technology, Chengdu, China, in 1991, and completed his postdoctoral fellowship in electromagnetic field with UESTC in 1993. He has been a faculty member since 1993, a professor since 1995, and the Dean of the School of Life Science and Technology, UESTC since 2001, the director of the Key Laboratory for NeurolInformation of Ministry of Education, since 2009. He was a visiting scholar with the University of Illinois at Chicago, USA, from September 1997 to August 1998, and a visiting professor with the McMaster University, Canada, from November 2000 to May 2001 and with the Aalborg University, Denmark, from November 2003 to February 2004. He has published more than 80 peer reviewed papers in international journals and conferences. His current research interests include EEG and IMRI with their applications in cognitive science and neurological problems.