CLASSIFYING EEG SIGNALS IN FISHER DISCRIMINANT SPACES BY RANDOM ELECTRODE SELECTION

Shiliang Sun

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China
slsun@cs.ecnu.edu.cn

ABSTRACT

This paper introduces an ensemble approach for electroencephalogram (EEG) signal classification, which aims to overcome the instability of the Fisher discriminant feature extractor for brain-computer interface (BCI) applications. Through the random selection of electrodes from candidate electrodes, multiple individual classifiers are constructed. In a feature subspace determined by a couple of randomly selected electrodes, principal component analysis (PCA) is first used to implement dimensionality reduction. Successively Fisher discriminant is adopted for feature extraction, and a Bayesian classifier with a Gaussian mixture model (GMM) is trained to carry out classification. The outputs from all the individual classifiers are combined to give a final label. Experiments with real EEG signals taken from a BCI indicate the validity of the proposed random electrode selection (RES) approach.

Index Terms— EEG signal classification, brain-computer interface (BCI), Fisher discriminant, Gaussian mixture model (GMM), random electrode selection (RES)

1. INTRODUCTION

The last decade has seen a rapidly increasing interest in the research of brain-computer interface (BCI) technology, due to its huge potential for applications, in particular, to provide a basic communication and control channel between the brain and the external devices for severely motor-disabled but cognitively intact people [1, 2, 3]. Although there exist a number of measurements for monitoring brain activities, such as magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI), electroencephalography that records electrical brain activities from scalp electrodes is deemed to be the most practical and recipient way for BCI applications. The reason is that it is non-invasive, relatively inexpensive, and possesses a high temporal resolution [1]. The classification of electroencephalogram (EEG) signals, an important component in EEG-based BCIs, is the focus of this paper.

Thanks to the National Natural Science Foundation of China and Shanghai Educational Development Foundation for funding respectively under Project 60703005 and Project 2007CG30.

Fisher discriminant is a classic feature extraction approach for describing multivariate data in the derived Fisher discriminant space [4, 5]. It plays an important role in the classification task of BCI research as well [6, 7, 8]. Nevertheless, the stability of Fisher discriminant is susceptible to the dimensionality of the original feature spaces. High dimensional feature vectors with relatively few training samples tend to cause its instability [7, 9]. To figure out this problem, some countermeasures have been proposed so far in the context of EEG signal classification. For example, Wang et al. reduce the dimension of original feature vectors by taking the absolute value average of their eight consecutive elements [7]. Principal component analysis (PCA) [5] for dimensionality reduction seems a plausible method to obtain robust scatter matrices for Fisher discriminant, however it may lose important discriminative information embedded in small eigenvalues. Moreover, there exists a family of electrode selection methods to carry out dimensionality reduction as well. Usually they optimize the electrode selection from the physiologically related locations to remove redundancy and noise [10].

The foregoing strategies for improving Fisher discriminant are either computationally demanding or lack a guarantee of retaining all the discriminative information. Recently the successful application of random subspace for classification ensembles [9, 11] gives us an inspiration, which constructs individual classifiers by sampling features randomly. Breiman demonstrates that the random subspace method benefits from accurate and diverse individuals, and does not overfit with increasing individual classifiers [12]. In this paper, an ensemble method to address the instability of Fisher discriminant for EEG signal classification, namely, random electrode selection (RES) is proposed with the effectiveness demonstrated empirically. It also has the potential to work in the scenarios of electrode malfunctions or poor electrode contacts.

2. METHOD

2.1. Data description and preprocessing

The analyzed data are provided by the IDIAP Research Institute of Switzerland [13, 14]. They are EEG recordings taken

from normal subjects during three mental imagery tasks. The mental tasks are imagination of repetitive self-paced left hand movements (class ω_1), imagination of repetitive self-paced right hand movements (class ω_2) and generation of different words beginning with the same random letter (class ω_3). Data from the first two subjects (denoted by S1 and S2 respectively) are used. For a given subject, there are four non-feedback sessions recorded. After spatial filtering and power spectral density estimation, the raw EEG signals are converted to 96-dimensional feature vectors with every 12 entries coming from one of eight centro-parietal electrodes. The numbers of samples in the four sessions for subjects S1 and S2 are respectively 3488/3472/3568/3504, and 3472/3456/3472/3472. In this paper all the samples are then normalized with respect to different electrodes. To be specific, each spectral component f of electrode e for sample $\tilde{x}_i$ is divided by the summation of the spectral components of $\tilde{x}_i$ on electrode e,

$$x_i^e(f) = \frac{\tilde{x}_i^e(f)}{\Sigma_{f=1}^{K_e} \tilde{x}_i^e(f)} \tag{1}$$

where x_i is the normalized sample, and K_e is the number of features on electrode e.

2.2. Fisher discriminant for multi-class feature extraction

Fisher discriminant attempts to seek a number of projection vectors $\{d_j^*\}_{j=1}^V$ efficient for discriminating between data from different classes. For a binary classification problem, it obtains a unitary projection vector d_1^* to maximize the ratio of the between-class scatter matrix S_b and the within-class scatter matrix S_w, that is, $d_1^* = \arg\max_{d_1} \frac{d_1^\top S_b d_1}{d_1^\top S_w d_1}$.

For a c-class problem, S_b and S_w have the following forms, $S_b = \Sigma_{i=1}^c n_i (m_i - m)(m_i - m)^\top$, $S_w = \Sigma_{i=1}^c \Sigma_{x_k \in \omega_i} (x_k - m_i)(x_k - m_i)^\top$, where m_i is the mean of n_i samples which belong to class ω_i, m is the mean of all samples [5]. The instability of Fisher discriminant mainly originates from the fact that S_w may be singular in case the ratio of sample number and sample dimensionality is relatively small. This situation often occurs for EEG signal classification especially when using multiple electrodes to record signals. Meanwhile, such a low ratio is also harmful to get a robust estimation of S_b and S_w due to the side effect of noise.

Although there are many variants for extending the traditional Fisher discriminant to the multi-class problem, the instability problem has not been really resolved. For evaluation purpose, we exploit the multi-class Fisher discriminant introduced in [4], which is to seek $d_n^* = \arg\max_{d_n} \frac{d_n^\top S_b d_n}{d_n^\top S_w d_n}$ with the orthonormal constraints $d_1^\top d_n = d_2^\top d_n = \ldots = d_{n-1}^\top d_n = 0,\ d_n^\top d_n = 1$. The number of projection vectors $\{d_j^*\}_{j=1}^V$ is fixed as the maximized possible rank of matrix S_b, that is $V = c-1$ if c classes are defined. For a sample x in the original feature space, its new coordinate y in Fisher discriminant space can be described as $y = [d_1^{*\top} x, d_2^{*\top} x, \ldots, d_V^{*\top} x]^\top$.

2.3. Bayesian classifiers

Let each class ω_i from the finite set of c classes $\{\omega_1, ..., \omega_c\}$ have prior probability $P(\omega_i)$ where $\sum_{i=1}^c P(\omega_i) = 1$. A Bayesian classifier [5] assigns the label of a test sample y in a Fisher discriminant space according to values of its posterior probabilities $\{P(\omega_1|y), ..., P(\omega_c|y)\}$,

$$y \in \omega_j \quad \text{if} \quad P(\omega_j|y) = \max_{i=1,\ldots,c} P(\omega_i|y) \ . \tag{2}$$

The posterior probability $P(\omega_i|y)$ can be computed using the class-conditional probability density $p(y|\omega_i)$ by Bayes formula: $P(\omega_i|y) = \frac{P(\omega_i)p(y|\omega_i)}{p(y)} = \frac{P(\omega_i)p(y|\omega_i)}{\sum_{i=1}^c P(\omega_i)p(y|\omega_i)}$.

The class-conditional probability density $p(y|\omega_i)$ is herein modelled as a GMM which is potential to catch subtle information for data distributions. Formally, $p(y|\omega_i)$ is supposed to be the weighted combination of N_i Gaussian probability density functions, that is,

$$p(y|\omega_i) = \sum_{k=1}^{N_i} \alpha_i^k G(y|\mu_i^k, \Sigma_i^k), s.t., \Sigma_{k=1}^{N_i} \alpha_i^k = 1, \alpha_i^k > 0, \tag{3}$$

where $G(y|\mu_i^k, \Sigma_i^k)$ is a Gaussian probability density function with mean μ_i^k and covariance Σ_i^k [15]. Here the parameters $\{N_i, \alpha_i^k, \mu_i^k, \Sigma_i^k\}$ $(k = 1, ..., N_i; i = 1, ..., c)$ in GMMs are estimated by the expectation maximization (EM) algorithm and the minimum message length principle [5, 16].

2.4. RES for EEG signal classification

In Fisher discriminant, it is hard to get a robust estimation for scatter matrices S_w and S_b when the training set is relatively small compared to the high dimensionality of feature vectors or data are badly contaminated by noise. Thus instability happens, which can be further aggravated if the within-class scatter matrix S_w is singular.

RES intends to conquer this instability. According to the electrophysiological knowledge, electrodes related to mental tasks are first enclosed as candidate electrodes. Features belonging to the candidate electrodes form the original feature space. The RES approach samples a number of electrodes at random from the candidate electrodes. The features from the selected electrodes make up of a electrode feature subspace. PCA is applied to carry out further dimension reduction whereby the eigenvectors answering for the zero eigenvalues are removed. This can diminish the singularity of matrix S_w and the negative influence of noise to a certain extent. Thus a somewhat stable Fisher discriminant may be expected in the lower dimensional electrode feature subspace. Consequently, an individual Bayesian classifier using GMMs is constructed in the Fisher discriminant space of the electrode feature subspace. The electrode selection process runs M times, and resultantly M individual Bayesian classifiers have been trained. The M outcomes are combined to give the final label

for a test sample x. Define the average posterior probability of x belong to class ω_i as $\overline{P}(\omega_i|x) = \frac{1}{M}\sum_{m=1}^{M} P(\omega_i|y_m)$ where y_m from the m^{th} Fisher discriminant space is the transformed coordinate of x in the m^{th} electrode selection. Similar to (2), the label of x is determined as

$$x \in \omega_j \quad \text{if} \quad \overline{P}(\omega_j|x) = \max_{i=1,\ldots,c} \overline{P}(\omega_i|x)\ . \tag{4}$$

The feasibility of RES is guaranteed by the efficacy of random subspace and other classification ensemble methods. It is widely acknowledged that an effective ensemble classification system should consist of individuals that are not only highly accurate, but are diverse as well [12]. In the RES, electrodes are selected from a set of candidate electrodes answering for the electrophysiological background of the desired mental tasks, thus the "accurate" requirement is met unless the recorded EEG signals are of bad quality. The "diverse" is also satisfied since each selected electrode feature subspace is distinct from others with a high probability.

Besides, RES can work without difficulty even if technical artifacts occur, such as electrode malfunctions or poor electrode contacts. These scenarios are likely to take place during signal recording, particularly for long-term usage of BCIs or clinical applications. In this case, conventional feature selection or electrode selection methods would lose their effects because they often require complicated computation or optimization to seek the best configuration from the retained intact features or electrodes. On the contrary, RES only needs a simple random sampling from the remaining electrodes after the erroneous electrodes are detected and removed for consideration. Therefore, RES almost does not delay the signal recording process, and thus favors long-term applications.

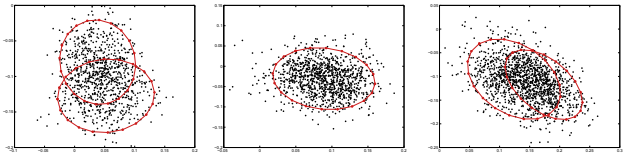
3. EXPERIMENT

3.1. Parameter configuration

The selected electrode number in RES is fixed as $40 \sim 60\%$ of the total number of electrodes closely related to the objective mental tasks, which is an appropriate balance to generate diverse and accurate individual classifiers. Hence, for the current problem of mental imagery, each time $3 \sim 5$ electrodes are randomly selected. The ensemble size M for RES is taken as 25. Later we will give a more economical estimation for ensemble size. In addition, two other methods we term FDRank and FDOptm are designed for comparison.

FDRank: FDRank uses PCA to carry out dimension reduction in the original feature space. Merely the eigenvectors for the sample covariance matrix with zero eigenvalues are removed. Then Fisher discriminant is used to project the data into a two-dimensional Fisher discriminant space in which a Bayesian classifier is constructed. This is the simplest improvement of Fisher discriminant to improve stability.

FDOptm: Although FDRank has a dimension reduction procedure, the retained dimensionality may still be relatively high. FDOptm tries to find the optimal dimension judged by the accuracy of 5-fold cross-validation on training data via thoroughly covering the possible dimensions. The dimension range is set as from 90% of the total sample energy to the rank of the sample covariance matrix. All the training data are reduced to the optimal dimension by PCA. Then Fisher discriminant follows and a Bayesian classifier is constructed.



(a) Class ω_1. (b) Class ω_2. (c) Class ω_3.

Fig. 1. The obtained GMMs for three mental imagery tasks.

3.2. Empirical results and comparisons

The component number N_i $(i = 1, ..., 3)$ in GMMs for class ω_i is initialized empirically to vary in $\{1, 2\}$. Fig. 1 provides an illustration of the obtained GMMs for data distribution of subject S1 on recording session 1. The corresponding Fisher discriminant space is derived by FDOptm. The component number of GMMs is selected automatically from the training sets. From this figure, we can also empirically see that the configuration of each ellipse is very reasonable.

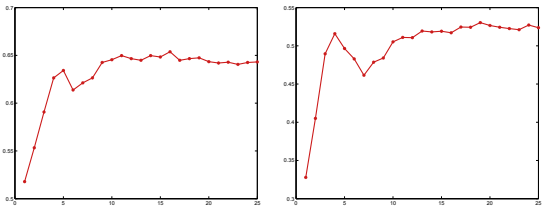
On top of the learned GMMs, FDRank, FDOptm and RES then respectively construct Bayesian classifiers for classifying test samples. To evaluate the RES approach sufficiently with the available data set, different combinations of the training set and the test set are exploited for classification. The classification results are given in Table 1. Therein "i(+j)∼k" means classification using session i (sessions i and j) as the training set and session k as the test set, and RES3, RES4 and RES5 respectively denote the RES approach using different electrode selection number 3, 4 and 5. The classification accuracy is calculated as correct classification number divided by the total number of samples in the specified test set.

Fig. 2 depicts the curve of accuracies for data set 1∼2 by RES4 with respect to different ensemble sizes. For other data sets, analogical tendency is obtained. The convergence of the generalized performance is thus revealed. Meanwhile, it also indicates that for electrode selection in our current classification task, less times (around 15) than 25 can obtain similar performance to that obtained by 25 times. This will alleviate the burden for training and storage requirements in RES.

From Table 1, the robustness of RES is manifested as RES3, RES4 and RES5 get similar performances. Also we can see that for subjects S1 and S2, RES and FDOptm significantly outperforms FDRank. Besides, the win-loss-tie scores between RES3, RES4, RES5 and FDOptm for subjects S1 and S2 are respectively 12-2-0, 11-3-0, 10-4-0. Statistical t-test shows that there are significant differences between RES3, RES4 and FDOptm at the 95% confident level. In addition,

Table 1. The classification accuracies (%) for subject S1 and S2 by using different methods

Subject	Method	Data set							Mean
		1∼2	2∼3	3∼4	1∼3	2∼4	1+2∼3	1+2+3∼4	
S1	FDRank	54.95	63.17	68.41	57.76	65.98	64.57	60.99	62.26
	FDOptm	66.39	69.76	69.66	66.82	69.32	71.02	73.12	69.44
	RES3	66.13	72.70	71.06	70.46	71.58	72.70	73.34	71.14
	RES4	64.31	70.35	70.69	69.23	71.12	71.80	72.77	70.04
	RES5	64.03	69.84	70.58	68.16	70.58	71.33	71.83	69.48
S2	FDRank	52.11	50.09	56.74	57.89	51.35	49.19	49.34	52.39
	FDOptm	52.11	55.36	56.77	57.89	51.53	53.28	57.00	54.85
	RES3	52.75	56.39	61.03	57.75	53.51	58.09	58.84	56.91
	RES4	52.37	55.79	61.35	57.86	53.46	57.55	57.86	56.61
	RES5	52.58	53.89	60.71	59.10	52.42	54.32	55.07	55.44



(a) Subject S1. (b) Subject S2.

Fig. 2. Accuracies of RES4 with different numbers of individual classifiers for data set 1∼2.

FDOptm is infeasible in real applications, since it involves heavy computation to find the optimal reduced dimension. The computational complexity of FDOptm is greatly higher than that of RES. Generally speaking, the computational times of RES and FDRank are of the same order of magnitude, while both are significantly lower than that of FDOptm. For example, the ratio of computational time for FDRank, FDOptm and RES4 (with selection times 15) for data set 1∼2 of subject 1 is around $1 : 130 : 5$.

4. CONCLUSION

A new classification ensemble approach RES is proposed in this paper. Through the random selection from electrodes answering for the objective mental tasks, RES effectively overcomes the instability involved in Fisher discriminant. Its superiority over FDRank and FDOptm is manifested empirically over a mental imagery classification problem. The simplicity characteristic of RES makes it suitable for stemming the torrent of technical artifacts (for example, electrode malfunctions or poor electrode contacts), which is very appealing especially for usage in long-term recordings. In the future, the issues of weighting electrodes according to their own discriminabilities can be investigated.

5. REFERENCES

[1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] T. Ebrahimi, J.M. Vesin, and G. Garcia, "Brain-computer interfaces in multimedia communication," *IEEE Signal Proc. Mag.*, vol. 20, no. 1, pp. 14–24, 2003.

[3] N. Birbaumer, "Brain-computer-interface research: coming of age," *Clin. Neurophysiol.*, vol. 117, no. 3, pp. 479–483, 2006.

[4] J. Duchene, and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 978–983, 1988.

[5] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2000.

[6] K.R. Müller, C.W. Anderson, and G.E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 165–169, 2003.

[7] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, and F. Yang, "BCI competition 2003-data set IV: an algorithm based on CSSD and FDA for classifying single-trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1081–1086, 2004.

[8] F. Galán, F. Oliva, and J. Guàrdia, BCI competition III, data set V: algorithm description. Available: http://ida.first.fraunhofer.de/projects/bci/competition_iii/results/martigny/ FerranGalan_desc.pdf

[9] X. Wang, and X. Tang, "Using random subspace to combine multiple features for face recognition," in *Proc. 6th Int Conf Automatic Face and Gesture Recognition*, pp. 284–289, 2004.

[10] T.N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, 2004.

[11] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[13] J.R. Millán, "On the need for on-line learning in brain-computer interfaces," in *Proc. Int. Joint Conf. Neural Networks*, pp. 2877–2882, 2004.

[14] S. Chiappa, and J.R. Millán, Data set V <mental imagery, multi-class>. Available: http://ida.first.fraunhofer.de/projects/bci/competition_iii/desc_V.html

[15] G. Mclachlan, and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[16] M.A.T. Figueiredo, and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, 2002.