Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Feature Extraction and Subset Selection for Classifying Single-Trial ECoG during Motor Imagery

Qingguo Wei[1,2], Xiaorong Gao[2], Shangkai Gao[2]
[1]Dept. of Electronic Engineering, Nanchang University, Nanchang 330029, China
[2]Dept. of Biomedical Engineering, Tsinghua University, Beijing 100084, China

Abstract—**The electrocorticogram (ECoG) recorded from subdural electrodes is a kind of BCI signal source that has the potential to achieve good classification results. The feature extraction and its subset selection are crucial for increasing classification accuracy rate. This paper proposes a new algorithm for classifying single-trial ECoG during motor imagery. The nonlinear regressive coefficients between signals on 10 leads are extracted in two frequency bands 0-3Hz and 8-30Hz as classification features. A genetic algorithm is used for the selection of the optimal feature subset and a support vector machine for their evaluation. The generalization error of 7% is achieved on Data set I of BCI Competition III.**

Keywords—**brain-computer interface, electrocorticogram, nonlinear regression, genetic algorithm**

I. Introduction

A brain-computer interface (BCI) is an assistive communication system that does not resort to the normal human output pathway consisting of periphery nerves and muscles [1]. People with severe disabilities such as amyotrophic lateral sclerosis and brainstem stroke may use this kind of technique to realize the controls of family facilities, wheelchairs or motor neuroprostheses, to improve their living quality.

The input signal with high quality is the basis of achieving good classification result. Electroencephalogram (EEG) and electrocorticogram (ECoG) are two different kinds of data recordings for BCIs. Since the ECoG signals, recorded from the surface of the cortex, posses the characteristics such as stability of location, freedom from muscle and movement artifacts, higher signal-to-noise ratio, and better spatial resolution [2], ECoG-based BCIs are therefore a promising BCI modality.

The method of feature extraction is crucial for good BCI communication. The commonly used features are extracted from band power, common spatial pattern (CSP) and autoregressive (AR) model. All these methods do not utilize the associate information between signals from different leads. In fact, such a kind of association does exist and might be a good kind of feature sources. Nonlinear regressive (NLR) coefficients represent the amplitude coupling of two signals and can be used as classification features. However, since the number of NLR coefficients increases squarely with the number of leads used, a high-dimensional feature vector is unavoidable. Given finite training samples, a high-dimensional feature space may lead to overfitting. In this paper, a wrapper method is used for dimensionality reduction by choosing a small number of meaningful features.

II. Method

Movement-related potentials (MRP) and event-related desynchronization (ERD) are two basic electrical physiological phenomena that are activated by limb movements or imagined movement [3] [4]. Wei *at al.* presented an algorithm for classifying single-trial ECoG by combining these physiological features extracted by common spatial subspace decomposition (CSSD) and achieved high classification accuracy on Data Set I of BCI Competition III [5]. In terms of classification rate, this algorithm is very good. However, it utilizes the whole 64 leads, and this is not feasible in practical settings.

To overcome this drawback, we try to perform the classification task by using a small number of leads. The NLR coefficients between signals on 10 leads are extracted in two frequency bands in line with MRP and ERD effects and are concatenated as one 200-dimensional feature vector. Then a genetic algorithm is used for feature selection and a support vector machine classifier for their evaluation. Fig. 1 is the proposed method for classifying single-trial ECoG during motor imagery of the left finger or tongue.

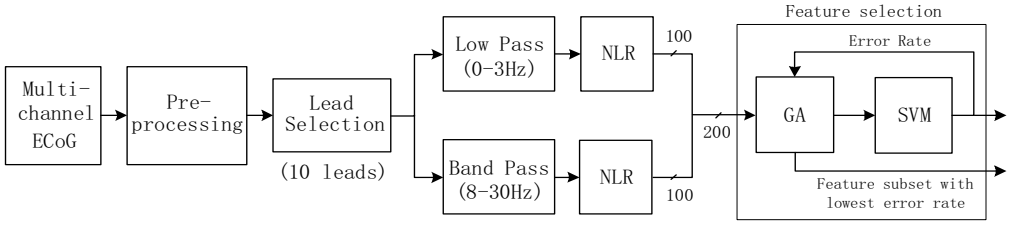

Fig. 1: The flowchart of the proposed approach for classifying single-trial ECoG during motor imagery of the left small finger or tongue

A. Data Acquisition and Preprocessing

The data in this study is the Data Set I of BCI Competition III. The data set includes a training set and a test set. Since the main goal of the study is to assess the methods of feature extraction, only training set is used for analysis in this paper. A detailed description of the data collection can be found in [6].

The given time series is 3000 sample points per trial. To save computational load and reduce the requirement for memory, all trials are downsampled to 300 points by picking each every ten points.

B. Lead Selection

For each of the two mental tasks, imagined movements of either small left finger (task A) or tongue (task B), MRP and ERD generate in a specific brain area. Thus, not signals on all leads contribute to classification accuracy rate. Moreover, a fewer number of leads are desired for a practical BCI system.

Lead selection is performed by the absolute band power difference of the two tasks in the frequency band 8-30 Hz. Since signals on these leads reflect the maximal difference of the two mental states, these leads should be located in motor area of the brain. The 10 leads with maximal band power distinctions (12, 21, 22, 29, 30, 31, 38, 39, 40 and 46) are chosen. The absolute band power differences between the signals of these two tasks on each lead are illustrated in Fig. 2.

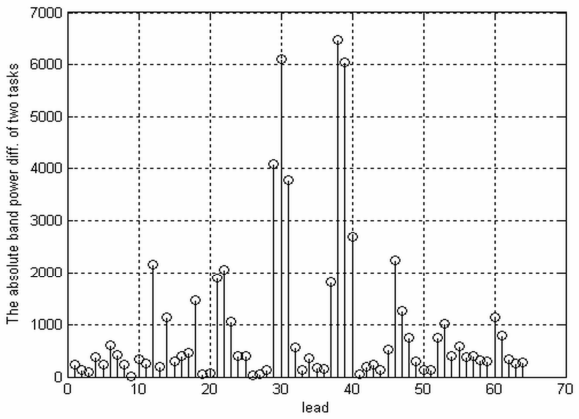

Fig. 2: The absolute band power differences between the signals of two tasks on each lead.

C. Feature Extraction

The interdependency of signals between leads is extensively used in epileptic prediction [7]-[9]. Amplitude coupling and phase coupling are two forms of such interdependency. In this paper, only amplitude coupling is utilized for feature extraction. Amplitude coupling is described by nonlinear regression (NLR) coefficient. Assume x and y are the signals on two leads respectively. Given signal x, the expectation of signal y, denoted as $\mu_{y|x}$, is the regression curve of y on x

$$\mu_{y|x}(x) = \int_{-\infty}^{+\infty} y\, p(y|x)dy \tag{1}$$

where $p(y|x)$ is the conditional probability of y given signal x.

The reduction of variance of y that can be obtained by predicting the y values from x according to the regression curve is the association measure given by

$$\eta^2_{y|x} = \frac{var(y) - E\left[(y-\mu_{y|x}(x))^2\right]}{var(y)} \tag{2}$$

$E\left[(y-\mu_{y|x}(x))^2\right]$ estimated from the regression curve is called the explained variance, i.e., it is explained or predicted on the basis of x. By subtracting the explained variance from the total variance $var(y)$, one obtains the unexplained variance. The basic idea is that if the amplitude of signal y is thought of as a function of the amplitude of signal x, given a certain value of x, the value of y can be predicted according to a NLR curve.

The estimation of this measure is called NLR coefficient h^2. To obtain an approximation of the regression curve, the x amplitude values are subdivided into M bins (M is determined experimentally and taken as 20 in this study.) For each bin, the x value in the midpoint (p_i) and the average of y values (q_i) are calculated, and the resulting points (p_i, q_i) are connected by segments of straight lines. Consequently, the NLR coefficient h^2 can be calculated according to the following expression

$$h^2 = \frac{\sum_{n=1}^{N}(y_n - \langle y\rangle)^2 - \sum_{n=1}^{N}(y_n - \hat{\mu}_{y|x}(x_n))^2}{\sum_{n=1}^{N}(y_n - \langle y\rangle)^2} \tag{3}$$

where $\hat{\mu}_{y|x}(x_n)$ is the linear piecewise approximation of the regression curve, and $\langle y\rangle$ denotes the average of y values over the N points of the time series. The estimator h^2 represents the strength of the association between the two signals and can take values between zero (y is totally independent of x) and one (y is totally dependent on x).

The preprocessed signals on 10 leads are first low pass (0-3Hz) and band pass (8-30Hz) filtered respectively in order to acquire MRP and ERD signals. Then two 100-dimensional feature vectors are separately obtained by calculating NLR coefficients of these signals. Finally these two feature vectors are concatenated to form one 200-dimensional feature vector.

D. Classification

The classifier used in this study is support vector machine (SVM). In order to obtain good generalization performance, a linear kernel function is utilized.

In order to assess classification performance, the generalization error is estimated by a 10×10-fold cross validation. Specifically, the original training set is randomly permuted for 10 times and each time the randomly permuted data is split into ten equal parts: each part is used for test and the remaining parts are used for training the SVM classifier. This cross validation procedure leads to 100 classification tests and the generalization error is decided by averaging the 100 test errors.

If the 200-dimensional feature vector is directly input into an SVM, the generalization error would be high. This is because many irrelevant or redundant features exist in the feature vector which could harm the performance of classification algorithm. Thus, it is necessary to select a small number of relevant features for the classification task.

E. Feature Subset Selection

The feature subset selection is performed by a wrapper method [10]. This method chooses features by using the classification algorithm as a subroutine for feature selection task. To select an optimal feature subset, a standard GA is utilized [11]. The basic structure of a standard GA is shown in Fig. 3. An initial population with 20 individuals is first randomly generated. At each generation, each individual is evaluated based on its fitness (or objective function). Those individuals with high fitness are selected for mating, and a new generation of individuals (offspring) with many of the attributes of their parents is produced by applying genetic operator crossover and mutation. This process leads to the evolution of populations of individuals with an improved fitness in terms of the given optimization task.

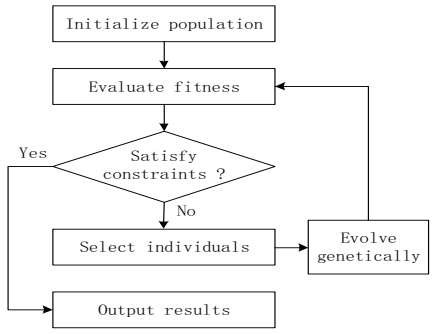

Fig. 3: The principle block diagram of a standard genetic algorithm.

When a GA is used for feature selection, individuals in a population are represented as a 200-bit binary string corresponding to the 200-dimensional feature vector. The GA operates on a population of such strings and chooses features by optimizing an objective function. When any bit is one, the corresponding feature is chosen; otherwise the corresponding feature is eliminated. The objective function is defined as the weighted sum of two decision variables, error rate $f_1(z)$ and relative feature number $f_2(z)$ [12]

$$f(z)=w_1\cdot f_1(z)+w_2\cdot f_2(z) \quad (4)$$

where the weights w_i are normalized such that $\sum w_i = 1$. $f_1(z)$ is produced by the classifier for a given feature subset denoted by the individual z and $f_2(z)$ is obtained by dividing the feature number selected in an individual z by the length of string (i.e. 200). The GA running parameters are chosen as follows: 1) Population size is 20; 2) Number of generation is 1000; 3) Generation gap is 0.9; 4) single point crossover is used and crossover probability is 0.7; 5) Mutation probability is 0.0035; 6) the selection function is stochastic universal sampling.

As shown in Fig. 1, the feature selection part includes a GA and an SVM. The GA is used for feature selection whereas the SVM is employed for the evaluation of chosen feature subset. The 200-dimensional feature vector is first subjected to feature selection through a GA, then the chosen feature subset is fed into an SVM classifier and the error rate of $10\times 10-$fold cross validation is input into the GA to produce the specified objective function. After 1000 generations of GA are run, a feature subset with lowest error rate is obtained.

III. RESULTS

Fig. 4 shows the evolution of the objective value, error rate and relative feature number of best individual with the generation of GA when the weights of error rate and relative feature number are taken as 0.9 and 0.1 respectively. After 1000 generations, an optimal subset of 48 features is acquired and corresponding error rate is 0.073. Although the changes of the error rate and the chosen feature number of the best individual at each generation are not monotonic as shown in the figure, their general tendency of decreasing with the generation of GA is clear.

The error rate and the chosen feature number of best individual after 1000 generations with respect to different values of weights are shown in Table I. As we can see from the table, when the weight of error rate decreases and the weight of feature number increases, the chosen feature number decreases monotonically, but the error rate does not change monotonically. The lowest error rate is achieved when the weights of the two decision variables take 0.8 and 0.2 respectively.

IV. DISCUSSION

The result presented in this study demonstrates that the proposed method can classify single-trial ECoG data of motor imagery with very good accuracy. NLR coefficients

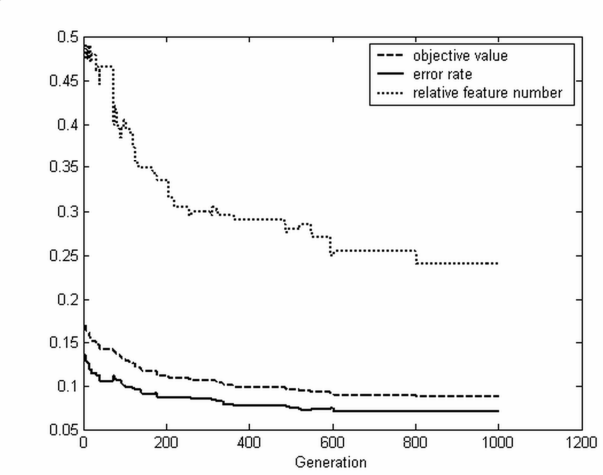

Fig. 4: When the weighted sum of error rate and relative feature number is taken as objective function, the objective value, error rate and relative feature number of the best individual each generation evolve with generation of the genetic algorithm.

TABLE Ⅰ

The relationship between the error rate and the feature number of best individual and the weights of two decision variables

Weights	Error Rate	Feature No.
$w_1 = 0.9,\ w_2 = 0.1$	**0.073**	**48**
$w_1 = 0.8,\ w_2 = 0.2$	**0.071**	**38**
$w_1 = 0.7,\ w_2 = 0.3$	**0.088**	**22**
$w_1 = 0.6,\ w_2 = 0.4$	**0.103**	**16**
$w_1 = 0.5,\ w_2 = 0.5$	**0.108**	**15**

reflect the amplitude coupling and interdependency of signals between different leads, and therefore is a kind of good features. However, the number of NLR coefficients is squarely increased with the number of selected leads. Hence, a suitable method of dimensionality reduction is needed to ensure good generalization ability of the classifier used.

According to the generalization error, the new method is a little inferior to that we proposed in BCI Competition III using 64 leads [5]. However, if only 10 leads are used in our previous algorithm, the generalization error on training set will be 9% that is worse than the result in this paper.

As for operating speed, the procedure of feature selection by GA is time consuming. But the work needs doing only once and can be done beforehand. After the optimal feature subset has been picked out, the GA can be removed from the classification algorithm. After the GA is excluded, the new algorithm is much simpler than the algorithm we used in competition and the resulting operating speed is higher.

In terms of the practicality, only 10 leads are utilized in the new algorithm. Actually, fewer leads are desired for a practical BCI system. For EEG-based BCIs, fewer leads can decrease the time of donning and doffing an electrode cap, whereas for ECoG-based BCIs, fewer leads can reduce the trauma of users.

The present algorithm uses only amplitude coupling as the source of feature. If both amplitude and phase coupling are combined for feature extraction, the classification performance is expected to be further improved.

ACKNOWLEDGMENT

This work was supported in part by Beijing Natural Science Foundation (#3051001), National Natural Science Foundation of China (#60318001), and Tsinghua-Yue-Yuan Medical Science Fund. The authors are grateful to the organizers of BCI Competition III for providing data.

REFERENCES

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

[2] E. E. Suter, "The brain response interface: communication through visually-induced electrical brain responses," Journal of Microcomputer Application, Vol. 15, pp. 31-45, 1992.

[3] C. Toro, G. Deuschl, R. Thatcher, S. Sato, C. Kufta, and M. Hallett, "Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG," Electroencephalography and Clinical Neurophysiology, Vol. 93, pp. 380-389, 1994.

[4] C. Babiloni, F. Carducci, F. Cincotti, P. M. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni, "Human movement-related potentials vs desynchronization of EEG alpha rhythm: A high-resolution EEG study," Neuroimage, Vol. 10, pp. 658-665, 1999.

[5] http://ida.first.fraunhofer.de/projects/bci/competition_iii/.

[6] T. N. Lal, T. Hinterberger, G. Widman, M. Schroder, J. Hill, W. Rosenstiel, C. E. Elger, B. Scholkopf and N. Birbaumer, "Methods towards invasive human brain computer interfaces," Advances in Neural Information Processing Systems 17, 737-744. (Eds.) Saul, L.K., Y. Weiss and L. Bottou, MIT Press, Cambridge, MA, USA (2005).

[7] M. Chavez, M. Le Van Quyen, V. Navarro, M. Baulac and J. Martinerie, "Spatio-temporal dynamics prior to neocortical seizures: Amplitude versus phase couplings," IEEE Transaction on Biomedical Engineering, Vol. 50, PP. 571-583, 2003.

[8] F. Bartolomei, F. Wendling, J. J. Bellanger, J. Regis and P. Chauvel, "Neural networks involving the medial temporal structures in temporal lobe epilepsy," Clinical neurophysiology, Vol. 112, pp. 1764-1760, 2001.

[9] M. Le Van Quyen, C. Adam, M. Baulac, J. Martinerie and F. Varela, "Nonlinear interdependencies of EEG signals in human intracranially recorded temporal lobe seizures," Brain Research, Vol. 792, pp. 24-40, 1998.

[10] G. John, R. Kohavi and K. Pfleger, "Irrelevant features and the subset selection problem," Proceedings of the Eleventh International Conference on Machine Learning, pp. 121-129, 1994.

[11] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, New York: Addison-Wesley, 1989.

[12] P. Hajela and C. Y. Lin, "Genetic search strategies in multicriterion optimal design," Structural Optimization, Vol. 4, pp. 99-107, 1992.