ERROR ENTROPY BASED ADAPTIVE KERNEL CLASSIFICATION FOR NON-STATIONARY EEG ANALYSIS

S. R. Liyanage[1], C.T. Guan[2], H. H. Zhang[2], K. K. Ang[2], J. -X. Xu[1] and T. H. Lee[1]

[1] National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
sidath@nus.edu.sg, {ctguan, hhzhang, kkang}@i2r.a-star.edu.sg, {elexujx, eleleeth}@nus.edu.sg

ABSTRACT

The performance of Brain-Computer Interface (BCI) applications are sometimes hindered by non-stationarity in the EEG data from sessions on different days. This paper proposes an algorithm for adaptive training of a SVM classifier to address the non-stationarity in EEG by adapting the kernel to data from subsequent sessions. The kernel width parameter of the kernel function of the SVM classifier is adapted using an information theoretic cost function based on minimum error entropy (MEE). An experiment is performed using the proposed method on EEG data collected without feedback from 12 healthy subjects in two sessions on separate days. The results using the proposed method yielded a mean accuracy of 75%, which is significantly better compared to the baseline result of 67% without kernel adaptation (P=0.00029).

Index Terms—Brain–computer interface (BCI), electroencephalography (EEG), classification, adaptation.

1. INTRODUCTION

Brain-Computer Interfaces (BCIs) are communication systems that enable subjects to send commands to computers by using only their brain activity [1]. Non-stationarity arising from high variability of EEG signals is a major obstacle in EEG-based BCI systems. Non-stationarity has been found to be linked to various factors such as, changes in the physical properties of the sensors, variability in neurophysiological conditions, psychological parameters, ambient noise and motion artifacts [2-4].

The importance of addressing session to session non-stationarity has been widely recognized in the BCI community. Various signal processing and learning methods such as, Bayesian transduction, active learning and distribution matching have been proposed [3-6]. Stationary Subspace Analysis (SSA) [4] is another unsupervised learning method that finds subspaces in which data distributions stay invariant over time.

Current research addressing non-stationarity also includes methods that adapt the classifiers using the knowledge from empirical data [7-9]. These methods include adaptation of LDA and SVM classifiers which are the commonly used classification methods in BCI [10]. Adaption of LDA involves updating the statistical parameters such as mean, covariance and bias [7]. Adaptive SVM methods include least square based methods with various penalty functions [8,9].

All these adaptive methods use minimization of error based on the classification output to optimize some parameter in the classifiers [7-10]. In this type of adaptations, the error is under the control of the parameters of the adaptive system because of the error depends on the true labels which is a function of the parameters been adapted.

Error entropy criterion takes into account the amount of information in the error distributions. Therefore minimization of error entropy considers the error distributions rather than error values. Error entropy based adaptive systems have been applied in designing adaptive filters [11-13]. However, the use of the error entropy to adaptation of kernel classification has not been attempted.

In this work we propose to use the error entropy to adapt the width of the Gaussian kernel of the SVM classifier. Adapting the classifier parameters have been found to produce faster adaptive systems than adapting the classifier models [7,14]. A subset of data from the later session is used as adaptation data to estimate an error entropy based cost function which is minimized by adapting the kernel width. Positive results were obtained for the proposed method on motor imagery EEG data collected on different days.

2. METHOD

The data from the initial session is used first to generate an initial model for the classifier after the basic preprocessing steps of bandpass and spectral filtering. Adaptation data from subsequent session is used to optimize the kernel width parameter. Figure 1 summarizes the proposed method. The pseudo code of the proposed method is shown in Figure 2 for further clarification.

The initial training data from the first session and adaptation data from later session are subjected to pre-processing steps of bandpass filtering at 8-30Hz. Initial

training data is spatially filtered by the Common Spatial Patterns (CSP) method [15,16]. Adaptation data on the other hand use the CSP projection matrix created on the initial data.

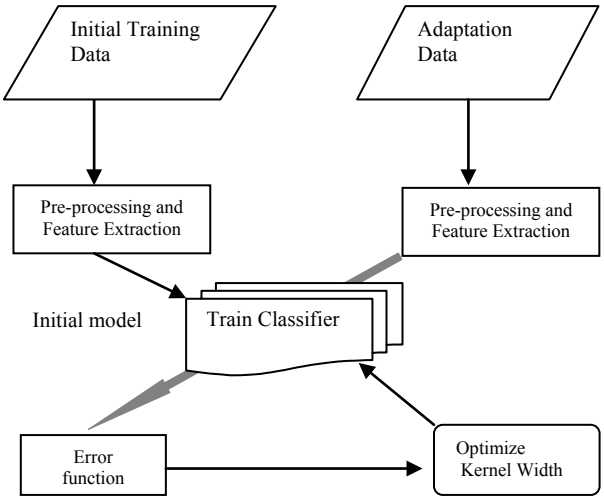

Figure 1: Proposed Method

The SVM classifier maximizes the margin of separation between two classes based on the assumption that it improves the classifier's generalization capability [10]. They map the input (x) into a high-dimensional feature space ($z = \phi(x)$) and construct an optimal classification hyperplane defined by $w \cdot z - b = 0$, where b is the bias. The optimal hyperplane is found by solving the primal problem,

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$s.t. \quad y_i(w \cdot z_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \ \forall i,$$

where x_i is the ith sample and $y_i \in [-1,+1]$ is the class label. This problem is solved in its dual form,

$$\max \quad \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \ \forall i, \quad \sum_i y_i \alpha_i = 0,$$

where $k(x_i,x_j)$ is the kernel function. The regularization parameter C, determines the tradeoff between minimizing the training error and minimizing model complexity. For the Gaussian kernel,

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right).$$

In this paper we adapt the kernel width parameter σ which defines the non-linear mapping from the input space to the high dimensional feature space. The initial classifier model is trained only on the training data from first session. Adaptation data is used to iteratively update the classifier kernel based on the error function. The error function indicates the error margin of the SVM classifier. For a given adaptation data sample, if the predicted class label is different from the actual class, the distance from the margin is multiplied by the predicted class value $y_p \in [-1,+1]$ to obtain the error function value as shown in Equation (1),

$$e = \begin{cases} \text{if } y_p \neq y_o, y_p \times d_p \\ \text{otherwise, } 0. \end{cases} \quad (1)$$

where y_p is the predicted label, y_o is the actual label and d_p is the distance from the classification hyperplane.

The KL divergence based cost function measures the difference in the estimated error and the actual error. We study the effect of adaptively training the classifier on the adaptation data from the second session by optimizing the kernel width of the parameter to minimize the KL divergence based cost function.

Inputs:

Initial Training data

Adaptation data

Output:

Classification model

Algorithm

1. Band-pass filter Initial Training data
2. Spatial filter Initial training data
3. Extract initial features
4. Train initial classification model
5. Band-pass filter Adaptation data
6. Extract features from Adaptation data
7. Feed features to classifier
8. Calculate error entropy
9. Calculate cost function value
10. Adapt the Kernel width of Kernel
11. Repeat step 7 until all adaptation data are used

Figure 2: Pseudo code of the proposed method

2.1. Error Entropy Criterion

The goal of adaptation using error entropy criterion (EEC) is to remove as much uncertainty as possible from the error signal [11]. This can be achieved by calculating the entropy of the error and minimizing it with respect to the free

parameters. The error entropy minimization can be achieved using information theoretic estimators.

Principe et al. developed estimators of information theoretic quantities based on Information Potential (IP), which is the mean of the probability density function of data and happens to be the integrand of Renyi's quadratic entropy [12]. Renyi's quadratic entropy of the error is defined as

$H_2(e) = -\log V(e)$, where $V(e) = E[p(e)]$ is the expected error.

Hence, Renyi's quadratic entropy is a monotonic function of the negative of $V(e)$. The logarithm is dropped as it does not change the location of the stationary point of the cost function for adaptation. The minimization of entropy corresponds to maximization of $V(e)$.

An efficient method to maximize $V(e)$ is to use estimators of information theoretic quantities. Minimizing the Kullback–Leibler divergence between the true and estimated probability distribution functions of error, denoted $f(e)$ and $\hat{f}_\sigma(e)$, as a function of the kernel width σ [13].

2.2. Minimizing Kullback–Leibler divergence for kernel Width Adaptation

The estimators of information theoretic quantities like entropy are based on Parzen kernels. Therefore, a kernel needs to be selected to estimate the pairwise interactions between samples.

In this criterion, kernel width controls the smoothing introduced by a kernel function used for non-parametric estimation of the probability density function from samples, as in Parzen windows [17].

The kernel width is considered as a parameter that can be adapted in a way that the discriminant information or the Kullback–Leibler loss between the estimated density (using the kernel) and the true density is minimized. In other words, the kernel width is adapted with its own cost function in a way that the estimated error distribution resembles the true error distribution as closely as possible, based on Kullback–Leibler divergence.

The objective is to minimize

$$D_{KL}(f||\hat{f}_\sigma) = \int f(e) log\left(\frac{f(e)}{\hat{f}_\sigma(e)}\right) de, \qquad (2)$$

where the subscript σ denotes the dependency of estimated probability distribution function $\hat{f}_\sigma$ on the kernel width. The equation (2) can be re-written as

$$D_{KL}(f||\hat{f}_\sigma) = \int f(e) log(f(e))\, de - \int log(\hat{f}_\sigma(e))\, f(e) de = \int f(e) log(f(e))\, de - E[log(\hat{f}_\sigma(e))]. \qquad (3)$$

where E is the expectation operator over the true distribution of e.

The first term in equation (3) is independent of the kernel width. Therefore, minimizing $D_{KL}(f||\hat{f}_\sigma)$ with respect to σ is equivalent to maximizing the second term $E[\log(\hat{f}_\sigma(e))]$. Which can be interpreted as the cross-entropy of the estimated probability distribution function, and the true probability distribution function. Using the sample estimator for the expectation operator for a Gaussian Kernel the objective function becomes

$$\hat{J}_{KL}(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N-1}\sum_{j=1,j\neq i}^{N} G_\sigma(e_i - e_j)\right). \qquad (4)$$

where N is the window of samples used to estimate density of the error, for a Gaussian kernel with width σ.

Taking the derivative of objective function in equation (4) with respect to kernel width σ results in,

$$\frac{\partial J_{KL}(\sigma)}{\partial\sigma} = E\left[\frac{\partial \hat{f}_\sigma(e)/\partial\sigma}{\hat{f}_\sigma(e)}\right] = E\left[\frac{\sum_{i=n-L}^{n-1}\exp\left(-(e-e_i)^2/2\sigma^2\right)\left((e-e_i)^2/\sigma^3 - 1/\sigma\right)}{\sum_{i=n-L}^{n-1}\exp\left(-(e-e_i)^2/2\sigma^2\right)}\right]. \qquad (5)$$

Using the equation (5) the update rule for kernel size can be formulated as,

$$\sigma_{n+1} = \sigma_n + \gamma\frac{\partial J_{KL}(\sigma)}{\partial\sigma}, = \sigma_n + \gamma E\left[\frac{\sum_{i=n-L}^{n-1}\exp\left(-(e-e_i)^2/2\sigma^2\right)\left((e-e_i)^2/\sigma^3 - 1/\sigma\right)}{\sum_{i=n-L}^{n-1}\exp\left(-(e-e_i)^2/2\sigma^2\right)}\right].$$

By evaluating the operand at the current sample of the error while dropping the expectation operator results in an approximation of the gradient which can be used as an efficient update rule,

$$\sigma_{n+1} = \sigma_n + \gamma E\left[\frac{\sum_{i=n-L}^{n-1}\exp\left(\frac{-(e_n-e_i)^2}{2\sigma_n^2}\right)\left(\frac{(e_n-e_i)^2}{\sigma_n^3} - \frac{1}{\sigma_n}\right)}{\sum_{i=n-L}^{n-1}\exp\left(\frac{-(e_n-e_i)^2}{2\sigma_n^2}\right)}\right]. \qquad (6)$$

The update rule in equation (6) is iteratively applied until all adaptation samples are considered. The updated kernel is applied for classification of test samples.

3. EXPERIMENT

The motor imagery data used for the analysis was collected using a Nuamps EEG acquisition hardware (http://www.neuroscan.com) with unipolar Ag/AgCl electrodes, digitally sampled at 250 Hz with a resolution of 22 bits for voltage ranges of ± 130mV. EEG signals from 22 scalp positions, mainly covering the primary motor cortices bilaterally were recorded. The sensitivity of the amplifier has been set to 100μV.

A total of 12 healthy subjects were recruited for the study. Ethics approval and informed consent were obtained. Two subjects chose to perform left hand motor imagery while the remaining 10 subjects chose to perform on the right hand. The subjects were instructed, in the form of visual cues displayed on the computer screen, to perform kinaesthetic

motor imagery of the chosen hand, and rest during the background rest condition.

EEG data were collected without feedback in two sessions from each subject on separate days. In the first session, two runs of EEG data were collected from a subject while performing motor imagery of the chosen hand and background rest condition. In the second session, three runs of EEG data were collected on another day while performing motor imagery of the chosen hand and background rest condition. Each run lasted approximately 16 minutes that comprised 40 trials of motor imagery and 40 trials of rest condition.

The motor imagery data collected during first session were used as training data for learning CSP spatial filters and the initial classifier, and first half of motor imagery data from the later session was used as adaptation data. The second half of motor imagery data from the later session was used as test data.

4. RESULTS AND DISCUSSIONS

The results obtained for the data collected are shown in Table 1. The twelve subjects are denoted as A1 to A12. The mean accuracies and standard deviations calculated for all the subjects are denoted as mean and S.D. in the table. The baseline classification uses a SVM classifier with a static Kernel and uses all the training data and the adaptation data for training the classifier. In the proposed method the data collected during first session and the adaptation data were used to train the classifiers. The second half of motor imagery data from the later session was used as test data.

Subject	Baseline	This Method	Increment
A1	60.5	68.3	7.8
A2	58.3	67.5	9.1
A3	51.1	55.9	4.7
A4	63.9	79.4	15.5
A5	64.2	74.3	10.1
A6	83.3	88.7	5.4
A7	79.4	79.4	4.5
A8	93.6	93.6	0.0
A9	65.5	79.6	14.1
A10	54.7	61.9	7.1
A11	50.5	65.9	15.3
A12	79.7	85.0	5.3
Statistics			
Mean	67.07	75.0	
S.D.	13.82	11.33	
P		**0.00029**	

Table 1: Comparative classification accuracy rate (%) results on the test data sets. P-value denotes the result of pairwise t-test against the baseline.

The observed mean baseline accuracy is 67%. The baseline result was compared against the results obtained using the proposed Kernel width adaptation method. Pairwise t-test was carried out between the baseline results and the proposed method. The mean accuracies from the proposed Kernel width adaptation method are found to be significantly higher than the baseline at a confidence level of 0.05.

The increments made by the proposed adaptive method over the baseline are shown in the fourth column of Table 1. Only one subject does not show any improvement in accuracy. All other subjects show substantial increments in accuracy.

5. CONCLUSION

In this study, a novel algorithm to adapt the Kernel width parameter of SVM classifier to improve classification of non-stationary EEG data is proposed. In the proposed algorithm, the width parameter of the Kernel of the classifier is iteratively adapted based on Information theoretic cost function to minimize the KL divergence between the estimated and the actual error distributions.

The proposed method is applied on EEG data collected without feedback from 12 healthy subjects in two sessions on separate days. The results using the proposed method yielded statistically significant improvement in classification accuracies on non-stationary EEG data across sessions compared to the baseline without kernel adaptation.

Future work based on this approach would include adaptation of Kernel mean and other parameters to optimize the adaptation. It would also be interesting to investigate how other cost functions would perform.

REFERENCES

[1] N. Birbaumer, T. Hinterberger, A. Kubler, N. Neumann, "The thought-translation device (ttd): neurobehavioral mechanisms and clinical outcome", *IEEE Trans. Neural Systems Rehab. Eng.* 11, pp. 120–123, 2003.

[2] A.Y. Kaplan, S.L. Shishkin "Application of the change-point analysis to the investigation of the brain's electrical activity", *Nonparametric statistical diagnosis: Problems and methods*, pp. 333-388, 2000.

[3] J. Pascual, C. Vidaurre, and M. Kawanabe, "Investigating EEG nonstationarities with robust PCA and its application to improve BCI performance", *International Journal of Bioelectromagnetism, 13*, pp. 50-51, 2011.

[4] P. von Bunau, F.C. Meinecke, S. Scholler, K.R.Mueller, "Finding stationary brain sources in EEG data", In: *Proceedings of the 32nd Annual Conference of the IEEE EMBS*, 2010, pp. 2810–2813.

[5] J.Quinonera-Candela, M.Sugiyama, A. Schwaighofer, N. D. Lawrence, *Dataset Shift in Machine Learning,* MIT Press, 2009.

[6] M. Kawanabe, W. Samek, P. von Bunau, and F. Meinecke, "An Information Geometrical View of Stationary Subspace Analysis", In T. Honkela, W. Duch, M.Girolami, and S. Kaski, (eds), *Artificial Neural Networks and Machine Learning - ICANN 2011,* Springer Berlin / Heidelberg , pp. 397-404, 2011.

[7] C. Vidaurre, M. Kawanabe, P. V. Bünau, B. Blankertz, K. R. Müller, “Toward Unsupervised Adaptation of LDA for Brain-Computer Interfaces”, *IEEE Transactions on Bio-medical Engineering*, 58(3):, pp 587 -597, 2011.

[8] Liu, Jingli, et al. "A weighted Lq adaptive least squares support vector machine classifiers–Robust and sparse approximation", *Expert Systems with Applications,* 38.3, pp. 2253-2259, 2011.

[9] G.L.Grinblat, "Solving nonstationary classification problems with coupled support vector machines", *IEEE Transactions on Neural Networks,* 22.1, pp. 37-51, 2011.

[10] Y. Li, K.K. Ang, and C.T. Guan, “Digital Signal Processing and Machine Learning”, Graimann, Bernhard, Brendan Allison, and Gert Pfurtscheller, eds., *Brain-computer interfaces: Revolutionizing human-computer interaction.* Springer, 2011.

[11] D.Erdogmus, and W. Liu. "Adaptive Information Filtering with Error Entropy and Error Correntropy Criteria", *Information Theoretic Learning* (2010): pp.103-140.

[12] J.Principe, J.Fisher, D.Xu, *Information theoretic learning, in: Unsupervised Adaptive Filtering*, Wiley, NewYork, 2000, pp.275–282.

[13] Abhishek Singh and Jose C. Principe, "Information theoretic learning with adaptive kernels", *Signal Proc.* , 91(2011)203–213

[14] A. Schlogl, C. Vidaurre, and K.-R. Mller, “Adaptive Methods in BCI Research - An Introductory Tutorial”, In: B.Graimann, B.Allison, G. Pfurtscheller*, Brain-Computer Interfaces Revolutionizing Human-Computer Interaction*, Springer, 331-355, 2010.

[15] H. Ramoser, J. Mueller-Gerking, and G. Pfurtscheller, “Optimal spatial filtering of single trial EEG during imagined hand movement”, *IEEE Transactions on Rehabilitation Engineering*, 8(4), 2000, 441–446.

[16] K. K. Ang, Z.Y. Chin, C.Wang, C.T.Guan and H.H. Zhang, “Filter Bank Common Spatial Pattern algorithm on BCI Competition IV Datasets 2a and 2b, Frontiers in Neuroscience”, 6, 2012.

[17] E. Parzen, "On the estimation of a probability density function and the mode", *Ann. Math. Stat.* , vol. 33, no. 2, pp. 1065-1076, Sept. 1962.