Journal of Fiber Bioengineering and Informatics 7:4 (2014) 627–636
doi:10.3993/jfbi12201415

A Batch-mode Active Learning Method Based on the Nearest Average-class Distance (NACD) for Multiclass Brain-Computer Interfaces ⋆

Minyou Chen [a], Xuemin Tan [a,b,*], John Q. Gan [b]
Li Zhang [a], Wenjuan Jian [a]

[a] *State Key Laboratory of Power Transmission Equipment & System Security and New Technology School of Electrical Engineering, Chongqing University, Chongqing 400044, China*

[b] *School of Computer Science and Electronic Engineering, University of Essex Colchester, CO4 3SQ, UK*

Received 6 June 2014; accepted (in revised version) 30 October 2014; available online 17 December 2014

Abstract

In this paper, a novel batch-mode active learning method based on the nearest average-class distance (ALNACD) is proposed to solve multi-class problems with Linear Discriminate Analysis (LDA) classifiers. Using the Nearest Average-class Distance (NACD) query function, the ALNACD algorithm selects a batch of most uncertain samples from unlabeled data to improve gradually pre-trained classifiers' performance. As our method only needs a small set of labeled samples to train initial classifiers, it is very useful in applications like Brain-computer Interface (BCI) design. To verify the effectiveness of the proposed ALNACD method, we test the ALNACD algorithm on the Dataset 2a of BCI Competition IV. The test results show that the ALNACD algorithm offers similar classification results using less sample labeling effort than Random Sampling (RS) method. It also provides competitive results compared with active Support Vector Machine (active SVM), but uses less time than the active SVM in terms of the training.

Keywords: Active Learning; Linear Discriminant Analysis (LDA); Nearest Average-class Distance (NACD); Brain-computer Interface (BCI)

1 Introduction

Brain–computer interfaces (BCI) provide a new non-muscular channel for sending messages and commands to the external world. In BCI literatures, many supervised methods have been pro-

⋆Project supported by the Fundamental Research Funds for the Central Universities in China (Project No. CDJZR13150010) and National "111" Plan Project (B08036).

*Corresponding author.
Email address: tanxuemin1987@gmail.com (Xuemin Tan).

posed for the classification of BCI data [1, 2]. The classification results of all these methods rely heavily on the number of labeled samples used for learning. However, collecting labeled data is often difficult, expensive and time-consuming. Two popular approaches, semi-supervised learning and active learning, have been proposed for dealing with this problem. Semi-supervised learning algorithms use a small set of labeled training data to build an initial classifier that can predict the labels of unlabeled data, and then add samples with predicted labels into the training set, resulting in more precise decision boundaries iteration by iteration. In active learning, a query function repeatedly queries the most uncertain samples from a pool of unlabeled data for annotating and updating the training set, and these samples have maximum ambiguity to belong to certain class. Usually, the most uncertain samples can be considered as the most informative ones. Thus, for active learning, redundant samples are avoided in training set, which greatly reduce both labeling cost and computational time. In recent years, active learning algorithms have been developed under the motivation of query strategies. Such strategies include uncertainty sampling, query-by-committee, expected model change, expected error reduction and so an.

The key issue of active learning is to find a good query function to reduce the number of samples needed to be labeled from a pool of unlabeled samples [3, 4]. Most query functions are for binary classification. For multiclass active learning, the binary classification is often extended to multiclass by One-against-all (OAA) or One-against-one (OAO) mechanism [5, 6]. Linear Discriminant Analysis (LDA) is a binary classification method and can be well extended for solving multi-class problems. As a popular classifier, LDA has been widely used in semi-supervised algorithms. Cai et al. [7] and Zhao et al. [8] proposed, respectively, a Semi-supervised Discriminant Analysis (SDA) method and a LDA-based self-training algorithm for face recognition. Another semi-supervised method was presented in [9], which combines linear discriminant analysis and manifold learning for improving the precision of hyperspectral imagery classification. However, little investigation on LDA-based active learning has been conducted, particularly in the BCI field.

In most existing active learning techniques, a single most uncertain sample is queried at each iteration [10]. This can be inefficient, because the classifier has to be re-trained for the arriving of each new sample. In this paper, our algorithm allows for batch-mode incremental learning.

In recent years, batch-mode active learning algorithms have been developed for the applications where labeled data is insufficient. Lewis and Gale [11] proposed an uncertainty sampling method which simply query the several instances for one iteration whose posterior probability is nearest to 0.5. The active learning technique proposed in [5] is to select n most uncertain samples, one closest to current separating hyperplane for each One-against-all (OAA) binary SVM. In [12], Guo proposed a novel batch-mode active learning approach that selects a batch of queries in each iteration by maximizing a natural mutual information criterion between labeled and unlabeled instances. Also, another discriminative batch-mode active learning approach was presented in [13], where information in unlabeled data is exploited and a batch of instances are selected by optimizing the target classification model.

In this paper, a novel batch-mode active learning method based on the nearest average-class distance (ALNACD) is proposed for solving multiclass BCI classification problems with LDA classifiers. The ALNACD uses the Nearest Average-class Distance (NACD) as query function which is used to query the most uncertain samples from unlabeled data. The proposed ALNACD is compared with Random Sampling (RS) and active SVM [5] on the Dataset 2a of BCI Competition IV with 9 subjects. Experimental results show the effectiveness of the proposed ALNACD

algorithm. In the article, our main contributions contain the following: first, a novel NACD query function is proposed for selecting most uncertain batch samples. Second, we show our algorithm, as the first active learning algorithm used in BCI field, is available. Third, the proposed algorithm are based on batch mode and used for solving multiclass BCI problems.

The rest of this paper is organized as follows. The related methods are reviewed and the ALNAZD algorithm based multiclass is proposed in section 2. The Dataset 2a of the BCI Competition IV is described in session 3, followed by experimental results. Conclusion is drawn in session 4.

2 Methods

2.1 Linear Discriminant Analysis (LDA)

The LDA [14] aims at finding a transformation matrix W which maximizes between-class scatter and minimizes within-class scatter, i.e.,

$$maximize \frac{tr(W^T S_b W)}{tr(W^T S_w W)} \tag{1}$$

where W is the LDA weight vector. Let $X = [x_1^1, x_1^2, \cdots, x_1^{N_1}, x_2^1, x_2^2, \cdots, x_2^{N_2}, \cdots, x_d^1, x_d^2, \cdots, x_d^{N_d}]$ be the data matrix of training samples and N_d be the number of samples in the dth class. The within-class scatter matrix (S_w) and between-class scatter matrix (S_b) are defined as follows:

$$S_w = \frac{1}{N} \sum_{d=1}^{K} \sum_{i=1}^{N_d} (x_d^i - m_d)(x_d^i - m_d)^T \tag{2}$$

$$S_b = \frac{1}{N} \sum_{d=1}^{K} N_d (m_d - m)(m_d - m)^T \tag{3}$$

where $m_d = \frac{1}{N_d} \sum_{i=1}^{N_d} x_d^i$ is the mean vector of the dth class and $m = \frac{1}{N} \sum_{d=1}^{K} \sum_{i=1}^{N_d} x_d^i$ is the total mean vector, N is the total number of samples, and $K = 2$ is the number of classes for the two-class problem.

The decision score function $f(x)$ is defined as

$$f(x) = Wx + b \tag{4}$$

where $b = -\sum_{d=1}^{K} N_d m_d / \sum_{d=1}^{K} N_d$ is the bias, and the sign of $f(x)$ is used to predict the class label for a given test sample. If $f(x) > 0$, the sample x belongs to the first class ($class1$), otherwise it belongs to the second class ($class2$).

2.2 Query Function Based on the Nearest Average-class Distance (NACD)

In this paper, the proposed ALNACD technique is based on the Nearest Average-class Distance (NACD) query function which is used for selecting most uncertain samples actively. In a binary

classification problem, if labeled samples are two-dimensional, as shown in Fig. 1, they are projected onto a lower-dimensional space (a line in this case) and two classes are separated by a separation line which makes the projection maximize the "separability" of the projected samples [14]. Because those unlabeled samples with the smallest distance to each class-centre ($mean1$ or $mean2$) have higher confidence to belong to certain class ($class1$ or $class2$), it is reasonable to assume that the most uncertain samples are the ones whose decision scores are near to the average of the two class-centres, thus adding them to training set is most likely to improve the performance of the classifier in the next iteration of re-training.

In this paper, we use NACD function to select most uncertain ones from unlabeled samples. The uncertain criterion is implemented by analyzing the decision scores of LDA classifier for unlabeled samples and the mean decision score of LDA classifier for two class-centres ($(mean1+mean2)/2$), which is adopted to query those samples with the smallest distance between the decision score of each unlabeled sample and the mean decision score of two class-centres. In addition, only the decision scores of the unlabeled samples in the range of [$mean1$, $mean2$] will be considered. It is clear that the smaller the distance of an unlabeled sample is, the easier the unlabeled sample is to be queried for expanding current training data set.

2.3 Multi-class ALNACD Algorithm Based One-against-all (OAA)

The One-against-all (OAA) strategy [5] involves a parallel architecture made up of n binary LDA classifiers, one for each class defined by one class against all the others. In the proposed multi-class ALNACD technique, we use the OAA strategy to train n binary LDA classifiers with an initial set of labeled samples. After the initial training, n decision scores $f_j(x)$ (j=*1, 2, . . . ,* n) are calculated separately based n binary LDA classifiers for each of unlabeled samples. Now we consider each binary LDA classifier separately to select m ($m \geq 1$) most uncertain samples at each iteration on the basis of the proposed query function NACD. The confidence of each unlabeled sample depends on the distance between the decision score of the unlabeled sample and the mean decision score of two class-centres. To select the most uncertain samples from unlabeled samples, m samples with the lowest confidence are selected by each binary LDA classifier at each iteration. This should be $h = m \times n$ samples selected for n binary LDA classifiers. However, if the decision scores of all unlabeled samples from at least one binary LDA are not in the range of [$mean1$, $mean2$] or at least one sample is selected by more than one binary LDA, thus only a total of $h \leq m \times n$ samples from n binary LDA classifiers are selected at each iteration. The below Steps describes the details of the proposed ALNACD algorithm. First, suppose that we have two raw data sets: D_I with labels as the initial labeled data set and D_F without labels as the unlabeled data set. D_I contains N_1 samples and D_F contains N_2 samples. Before describing the steps, we clearly give the class-centre definition for each binary classification:

$$mean1 = mean(f_j(x_i))(x_i \in D_{I1}) \tag{5}$$

$$mean2 = mean(f_j(x_i))(x_i \in D_{I2}) \tag{6}$$

where $mean1$ and $mean2$ denote the first class-centre and the second class-centre in the initial labeled data set D_I. D_{I1} and D_{I2} respectively represents the data set belonging to the first and the second class of D_I. Obviously, $D_I = D_{I1} \cup D_{I2}$.

Step 1: Train n binary LDA classifiers with D_I. Let $f_j(.)$ be the decision scores of the jth binary LDA classifier and set $k = 0$.

Step 2: The kth iteration (k=1, 2, ..., K$_0$) follows Step 2 to 6. Set $h = 0$.

Step 3: For j=1 to n, if ($number(f_j(x) \in [mean1, mean2]) > m$) ($x \in D_F$, the $number(.)$ represents the number of samples satisfying the condition in the parentheses), for the jth binary LDA classifier, select the m samples from D_F, whose decision scores are closest to $(mean1 + mean2)/2$, (see the equation (7) and (8)), $h = h + m$, otherwise, select the samples from the unlabeled data set D_F, whose decision scores $f_j(x) \in [mean1, mean2]$, $h = h + number(f_j(x) \in [mean1, mean2])$.

$$d_j(x_i) = |f_j(x_i) - (mean1 + mean2)/2| \quad (x_i \in D_F) \tag{7}$$

$$newd_j(x_i) = ascend(d_j(x_i)) \tag{8}$$

where the difference $d_j(x_i)$ between the LDA's decision score $f_j(x_i)$ of the unlabeled sample predicted with the jth class and the LDA's decision score of the mean of binary class-centre, represents the uncertainty of each unlabeled sample. $newd_j(x_i)$ denotes the results in the ascending order for $d_j(x_i)$, which means the uncertainty of the unlabeled samples predicted is ranked from the most uncertain members to the most certain members. Then m most uncertain samples are selected from the unlabeled samples

Step 4: Assign true class labels to the h selected samples D_S^k and add D_S^k into the labeled data set D_I and get rid of D_S^k from unlabeled data set D_F in the kth iteration. In here,

$$D_I = D_I + D_S^k \tag{9}$$

$$D_F = D_F - D_S^k. \tag{10}$$

Step 5: Retrain the n binary LDA classifiers with the updated data set D_I.

Step 6: (termination criterion) If $k \geq (N_2 \times \beta)/(m \times n)$, the algorithm terminates after the kth iteration, where β is a pre-determined percentage of the number N_2 of initial unlabeled data set D_F. Otherwise, go to Step 2 to perform the $(k + 1)$th iteration.

3 Experiments and Results

3.1 Description of the Electroencephalographic (EEG) Data

In this paper, the Dataset 2a of BCI Competition IV [15] is used to test the proposed ALNACD active learning algorithm, which consists of EEG data recorded from 9 subjects who performed imagined movements of left hand (class 1), right hand (class 2), feet (class 3), and tongue (class 4). The subjects were sitting in a comfortable armchair in front of a computer screen. At the beginning of a trial ($t = 0$ s), a fixation cross appeared on the black screen. In addition, a short acoustic warning tone was presented. After two seconds ($t = 2$ s), a cue in the form of an arrow pointing either to the left, right, down, or up (corresponding to one of the four classes: left hand, right hand, foot, or tongue) appeared and stayed on the screen for 1.25 s. This prompted the subjects to perform the desired motor imagery task. No feedback was provided. The subjects were asked to carry out the motor imagery task until the fixation cross disappeared from the screen at $t = 6$ s. There was a short break with black screen between trials. Two sessions of motor imagery EEG data were recorded from each subject on different days using 22 electrodes as shown in Fig. 2. The signals were sampled at frequency 250 Hz and bandpass-filtered between 0.5

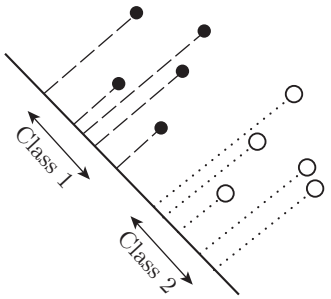

Fig. 1: The intuition behind LDA. The black dots and white dots, respectively, represent *class* 1 and the *class* 2

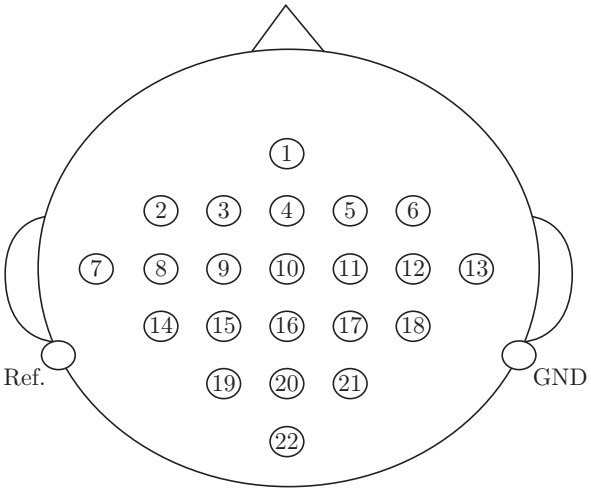

Fig. 2: Electrode montage corresponding to the international 10-20 system

Hz and 100 Hz. Each session is comprised of 6 runs separated by short breaks. One run consists of 48 trials (12 for each of the four possible classes), yielding a total of 288 trials per session and a total of 576 trials for two sessions.

For each subject, the first 400 trials are considered as the training set T and the remaining 176 trials are used as the independent testing set *TS*. We only select randomly a small number of samples from T as the initial labeled date set D_I, with the same proportion for each class, and the rest is treated as unlabeled data setD_F for expanding D_I. The performance of the classifiers is tested on *TS*. The above process is repeated 30 runs with all the samples in T randomly shuffled in each run. We perform algorithms on T for subject 1, 3, 6, 7 and 9 (20 trials for D_I and 380 trials for D_F), and subject 2, 4, 5 and 8 (40 trials for D_I and 360 trials for D_F).

3.2 Preprocessing and Feature Extraction

The EEG signals are further band-pass filtered between 8Hz and 30Hz and zero centered. The well-known common spatial patterns (CSP) algorithm is adopted to extract features between 3 s and 6 s for each trial. More details on the application of common spatial patterns to BCI are described in the literatures [16, 17]. In this paper, we use OAA for CSP feature extraction of four classes. For each binary problem, the choice of pairs of CSP features is set to 2 for Dataset 2a, which means after CSP transformation only the first two rows and the last two rows of the projected signals are used for extracting four features that are the normalized logarithmic variances of the four projected signals, which result in a total of 16 features for each sample. This is because a greater choice of pairs did not significantly improve classification accuracy [18].

3.3 Experimental Results

To verify the effectiveness of ALNACD on the BCI data with a four-class classification problem, we compare the performance of ALNACD on 9 subjects with the Random Sampling (RS) and active SVM [5] that is based on the OAA formulation of binary classifiers. In the RS approach,

at each iteration, a batch of h samples ($h = m \times n$) are randomly selected from the unlabeled data set D_F, assigned with true labels, and then added into labeled data set D_I for updating the classifiers. In our implementation of active SVM, we adopt SVM classifiers with RBF (Gaussian) kernel and use the LIBSVM library to implement the algorithm [19]. The parameter pair (C, g) is searched with the one with the best cross-validation accuracy. The C and g, respectively, denotes the regularization parameter and the RBF (Gaussian) kernel. We implement all algorithms in MATLAB on a 3.2 GHz 2 GB PC.

In here, we set $m = 2$ and $\beta = 60\%$, which respectively denotes 2 samples are selected to update each binary classifier and only 60% number of initial unlabeled samples with most informative are queried by NACD method for improving classification accuracy. Fig. 3 shows the average classification accuracies provided by different methods versus the number of iterations. One can see that the proposed ALNACD always produces better classification accuracy than the RS. However, our algorithm shows lower classification accuracy than active SVM in the previous

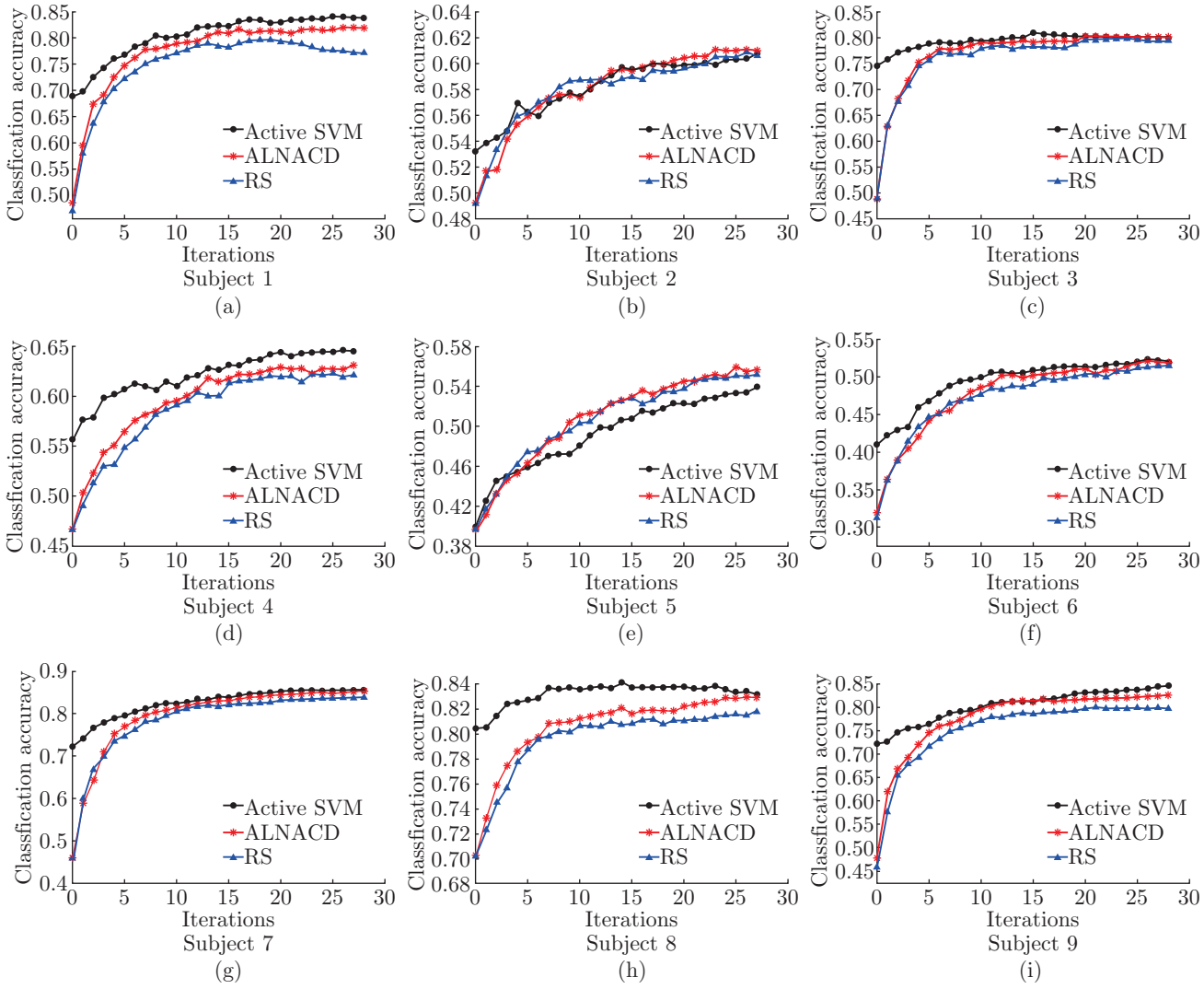

Fig. 3: Average classification accuracies provided by the proposed ALNACD, active SVM and RS for each of 9 subjects. Fig. 3 (a) to Fig. 3 (i) respectively represents the performance from Subject 1 to Subject 9

iterations and obtains similar accuracy with active SVM in later iterations for most subjects. From Fig. 3, we also can see that our algorithm and active SVM can yield similar accuracies with the RS but the former uses less samples than latter. However, subject 2 and 5 on classification accuracy are the exception.

In Table 1, we report the number of examples required for our algorithm and active SVM which attain similar accuracy as the RS after 60% number of unlabeled samples are queried. The results, tabulated in Table 1, show that our algorithm yields similar accuracy with the RS using 72.00%, 18.52%, 32.14%, 37.04%, 11.11%, 14.29%, 42.86%, 40.74% and 64.29% less samples, and active SVM can yield similar accuracy compared with the RS using 82.14%, 0.00%, 64.29%, 55.56%, 0.00%, 25.00%, 57.14%, 92.59% and 67.86% less samples for each of 9 subjects. The ultimate goal of active learning is to attain the similar accuracy with as little labeled data as possible. The average reduction number compared with RS in Table 1 for ALNACD is quite close to that for Active SVM, it needed 137.78 samples as compared with about 111.11 samples for the latter. This is not too much more. From Table 1, we also can see that the active SVM is not more effective than the RS and ALNACD for subject 2 and 5 and even use more or the same number of samples for attaining similar classification accuracy with the ALNACD or the RS respectively. Nevertheless, for most subjects, the performance of the ALNACD and active SVM still meets the ultimate goal of active learning which is to attain high classification accuracy with as little labeled data as possible.

Table 1: The reduction in the number of training samples needed for ALNACD and active SVM

Subject	ALNACD	active SVM	RS	Reduction needed for ALNACD (%)	Reduction needed for active SVM (%)
1	56	40	224	75.00	84.14
2	176	216	216	18.52	0.00
3	152	80	224	32.14	64.29
4	136	96	216	37.04	55.56
5	192	216	216	11.11	0.00
6	192	168	224	14.29	25.00
7	128	96	224	42.86	57.14
8	128	16	216	40.74	92.59
9	80	72	224	64.29	67.86
Average	137.78	111.11	220.44	37.33	49.62

It is worth noticing that the difference between the number of samples needed by our ALNACD algorithm and active SVM is not too much for most subjects. However, the time needed for the training is much less for our ALNACD algorithm as compared with the active SVM for each of 9 subjects as shown in Table 2.

4 Conclusion

In this paper, we propose a novel active learning method based the nearest average-class distance (ALNACD) for solving multi-class problems with LDA classifiers, which initially only needs a

Table 2: Comparison of training time over 30 runs for ALNACD, active SVM and the RS

Subject	ALNACD (s)	active SVM (s)	RS (s)
1	5.82	16.53	4.60
2	6.50	18.78	4.49
3	5.80	15.58	4.61
4	5.59	20.03	4.47
5	5.66	27.72	4.53
6	6.00	19.27	4.65
7	5.80	17.56	4.68
8	5.89	17.60	4.52
9	5.84	19.73	4.60
Average	5.88	19.20	4.57

small set of labeled samples to train classifiers and effectively reduces the expenses and time of obtaining labeled data. A query function called Nearest Average-class Distance (NACD) is proposed to identify and use only those samples with high uncertainty in the learning process. Experimental results show that the proposed method always provides better accuracies than the RS at each iteration and obtains similar classification accuracies as the RS by using less samples. It also can provide competitive results compared with active SVM, but use less time than active SVM in terms of the training. As we known, SVM is an advanced method and can usually achieve better classification results than LDA when the appropriate parameters are used [20]. However, the proposed NACD strategy based LDA can select more informative samples and even use much less time than active SVM, which further demonstrates the effective of the NACD.

Although the performance of the proposed method is satisfactory, the method does not include any diversity criterion for selecting multiple samples. This should be done by defining a diversity criterion that can be implemented for avoiding losing one of the most important properties of the proposed method in future research.

Acknowledgements

The authors would like to thank the Graz BCI group for sharing their data sets.

References

[1] Ang KK, Chin ZY, Wang C, Guan C and Zhang H. Filter bank common spatial pattern algorithm on BCI competition iv datasets 2a and 2b. Front Neurosci: 2012; 6-39.

[2] Liu C, Wang H, Lu Z. EEG classification for multi-class motor imagery BCI. In Proc. of the 25th Chinese Control and Decision Conference (CCDC): 2013; 4450-4453.

[3] Nguyen TT, Binh ND and Bischof H. Efficient boosting-based active learning for specific object detection problems. International Journal of Electrical, Computer and Systems Engineering: 2009; 3: 150-155.

[4] Yang T, Li J, Pan Q, Zhao C and Zhu Y. Active learning based pedestrian detection in real scenes. In Proc. of the 18th International Conference on Pattern Recognition: 2006; 904-907.

[5] Mitra P, Uma Shankar B, Pal K. Segmentation of multispectral remote sensing images using active support vector machines. Pattern Recogn Lett: 2004; 25: 1067-1074.

[6] Pasolli E, Melgani F. Active learning methods for electrocardiographic signal classification. IEEE T Inf Technol B: 2010; 14: 1405-1416.

[7] Cai D, He X and Han J. Semi-supervised discriminant analysis. In Proc. of the IEEE 11th International Conference on Computer Vision: 2007; 1-7.

[8] Zhao X, Evans N and Dugelay J. Semi-supervised face recognition with LDA self-training. In Proc. of the 18th IEEE International Conference on Image Processing: 2011, 3041-3044.

[9] Zheng Z and Peng Y. Semi-supervised based hyperspectral imagery classification. Sensors & Transducers: 2013; 156: 298-303.

[10] Tong S and Koller D. Support vector machine active learning with applications to text classification. J Mach Learn Res: 2002; 2: 45-66.

[11] Lewis DD and Gale WA. A sequential algorithm for training text classifiers. In Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 1994; 3-12.

[12] Guo Y. Active instance sampling via matrix partition. NIPS: 2010; 802-810.

[13] Guo Y, Schuurmans D. Discriminative Batch Mode Active Learning. NIPS: 2007.

[14] Xanthopoulos P, Pardalos PP and Trafalis TB. Robust Data Mining. Springer: 2013.

[15] Naeem M, Brunner C, Leeb R, Graimann B and Pfurtscheller G., Seperability of four-class motor imagery data using independent components analysis. J Neural Eng: 2006; 3: 208.

[16] Wang H, Tang Q and Zheng W. L1-norm-based common spatial patterns. IEEE T Biomedical Enginee: 2012; 59: pp. 653-662.

[17] Fanga Y, Chena M, Zhengc X and Harrisond RF. Extending CSP to Detect Motor Imagery in a Four-class BCI. Journal of Information & Computational Science: 2012; 9: 143-151.

[18] Muller-Gerking J, Pfurtscheller G, and Flyvbjerg H. Designing optimal spatial filters for single-trial EEG classification in a movement task. Clin. Neurophysiol: 1999; 110: 787-798.

[19] Chang CC and Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST): 2011; 2: pp. 27.

[20] Xiong T and Cherkassky V. A combined SVM and LDA approach for classification. In Proc. of the IEEE International Joint Conference on Neural Networks: 2005; 1455-1459.