Relevant Feature Extraction by Combining Independent Components Analysis and Common Spatial Patterns for EEG Based Motor Imagery Classification

Aarti Bhalla
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi India – 110067

R. K. Agrawal
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi India - 110067

Abstract – **Brain Computer Interfaces (BCI) use brain's signal such as electro-encephalogram (EEG), reflecting the user's intention. Event related desynchronisation or synchronisation correspond to attenuation or enhancement of power in certain frequency bands over the sensorimotor cortex of the brain, due to actual or imagined movement of body parts. Such aspects of the signal can be better represented in terms of relevant features to achieve accurate mental task classification for higher performance of BCI. The existence of inflicted noise and other artifacts in EEG signals and high dimensionality of the dataset hinder the classifier's performance in BCI. In order to overcome these obstructions, it is impertinent to derive independent noise free components from raw signal and diminish the length of the extracted feature vector. For optimising the classification task, the amalgamation of ICA technique and CSP is explored to obtain the most relevant and discriminative subset of features. The effectiveness of this methodology, in comparison to other methods found in literature, is demonstrated through experimental results on BCI Competition datasets.**

Keywords- Brain Computer Interfaces, Mental tasks, Feature Extraction, Feature Selection, Common Spatial Patterns, BandPower, Independent Component Analysis.

I. INTRODUCTION

Controlling the world with the mind has always been a dream of humanity. Establishing direct communication between the brain and computer has been an agenda of scientific research for a long time. Brain is one of the most vital organs of human body, which controls the co-ordination of human muscles and nerves (Wang *et al.,* 2011). The brain consists of complex structure of billions of neurons for carrying out various body organ functions, movements, control and communication. It also receives stimulus from various sense organs and sends responses through neural pathways to these sense organs. Communication is a basic human need which involves more than just speaking and listening. But severe neurological diseases such as amyotrophic lateral sclerosis (ALS), brain stem stroke, locked-in condition etc. restrict a person's ability to communicate emotions, thoughts and basic needs. Such patients usually have an active brain with normal brain activities. These people rely on alternative ways of communication. Processing and analysing bio-signals using software techniques are playing role since 1960s to provide physicians with fast and accurate means of diagnosis (Gandhi *et al.*, 2011). The research works (Licklide, 1960, Engelbart, 1962) have emphasized the potential of a symbiotic relation between human and computer. A **brain computer interface (BCI)** is a communication system by which a person can send messages or request for basic necessities via her brain signals without using peripheral nerves and muscles. The electrical signals occurring in the brain due to neuron activity carry information for the purpose of communication with a computing device. Thus BCI provide an augmentation to motor disabled people. For establishing communication between a computer and a brain, one has to consider their differences too i.e. firstly the functioning of the brain is slower than a computer, secondly, the brain performs in parallel, whereas a sequential computer performs sequentially and is efficient for the most complicated functions with very high precision. BCI is an interesting area of research which requires interdisciplinary knowledge of subjects such as Biology, Mathematics, Engineering, Physiology, Psychology and Computer Science. Due to improved understanding of functioning of the brain, low-cost computing devices, and advancement in signal processing techniques, the field of BCI has received a great interest in the past 20 years.

Among various existing techniques of brain signal acquisition, electro- encephalo graph (EEG) is the most commonly used. EEG signal is a complex mixture of brain signals emitted from different cortices of the brain (Wolpaw et al, 2012) and is often corrupted with artifacts i.e. EOG, EMG etc. and external noise. These signals taken raw are unable to capture a person's mental state.

Raw EEG signals have a weak spatial resolution due to volume conduction. This becomes a problem when the relevant signals are weak while in the same frequency band, other sources produce strong signals. For single trial EEG analysis, BCI system is calibrated to the specific characteristics of each user by calculating subject specific spatial filters. These spatial filters are designed in a way that the variances of the out-coming signals carry the most discriminative information. Common Spatial Pattern (CSP) technique is one of highly successful spatial technique which helps in estimating spatial filters to analyse multichannel data.

Common Spatial Pattern(s) (CSP) is used to quickly estimate relevant information from the data related to oscillatory processes. It has been applied for detection of major brain rhythm modulations (e.g. mu, beta), e.g. related to stress/ relaxation, sensori-motor imagery, workload aspects, visual processing vs. idling and other idle-rhythm-related problems, or thought recognition. CSP is a supervised signal enhancement method that detects patterns in the EEG by incorporating the spatial information of the EEG signal. The CSP algorithm exploits features such as event-related synchronization and desynchronization localized in the (sensori-) motor cortex.

Given the recordings from two class distributions, the aim of CSP algorithm is to find spatial filters (directions) which maximize variance for one class and minimize variance for the second class simultaneously. It is stepwise implemented as signal (pre-) processing (spatial/spectral filtering) followed by feature extraction and machine learning. A frequency filter is applied first, followed by spatial filter, followed by log-variance feature extraction and lastly a classifier applied to the extracted features. With spatial filtering, the original channels are mapped down to a small number of channels (usually 4-6) having maximally informative variance w.r.t. to the prediction task. The CSP filters can be computed from the covariance matrices of each class by solving a generalized eigenvalue problem.

Some of CSP's features and limitations include the following. It is simpler to implement, faster to execute and robust. A *priori* selection of subject-specific frequency bands is not required. However, these bands must be known for methods like band-power and frequency-estimation (Guger *et al* 2000b). CSP achieves satisfactory result for synchronous (cue based) BCIs but is less effective for asynchronous BCIs (Nicolas-Alonso *et al.*, 2012). The time related variations and correlation among frequency bands in the signal are not captured by the CSP method. The working of CSP depends on spatial resolution since it utilizes many electrodes for enhanced performance (Pfurtscheller *et al* 2000, Guger *et al* 2000b). The electrode positions must remain unchanged across all trials and sessions for CSP method to give genuine results. (Ramoser *et al* 2000).

A major drawback in the CSP application is that it is highly sensitive to artifacts and noise in the EEG. The spatial filters are calculated from the covariance matrices of

trials having large dimesions with contrastingly small samples. A trial containing artifacts can severely modify the filters (Guger *et al* 2000b, Ramoser *et al* 2000).

Independent Component Analysis (ICA) generates source components that are independent of each other, out of a given mixture of signals. The artifact/noise component can be subtracted from such source components. Thus the effect of unwanted artifacts is diminished in the processed signal which is devoid of any artifacts/ noise. The model obtained from these source signals will be robust to changes since it is built from data which has been purified by ICA technique. ICA removes the artifacts in a BCI system by first segregating the observed signal into its source signal and artefacts and then eliminating the artefact component. However, suppression of artifacts may distort the power spectrum of the underlying cognitive function. Besides, ICA requires the artefacts to be independent with respect to the EEG signal which is not true in most of the cases.

All such source signals may not be relevant for the classification task. Also having multiple source signals result in a large number of features which is not desirable due to the 'curse of dimensionality' (Bellman, 1961) which taxes the performance of a classifier. Thus, dimensionality reduction is necessary to keep only the most important and informative source signals. We can apply CSP method to independent components to evaluate their relevance or for improved robustness against artifacts. With CSP, only the most important components would then be retained. In this way, both goals of signal processing – artifact removal and dimensionality reduction are achieved using the combination of CSP and ICA technique.

In this paper, we first apply ICA to obtain independent components from the non stationary raw EEG signal and then apply CSP to find a projection of channels that maximizes the variance for one class and minimizes the variance for the other. This approach effectively deals with the 'curse of dimensionality' problem often found in high dimension datasets. Besides artifact removal, this approach avoids over fitting the classifier for a given test dataset and gives near optimal results in general. This study attempts to validate this approach and compare with existing CSP based approaches through statistical tests. The criterion is to find out the approach giving comparable accuracy but employing lesser features than other approaches. Not only such approach be less complex but more immune to overfitting for a given dataset. To the best of mine knowledge, an elaborative comparison among various CSP based techniques from that aspect has not been done on BCI data till date. Section 2 describes the above techniques and related methods. In section 3 the datasets and experimental setup are examined. Section 4 validates and compares the performance among various approaches. Lastly, a brief conclusion is drawn.

II. CSP AND RELATED WORKS

A matrix notation is suitable to represent the EEG signal. Let X_1^i and X_2^i be the raw EEG data of trial i having dimension $N \times T$ where N denotes number of channels and T denotes number of samples in time for class I and II respectively. The covariance matrices of class I and II are given as

$$\sum_1^i = X_1^i X_1^i / trace(X_1^i X_1^{i\,T}) \quad \text{and} \quad \sum_2^i = X_2^i X_2^i /$$
$$trace(X_2^i X_2^{i\,T}) \text{ respectively.}$$

The normalized covariance matrices averaged over trials of class I and II are given as

$$\sum_1 = <\sum_1^i> \text{trials} \quad \text{and} \quad \sum_2 = <\sum_2^i> \text{trials}$$
$$(1)$$

A matrix W and diagonal matrix D with elements in [0, 1] is determined to maximise

$$W\sum_1 W^T = D \quad \text{such that} \quad W(\sum_1 + \sum_2)W^T = I \quad (2)$$

The rows of matrix W are the spatial filters, whereas the columns of matrix W^{-1} are Common Spatial Patterns. Using this projection matrix W the EEG recordings X_1^i are decomposed into

$$Z_i = W X_1^i \quad (3)$$

A large corpus of CSP-based approaches aim at achieving enhanced control over spectral filtering. Several other methods exist to adapt the spectrum to a process of interest, among others common spatio-spectral patterns, common sparse spectral spatial pattern, r^2-based heuristics, automated parameter search, and manual selection based on visual inspection. Several of these methods have been shown to give approx. comparable results. An alternative and competitive method, especially when there are complex interactions between frequency bands and time periods are to be modelled is the dual-augmented Lagrange paradigm which learns both spatial filters and their relative weightings in a unified cost function.

The common spatio-spectral pattern (CSSP) filter is an extension of the CSP filter (Lemm *et al.*, 2005) that involves time delay embedding. The CSSP's transform is given by:

$$Z = WX + W_\tau X_\tau = \tilde{W} \begin{bmatrix} X \\ X_\tau \end{bmatrix} \quad (4)$$

Where, $\tilde{W} = [W \ W_\tau]$ is a CSSP matrix in which the number of channels get doubled. However, it requires choice of a frequency band and hyper-parameter τ.

The success of BCI based on CSP depends on the choice of frequency bands and hyper-parameter τ which is difficult to adjust. In order to overcome this, Novi *et al.* (2007) used sub-band CSP filters with different non

overlapping frequency bands and combined their output linearly. This model does not require prior knowledge of frequency bands and fine tuning of hyper-parameters.

FBCSP

In filter-bank CSP (Ang *et. al*, 2008), a set of CSP filters is learned for each of several time/frequency filtering methods, followed by log-variance feature extraction. The extracted features are concatenated over all selected spectral filters before machine learning. Due to the problem of over fitting, even though FBCSP cannot replace CSP yet it is beneficial for oscillatory processes having different spatial topographies, jointly active in different frequency bands.

For a given prediction task i.e. recognizing complex event-related dynamics in response to a stimulus, their concerted behaviour must not be ignored. With filter-bank CSP capturing oscillations in various time windows rather than frequency windows is possible. In a scenario involving workload measurements, FBCSP can deduce relevant interactions between frequency bands e.g. mu/alpha.

SpecCSP is used when the frequency and location of some (conjectured) oscillatory process is not known beforehand. While CSP uses "a priori" fixed bands, this method can learn subject specific frequency bands that exhibit the oscillatory processes of interest. However, it requires a suitable wide ranged spectral filter to give improved results.

This method iterates to perform optimization of spectral and spatial filters alternatively and subsequently extracts log-variance features from the processed signal. These features are then fed to a classification algorithm such as, LDA. By focussing on certain frequency bands i.e. alpha band, the parameters such as frequency prior and the spectral filter can be tuned to extend the considered spectrum to high-gamma oscillations. One can also adapt the time window of interest and the learner component (e.g., a good alternative choice being logistic regression).

Band Power

This method exploits event-related synchronization and desynchronization, localized in the motor cortex via per-channel logarithmic band power. (Pfurtscheller and da Silva, 1999) It considers log-variance as features of the signal. The resulting feature vectors are then passed along to the learner component. Although a primitive method by modern standards, log-BP is simple enough for oscillatory processes.

It typically creates a relatively low-dimensional feature space that can be operated by almost any classifier, including non-linear ones such as SVMperf and QDA. One needs to tune parameters such as the length of the data epoch and the choice of a frequency band (defaulting to

motor imagery time scales and frequency ranges), both of which can also be found via a small parameter search.

However, neither any complex temporal variation gets detected in the oscillations nor the interaction between multiple frequency bands. Band Power does not include data-adaptive signal processing as it uses a non-adaptive spatial filter, the surface Laplacian, and a non-adaptive spectral filter. This limitation can be overcome by selecting an adaptive machine learning/ spectral filtering component.

ICA

Another spatial technique called the **Independent Component Analysis** (ICA) statistically disintegrates a given mixture of signals into a set of independent source signals without any prior information about the signal. ICA is based on the assumption that the unknown source signals are mutually independent and these are generated by different cognitive activities or artefacts in the brain. The observed EEG signal $x(t)$ can be written in terms of its source signal $s(t)$ as:

$$\mathbf{x}(t) = \mathbf{A}(\mathbf{s}(t)) + \mathbf{n}(t) \qquad (5)$$

where A is the mixing matrix, and $n(t)$ is random noise. The number of source signals is less than the number of observed signals.

The ICA algorithm comprises the computation of estimate of $s(t)$ by inverting A thereby mapping the mixture $x(t)$ to the source space. Furthermore, if we assume that the observed data is noiseless then the noise term $n(t)$ can be eliminated to simplify the above equation. Certain algorithms exist that can derive source signal $s(t)$ and matrix A from $x(t)$ e.g. Infomax, FastICA, SOBI.

The channels constructed from ICA are usually better than the original signal. The components have less correlation than the raw channels, which implies improved covariance matrices. Sparse features can be derived from ICA due to assumption that only a few components carry the relevant localized information while the rest carry irrelevant data for the optimization task. A high level of semantic meaning is associated with components than channels i.e. line noise, eye blinks, muscle activity, heart-beat etc. Prior knowledge of specific brain locations can facilitate in deriving the semantics of the components localized using techniques such as, beam forming, dipole fitting, sparse Bayesian learning etc.

There exist several variants of ICA, with AMICA (Makeig *et al.*, 2008) being the most applicable one. Its striking feature is that it allows the source signal densities be based on a flexible model. Thus, for EEG data, this method obtains more statistically independent solutions than the other existing approaches. Also, AMICA can efficiently capture the non-stationarities in the signal by the usage of multiple models. Another commonly used variant for EEG data is Infomax (Bell and Sejnowski, 1995). It executes in lesser time than AMICA and is easier to function.

A simple version of ICA implementation is FastICA (A. Hyvaerinen, 1999). It converges relatively faster, but the solution is not that good in terms of quality in comparison to extended Infomax or AMICA. However, it can be a good alternative for iterative computations (cross-validation) due to its enhanced speed. For small data sets, KernelICA can be relied upon since it follows a kernel approach that requires high execution time but uses a flexible model of source densities as in AMICA.

III. DATASETS AND EXPERIMENTAL SETUP

The datasets are taken from BCI competition 3 dataset IVa [13] that comprises brain signal values from five healthy subjects namely (aa, al, av, aw, ay). The subjects were given visual cues for 3.5 s to perform three types of motor imagery i.e. left hand, right hand and right foot movement. For each subject, 280 trials were recorded using Brain Amp amplifiers and a 128 channel Ag/AgCl electrode cap from ECI. 118 EEG channels were measured at positions of the extended international 10/20-system. Signals were band-pass filtered between 0.05 and 200 Hz and then digitized at 1000 Hz with 16 bit (0.1 uV) accuracy. The actual cognitive states for only some right hand and right foot trials are available in a vector whereas the cognitive states for other trials are to be determined by the proposed model.

The continuous dataset is converted to epoched dataset with each epoch having certain number of trials. The length of the data epoch and the choice of a frequency band (defaulting to motor imagery time scales and frequency ranges) are the parameters that are most commonly tuned to the task, both of which can also be found via k-fold cross validation search. The main user-configurable parameters are the selection regions in time and frequency and the machine learning component.

Since the dimensionality of the feature space is large and since complex interactions may be present, a more complex classifier than the default LDA may be necessary to learn an appropriate model. However, with only little calibration data, the risk of over fitting is amplified, so that the performance should always be compared to standard CSP (and Spec-CSP). Another reason is that complex (relevant) interactions between different frequency bands are seemingly rarely observed in practice.

The classifier shall be sparse, since we assume that only few of the independent components in the data will carry relevant information. A Bayesian variant of the classifier will be used (using automatic relevance determination) to avoid the need for time-consuming regularization. As an alternative, the l1-regularized variant of the classifier could be used.

To sustain the frequency domain characteristics of EEG signals, fast Fourier transform (FFT) was taken to convert the signal from time domain to frequency domain. ICA step was then applied to remove artefacts and enhance the signal to noise ratio, and EEG was filtered through an 8-30 Hz bandpass filter. This is because the bands corresponding to the Mu and Beta (sensorimotor) rhythms lie within this frequency range. CSP algorithm is applied to

obtain the most optimal and discriminative spatial patterns. The number of features for classification were obtained through 5 fold cross validation for each value of r.

Linear Discriminant Analysis (LDA) was used to classify the mental tasks with the help of extracted features.

IV. RESULTS

Table 1 shows comparison results of various CSP based techniques obtained by 10 fold cross validation on test data. The values denote classification accuracy with number of features utilised (in parenthesis). To obtain a fair evaluation statistical test is applied to compare various models. The methods which perform statistically significantly different are highlighted in bold. * denotes values with sparse classifier.

It can be easily observed that the combination of CSP-ICA is superior to other methods as far as the result of statistical t-test is concerned. This method utilises least number of features while providing comparable accuracy to the best in 3 out of five cases. The method ExpCSP (experimental CSP) has 10 features and uses sparse Logistic regression for classification.

Other techniques like spectral CSP and FBCSP give higher accuracy but at the same time require more number of features, which is not the objective here. ICA-CSP uses only two features and gives the best accuracy in two out of five subjects 'al' and 'av'. It also gives good results for subjects 'aw' and 'ay'. The objective of having comparable accuracy but, using fewer features is clearly attained by ICA-CSP making it an effective method for BCI datasets.

The Band power and ExpCSP methods perform poorly on most datasets thereby indicating their non applicability to non-stationary EEG dataset. As an exception, Band power gives the best result for subject 'aw'. This can be owed to its suitability to the frequency bands of 'aw'.

V. CONCLUSION

Brain Computer Interface assists severely physically challenged people to communicate with the help of electroencephalogram (EEG) signal. Features derived from multiple channels result into a large sized feature vector but the available number of samples is small. It may hinder the classifier's performance for mental task classification. In this paper, we have investigated and compared seven well-known multivariate spatial filter based feature extraction techniques to determine a minimal subset of relevant and discriminative features. A reduced set of significant features not only decreases the time complexity to learn a model but also reduces the bias of a classifier to estimate error on a specific dataset. The performance of various methods based on CSP paradigm is evaluated in terms of classification accuracy and those methods

that perform statistically significantly different with fewer features for most of the subjects are identified. Experimental results demonstrate that classification accuracy improves considerably with the use of linear classifier that utilizes temporal information of the signal.

Table 1. Classification accuracy of different spatial filters for EEG Data

Subjects	CSP	FBCSP	Band Power	SpecCSP	ExpCSP	ICA-CSP (epoched)
aa	0.84(2)	0.85(6)	0.70*(8)	**0.92**(6)	0.69(8)	0.83(2)
al	**0.99**(6)	**0.99**(6)	0.97(10)	**0.99**(6)	0.96(6)	**0.99**(2)
av	**0.87**(6)	0.79(6)	0.71(8)	0.77(8)	0.70(10)	**0.87**(2)
aw	0.91(6)	0.85(6)	**1.00***(10)	0.93(6)	0.75(10)	0.97(6)
ay	**1.00**(10)	**1.00**(8)	0.72(6)	0.95(10)	0.43(10)	0.95(6)

REFERENCES

[1] Blankertz, B., Dornhege, G., Krauledat, M., Mïller, K., and Curio, G. "The non-invasive Berlin Brain-Computer interface: Fast acquisition of effective performance in untrained subjects." NeuroImage 37, 2 (Aug. 2007), 539-550.

[2] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus- Robert Mueller. "Optimizing spatial filters for robust EEG single-trial analysis." IEEE Signal Process Mag, 25(1):41-56, January 2008

[3] Pfurtscheller, G., and da Silva, L. "Event-related EEG/MEG synchronizaion and desynchronization: basic principles." Clin Neurophysiol 110 (1999), 1842-1857.

[4] Buzsaki, G., "Rhythms of the brain" Oxford University Press US, 2006

[5] Fukunaga K., "Introduction to Statistical Pattern Recognition" Academic Press, Computer Science and Scientific Computing Series, 1990

[6] Ramoser, H., Gerking, M., and Pfurtscheller, G. "Optimal spatial filtering of single trial EEG during imagined hand movement." IEEE Trans. Rehab. Eng 8 (2000), 446, 441.

[7] Tomioka, R., Dornhege, G., Aihara, K., and Mueller, K.-R. "An iterative algorithm for spatio-temporal filter optimization." In Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course 2006.

[8] Ryota Tomioka, Guido Dornhege, Guido Nolte, Benjamin Blankertz, Kazuyuki Aihara, and Klaus-Robert Mueller "Spectrally Weighted Common Spatial Pattern Algorithm for Single Trial EEG Classification", Mathematical Engineering Technical Reports (METR-2006-40), July 2006.

[9] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Mïller. "Spatio-spectral filters for improving classification of single trial EEG." IEEE Trans Biomed Eng, 52(9):1541-1548, 2005.

[10] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Mïller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," IEEE Transactions on Biomedical Engineering, vol. 53, no. 11, pp. 2274-2281, 2006.

[11] Kai K. Ang, Zhang Y. Chin, Haihong Zhang, Cuntai Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface" , In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 2390-2397, June 2008.

[12] Onton J & Makeig S. "Broadband high-frequency EEG dynamics during emotion imagination." Frontiers in Human Neuroscience, 2009.

[13] http://sccn.ucsd.edu/wiki/BCILAB, Open Source Matlab Toolbox for Brain-Computer Interface research.

[14] Bell, A. J., and Sejnowski, T. J. "An information-maximization approach to blind separation and blind deconvolution." Neural Comput. 7, pp 1129-1159, June 1995.

[15] A. Hyvaerinen. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis." , IEEE Transactions on Neural Networks 10(3), pp 626-634, 1999.

[16] S. Makeig, J. A. Palmer, K. Kreutz-Delgado, and B. D. Rao, "Newton Method for the ICA Mixture Model", In Proceedings of the 33rd IEEE International Conference on Acoustics and Signal Processing (ICASSP 2008), Las Vegas, NV, pp. 1805-1808, 2008.

Author's profile

Aarti Bhalla is currently doing Ph.D at School of Computer and Systems Sciences, Jawaharlal Nehru University. Her research areas of interest are BCI and Pattern Recognition. She is carrying out her research work under the guidance of Prof. **R.K. Agrawal**, School of Computer and Systems Sciences, Jawaharlal Nehru University.