2010 International Conference on Pattern Recognition

A Covariate Shift Minimization Method to Alleviate Non-Stationarity Effects for an Adaptive Brain-Computer Interface

Abdul Satti[1], Cuntai Guan[2], Damien Coyle[3] and Girijesh Prasad[4]
[1,3,4]*Intelligent Systems Research Centre, University of Ulster, UK*
[2]*Institute for Infocomm Research (I^2R), Singapore*
satti-ar@email.ulster.ac.uk, ctguan@i2r.a-star.edu.sg, dh.coyle@ulster.ac.uk, g.prasad@ulster.ac.uk

Abstract

The non-stationary nature of the electroencephalogram (EEG) poses a major challenge for the successful operation of a brain-computer interface (BCI) when deployed over multiple sessions. The changes between the early training measurements and the proceeding multiple sessions can originate as a result of alterations in the subject's brain process, new cortical activities, change of recording conditions and/or change of operation strategies by the subject. These differences and alterations over multiple sessions cause deterioration in BCI system performance if periodic or continuous adaptation to the signal processing is not carried out. In this work, the covariate shift is analyzed over multiple sessions to determine the non-stationarity effects and an unsupervised adaptation approach is employed to account for the degrading effects this might have on performance. To improve the system's online performance, we propose a covariate shift minimization (CSM) method, which takes into account the distribution shift in the feature set domain to reduce the feature set overlap and unbalance for different classes. The analysis and the results demonstrate the importance of CSM, as this method not only improves the accuracy of the system, but also reduces the classification unbalance for different classes by a significant amount.

1. Introduction

The goal of a brain-computer interface (BCI) is to provide a communication pathway between the brain and an external device by converting electrophysiological or metabolic brain activity into control signals for applications and devices. Commands directly encoded in neurophysiological signals provide an effective communication corridor for patients suffering from motor impairments, severe cerebral palsy and spinal cord injuries [12].

A major challenge for BCI research is the alteration in brain activity occurring continuously in association with diverse behavioral and mental states [3, 8]. The period of time within which spectral properties of the EEG waveform can be considered stable is reasonably small, with estimates ranging from 4s to 1 min [2]. Explicitly, the two main sources of non-stationarity reported in [4, 10] are; i) differences between the samples extracted from a training session and the samples extracted during an online session, ii) alteration in the users brain activity during online operation. As a result, the general hypothesis that the signals sampled in the training set follow a similar probability distribution to the signals sampled in a future test phase, or in an online situation, is violated [11]. This covariate shift is even more severe, as the number of sessions increase, resulting in unpredictable and deteriorating system performance.

Here, we present an unsupervised adaptation method in which the feature set distribution is analyzed to estimate the covariate shift between the feature distribution of the training and test data. A careful examination reveals an increase in the classification unbalance between classes as a result of the covariate shift, thus increasing the bias towards one of the classes and a drop in classification accuracy (CA). The classification unbalance is quantified using a confusion matrix index (CI), cf. section 3.2. The shift in distribution is estimated using a least squares fitting polynomial and removed from the feature set. As a result, the CI decreases and the CA improves.

The remainder of the paper is organized into three sections. Section 2 contains details on the datasets and acquisition procedure. Section 3 explains the methodology for preprocessing, feature extraction, feature selection, covariate shift minimization (CSM) and classification. Results are presented and discussed in section 4.

2. Data Acquisition and Configuration

The datasets used in this analysis is dataset IIb (9 subjects, 5 sessions each) provided for the BCI competition IV [1, 6]. Three bipolar recordings (C3, Cz, and C4) were made with a sampling frequency of 250 Hz. The cue-based paradigm consisted of two classes, namely left hand (class 1) and right hand (class 2). A detailed description

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.34

of the data acquisition procedure is available at [1,6]. For this analysis, we have used a constant time window of 0.5 to 2.5 seconds after the presentation of the instructional cue in a trial for feature extraction and classification.

3. Methodology

Firstly subject specific-frequency bands are selected using particle swarm optimization [7,9] to spectrally filter the raw EEG. Common spatial pattern (CSP) filtering [4] is then applied to the raw EEG datasets to spatially filter the spectrally filtered data. An optimally designed CSP filter is used to produce a surrogate data space which improves the separability of two populations. The variances of the CSP filtered channels are analysed for covariate shift. A CSM method is then applied to minimize the covariate shift in the features and classified using linear discriminant analysis (LDA).

3.1. Common Spatial Patterns

The CSP filtering involves linearly projecting the multichannel EEG data into a new space by a weighted summation of the appropriate channels. This projection is based on the simultaneous diagonalization of the covariance matrices a and b from both classes so that the eigenvalues of covariance matrices sum to 1 [4].

$$\Sigma_a.W = (\Sigma_a + \Sigma_b).W.D \tag{1}$$

Here, the diagonal matrix D contains the (generalized) eigenvalues of Σ_a and the column vectors of W are the filters for the CSP projections. Each EEG trial X can be projected using the mapping matrix W as $Z = WX$. By construction, the variance for a left movement imagination is largest in the first row of $Z = [z_1, z_2, z_n]$ and decreases for the subsequent rows. The opposite is the case for a trial with right motor imagery. The appropriate number of eigenvectors p from both sides is chosen as filters; generally a value of p between 1 and 3 from either side of the eigenvector matrix is optimal [4]. For feature extraction, the normalized variance of the surrogate data channels is used, i.e.,

$$f_i = v_j / \sum_{j=1}^{2p} v_j \; and \; v_j = var(z_j) \tag{2}$$

Here, we would like to draw the reader's attention to the redundancy in the normalized features. As for simplicity, to focus on non-stationarity, this analysis is based on only two features for each trial, i.e., p=1, therefore, only one normalized feature is sufficient as the other feature is entirely redundant. This is because one of the features is the mirror image of the other pivoted around the mean of 0.5 (e.g., if one feature value is 0.1 then the other is 0.9, therefore, both are at a distance of 0.4 from the mean of 0.5 i.e. mirrored across the mean).

3.2. Confusion matrix index

Here, we define a confusion matrix index (CI), for a binary class problem, which is derived from the confusion matrix to explain the bias or the unbalance in the classification. This is computed as:

$$CI = k^{-1}.(a_1 - a_2) \tag{3}$$

where, a_1 and a_2 are the correctly classified trials for class 1 and 2, respectively, and k is the total number of trials. The value of CI lies between -1 and +1. CI close to -1 reflects a bias towards class 2, i.e., CA of class 2 is much greater than class 1, and vice versa for the other class.

3.3. Covariate Shift Minimization

In this section we investigate the shift in distribution and effects of this alteration in distribution on the classification accuracy and CI. We use a least square fitting polynomial to estimate and predict the amount of covariate shift. This polynomial model determines which input features of the distribution drives responses and in what direction. The equation for a polynomial fit on features f, of order h, over N trials, is defined as,

$$y_i = a_0 + a_1 f_i + a_2 f_i^2 + ... + a_h f_i^h \tag{4}$$

i.e., $Y = F.A$. The coefficient matrix A is designed to satisfy the linear equation; $A = (F^T.F)^{-1}.F^T.Y$, where T denotes transpose of a matrix.

The matrix A is continuously updated over N previous trials. As a result, the polynomial value at a current trial n represents the shift using only a finite causal subset of the previous N trials. This confines the adaptation to only activate after N-1 trials, i.e., no adaptation is performed on the initial N-1 trials, which are only used to estimate the distribution trend. Here, we use a constant subset of N=10 trials as a baseline to show the effectiveness of the method. However, particular care should be taken in choosing the trials subset as a very small N can cause the polynomial to overfit as it can exaggerate minor fluctuations in the feature distribution, conversely, a much larger subset can make the polynomial estimate excessively resistant to major shifts in distribution.

Figures 1B and 1D show feature distributions for the training and the testing sessions. It is clearly evident that the feature distribution has considerably changed from the training to testing session. This is reflected in the CI value of -0.67, i.e., a bias towards class 1. Figures 1A and 1C show the polynomial fitting of the order 3 to the training and testing features distribution. A polynomial approximation when applied to the training feature distribution shows very little deviation from the optimum linear

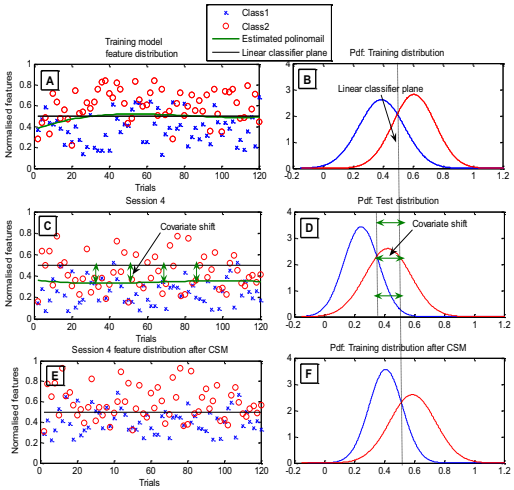

Figure 1. Feature distribution and pdf for subject 5.

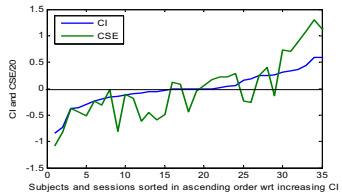

Figure 2. Covariate shift estimate (CSE) and CI relational plot.

classifier plane indicating that the model generated on the training samples fits the training data precisely. However, the polynomial approximation on the test session shows a much higher deviation, explaining a significant alteration in the test feature distribution. The covariate shift estimate (CSE), Ψ, for a session with k trials is calculated as the area enclosed by the polynomial, y, and the linear classifier plane l:

$$\Psi = \int_0^k y(x)dx - \int_0^k l(x)dx \tag{5}$$

Figure 2 shows a strong correlation between the CSE and the CI. As the covariate shift increases, the CI increases. Based on this analysis, we aim to reduce the covariate shift, which in turn is expected to reduce the confusion index and improve the accuracy and reliability for the proceeding multiple test sessions.

The estimated polynomial y, at trial n, which is an indication of the shift, is removed and the feature for the respective trial is readjusted by adding the common mean, μ_0, of the training feature distribution:

$$f_n^{'} = f_n - y_n + \mu_0 \tag{6}$$

Figure 1E shows the feature distribution after the shift is removed. As a result the LDA decision boundary of the adjusted test session feature distribution now coincides with the training distribution (cf. figure 1B and 1F).

The LDA decision boundary for feature vector x is calculated as,

$$x.\Sigma^{-1}(\mu_1 - \mu_2) + \Sigma^{-1}(\mu_1 - \mu_2).\mu_0 \tag{7}$$

Here, μ_1, μ_2 , μ_0 and Σ are the means for classes 1 and 2, the joint distribution mean and the joint covariance of the feature vectors respectively. As Σ, μ_1, μ_2 and μ_0 are scalars in (7) due to the use of single feature per trial for simplicity, the decision boundry can be simplified to $x + \mu_0 = 0$, i.e., the joint mean of the training distribution acts as the discrimination threshold. This justifies the CSM method described by (6). Just to note this method can be extended to any number of features per trial where CSM can be applied to each feature vector independently to remove the shift in distribution for each feature vector explicitly.

4. Results and discussion

Table 1 compares the CA and CI for all the subjects over 5 sessions with and without CSM. Out of 36 test sessions for 9 subjects, CSM has helped to improve the CA in 31 cases ($p \ll 0.001$ achieved using a Repeated Measures One-way Analysis of Variance [5] RM-ANOVA). The mean accuracy for all the subjects except the subject 3 has improved (no significant difference for subject 3). An insight into the feature distribution of subject 3 reveals a high feature set overlap for different classes. Even though the CI has decreased from 0.59 to 0.17, the distribution overlap is irremovable as the features severely overlap (pdf overlap of 91% for two classes, cf. Figure 3).

The mean CI for 9 subjects has been decreased significantly using CSM for all the subjects from 0.23 to 0.08 on average (RM-ANOVA: $p \ll 0.001$). A notable improvement is the decrease in the CI for subject 5 from 0.51 to 0.02, therefore, increasing the CA by 8.8%. Here, the reader's attention should be drawn to the analysis of the spectrally filtered variances of the original channels. Figure 4 shows the relative increase in the subject's performance as the normalized difference between the mean variance on channel C3-C4 and Cz increases. The plot

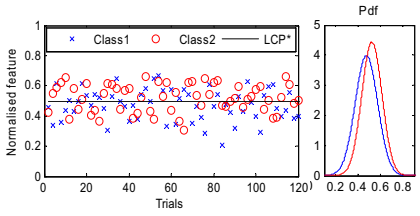

Figure 3. Feature distribution for subject 3 session 3.

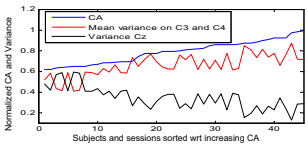

Figure 4. Normalized CA and variance relational plot.

Table 1. CA comparison

	CA (No CSM) [%]					CA (CSM) [%]				
	Sessions					Sessions				
Sub*	2	3	4	5	M**	2	3	4	5	M**
1	65.8	76.9	71.5	65.7	**69.5**	68.2	79.4	73.2	68.8	**72.4**
2	54.7	56.3	66.8	58.8	**58.2**	57.5	63.8	69.4	65.7	**64.1**
3	65.8	56.3	61.5	65	**61.9**	66.3	55	58.8	66.3	**61.6**
4	89.3	98.2	95.7	95.7	**94.9**	91.3	100	98.2	96.3	**96.4**
5	72.2	74.4	73.2	71.3	**72.5**	78.8	80.7	86.9	78.8	**81.3**
6	76.5	75	75.8	80.7	**76.2**	75	80	81.3	81.3	**79.4**
7	63.5	87.5	79.9	84.4	**78.8**	65.7	89.4	80	85	**80.1**
8	70.2	83.2	80.2	86.9	**79.5**	69.8	83.8	83.2	91.9	**82.2**
9	64.9	80	85	76.9	**76.5**	67.8	81.9	88.8	83.8	**80.6**
M**	**70**	**76.4**	**77**	**76.2**	**73.9**	**71.2**	**79.4**	**80**	**79.8**	**78.6**

$*Subjets, **Mean$

Table 2. CI comparison

Subjects	1	2	3	4	5	6	7	8	9	**Mean**
No CSM	0.27	0.48	0.59	0.07	0.51	0.24	0.09	0.21	0.15	**0.23**
CSM	0.03	0.07	0.17	0.02	0.02	0.09	0.02	0.20	0.07	**0.08**

indicates that the performance of the subjects is near random (CA∼=50-60%) when the variance on Cz is almost equal to or greater than the variance on C3 and C4. The performance of the BCI system is superior when this variance relationship factor increases. This information can be used not only to predict the subject's performance in screening sessions, but also for channel selection for a multiple channel system. This is a topic of future research.

Using CSM, a more accurate system with low CI has resulted in a balanced BCI performance, which is important for the application of un/semi-supervised adaptation methods. Future work will also be aimed at analyzing the semi-supervised learning approach where confidence bounds would be used to label the online trials to adapt and reconfigure the classifier model continuously. The semi-supervised learning approach would be applied in series with the CSM as, importantly, CSM provides an unbiased and a balanced classification output to the learning module for more accurate and unbiased label estimates for learning.

References

[1] http://www.bbci.de/competition/iv/. 2008.

[2] T. Dikanev, D. Smirnov, R. Wennberg, J. L. Velazquez, and B. Bezruchko. EEG nonstationarity during intracranially recorded seizures: statistical and dynamical analysis. *Clinical Neurophysiology*, 116(8):1796–1807, 2005.

[3] D. Gribkov and V. Gribkova. Learning dynamics from nonstationary time series: analysis of electroencephalograms. *Physical Review E*, 61(6):6538–6545, 2000.

[4] C. Guger, H. Ramoser, and G. Pfurtscheller. Real-time EEG analysis with subject-specific spatial patterns for abrain-computer interface (BCI). *IEEE transactions on rehabilitation engineering*, 8(4):447–456, 2000.

[5] Y. Hochberg and A. Tamhane. *Multiple comparison procedures*. Wiley New York, 1987.

[6] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller. Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):473–482, 2007.

[7] A. Satti, D. Coyle, and G. Prasad. Optimal frequency band selection with particle swarm optimization for a brain computer interface. In *Proceedings of the Workshop on Evolutionary Computing Lecture Series by pioneers*, pages 75–78, 2008.

[8] A. Satti, D. Coyle, and G. Prasad. Continuous eeg classification for a self-paced bci. In *NER '09, 4th International IEEE/EMBS Conference*, pages 315 –318, 2009.

[9] A. Satti, D. Coyle, and G. Prasad. Spatio-spectral & temporal parameter searching using class correlation analysis and particle swarm optimization for a brain computer interface. In *IEEE international conference on Systems, Man and Cybernetics*, pages 1731–1735, 2009.

[10] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K. R. Muller. Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13–R23, 2006.

[11] M. Sugiyama, M. Krauledat, and K. Muller. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research*, 8:985–1005, 2007.

[12] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.