Bio-Medical Materials and Engineering 26 (2015) S1019–S1025
DOI 10.3233/BME-151397
IOS Press

EEG feature selection method based on decision tree

Lijuan Duan[a, c], Hui Ge[a ,c], Wei Ma[a, c] and Jun Miao[b,*]

[a] *Key Laboratory of Trusted Computing, Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, College of Computer Science and Technology, Beijing University of Technology, Beijing, 100124, China*

[b] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China*

[c] *National Engineering Laboratory for Critical Technologies of Information Security Classified Protection, Beijing 100124, China*

Abstract. This paper aims to solve automated feature selection problem in brain computer interface (BCI). In order to automate feature selection process, we proposed a novel EEG feature selection method based on decision tree (DT). During the electroencephalogram (EEG) signal processing, a feature extraction method based on principle component analysis (PCA) was used, and the selection process based on decision tree was performed by searching the feature space and automatically selecting optimal features. Considering that EEG signals are a series of non-linear signals, a generalized linear classifier named support vector machine (SVM) was chosen. In order to test the validity of the proposed method, we applied the EEG feature selection method based on decision tree to BCI Competition II datasets Ia, and the experiment showed encouraging results.

Keywords: Decision tree, feature selection, optimal features, EEG, brain-computer interface

1. Introduction

Electroencephalograph (EEG) is a neuro-electricity activity collected by the conductive medium that contains a large number of information representing physiological and psychological state of human. A brain-computer interface (BCI) is a direct communication pathway between the brain and an external device [1]. The study of EEG signals processing and analysis is crucial to gain understanding in scientific endeavours. It is of great significance in many fields, such as, functional rehabilitation, brain and cognitive science, and entertainment. Nowadays, the study of EEG signals is applied to many signals processing and analyzing methods [2], and some of them have had a good response to the performance [3]. However, automated EEG analysis is still a challenging problem due to the poor

*Address for correspondence: Jun Miao, Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China. Tel.: 62600556; Fax: 62600523; E-mail: jmiao@ict.ac.cn.

0959-2989/15/$35.00 © 2015 – IOS Press and the authors.

This article is published with Open Access and distributed under the terms of the Creative Commons Attribution and Non-Commercial License.

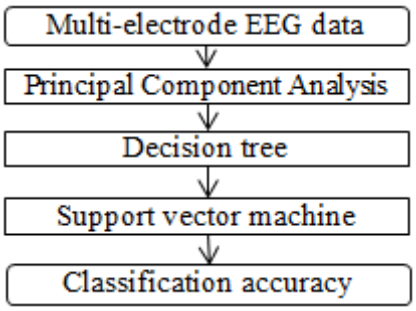

Fig. 1. The framework of the proposed EEG feature selection method.

resolution of EEG [4]. Automated feature selection is an important part of automated EEG analysis. In this paper, we will propose a solution for the automated feature selection.

In order to improve the performance of the BCI system, feature selection aims at removing redundant or irrelevant features. Mahnaz Arvaneh *et al.* proposes the EEG channel selection using decision tree to select appropriate subset of the channels [5] and decision tree is used in feature selection for the first time. The experiment showed that the EEG channel selection using decision tree removes irrelevant or correlated electrodes and reduces the average number of electrodes. However, the method does not remove redundant information of each channel, and the classification accuracy remains low for some subjects. In feature selection, we made some efforts previously. Optimal electrodes recombination method is proposed in the experimental paper [6]. First, we reduced the data dimensionality of a single electrode using Principal Component Analysis (PCA). Then, we calculated the classification accuracy of each single electrode, respectively, with the data obtained from the first step in the experiment. Furthermore, we selected electrodes, which have relatively high classification accuracy as optimal electrodes. Finally, we recombined optimal single electrodes to obtain different electrode combinations and calculate the classification accuracy of different electrode combinations. Optimal electrodes recombination method obtained high classification accuracy, but it was not automated during feature selection.

In this study, we proposed a novel EEG feature selection method based on decision tree. Compared with the EEG channel selection using decision tree, the proposed method improved classification accuracy. And compared with optimal electrodes recombination method, the proposed method automatically selected optimal features of each channel. We applied the method to BCI Competition II datasets Ia, and the experiment showed encouraging results.

This paper is organized as follows: Section 2 illustrates the proposed method. In section 3, the experiment is presented. Finally, section 4 gives conclusions.

2. Methodology

The method involved four processes: (1) input multi-electrode EEG data, (2) extract feature by using PCA, (3) select optimal features by using Decision tree (DT) and (4) calculate the classification accuracy by using Support vector machine (SVM). The framework of the proposed EEG feature selection method is shown in Figure 1.

2.1. Extracting feature

EEG signals have redundant information in high dimensional space. Therefore, a method to reduce the dimensionality should be chosen to remove redundant information. Principal component analysis (PCA) is a well-established method for feature extraction and dimensionality reduction [7]. Moreover, it is the linear projection from an original m-dimensional space to a q-dimensional space (m>q), relying on the maximization of the total scatter matrix of projected samples [8]. Therefore, PCA is used to reduce EEG data's dimensionality and obtain EEG features for feature selection. The primary idea is described as follows:

For training data Z including N samples, the mean vector is defined as $\overline{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$. The difference between each sample vector and sample mean vector is given by $\Delta z_i = z_i - \overline{z}$ (i= 1,2,…, N). The covariance matrix is computed using $\mathrm{C} = \frac{1}{m}\sum_{j=1}^{m} \Delta z_i{}^T \, \Delta z_i$, where m is the dimensionality of the sample vector. Then eigenvalue λ and unit eigenvector ξ are calculated according to $|\lambda \mathrm{I} - \mathrm{C}| = 0$ and $(\lambda_\mathrm{j}\mathrm{I} - \mathrm{C})\xi_\mathrm{j} = 0$,where $\lambda = \{\lambda_\mathrm{j}\}$ (j=1,2,..., m and $\lambda_1 > \lambda_2 > \cdots > \lambda_\mathrm{m}$), $\xi = \{\xi_\mathrm{j}\}$ and I is a unit matrix. Suppose that EEG data's dimensionality is reduced from m to q. The transformation matrix M is given by $\mathrm{M} = (\xi_1, \xi_2, \ldots, \xi_q)^T$. Finally, a new projected vector X is obtained according to $\mathrm{X} = \mathrm{MZ}$, where $\mathrm{Z} = \{\mathrm{z_i}\}$ (i=1,2,…, N) and the dimensionality of X is q. X will be regarded as training dataset for selecting optimal feature.

2.2. Selecting optimal feature

Decision tree has been widely used in the classification because of its fast speed and high precision. However, the main factor that influences the performance of decision tree classification is the selection problem [9]. For EEG signals, attribute selection problem is spatial feature selection problem. There are two phases in EEG feature selection method based on decision tree:

1. Constructing the tree to reduce features
2. Pruning the tree to avoid over-fitting

2.2.1. Constructing the tree

C4.5 [10], one of the widely used DT algorithms, is used. C4.5 is a modified algorithm of ID3 and adopts a divide and conquer strategy to construct a decision tree.

For the EEG data that are extracted features, training dataset X includes N samples, namely $\mathrm{X} = \{\mathrm{x_1}, \mathrm{x_2}, \ \ldots \ , \ \mathrm{x_N}\}$. Each sample $\mathrm{x_i}$ includes q features, namely $\mathrm{x_i} = \{\mathrm{v_{i1}}, \mathrm{v_{i2}}, \ldots, \mathrm{v_{iq}}\}, \mathrm{i} = 1,2, \ldots, \mathrm{N}$. The q features are regarded as q attributes. Therefore, the attribute set A is defined as $\mathrm{A} = \{\mathrm{A_1}, \mathrm{A_2}, \ldots, \mathrm{A_q}\}$. Each attribute $\mathrm{A_k}$ includes N values, namely $\mathrm{A_k} = \{\mathrm{v_{1k}}, \mathrm{v_{2k}}, \ldots, \mathrm{v_{Nk}}\}, \mathrm{k} = 1,2, \ldots, \mathrm{q}$. The process of constructing tree begins with the root (first internal node) with whole training dataset X. Internal nodes store different test attributes. Leaves store the labels of the class. From the attribute set A, the attribute that best divides the samples into their classes is chosen as the split attribute in the internal node. Suppose that the attribute $\mathrm{A_t}$ is the split attribute in internal node t, because the split attribute $\mathrm{A_t}$ includes w different values, the internal node t will then split the decision tree into w branches. In other words, the split attribute $\mathrm{A_t}$ divides training dataset X into w subsets, namely $\{\mathrm{X_1}, \ \mathrm{X_2}, \ \ldots, \ \mathrm{X_w}\}$.

For given dataset X, the entropy is computed using:

$$Entropy(X) = -\sum_{j=1}^{c} p_j \log_2(p_j) \tag{1}$$

where p_j and c are the probabilities that samples in dataset S belong to class j and the total number of classes, respectively.

For given attribute A_k, the information gain of dataset X is computed using:

$$Gain(X, A_k) = Entropy(X) - \sum_{v \in Values(A_k)} \frac{|X_v|}{|X|} Entropy(X_v) \tag{2}$$

where values(A_k) are a set of different values of A_k, X_v is X's subset, and X_v's samples' value in attribute A_k is v.

For given attribute A_k, the split information of dataset X is computed using:

$$SplitI(X, A_k) = -\sum_{j=1}^{c} \frac{|X_j|}{|X|} \text{lb}\left(\frac{|X_j|}{|X|}\right) \tag{3}$$

where X_j is X's subset whose samples belong to class j.

For given attribute A_k, the information gain ratio of dataset X is computed using:

$$GainRatio(X, A_k) = \frac{Gain(X, A_k)}{SplitI(X, A_k)}. \tag{4}$$

In remaining attributes that are not used as test attributes, we choose an attribute as a new test attribute. The new test attribute has the maximum information gain ratio value and its information gain value is not less than the average information gain value of the whole attributes.

The tree constructing continues until all the remaining samples belong to the same class, or there is no remaining attribute left.

2.2.2. Pruning the tree

C4.5 adopts post-pruning method to avoid the over fitting problem. Pruning arbitrary node L includes the following steps. First, the subtree whose root is node L is deleted. Then, node L is changed to a leaf. Finally, the class of leaf L is decided by training samples associated with L, following the principle of the minority subordinating to the majority. Node L is then deleted, if and only if, the performance of pruned tree is not worse off than the performance of the tree without pruning.

After the above steps, a simplified tree is built. Whole internal nodes' split attributes in simplified tree are regarded as optimal attributes, namely optimal features. Then optimal features are combined for classification.

2.3. Calculating the classification accuracy

EEG signals are a series of non-linear signals. The classification of EEG signals has drawn attention in recent years [11]. Support Vector Machine (SVM) is fit for processing non-linear data and has contributed great generalization performance [12]. In SVM, non-linear input vectors are mapped to a very high-dimension feature space and are divided into different classes by the hyperplane [13]. The SVM algorithm has already been proved effective for different kinds of pattern recognition tasks in various

researches [14]. Therefore, SVM is used for classification. The main idea of SVM is described as follows:

Suppose that training dataset with labels is $(x_i, y_i)(i = 1,2, \ldots, N)$, where $x_i \in R^N$ and $y_i \in \{-1,1\}$. If the classification hyperplane is $wx_i + b = 0$, $wx_i + b$ satisfy:

$$y_i\{wx_i + b\} - 1 + \beta_i \geq 0 \tag{5}$$

where β_i is a slack variable ($\beta_i \geq 0$), w is the weight vector. $\beta_i = 0$ denotes that samples are linearly separable. $\beta_i > 0$ denotes that samples are not linearly separable. Samples which satisfy $y_i\{wx_i + b\} - 1 + \beta_i = 0$ are support vectors. The optimal hyperplane is the hyperplane that satisfies Eq. (5) and minimizes $\|w\|^2$. If samples are linearly separable, the optimal classification discriminant function is $f(x) = \mathrm{sgn}\left[\sum_{i=1}^{N_s} \alpha_i\, y_i (x_i \cdot x) + b\right]$, where $(x_i \cdot x)$ denotes the inner product of support vector x_i and input vector x, N_s is the number of support vectors, α_i is the corresponding coefficient of support vector x_i. If samples are not linearly separable, the optimal classification discriminant function is $f(x) = \mathrm{sgn}\left[\sum_{i=1}^{N_s} \alpha_i\, y_i K(x_i, x) + b\right]$, where $K(x_i, x)$ is the kernel function. Specially, sigmoid function is $K(x_i, x) = \tanh[v(x_i \cdot x) + c]$.

3. Experiments

3.1. Data description

The EEG data from publicly available BCI Competition II datasets Ia [15] was used in this study. The datasets were acquired from a healthy subject. The task of the subject was to move a cursor up and down on a computer screen by imagination. Simultaneously, the subject's cortical potentials were acquired. When the cortical potentials of the subject were recording, the subject received visual feedback of their slow cortical potentials. Central parietal region electrode (Cz -Mastoids) was regarded as the reference electrode, and also recorded their slow cortical potentials. If the subject's slow cortical potentials were positive, the cursor on the screen was moved up. If the subject's slow cortical potentials were negative, the cursor was moved down. There were six electrodes used to collect the subject's EEG signals. They are, respectively, A1-Cz (A1=left mastoid), A2-Cz, two cm frontal of C3 (F3), two cm parietal of C3 (P3), two cm frontal of C4 (F4), and two cm parietal of C4 (P4). The sampling rate was 256 Hz. The experiment included 268 trials and each trial lasted six seconds without inter-trial intervals. In addition, there were 561 samples for EEG signal analysis. From 0.5 second to 6 seconds, there were a highlighted goal appeared at the top or the bottom of the screen to indicate negativity or positivity. From 2 second to 5.5 seconds, EEG signals were recorded for training and testing.

3.2. Experimental results

In order to test the validity of the proposed method, we compare our method with optimal electrodes recombination method on feature's number and classification accuracy. In feature exaction, EEG data's dimensionality was reduced from 896 to 3 for each electrode by using PCA. In the classification, its accuracy was obtained by using the averaged 10×10-fold cross-validation. The specific number of features for 10 times experiments is shown in Table 1. As can be seen in Table 1, the proposed method

reduced the number of features from 18 to 13.5. The kernel function in SVM is the sigmoid function. The experimental results for 10×10-fold cross-validation are shown in Table 2. For each experiment, 505 samples were used for training and 56 samples for test. Except the sixth-time experiment, the left nine times experiments' classification accuracies are approximately 90%. From the last row of Table 2, the average classification accuracy obtained by using the proposed method was about 0.9% higher than optimal electrodes recombination method.

4. Conclusions

In this paper, the EEG feature selection method based on decision tree was proposed. The proposed method was made up of four steps. Firstly, multi-electrode EEG signals are input. Then, features are extracted by using PCA to reduce data's dimensionality. Subsequently, optimal features were selected by DT, and optimal features were recombined for classification. Finally, the classification accuracy was calculated.

The main innovation was making use of the advantage that DT automatically selects optimal features. Different from EEG channel selection using decision tree, the proposed method not only selects appropriate channels, but also selects appropriate features of time domain in each channel. Moreover, the EEG feature selection method based on decision tree was applied to BCI Competition II datasets Ia. Experimental results showed that the average classification accuracy obtained by using our method is about 0.9% higher than using optimal electrodes recombination method.

Table 1
Comparison of the features' number of two methods

Time	Decision Tree The Number of Features	Optimal Electrodes Recombination The Number of Features
1	14	18
2	14	18
3	13	18
4	13	18
5	14	18
6	14	18
7	13	18
8	14	18
9	15	18
10	11	18
Average	13.5	18

Table 2
Performance comparison of two methods

Time	Decision Tree Classification Accuracy (%)	Optimal Electrodes Recombination Classification Accuracy (%)
1	91.0714	87.5000
2	89.2857	85.7143
3	89.2857	85.7143
4	92.8571	89.2857
5	91.0714	85.7143
6	76.7857	80.3571
7	89.4737	91.2281
8	91.0714	96.4286
9	89.2857	91.4286
10	91.0714	89.2857
Average	89.1259	88.2300

Acknowledgments

This research is partially sponsored by Natural Science Foundation of China (Nos. 61003105, 61175115, 61370113 and 61272320), Beijing Municipal Natural Science Foundation (4152005 and 4152006), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (CIT & TCD201304035), Jing-Hua Talents Project of Beijing University of Technology (2014-JH-L06), and Ri-Xin Talents Project of Beijing University of Technology (2014-RX-L06), the Research Fund of Beijing Municipal Commission of Education (PXM2015_014204_500221) and the International Communication Ability Development Plan for Young Teachers of Beijing University of Technology (No. 2014-16).

References

[1] T. Ebrahimi, J.M. Vesin and G. Garcia, Brain-computer interface in multimedia communication, Signal Processing Magazine **20** (2003), 14–24.
[2] A. Bashashati, M. Fatourechi, R.K. Ward, et al., A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals, Journal of Neural Engineering **4** (2007), R32–R57.
[3] C. Liu, H. Wang and Z. Lu, EEG classification for multiclass motor imagery BCI, 25th Chinese Control and Decision Conference (CCDC), IEEE, 2013, 4450–4453.
[4] Ahmed Al-Ani and Akram Al-Sukker, Effect of feature and channel selection on EEG classification, 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06), 2006, 2171–2174.
[5] M. Arvaneh, C. Guan, K.K. Ang, et al., EEG channel selection using decision tree in brain-computer interface, Proceedings of the Second APSIPA Annual Summit and Conference, 2010, 225–230.
[6] Lijuan Duan, Xuebin Wang, Zhen Yang, Haiyan Zhou, Chunpeng Wu, Qi Zhang and Jun Miao, An emotional face evoked EEG signal recognition method based on optimal EEG feature and electrodes selection, Lecture Notes in Computer Science **7062** (2011), 296–305.
[7] A. Subasi and M. Ismail Gursoy, EEG signal classification using PCA, ICA, LDA and support vector machines, Expert Systems with Applications **37** (2010), 8659–8666.
[8] I.T. Jolliffe, Principal component analysis, Proceedings of the 5th WSEAS Int. Conf. on Signal, Speech and Image Processing, Springer-Verlag, Corfu, 1986.
[9] X. Chen, J. Wu and Z. Cai, Learning the attribute selection measures for decision tree, Fifth International Conference on Machine Vision (ICMV 12), International Society for Optics and Photonics, 2013, 87842S–87842S-8.
[10] J.R. Quinlan, C4. 5: Programs for Machine Learning, Morgan kaufmann, San Mateo, CA, 1993.
[11] Y. Xiong, Y. Luo, W. Huang, et al., A novel classification method based on ICA and ELM: A case study in lie detection, Bio-Medical Materials and Engineering **24** (2014), 357–363.
[12] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery **2** (1998), 121–167.
[13] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning **20** (1995), 273–297.
[14] X. Jie, R. Cao and L. Li, Emotion recognition based on the sample entropy of EEG, Bio-Medical Materials and Engineering **24** (2014), 1185–1192.
[15] N. Birbaumer, Data Sets Ia for the BCI Competition II, http://www.bbci.de/competition/ii/#datasets, 2015.