An Extensive Review on Feature Extraction for Designing Brain Computer Interface

Bambam Kumar Choudhary[1], **Prof. Anshul Sarawagi**[2]

[1]M.Tech Scholar, Department of Computer Science and Engineering, IES College, Bhopal, M.P., India

[2]Research Guide, Professor, Department of Computer Science and Engineering, IES College, Bhopal, M.P., India

Abstract: ***Electroencephalograph (EEG) is useful modality nowadays which is utilized to capture cognitive activities in the form of a signal representing the potential for a given period. Brain Computer Interface (BCI) systems are one of the practical application of EEG signal. Response to mental task is a well-known type of BCI systems which augments the life of disabled persons to communicate their core needs to machines that can able to distinguish among mental states corresponding to thought responses to the EEG. The success of classification of these mental tasks depends on the pertinent set formation of features (analysis, extraction and selection) of the EEG signals for the classification process.This paper presents a review study of various features transformation techniques for EEG signal, which can be very useful in designing of any type of BCI system.***

Keywords: Brain Computer Interface, Response to Mental Tasks, Feature Extraction, Empirical Mode Decomposition, Electroencephalograph

1. Introduction

The Brain-Computer Interface (BCI) is one of the regions which has sponsored up in developing techniques for assisting neurotechnologies for ailment prediction and manage motion [1, 2, 11]. BCIs are rudimentary geared toward availing, augmenting or rehabilitating human cognitive or motor-sensory characteristic [10, 15]. To capture brain activities, EEG is one of the prevalent technology as it provides signal with high temporal resolution in a non-invasive way [10, 11]. Mental task classification (MTC) based BCI is one of the famed categories of BCI technology which does no longer involve any muscular activities [3] i.e. EEG responses to mental tasks.

Conventional pattern recognition is mainly divided in two main component, feature analysis and pattern classification. Feature analysis is achieved by two step, parameter extraction and feature extraction. In the parameter extraction step, information relevant for pattern classification is extracted from the input in the form of parameter vector. In the parameter extraction step, the parameter vector is transformed to a feature vector. Feature extraction can be conducted independently or jointly with either parameter extraction or classification. When an input data is represented by high dimension, there needs a technique to reduce the dimension of the input data so that only relevant data is used for pattern recognition. Feature extraction is transformation of the data from one space to other space .This may leads to a better representation or reduction of dimension from higher to lower. It makes easy to reduce the complexity of the learning algorithm or classifier. Pattern extraction techniques can further be categorized as

- Linear
- Non-linear

The main linear technique for dimension reduction is principle component analysis based on linear mapping of the data to a lower dimension space in such a way that the variance of data in lower dimension is maximized One approach to cope with the problem of high dimensionality is to reduce the dimensionality by combining feature. There are four aspect of feature extraction [7, 4].

- Feature construction
- Feature subset generation (search strategy)
- Evaluation criterion definition
- Evaluation criterion estimation

These four aspects can be divided of into two categories: feature construction and feature selection. The last three aspects are closely related to feature selection. Feature selection is primly performed to select most relevant and informative feature and remove noise, irrelevant and redundant feature. It can also used for following.

- Data reduction
- Feature reduction
- Improve performance
- Data understandability

2. Principle Component Analysis

The main feature extraction in linear domain is Principle Component Analysis [9]. Principal component analysis (PCA) involves mathematical operation that linear transforms a number of uncorrelated variables to a number of correlated variables, principal component. PCA seeks a projection that best represents the data in least square sense [22, 7]. PCA is simplest form of true -vector based multivariate analysis. PCA is closely related to factor analysis; , some statistical packages deliberately conflate the two techniques. True factor analysis makes different assumptions about the underlying structure and solves vectors of a slightly different matrix. PCA is based on some statistical properties of random variables. Suppose X is given random vector population $(x_1, x_2, \dots x_n)^t$ and spouse the mean vector of the population is denoted by

$$\mu = E[X] \tag{1}$$

And the covariance matrix of the same data set is given by

$$C_x = E\{(x - \mu_x)(x - \mu_x)^T\} \tag{2}$$

The component of C_x is denoted by C_{ii} indicates the covariance between x_i and x_j. The component c_{ij} denoted the variance of x_i. The covariance matrix is always symmetric. From a symmetric matrix one can calculate the orthogonal basis by finding its vectors and values. By ordering the vectors in the order of descending one can create an ordered orthogonal basis with the first vector having largest variance of the data. In this way one can find directions in which dataset have sufficient amount of energy. Let A be a matrix consisting vectors of the covariance as row vector. By transforming a data vector x we get

$$y = A(x - \mu_x) \tag{3}$$

The component of y may be treat as coordinate of orthogonal basis. One can reconstruct original data vector x from y as

$$x = A^T y + \mu_x \tag{4}$$

The data can represent by in terms of only few basis vectors of orthogonal matrix. Let A_k be a row vector of first k vectors of given covariance matrix then above two transforms can be written as

$$y = A_K(x - \mu_x) \tag{5}$$

And x can be represent as

$$x = {A_K}^T y + \mu_x \tag{6}$$

This means that one can projects the original data vector on the coordinate axes having the dimension K and transforming the vector back by a linear combination of the basis vectors. This minimizes the mean-square error between the data.

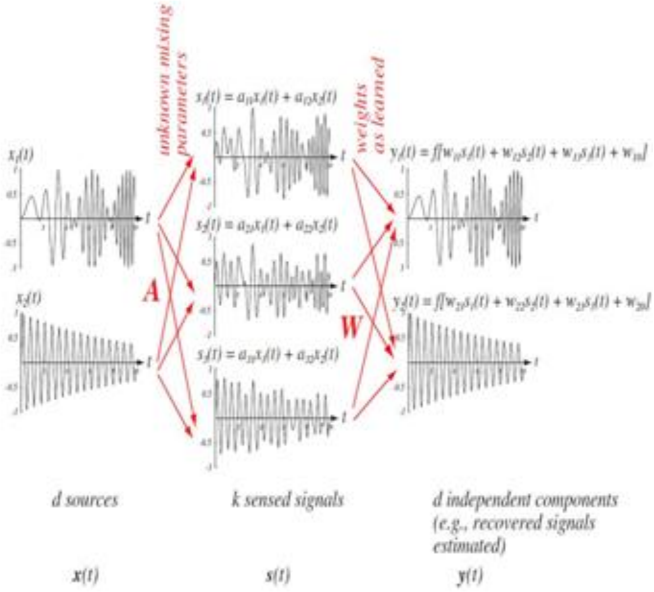
unknown mixing parameters
weights as learned
$x_1(t)$
$x_2(t)$
A
W
$s_1(t) = a_{11}x_1(t) + a_{12}x_2(t)$
$s_2(t) = a_{21}x_1(t) + a_{22}x_2(t)$
$s_3(t) = a_{31}x_1(t) + a_{32}x_2(t)$
$y_1(t) = f[w_{11}s_1(t) + w_{12}s_2(t) + w_{13}s_3(t) + w_{10}]$
$y_2(t) = f[w_{21}s_1(t) + w_{22}s_2(t) + w_{23}s_3(t) + w_{20}]$
t
d sources
k sensed signals
d independent components (e.g., recovered signals estimated)
x(t)
s(t)
y(t)

Figure 1: Geometric object rotation

3. Independent Component Analysis

Independent component analysis (ICA) seeks the directions of feature space where data are most independent from each other, where as other methods of feature extraction like PCA and non linear component analysis seek directions in feature space that best represent the data in sum-square error sense. This goal of ICA can be understood in the domain of Blind Sources pattern [5, 8, 7].

To rigorously define ICA a model is used calla statistical latent variable. Let x be given observed random vector $(x_1, x_2, \dots x_n$ and the component of random vector $s = (s_1, s_2, \dots s_n$ then there should be a linear static transform Was $s = wx$ Into maximally independent component by using some function $F(s_1, s_2, \dots s_n$ of independence [7]. Let us assume that we have a signal $x_i(t)$ of d-components at given time t. In absence of noise, we can write the multivariate density function as

$$p[x(t)] = \prod_i^d p[x_i(t)] \tag{7}$$

If there is k-dimension vector is observed at each moment,then,

$$s(t) = Ax(t) \tag{8}$$

where A is $k * d$ matrix, x is source of signal, s is d-components signal. The main goal of ICA is to extract d component in s that are independent as much as possible. The distribution in the output is related to the distribution

$$p_y(y) = \frac{p_s(s)}{|J|} \tag{9}$$

where J Jacobean matrix[7].

4. Linear Discriminate Analysis

PCA is suitable for finding the components which represents data in feature space but may not useful for discriminating data in different class. PCA only seeks directions that are efficient for representation where as linear discriminating analysis (LDA) seeks directions that are efficient for discrimination. LDA is basically based on ANOVA (analysis of variance) and regression analysis [13, 24, 7]. The ultimate goal of LDA is to maximize this function Spouse that given a set of n- dimensional data sample $x_1, x_2, \dots x_n$. If there is formation of linear combination of the component $y = w^t x$ and a corresponding set of n samples $y_1, y_2, \dots y_n$ now define a function $J(w)$ such that

$$J(w) = \frac{w^t s_B w}{w^t s_W w} \tag{10}$$

Where s_B is the scatter matrix between classes and s_W is the scatter matrix within class and are given by

$$s_B = \sum_{\substack{c\\ \square}} (\mu_c - \bar{x})(\mu_c - \bar{x})^t \tag{11}$$

$$s_W = \sum_{\substack{c\\ \square}} \sum_{\substack{i\in c\\ \square}} (x_i - \mu_c)(x_i - \mu_c)^t \tag{12}$$

is mean of class c Where is the overall mean of data set.

SB and SW are symmetric and positive semi definite but SW is usually nonsingular if $n > d$. In mathematical physics above expression is also known as Rayliegh quotient. We chose W in such a way that maximize J by satisfying this equation. SW=λSB where λ some constant

5. Feature Extraction from Brain Signal

The EEG data is source of measuring brain activities. But capturing EEG data through electrode provide large amount of data. In order to obtain desire output of BCI it is necessary to reduce large number of EEG data values into manageable values [14, 12]. The values can be represented as feature of brain. There are many feature extraction techniques in the literature categories as:

1) Temporal information
2) Frequency based information
3) Hybrid method

5.1 Temporal method

These methods use as feature the temporal variation of signal. These methods can be categories based on various parameters. **Signal Amplitude** This is the simplest form of temporal information. It could be extracted from amplitude of EEG signal. **Autoregressive parameters** In this method it is assume that a signal at time t S (t) is modeled by sum of its previous values of at (t-1), (t-2) â€¦ 1 second.

$$X(t) = a_1X(t-1) + a_2X(t-2) + \cdots + a_pX(t-p) + E_t \quad (13)$$

$$Activity\left(X(t)\right) = VAR\left(X(t)\right) \quad (14)$$

$$Mobility\left(X(t)\right) = \sqrt{\frac{Activity\frac{dX(t)}{dt}}{Activity\left(x(t)\right)}} \quad (15)$$

Where ai is the autoregressive parameter which has to be calculated by various method such as Burg method, Waltch method [6, 11]. **Hjorth parameter** The Hjorth parameter explores the dynamic model of temporal signal X(t) by three measures those are activity, mobility and complexity [12]. Activity(X(t))=VAR(X(t))

$$Complexity = \frac{Mobility\left(\frac{dX(t)}{dt}\right)}{Mobility\left(x(t)\right)} \quad (16)$$

$$A(f) = \sum_{k=0}^{m} a_k e^{-j2\pi fk} \quad (17)$$

5.2 Frequency based Method

As it is known, that EEG signal is made from different type of rhythm (frequencies). Performing different type of mental task makes amplitude of this rhythm very. General frequency method divided into following categories [12].

5.2.1 Band Power Feature

Band power feature are generally computed for several frequencies bands previously determined by according to mental state recognized. Such features are used with success of motor imagery and classification of tasks [12, 23, 20].

5.2.2 Power spectral density method

Power Spectra Density (PSD) depicts distribution of power of signal between frequencies. The square of the signal of Fourier transform gives PSD feature or the feature can be obtained by computing Fourier transform of autocorrelation function of the signal. These methods are used to estimate frequencies and power of signals from noise corrupted measurement called vector method. These are based on an -decomposition of a correlation matrix of the noise corrupted signal[18, 19]. Eigen vector methods give high frequency spectra resolution even signal â€“to noise (SNR) is very low. Those methods are best suited for that signal which is assumed composed from several specific sinusoids signals with noisy. In the literature there are three methods decomposing vector â€“Pisarenko, multiple signal classification (MUSIC), and minimum norm were used for generate power spectral density.

Pisarenko Method

The Pisarenko method is a noise subspace frequency estimator [17, 19]. The method was given by Pisarenko in 1973. The method is used to estimate power spectral density, when expected frequencies contain sharp peaks. The polynomial which contains zeros on the unit circle denoted by A(f). Where

$$S^{\#}a = 0 \quad (18)$$

Where is the coefficient of desired polynomial and m is the order of the filter, A(f).The method uses only the vector corresponding to the minimum value and to calculate spectrum. The method therefore find a such that

$$P_{PISARENKO}(f) = \frac{1}{|A(f)|^2} \quad (19)$$

Where S is signal matrix, # represent conjugate complex transpose, a is Eigenvector of estimated auto correlation matrix [] From the vector corresponding to minimum value, the Pisarenko method determines the signal power spectrum density from the given polynomial and is given by

MUSIC Method

The MUSIC method gives PSD value in a noise subspace as in the Pisarenko method. The difference between the two is that the Pisarenko method gives PSD value corresponding to minimum value whereas the MUSIC method gives PSD correspond to all values in given noise subspace. The method was given by Schmidt (1986) [19]. It reduces the effects of spurious zeros by using the averaged spectra of all the vectors corresponding to the noise subspace. The resultant PSD is given by

$$P_{MUSIC}(f) = \frac{1}{\frac{1}{K}\sum_{i=0}^{K-1}|A_i(f)|^2} \quad (20)$$

Where K is the dimension of noise subspace, is the desired polynomial that corresponds to all the vectors of noise subspace.

5.2.3 Minimum-Norm method

The method was given by Ubeyli and Guler [18, 19]. The Minimum-Norm method causes spurious (imaginary) zeros inside the unit circle and estimates a desired noise subspace vector a from either the noise or signal subspace vectors. The difference between Pisarenko method and the Minimum-Norm method is that the Pisarenko method uses only the noise subspace corresponding to the minimum value whereas the Minimum-Norm method uses linear combination of all noise subspace vectors. The polynomial A(f) is given by

$$A(f) = A_1(f)A_2(f) \quad (21)$$

$$\text{where, } A_1(f) = \sum_{k=0}^{L} b_k e^{-j2\pi fk}, b_0 = 1,$$

$$A_2(f) = \sum_{k=0}^{m-L} c_k z^{-k}, c_0 = 1,$$

Where and are the coefficients of the two polynomial components of A(f). The polynomial has L real zeros on circle of unit radius while has m-L imaginary zeros. So our aim is to minimize the imaginary zeros in , for that we minimize the following parameter Q, given by

$$Q = \sum_{k=0}^{M} |a_k|^2, a_0 = 1. \quad (22)$$

The Minimum-Norm PSD can be measured from

$$P_{MIN}(f,K) = \frac{1}{|A(f)|^2} \quad (23)$$

Where *K* is the dimension of noise subspace.

5.3 Time Frequency representation

Generally BCI has been used signal properties which are lied in both time and frequency domain, both has their own merit and demerit. A method which can be hybrid form of temporal and frequency domain have been used to design of BCI. The advantage of this hybrid time- frequency representation is that as they can capture relatively sudden temporal variation of signal, while still keeping frequency based information [12]. On the other hand in pure frequency based, it is assumed that signal is in stationary state. It can be categories as following

Short Time Fourier Transform (STFT)
Let w be a given windowing function which is non-zero only for short period of time then first w is first multiplied by input signal x(n) and then the Fourier transformation in discrete time STFT X(n,w) of given signal x(n) is as follows

$$X(n,\omega) = \sum_{n=-\infty}^{+\infty} x(n)\,w(n)\,e^{-j\omega n} \qquad (24)$$

Where ω is frequency.

Wavelets
The other method is decomposing of signal into small basic function is wavelet transform. The wavelet transform is a set of small wavelet where a and b are scaled. If is mother wavelet, wavelet transform is given by

$$\Phi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\Phi\left(\frac{t-b}{a}\right) \qquad (25)$$

The wavelet transform of given signal is given by

$$W_x(s,u) = \int_{-\infty} x(t)\,\Phi_{u,s}(t)\,dt \qquad (26)$$

Where is wavelet transformed and x is given signal, s and u are scaling and translating factors respectively. The main advantage of wavelets is that it allows us to analyze the signal at different scales [16, 12]. **Amplitude-phase coupling measure** There exist a coupling phenomena due to different cognitive acts need integration and communication of different regions of mind. So there is a need a method which can couple these different areas according to the phase and amplitude of the signal of the different regions. There are two methods for signal coupling, one of them is linear and the other one is non-linear. Linear signal coupling is described by cross correlated in time domain coherence in frequency domain. The nonlinear coupling can be described by a nonlinear regressive (NLR) coefficient and phase locking value (PLV). These values differentiates phase and amplitude in there nonlinear coupling and could be generate classification of mental task [21]. The phase coupling measure extract relevant feature for the classification of BCI framework. The combination of PLV and power spectral density can be successfully recognize up to three cognitive tasks. Amplitude coupling measure is given by NLR coefficient which is as follow

$$h^2 = \frac{\sum_{n=1}^{N}(y_n-\langle y\rangle)^2-\sum_{n=1}^{N}(y_n-\hat{\mu}_{y|x}(x_n))^2}{\sum_{n=1}^{N}(y_n-\langle y\rangle)^2} \qquad (27)$$

where is the is the linear piecewise approximation of the regression for signal x and y, curve <y> denotes the average values over the N points. The value of gives the strength of coupling between two signals which lies between 0 and 1. This can be linear and nonlinear, for liner relationship and for nonlinear .

Phase coupling measure (PLV) For given signal x(t) the analytic signal is given by

$$\eta_x(t) = x(t) + i\tilde{x}(t) = A_x(t)\,e^{i\theta_x(t)} \qquad (28)$$

Where is the Hilbert transform of x(t) and is given by

$$\tilde{x}(t) = \frac{1}{\pi}\,p.v.\int_{-\infty}\frac{x(\tau)}{t-\tau}\,d\tau \qquad (29)$$

p.v. is Cauchy integral principle value. The instantaneous phase can be calculated by

$$\theta_x(t) = \arctan\left(\frac{\tilde{x}(t)}{x(t)}\right) \qquad (30)$$

For discrete signals, the phase locking value is given by

$$PLV = \left|\frac{1}{N}\sum_{n=1}^{N} e^{i\Delta\theta(n)}\right| \qquad (31)$$

PLV is equal to average length of unit vector in one window. PLV lies between 0 and 1 [21].

6. Conclusion

A suitable set of features from EEG signal is an utmost requirement of designing of any BCI system. This set can be found with the proper feature extraction method. Feature extraction can be carried out either transformation of the signal one domain to another domain only, or parametric features formation after transforming of the signal one domain another. This paper presented a review study of feature extraction in BCI field.

References

[1] Charles W Anderson, Erik A Stolz, and Sanyogita Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *Biomedical Engineering, IEEE Transactions on*, 45(3):277–286, 1998.

[2] F Babiloni, F Cincotti, L Lazzarini, J Millan, J Mourino, M Varsta, J Heikkonen, L Bianchi, and MG Marciani. Linear classification of low-resolution eeg patterns produced by imagined hand movements. *Rehabilitation Engineering, IEEE Transactions on*, 8(2):186–188, 2000.

[3] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering*, 4(2):R32, 2007.

[4] K Chen, V Kvasnicka, PC Kanen, and Simon Haykin. Supervised and unsupervised pattern recognition: Feature extraction and computational intelligence [book review]. *IEEE Transactions on Neural Networks*, 12(3):644–647, 2001.

[5] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[6] Guido Dornhege, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Combining features for bci. In *Advances in Neural Information Processing Systems*, pages 1139–1146, 2003.

[7] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[8] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[9] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[10] Laura Kauhanen, Tommi Nykopp, Janne Lehtonen, P Jylanki, Jukka Heikkonen, Pekka Rantanen, Hannu Alaranta, and Mikko Sams. Eeg and meg brain-computer interface for tetraplegic patients.

Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 14(2):190–193, 2006.

[11] Zachary A Keirn and Jorge I Aunon. A new mode of communication between man and his surroundings. *IEEE transactions on biomedical engineering*, 37(12):1209–1214, 1990.

[12] Fabien Lotte. *Study of electroencephalographic signal processing and classification techniques towards the use of brain-computer interfaces in virtual reality applications*. PhD thesis, INSA de Rennes, 2008.

[13] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999.

[14] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

[15] Gert Pfurtscheller, Christa Neuper, Alois Schlogl, and Klaus Lugger. Separability of eeg signals recorded during right and left motor imagery using adaptive autoregressive parameters. *Rehabilitation Engineering, IEEE Transactions on*, 6(3):316–325, 1998.

[16] Lei Qin and Bin He. A wavelet-based time–frequency analysis approach for classification of motor imagery for brain–computer interface applications. *Journal of neural engineering*, 2(4):65, 2005.

[17] Petre Stoica and Arye Nehorai. Study of the statistical performance of the pisarenko harmonic decomposition method. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, volume 135, pages 161–168. IET, 1988.

[18] Elif Derya Übeyli and İnan Güler. Features extracted by eigenvector methods for detecting variability of eeg signals. *Pattern Recognition Letters*, 28(5):592–603, 2007.

[19] Elif Derya Übeyli. Implementing eigenvector methods/probabilistic neural networks for analysis of eeg signals. *Neural networks*, 21(9):1410–1417, 2008.

[20] Tao Wang, Jie Deng, and Bin He. Classifying eeg-based motor imagery tasks by means of time–frequency synthesized spatial patterns. *Clinical Neurophysiology*, 115(12):2744–2753, 2004.

[21] Qingguo Wei, Yijun Wang, Xiaorong Gao, and Shangkai Gao. Amplitude and phase coupling measures for feature extraction in an eeg-based brain–computer interface. *Journal of Neural Engineering*, 4(2):120, 2007.

[22] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[23] [23] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.