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. CD-JZR13150010) 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