METRIC LEARNING FOR EVENT-RELATED POTENTIAL COMPONENT CLASSIFICATION IN EEG SIGNALS

Qi Liu, Xiao-guang Zhao and Zeng-guang Hou

The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing 100080, PRC

Abstract - **In this paper, we introduce a metric learning approach for the classification process in the recognition procedure for P300 waves in electroencephalographic (EEG) signals. We show that the accuracy of support machine vector (SVM) classification is significantly improved by learning a similarity metric from the training data instead of using the default Euclidean metric. The effectiveness of the algorithm is validated through experiments on the dataset Ⅱ of the brain-computer interface (BCI) Competition Ⅲ(P300 speller).**

Index Terms – Metric learning, SVM, P300

1. INTRODUCTION

The P300 is the event related potential (ERP) component most commonly used as a metric of cognitive function in decision making processes on account of the presence, magnitude, topography and timing of its signals. The P300 has been applied to multiple brain-computer interface (BCI) systems, such as the P300 speller [1, 2], psychological tests [3, 4], and clinical medicine [5]. The issue critical to the design and effectiveness of all its applications is the ability to correctly recognize the P300 component from the electroencephalographic (EEG) signals collected from the brain.

In recent years, many methods have been developed for the effective recognition of P300 waves. However, due to the poor signal-to-noise ratio (SNR) of raw EEG signals, improving recognition performance in signal processing remains a live problem. The recognition procedure mainly includes feature extraction and classification of signals. Approaches to feature extraction fall roughly into one of four categories [6]: time or frequency methods [7], conventional time-frequency methods [8], model parameter methods [9], and wavelet decomposition-based methods [10, 11, 12]. Among these, wavelet decomposition-based methods can best represent non-stationary EEG signals. In such methods, information is described in various time windows and frequency bands. Also, many effective classification algorithms have been used for the recognition of P300 such as neural networks [13], hidden Markov models (HMM) [14], k-means clustering [13, 15], k-nearest neighbors (kNN) [16], support vector machines (SVM) [1, 17], etc.

This paper focuses primarily on the classification process in the recognition procedure for P300 waves. In most of the abovementioned classification methods, a Euclidean metric is commonly used to measure similarities among raw data points, and often fails to generate discriminative representations for a given problem. To solve this problem, we present a similarity metric-learning algorithm for the classification of P300 waves in EEG signals. We focus on the measurement of the distance between samples for an SVM classifier. The performance of the classifier is significantly improved by learning a global Mahalanobis distance metric from labeled samples.

2. SUPPORT VECTOR MACHINE

An SVM is a learning strategy based on statistical theory. In this strategy, the data entered is mapped into a high-dimensional feature space where samples belonging to different categories are separated by an optimal hyperplane.

We are given the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in R^d$, $y_i \in \{1, -1\}$. To obtain the optimal hyperplane to separate the samples, the following quadratic programming problem (1) is solved:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i^p \tag{1}$$

$$\text{s.t.} \quad y_i\left(\mathbf{w}^T\varphi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, n$$

where $\mathbf{w}$ and b determine the hyperplane in feature space, C is the punishment factor to emphasize the loss caused by outliers, ξ_i is a slack variable, $p > 0$ ($p = 1$ and $p = 2$ are two common choices, called L_1 loss and L_2 loss, respectively), and φ maps the input vectors into high-dimensional space. Usually, the above quadratic programming problem is solved using its dual:

$$\max_{\alpha_i} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{2}$$

$$\text{s.t.,} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \cdots, n$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \right\rangle$ is the kernel function to generalize the SVM to non-linear decision functions. The unknown input pattern $\mathbf{x}$ is classified

This work was Supported by the National Natural Science Foundation of China (Grant No.61271432 and 61333016) and the National CNC machine tool special Foundation of China (Grant No.SK201201A35-02).

according to the following decision function:

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K\left(\mathbf{x}_i, \mathbf{x}\right) + b\right) \tag{3}$$

3. METHOD

As shown in Fig 1, EEG signal processing mainly includes data collection, pre-processing, feature exaction and pattern classification.

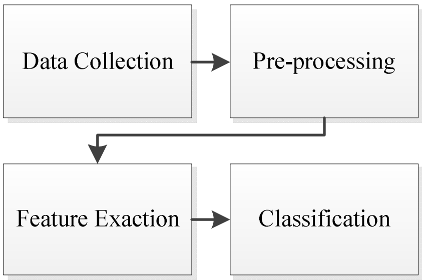

Fig. 1 Flowchart of EEG signal processing.

3.1 Preprocessing

To construct high-level signal characteristics suitable for classification, signal preprocessing is required. The process consists of signal segmentation, selection of electrodes, superposition, filtering, and data normalization.

It is well known that the P300 is a kind of late positive component [22]. For the time-locked assumption between the stimulus and the response, we take the values of the signal from 0 to 700ms after stimulus onset from the electrode channels Fz (34), Cz (11), Pz (51), Oz (62), C3 (9), C4 (13), P3 (49), P4 (53), PO7 (56), and PO8 (60), where channel assignment numbers are indicated in parentheses. Data from other electrode channels is also used but has no obvious influence on the result.

Since the SNR of EEG signals is usually very low, the application of the superposition operation to the recorded reactions of the repetitive stimulations is necessary to reduce interference signals and enhance the desired signals. Otherwise, since the frequency of P300 waves is mainly distributed in low-frequency regions, a six-order band-pass Chebyshev Type I filter – the cutoff frequencies for which are 0.5Hz and 30 Hz – is designed to filter each extracted signal. Following this, signals from each selected channel are normalized to zero mean and unit variance.

3.2 Feature exaction

The purpose of feature exaction is to find features in EEG signals that effectively characterize the cognitive task in question. The extracted feature vectors representing different tasks are expected to have obvious differences, which forms the basis of the recognition of cognitive components.

Due to the temporally varying and non-stationary character of EEG signals, traditional analysis methods cannot clearly distinguish frequency components representing cognitive processes contained in a certain timeframe from some minute, transient feature. Therefore, discrete wavelet transform, which is a typical time-frequency analysis method, is suitable to analyze EEG signals.

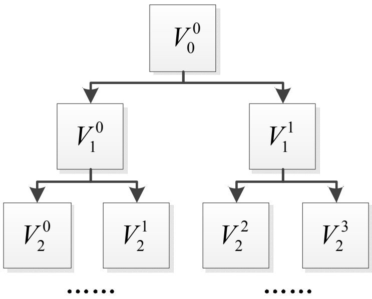

Fig. 2 Structure of WPD. V_i^j indicates the space expanded by the jth node of the ith-layer WPD.

However, wavelet decomposition (WD) partitions only the frequency axes in the low-frequency band, which may reduce analytical precision. To solve this problem, wavelet packet decomposition (WPD) is adopted [12]. Compared to WD, WPD implements equal width decomposition not only in low-frequency bands but also in the high-frequency ones, which provides a more precise way to analyze non-stationary EEG signals. The structure of WPD is shown in Fig 2.

In this study, we decompose each epoch into three levels using wavelet packet transform. Quadratic B-Spline functions are used as mother wavelets because of their similarity to evoked responses. Since the signal has already been filtered, eight sets of coefficients from the following frequency bands are obtained: 0.5-4Hz, 4-8Hz, 8-12Hz, 12-16Hz, 16-20Hz, 20-24Hz, 24-28Hz and 28-30Hz. The wavelet packet decomposition of EEG signals is implemented by using the MATLAB software package.

Following this, the wavelet packet energy and entropy of each node are calculated to construct feature vectors. The wavelet packet energy indicates the strength of the signal as it gives the area under the curve of the power. The energy of an EEG signal of finite length is given by (4).

$$En\left(s\right) = \sum_i s_i^2 \tag{4}$$

where s_i represents the projection coefficient of signal s in an orthonormal basis. The energy feature of each epoch is:

$$\begin{gathered}\mathbf{Enx}_i = \left[En_{ij}(\mathbf{l}_3^0), En_{ij}(\mathbf{l}_3^1), \cdots, En_{ij}(\mathbf{l}_3^7)\right], \\ i = 1,2,\cdots n, j = 1,2,\cdots,10\end{gathered} \tag{5}$$

where $\mathbf{l}_3^0 \sim \mathbf{l}_3^7$ represents the segments in nodes of the third layer. n represents the number of epochs and j represents the 10 selected channels.

Wavelet packet entropy is a measure of the disorderliness of EEG signals. According to Quiroga [18], with the emergence of the P300, the use of wavelet packet entropy for signal analysis has witnessed a marked decline. There are several entropy types, such as Shannon, log energy, sure, threshold, etc. Here, Shannon entropy is employed and calculated according to (6) to measure the complexity of EEG signals.

$$Ent(s) = -\sum_i s_i^2 \log(s_i^2) \tag{6}$$

where s_i also represents the projection coefficient of signal s in an orthonormal basis. Therefore, the entropy feature vector of each epoch is:

$$\begin{gathered}\mathbf{Entx}_i = \left[Ent_{ij}(\mathbf{l}_3^0), Ent_{ij}(\mathbf{l}_3^1), \cdots, Ent_{ij}(\mathbf{l}_3^7)\right], \\ i = 1,2,\cdots n, j = 1,2,\cdots,10\end{gathered} \tag{7}$$

Finally, the feature vector of each epoch is constructed as

$$\mathbf{x}_i = [\mathbf{Enx}_i, \mathbf{Entx}_i], i = 1,2,\cdots,n \tag{8}$$

According to the above procedure, a 160-dimensional (10 channels × 8 frequency bands × 2 (energy, entropy)) feature vector for each epoch is extracted from the EEG signals. Together with their category labels, the instances from a training set are used to design the classifier.

3.3 Classification

A conventionally used kernel function in SVM is the category in which the Euclidean distance is contained in order to determine the degree of similarity between data points in the kernel space. The kernel function is formulated as:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = f\left(d\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \tag{9}$$

There are several widely used kernel functions, such as the radial basis function (RBF), the negative distance kernel, the logarithmic kernel and the B-spline kernel. SVMs using these kernels are successfully used in many applications. However, the Euclidean distance ignores structure information which can provide more clues for classification.

In many cases, the samples are still inseparable after having been mapped into the high-dimensional kernel space. In traditional SVMs, by maximizing the margin between two classes, the learned transformation controls the distance between classes while ignoring the distance between data points within a class. Therefore, since it keeps instances of the same class close while pushing instances of different classes farther away, a metric learning algorithm is a reasonable way to improve the classification accuracy of the SVM [19].

Inspired by the metric learning theory, a Mahalanobis similarity metric learnt from labeled instances is used in this paper in place of the Euclidean metric in the kernel of the SVM, in order to improve classification performance. The principle of metric learning is illustrated in Fig 3 [20].

For SVM, we choose the exponential radial basis function (RBF) as the kernel function:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma d\left(\mathbf{x}_i, \mathbf{x}_j\right)\right), \gamma > 0$$

where $d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T M\left(\mathbf{x}_i - \mathbf{x}_j\right)$ is the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\gamma > 0$ is the width of the Gaussian distribution. $M = L^T L$ is a positive semi-definite matrix. To distinguish it from Euclidean distance, the Mahalanobis distance is indicated by (10).

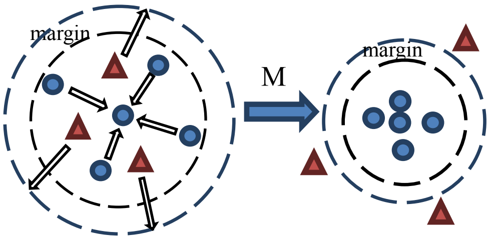

Fig. 3 Illustration of the distribution of instances belonging to two classes [19]. By applying the metric learning algorithm (minimizing the distances between instances in the same class while keeping instances from the other class far away), the input instance is surrounded by training instances of the same class.

$$d_M\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T M\left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{10}$$

In order to separate instances belonging to different classes, the desired metric should satisfy the following constraints:

$$\begin{gathered}d_M\left(\mathbf{x}_i, \mathbf{x}_k\right) - d_M\left(\mathbf{x}_i, \mathbf{x}_j\right) > 0, \\ \forall i, j, k \ and \ y_i = y_j, y_i \neq y_k\end{gathered} \tag{11}$$

Usually, the above constraints are too strict to satisfy. Therefore, slack variables are introduced as follows:

$$\begin{gathered}d_M\left(\mathbf{x}_i, \mathbf{x}_k\right) - d_M\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 1 - \xi_{ijk} \\ \forall i, j, k \ and \ y_i = y_j, y_i \neq y_k\end{gathered} \tag{12}$$

Finally, the optimization problem is formulated as:

$$\begin{gathered}\min_{M, \xi_{ijk}} \sum_{ij} \eta_{ij} d_M\left(\mathbf{x}_i, \mathbf{x}_j\right) + C \sum_{ijk} \eta_{ij} (1 - y_{ik}) \xi_{ijk} \\ \text{s.t.} \begin{cases} d_M\left(x_i, x_k\right) - d_M\left(x_i, x_j\right) \geq 1 - \xi_{ijk} \\ \xi_{ijk} \geq 0 \\ M \geq 0 \end{cases}\end{gathered} \tag{13}$$

where $y_{ij} \in \{1,0\}$ indicates whether $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same class, and $\eta_{ij} \in \{1,0\}$ indicates whether $\mathbf{x}_j$ is a selected neighbor of $\mathbf{x}_i$ within the same class. Following this, the algorithm proposed in [19], – called the large margin nearest neighbor (LMNN) method – is used to solve (13).

4. EXPERIMENTS

4.1 Dataset description

To make the results comparable, the SVM-based metric learning approach for the classification of EEG signals is carried out on a public dataset: dataset Ⅱ of the BCI Competition Ⅲ (P300 speller). We briefly introduce the experiment and the data processing.

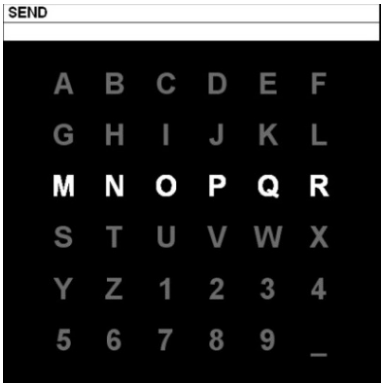

Fig. 4 Adopted P300 speller paradigm. The highlighted row is the one intensified.

The row-column P300 speller paradigm is adopted where a 6×6 matrix (shown in Fig. 4) containing 36 symbols is presented to subjects. In order to spell each character, all rows and columns of this matrix are randomly intensified. Sets of 12 intensifications are repeated 15 times for each character epoch. P300 exist in the EEG signals associated with the rows or columns containing the desired characters.

The 64-channel EEG signals are collected from two subjects, A and B. For each, the recorded data contains one training set (85 characters) and one test set (100 characters).

The collected signals are digitized at 240Hz. According to the process described in section 3, signals from the 10 selected channels are band-pass filtered from 0.5Hz to 30Hz. For each channel, all data samples obtained between 0 to 677ms after intensification begins are extracted. At this stage, an extracted signal from a single channel is composed of 160 points. The training set is composed of 1020 (85×12) post-stimulus feature vectors and each feature vector contains 160 elements. The test set is composed of 1200 (100×12) 160-dimensional feature vectors.

4.2 Results

In this section, we test the classification performance of the metric learning-based algorithm on the dataset described above. The accuracy of the predicted characters is used to evaluate the classification accuracy. The input character is detected by the intersection of the row and the column associated with the P300 wave. For this purpose, the classifier is trained for binary classification, and instances are labeled "1"/"-1" for P300 presence/absence, respectively. The maximum score of the discriminant function (3) indicates the presence of a P300 wave. Parameters such as C are determined by a five-fold cross validation [21].

Tables 1 to 3 show the results. Specifically, Table 1 shows the spelling accuracy of our method on the test datasets of the two subjects with respect to the number of repetitions used in superposition. Table 2 compares several effective methods. 15 repetitions are used for all of them. Table 3 shows the spelling accuracy of our method for different features using the test datasets.

Table 1 Classification performances (in %) of correctly recognized characters

Repetitions	1	3	5	7	10	15
Subject A	12	48	70	82	87	95
Subject B	27	57	75	86	92	96

Table 2 Comparison of classification performances (in %) of correctly recognized characters

Method	Standard SVM	Ensemble SVM	Ours
Subject A	92	97	95
Subject B	89	96	96

Table .3 Classification performances (in %) of correctly recognized characters for different features

Feature	Energy(En)	Entropy(Ent)	En&Ent
Subject A	87	85	95
Subject B	89	90	96

As we can see, the accuracy improves with increasing number of repetitions. As shown in Table 2, the performance of our algorithm is superior to the standard SVM algorithm and is comparable to the winning algorithm as reported in the competition in [1]. Furthermore, as shown in Table 3, the combined use of energy and entropy features obtains more accurate classification results than the respective use of each.

From the results, we can see that the proposed algorithm performs well in the recognition of P300. The combined use of the energy and entropy features and the strategy of metric learning both make great contributions to the improvement of the classification performance.

4.3 Discussion

An advantage of our approach is that no time-consuming operations, such as artifact removal or bootstrapping, are required during preprocessing. This makes it possible for our method to be applied to online tasks. However, for many practical applications, EEG data may be seriously contaminated by various disturbances. In such cases, several operations—especially artifact rejection—have to be applied to the data to obtain a reasonable SNR. Hence, no automatic method developed thus far is good enough for all practical applications.

In the feature-extraction phase of our method, each epoch is decomposed into three levels by wavelet packet transform. We also tried to decompose signals to five or seven levels, which had no significant effects on the results. However, the optimal basis selection of wavelet packet decomposition as well as feature selection features such as Principal Component Analysis (PCA) can be added to our method for data compression in future research.

With regard to future research, exploring wavelet packet energy and entropy as well as other features that can better represent the cognitive component in EEG signals is a feasible way to improve the performance of the classification system. Moreover, for practical applications, since labeled data are usually insufficient for effective classification, a semi-supervised version of the algorithm should be considered. That is to say, both labeled and unlabeled data should be used in the train of the classifier. Finally, to improve the adaptability of the algorithm and reduce training time, online updates of the model should be considered.

5. CONCLUSIONS

In this paper, a metric learning approach was proposed to address the cognitive component classification problem of EEG signals. By learning a Mahalanobis distance metric that captures discriminating information of features from labeled instances, the kernel function of the SVM is redefined. That is to say, before the SVM classifier is applied, instances of different classes are transformed to a new space in which the separation between data points is significantly improved. The overall implementation of the algorithm is easy to understand, and the computational burden is low. Experiments on the BCI speller dataset showed the improved performance of the proposed algorithm. The results showed that the metric learning-based method is suitable to address EEG signal-processing problems.

Nevertheless, there remains scope for improvement in various aspects of the algorithm by way of future work such as data compression and feature selection. To expand the application scope of our algorithm, more experiments on the recognition of P300 or other ERP components are required. In addition, if we consider practical applications, such as crime information identification based on EEG signals, the complex application environment and unpredictable interference will certainly make challenging demands of our algorithm.

REFERENCES

[1] Alain Rakotomamonjy, Vincent Guigue, "BCI Competition III: Dataset II- Ensemble of SVMs for BCI P300 Speller," *IEEE Trans. Biomed. Eng*, vol. 55, no.3, pp. 1147–1154, 2008.

[2] B. Blankertz, K.-R. Mueller, G. Curio, T. Vaughan, G. Schalk, J.Wolpaw, A. Schloegl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M.Schroeder, and N. Birbaumer, "The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials," *IEEE Trans. Biomed. Eng*, vol. 51, pp. 1044–1051, 2004.

[3] J. F. Gao, X. G. Yan, J. C. Sun, and C. X. Zheng, "Denoised P300 and Machine Learning-based Concealed Information Test Method," *Comput. Method Prog. Biol,* vol. 94, pp. 410-417, 2011.

[4] V. Abootalebi, M.H. Moradi, M.A. Khalilzadeh, "A Comparison of Methods for ERP Assessment in a P300-based GKT", *International Journal of Psychophysiology*, vo. 62, pp. 309–320, 2006.

[5] F. Nijboer, E. W. Sellers, J. Mellinger et al., "A P300-based Brain Computer Interface for People with Amyotrophic Lateral Sclerosis," *Clinical Neurophysiology*, vol. 119, no. 8, pp. 1909–1916, 2008.

[6] B. H. Yang, L. Liu, P. Zan, and W.Y. Lu, "Wavelet Packet Based Feature Extraction for Brain-Computer Interfaces", *Int. Conf. Life System Modeling and Simulation (LSMS)*, pp. 19–26, 2010.

[7] X.J. Guo, X.P. Wu, D.J. Zhang, "Motor Imagery EEG Detection by Empirical Mode Decomposition," *International Joint Conference on Neural Networks*, pp. 2619–2622, 2008.

[8] X.Q. Yu, M.S. Xiao, Y. Tang, "Research of Brain-Computer Interface based on the Time-Frequency-Spatial Filter," *Bioin--formaics and Biomedical Engineering*, 2009.

[9] M.Y. Zhao, M.T. Zhou, Q.X. Zhu, "Feature Extraction and Parameters Selection of Classification Model on Brain-Computer Interface," Bioinformatics and Bioengineering, pp.1249–1253, 2007.

[10] J. Sherwood, R. Derakhshani, "On Classifiability of Wavelet Features for EEG-Based Brain-computer Interfaces," *International Joint Conf. Neural Networks*, *Atlanta, Georgia, USA*, pp.14-19, 2009.

[11] V. Bostanov, "BCI Competition 2003—Data Sets Ib and IIb: Feature Extraction From Event-Related Brain Potentials With the Continuous Wavelet Transform and the t-Value Scalogram," *IEEE Trans. Biomed. Eng*, vol. 51, no. 6, 1057-1061, 2004.

[12] T. Wu, G.Z. Yan, B.H. Yang, H. Sun, "EEG Feature Extraction Based on Wavelet Packet Decomposition for Brain Computer Interface," *Measurement*, vol.41, no. 6, pp. 618–625, 2008.

[13] U. Orhanu, M. Hekim, M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Systems with Applications,* vol.38, no.10, pp.13475 一 13481, 2011.

[14] S. GHelmy, T. Al-ani, Y. Hamam, et al. "P300 based brain-computer interface using Hidden Markov Models," *Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing*, IEEE, pp. 127-132, 2008.

[15] S. Yu wen, et al, "Programmable neural processing on a Smart dust for brain-computer interfaces," *IEEE Trans. Biomedical Circuits and Systems*, vol. 4, no. 5, pp. 265-273, 2010.

[16] A. Yazdani, T. Ebrahimi, U. Hoffmann, "Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier," *EMBS*, pp. 327-330, 2009.

[17] M. Kaper, P. Meinicke, U. Grossekathoefer, et al. "BCI competition 2003-data set IIb: Support vector machines for the P300 speller paradigm," *IEEE Trans. Biomed. Eng*, vol. 51, no.6, pp. 1073-1076, 2004.

[18] R. Q. Quiroga, "Quantitative analysis of EEG signals: time-frequency methods and chaos theory," *Institute of Physiology-Medical University Lubeck and Institute of Signal Processing-Medical University Lubeck*, 1998.

[19] K. Q. Weinberger, L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journey of Machine Learning Research*, vol. 10, pp. 207-244, 2009.

[20] X. Q. Zhu, P. H. Gong, Z. S. Zhao and C. S. Zhang, "Learning Similarity Metric with SVM", *International Joint Conference on Neural Networks*, pp. 3342-3349, 2012.

[21] B. D. Ripley, *Pattern Recognition and Neural Networks,* Cambridge, U.K.: Cambridge University Press, 1996.

[22] Farwell LA, Donchin E, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials", *Electroenceph clin Neurophysiol*, vol. 70, pp. 510-23,1988.