CLUSTER IMPURITY AND FORWARD-BACKWARD ERROR MAXIMIZATION-BASED ACTIVE LEARNING FOR EEG SIGNALS CLASSIFICATION

Huijuan Yang, Cuntai Guan, Kai Keng Ang, Yaozhang Pan and Haihong Zhang

Institute for Infocomm Research,
Agency for Science, Technology and Research (A*STAR), Singapore 138632
Email: {hjyang, ctguan, kkang, yzpan and hhzhang}@i2r.a-star.edu.sg

ABSTRACT

This paper investigates how to apply active learning for the classification of motor imagery electroencephalography (EEG) signals to boost the performance for small training size. A new criterion is proposed to select the most representative and informative queries. The candidates are firstly chosen from the samples close to the center of the cluster that has the highest impurity of classes. A predefined number of such candidates and classifiers are forwardly buffered. Subsequently, the query is chosen such that the buffered classifiers can backward maximize the classification errors on labeled data. Experimental results conducted on the BCI competition IV data set IVb show the superior performance of the proposed active learning scheme, which is on average 5.12% higher in accuracy than that of the passive method by choosing the training size from 28 to 112.

Index Terms— Active learning, clustering, motor imagery EEG signals, cluster impurity, forward-backward error maximization.

1. INTRODUCTION

Machine learning techniques have been successfully applied to classify brain signals in brain computer interface, which provides an effective communication channel between the paralyzed people and the outside world. The evoked potential changes of motor imagery can eventually be translated into commands to operate the external devices [1]-[4]. How to design an adaptive learning algorithm to fully utilize the dynamics of the data and the abundant unlabeled data is a major issue for online implementation [3]. The capability of achieving good performance for small training set makes active learning attractive for EEG signal classification. The goal of active learning is to choose the most informative and representative samples to boost the performance [5]-[10]. A general active learning process starts with an initially labeled data set L=$\{x_1, x_2, ..., x_n\}$ with labels Y=$\{y_1, y_2, .., y_n\}$, where $Y \in \{0, 1\}$ for two-class classification. Further, there is an unlabeled data set U=$\{x_{n+1}, x_{n+2}, ..., x_{n+m}\}$, generally, $n \ll m$. In each iteration, the learning algorithm will pick a sample x_k, where $x_k \in U$ based on certain criteria and asked the oracle to label the sample. The labeled sample is then put in set L and removed from set U, i.e., L=$L \bigcup \{x_k\}$ and U=$U \backslash \{x_k\}$, where "$\bigcup$" and "$\backslash$" denote the "set union" and "set minus", respectively. Basically, there are several ways to select the query samples. *Query by uncertainty* selects those samples close to the decision hyperplane of the current classifier [6][7][8]. *Query by committee* chooses the samples which are assigned to different classes by a committee of classifiers [5][8]. While *Query by error reduction* chooses the samples to minimize errors of the new classifiers [9]. To build an effective classifier for small training set and to fully utilize the informative unlabeled data, we propose a novel criterion to select the query samples by combining *query by committee* and *query by uncertainty*. Specifically, a predefined number of candidates are selected from the centers of those clusters that have the highest impurity in the forward direction. The same number of the most recently used classifiers are buffered in the backward direction. Subsequently, the query sample is chosen such that the classification errors for the labeled samples by employing the buffered classifiers can be maximized.

2. PROPOSED METHODS

2.1. Preprocessing: Filtering and Feature Extraction

Let's denote the EEG signals as S=$(s_{ijk})^{n_c \times n_t \times n_r}$, where n_c, n_t and n_r denote the number of channels, samples and trials, respectively. The time segment of 0.5s to 1.5s from onset of the visual cue is used. The signal is divided into n_s overlapping sub-bands ranging from 4Hz to 36Hz, which is filtered by an mth order low-pass digital Chebyshev Type II filter to obtain the band-pass filtered signal: E=$(e_{sijk})^{n_s \times n_c \times n_t \times n_r}$. The filtered signal is then used to compute Common Spatial Pattern (CSP) features. CSP decomposes the EEG signal such that the variances of the new time series are optimal to discriminate the two classes [1][2][4]. Let's denote Σ_1 and Σ_2 as the covariance matrices of the band-pass filtered EEG signal E for the respective motor imagery action.

$$W^T \Sigma_1 W = \varpi_1 \;\; and \;\; W^T \Sigma_2 W = \varpi_2 \qquad (1)$$

where ϖ represents the diagonal. Scaling W such that $\varpi_1+\varpi_2=I$, which can be achieved by solving the generalized eigenvalue problem.

$$\Sigma_1 w = \lambda \Sigma_2 w \tag{2}$$

The large λ_j corresponds to spatial filter w_j that yields high variance in one motor imagery action and low variance in another action. Hence, the two task-specific activations can be differentiated. The resultant filter $(w_{jj})^{n_c \times n_c}$ is used to filter the data in each trial for different sub-bands, which gives $\widetilde{Z}=w_{jj}^T E^{n_c \times n_t}$. The final CSP features are obtained by $F=log(var(Z_p))$, where Z_p are the first and last m rows of $\widetilde{Z}$, m=1 is selected in implementation. This results in a feature length of $l_f=n_s \times 2m$ for each trial. Hence, the CSP feature vectors for labeled (L) and unlabeled (U) sets are ${F_{jl}}^{(n_l \times l_f)}$ and ${F_{jl}}^{(n_u \times l_f)}$, respectively, where n_l and n_u denote the number of trials for labeled and unlabeled sets and l=1,2,...,l_f. It is worth noting that the samples in this paper refer to the trials which are represented by feature vectors.

2.2. Proposed Active Learning Scheme

How to choose the query samples is critical in active learning. In order to build a classifier to include the most informative and representative samples, we propose a criterion to firstly choose the candidates based on the cluster impurity and the distances from the samples to the centers of the selected cluster. Secondly, we buffer a predefined number of candidates (N_s) that satisfy the criterion, and the most recent N_s classifiers for each candidate in the forward direction. Finally, the query sample is chosen such that the classification errors using the buffered classifiers for the labeled set (including the buffered candidates) are maximized in the backward direction. The proposed scheme is hence named as "cluster impurity and forward-backward error maximization-based active learning (CIFBEM-AL)", which is illustrated in Fig. 1 and described as follows.

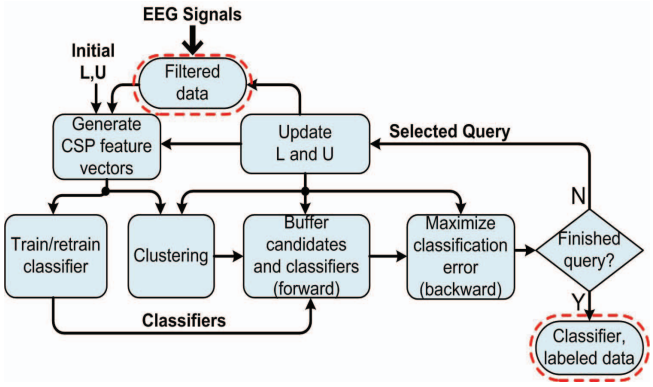


Fig. 1. Our proposed active learning scheme.

1) Cluster the whole set of samples (F_{csp}) into N_c clusters such that the within cluster errors are minimized. *K-means clustering* with N_c=4 is chosen. The indexes of the ith cluster ($I_c(i)$) are obtained by

$$I_c(i) = \arg \min_{C_s} \sum_{i=1}^{N_c} \sum_{F_{csp}(j) \in C_s(i)} ||F_{csp}(j) - u_s(i)||^2 \tag{3}$$

where C_s={$C_s(1)$,$C_s(2)$,...,$C_s(N_c)$} and $u_s(i)$ denote the total set of clusters and the sample mean of the ith cluster; j is the index of the feature vector F_{csp}.

2) Calculate *cluster impurity* using the labeled samples. The *cluster impurity* for the ith cluster is calculated by

$$P_u(i) = \frac{min(N_{c0}(i), N_{c1}(i))}{max(N_{c0}(i), N_{c1}(i))} \tag{4}$$

where $N_{c0}(i)$ and $N_{c1}(i)$ are defined as the number of samples that belong to the class with label "0" and the class with label "1" for the ith cluster, which are given by

$$N_{c0}(i) = \sum_{j \in I_c(i)} (Y|Y(j)=0)) \tag{5}$$

$$N_{c1}(i) = \sum_{j \in I_c(i)} (Y|Y(j)=1)) \tag{6}$$

where Y is the class label and obviously $P_u(i) \in [0\ 1]$ always holds.

3) Query cluster selection. The criterion for query cluster selection is defined as

$$\hat{i} = \arg \max_i (P_u(i)) \tag{7}$$

The cluster is chosen such that the cluster impurity for the labeled samples is the highest. When this is true, the number of features from two classes are close to each other for labeled set. Considering the similarity in the feature vectors distributions of the same cluster for labeled and unlabeled sets, it is reasonable to assume that the unlabeled samples in the cluster would be more uncertain.

4) Query sample selection. The sample that is close to the center of chosen cluster ($C_c(\hat{i})$) is chosen as the candidate, which is given by

$$\hat{m} = \arg \min_{m, m \in U} \mathcal{D}_s(C_c(\hat{i}), m) \tag{8}$$

where $\mathcal{D}_s(C_c(\hat{i}), m)$ is the distance between the unlabeled samples (index: m) to the center of the chosen cluster $C_c(\hat{i})$, which is given by

$$\mathcal{D}_s(C_c(\hat{i}), m) = K_g(F_{csp}(C_c(\hat{i})), F_{csp}(m)) \tag{9}$$

where $K_g()$ is the Gaussian kernel function, which is given by $K_g(x,y) = e^{-||x-y||^2/2\sigma^2}$, where σ=0.25 is chosen in implementation.

5) Forward-backward classification error maximization. A total of N_s candidates are forwardly buffered based on

steps 1 to 4. The same number of classifiers from the most recent classifiers for each candidate will be buffered as well. The idea is to choose the most uncertain samples using the buffered classifiers. Assume the jth buffered classifier for sample $s(k)$ is $f(S_v(i,j), s(k))$, which is given by [10]

$$f(S_v(i,j), s(k)) = sign(\sum_{j=1}^{n} y_j \alpha_j K(S_v(i,j), s(k))) \quad (10)$$

where $K(S_v(i,j), s(k))$ is the kernel matrix defining similarity between the candidate $s(k)$ and j-th support vector $S_v(i,j)$; α_j and y_j are the coefficients and labels of support vectors in the form of $\{\pm 1\}$ for ith classifier, respectively. The query sample $q(\hat{k})$ is chosen from the forwardly buffered N_s samples so that the total classification errors on the labeled samples including the buffered candidates ($L(k)$) using the buffered N_s classifiers are maximized, which are given by

$$q(\hat{k}) = \arg\max_{k} || \sum_{i=1}^{N_s} \sum_{m=1}^{L(k)} f(S_v(i,j), s(m)) - Y^*(m) || \quad (11)$$

where $||x||$ gives the absolute value of x, $Y^*(m)$ is the label in the form of $\{\pm 1\}$. In this way, the chosen sample is considered to be the most informative and uncertain.

3. EXPERIMENTAL EVALUATION

Experiments are conducted using BCI competition IV data set IVb, which contains three bipolar recordings (C3, Cz, and C4) with a sampling frequency of 250Hz. The cue-based screening paradigm consists of 160 trials for two classes of motor imagery of left hand and right hand. A small initially labeled set of size L=6 is chosen as a starting point and the querying process is iterated for 10 times, the averaged accuracies are shown in Fig. 2. It can be observed from the figure, the accuracy slowly approaches 100%, 92% and 88% for subjects 4, 7 and 5, respectively. It is close to 85% for subject 8, and 80% for subjects 1, 6 and 9. While the performance for subjects 2 and 3 is not good which is also true for existing approaches.

To show the efficacy of the proposed CIFBEM-AL learning method, it is compared with a baseline passive learning method which is similar to our scheme, i.e., a simplified version of FBCSP [1] and SWDCSP [2], namely "sFBSWD". However, feature selection in FBCSP and discriminant frequency band selection in SWDCSP are not implemented to have a fair comparison, considering the facts that no selection of features, frequency bands, channels and time-segments is employed in our scheme. A total of 10 runs are conducted to randomly choose the predefined numbers of training samples to train the classifier, which is subsequently used to classify the unlabeled samples for sFBSWD. Similarly, 10 runs are conducted for our proposed CIFBEM-AL, with the comparison of the average accuracies at pre-defined training sample sizes for sFBSWD and CIFBEM-AL shown in Table 1. Note that Support Vector Machines (SVMs) with linear kernel is used as the classifier. Considering the facts that the classifiers are trained using the same number of samples (L), it is easily seen from the table that active learning can pick up more informative and representative samples to build the classifier. This has led to an increase in accuracy of 1.81%, 4.10%, 5.77%, 5.46%, 5.15% and 2.17% for L=14, 28, 56, 84, 112 and 140, respectively, compared with that of passive learning methods. The improvement is more significant when training sample size is small, e.g., an average increase in accuracy of 5.12% is achieved for L=28 to L=112. With the increase of the training samples, the performance of passive and active learning algorithms tends to be similar, e.g., when L=140. A paired sample t-test is conducted on the null hypothesis that difference in the accuracies of active and passive learning methods is a random sample from a normal distribution with mean 0. The null hypothesis is rejected for all the training sample sizes with p=0.0021, indicating the significance of the accuracy increase using active over passive learning. This not only demonstrates the efficacy of the proposed active learning scheme but also shows its advantages in boosting the performance when the training data size is small.

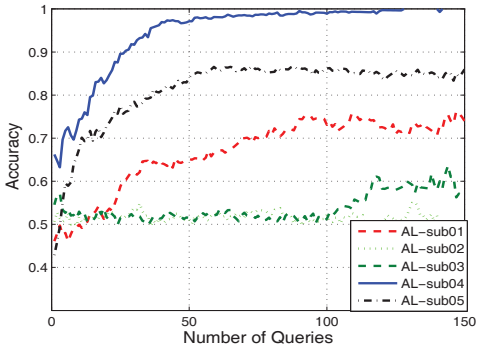
4. CONCLUSIONS

In this paper, we investigate the problem of applying active learning for the classification of the motor imagery EEG signals in brain computer interface. Specifically, we address the problem on how to select the most representative and informative samples to build the classifier such that the performance can be boosted when the training sample size is small. The samples are firstly clustered and the cluster that has high impurity is selected. The samples that are close to the center of chosen clusters are selected as the candidates. A predefined number of candidates and the most recent classifiers are forwardly buffered, subsequently, the candidate that can backward maximize the classification errors on the labeled data is chosen as the query sample. This ensures the uncertainty and informativeness of the selected query. Experimental results conducted using BCI competition IV data set IVb show that on average the resultant accuracy is 5.12% higher than that achieved using the passive learning algorithm, when the training data size varying from 28 to 112. This further demonstrates the effectiveness of proposed active learning method in boosting the performance for small training data set.

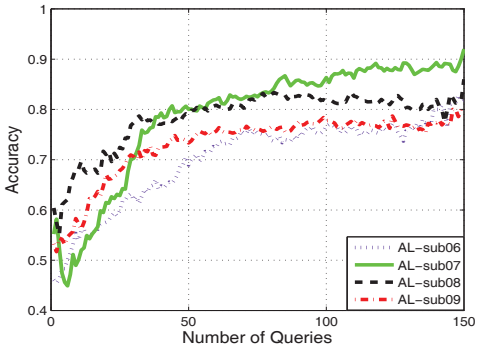
5. REFERENCES

[1] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," *30th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) 2008*, pp. 4178-4181.

[2] G. Sun, J. Hu and G. Wu, "A Novel Frequency Band Selection Method for Common Spatial Pattern in Motor Imagery Based



(a) Subjects 01 to 05.



(b) Subjects 06 to 09.

Fig. 2. The accuracies achieved by proposed CIFBEM-AL algorithm using BCI Competition IV data set IVb for all the subjects.

.

Table 1. Comparisons of Accuracies for Different Training Sizes Using Passive (sFBSWD) and Active Learning (CIFBEM-AL)

Methods	Subj.	L(14) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(28) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(56) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(84) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(112) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(140) ($\mathcal{A}_c \pm \mathcal{V}_r$)
Passive (sFBSWD)	S1	52.81±4.98	58.64±6.13	67.31±3.48	67.37±5.14	69.79±5.66	73.50±13.13
	S2	50.89±3.43	52.95±3.47	52.60±4.95	55.00±4.84	52.50±6.72	57.00±7.15
	S3	53.36±5.20	52.42±3.66	54.71±4.63	55.53±3.27	52.08±5.97	62.50±13.18
	S4	63.01±3.86	73.41±7.98	86.15±3.74	90.00±2.72	90.63±3.96	92.00±4.83
	S5	58.49±4.82	61.59±5.65	65.77±6.82	68.55±5.02	76.88±8.64	80.50±10.92
	S6	52.53±4.54	54.92±4.03	60.58±3.71	64.87±2.41	67.92±6.45	69.00±8.10
	S7	56.10±4.70	66.14±6.53	72.60±6.15	80.26±2.40	82.92±4.99	82.00±7.15
	S8	56.51±5.45	60.68±6.21	72.98±4.65	75.79±4.39	79.58±5.88	78.50±8.18
	S9	51.92±6.62	58.56±6.14	65.58±5.32	68.68±6.53	67.50±6.75	73.50±7.09
	A_{ac}	**55.07±4.84**	**59.92±5.53**	**66.48±4.83**	**69.56±4.08**	**71.09±6.11**	**74.28±8.86**
Proposed Active CIFBEM -AL	S1	49.12±0.20	53.16±0.47	65.14±0.28	70.26±0.19	72.65±0.40	72.86±1.32
	S2	50.41±0.16	52.71±0.22	51.14±0.19	52.99±0.36	51.63±0.16	51.90±0.56
	S3	52.18±0.10	52.41±0.15	51.52±0.27	51.04±0.13	54.29±0.38	58.57±0.32
	S4	69.66±0.38	85.79±0.24	97.14±3.45e-4	98.57±1.16e-4	99.39±8.75e-5	100±0
	S5	63.95±0.60	74.74±0.19	84.10±8.54e-4	85.58±7.24e-4	86.12±0.18	84.76±0.63
	S6	53.20±0.20	58.87±0.18	68.95±7.66e-4	74.94±0.25	76.33±7.66e-4	76.19±0.36
	S7	51.16±0.41	62.63±0.42	80.00±0.28	83.51±0.51	87.76±5.83e-4	87.62±0.37
	S8	66.53±0.30	69.62±0.24	78.95±0.12	82.47±0.13	81.63±0.23	80.00±0.58
	S9	55.71±0.12	66.24±0.56	73.33±0.16	75.84±0.18	76.33±0.23	76.19±1.27
	A_{ac}	**56.88±0.27**	**64.02±0.30**	**72.25±0.14**	**75.02±0.19**	**76.24±0.18**	**76.45±0.60**

Note: A_{ac}: Average Accuracy (%) over all subjects (shown in bold in the last row for each method). $\mathcal{A}_c$: Accuracy (%), $\mathcal{V}_r$: Variance.

Brain Computer Interface," *WCCI 2010 IEEE World Congress on Comp. Inte.*, pp. 335-340, July 2010, Barcelona, Spain.

[3] C. Vidaurre, C. Sannelli, K.-R. Mller and B. Blankertz, "Machine-Learning Based Co-adaptive Calibration: Towards a Cure for BCI illiteracy," *Neural Comput.*, vol. 23, no. 3, pp. 791-816, 2011.

[4] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Mller, "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis," *IEEE Signal Processing Magazine*, vol. 41, pp. 41-56, January 2008.

[5] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by Committee," *Proc. ACM Workshop Comput. Learn Theory,* pp. 287-294, 1992.

[6] G. Schohn, D. Cohn, "Less is More: Active Learning with Support Vector Machines," *Proc. of the 17th Int. Conf. on Machine Learning,* 2000.

[7] S. Tong and D. Koller, "Support Vector Machine Active Learning With Applications to Text Classification," *Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2001.

[8] E. Pasolli and F. Melgani, "Active Learning Methods for Electrocardiographic Signal Classification," *IEEE Trans. on Information Technology In Biomedicine,* pp. 1405-1415, vol. 14, no. 6, November 2010.

[9] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," *In ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pp. 58-65, 2003.

[10] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski and W. J. Emery, "Active Learning Methods for Remote Sensing Image Classification," *IEEE Trans. Geoscience and Remote Sensing,* vol. 47, no. 7, pp. 2218-2232, July 2009.