Sparse Uncorrelated Cross-Domain Feature Extraction for Signal Classification in Brain-Computer Interfaces

Honglei Shi, Shiliang Sun
Shanghai Key Laboratory of Multidimensional Information Processing,
Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, China
Email: lhshi12@gmail.com, slsun@cs.ecnu.edu.cn

***Abstract*—In this paper, we present a novel dimensionality reduction method, called sparse uncorrelated cross-domain feature extraction (SUFE), for signal classification in brain-computer interfaces (BCIs). Considering the differences between the source and target distributions of signals from different subjects, we construct an optimization objective which aims to find a projection matrix to transform the original data in a high-dimensional space into a low-dimensional latent space. In the low-dimensional space, both the discrimination of different classes and transferability between the source and target domains are preserved. To make sure the minimum information redundancy, the extracted features are designed to be statistically uncorrelated. Then, by adding the l_1-norm penalty, we incorporate sparsity into the uncorrelated transformation. In the experiments, we evaluate the method with multiple datasets, and compare with the state-of-the-art methods. The results show that the proposed approach has better performance and is suitable for cross-domain signal classification.**

I. Introduction

Brain-computer interfaces (BCIs) are systems used for assisting, augmenting, or repairing human cognitive or sensory-motor functions by analyzing electro-physiological signals of the brain and then translating the signals into physical commands [24], [31]. They are direct communication and control pathways between the brain and external devices that do not require any peripheral devices. Thus BCIs are very helpful to predict the movement intentions, e.g., left or right hand movement, of subjects (or users) who suffer from motor disabilities and help them to interact with the environment. In the past ten years, the development of BCIs has grown substantially to which many machine learning methods have been successfully applied [2], [29], [32]. In BCIs, it is known that the patterns of the recorded signals are considerably different among subjects due to the diversity among persons, especially at the early training stage. However, the latent characteristics among signals may not change vastly according to the assumption in [18]. Therefore, cross-domain learning methods [23] can be applied, and an inter-subject/inter-section knowledge transformation model can be trained to learn some common knowledge from some subjects/periods to accelerate the training procedures for other subjects/periods [1], [10], [26], [27]. In such a model, both the differences and the relations among the signals recorded from different subjects or periods are considered to gain a better performance when applying to the target subjects. In this paper, we focus on the knowledge transfer in the feature extraction stage and propose a novel sparse cross-domain feature extraction method for signal classification in BCIs to quicken the training session in the systems.

For many applications in data mining and machine learning, a very important problem is feature extraction. There are many feature extraction methods proposed for the single domain scenario where the training and test samples are obtained from the same distribution, such as principal component analysis (PCA) [17] and linear discriminant analysis (LDA) [11], [12]. Given a data set, LDA seeks an optimal linear transformation by maximizing the ratio of the between-class distance to the within-class distance which are encoded in the between-class scatter matrix and the within-class scatter matrix respectively. The singularity problem is a major disadvantage of LDA and for computational purpose the scatter matrices must be nonsingular. But in many applications, such as face recognition [15], [25], pattern recognition [3], [12], and microarray analysis [8], the between-class and within-class scatter matrices are usually singular for undersampled data, that is, the dimension of the feature space is larger than the number of data points. Another disadvantage is the lack of sparsity in the LDA solution [35].

To address the singularity problem, there are many extensions of LDA proposed, including regularized linear discriminant analysis [7], orthogonal LDA (OLDA) [34], subspace LDA [25], etc. One of the other important extensions of LDA, i.e., uncorrelated LDA (ULDA), is motivated by extracting features with uncorrelated attributes [6], [15], [16], [33]–[35]. It is shown that the feature vectors extracted via ULDA are statistically uncorrelated, indicating the minimum information redundancy, which is greatly desirable for many applications. In [15], Jin et al. proposed an algorithm whose solution involves d generalized eigenvalues, if there exist d optimal discriminant vectors. However, this algorithm is computationally expensive for high-dimensional and large datasets, and does not address the singularity problem either. In [33], Ye et al. addressed the singularity problem in classical ULDA. By introducing an optimization criterion which combines the key ingredients of ULDA (which is based on QR-decomposition) and regularized LDA, they employed the generalized singular value decomposition (GSVD) tool [13] to solve the singularity problem directly. In this way, they avoided the information loss in the subspace. In many numerical experiments [6], [15], [33],

[35], the effectiveness of ULDA has been demonstrated.

Sparsity in the LDA solution has a great advantage for the high-dimensional data analysis to make the interpretation of the extracted features much easier. In LDA, every extracted feature in the low-dimensional space is a linear combination of all the original features. However, the coefficients of such linear combination are general nonzero. So the interpretation of the extracted features is difficult. There are some works established to overcome this obstacle, such as the penalized LDA (PLDA) [30], the sparse discriminant analysis (SLDA) [5], and sparse Fisher linear discriminant [9], etc. Almost all existing sparse LDA methods deal with sparsity by using l_1 penalty (i.e. Lasso penalty [28]) or its variants of the transformation matrix to objective functions [35].

In BCIs, though they may share some common knowledge, the signals recorded for one subject may be largely different from the signals for another subject, and the differences cannot be ignored. Thus, former feature extraction methods in single domain may not generalize well in these applications. In this case, researchers can develop cross-domain learning methods, and learn classification models on the source domain (data from one subject or one period) and apply them to the target domain (data from another subject or period) for classification tasks. In this paper we aims to design a new sparse cross-domain dimensionality reduction method for BCIs. We refer the proposed method as sparse uncorrelated cross-domain feature extraction, SUFE for short. Inspired by the previous framework of uncorrelated discriminative dimensionality reduction methods constructed in one domain, we find a transferable feature extraction method which can be used from the source domain to the target domain. In the proposed SUFE approach, a transformation $G^\top$ is learned which projects the original data space into a low-dimensional space. After projection, we make sure the following characteristics of the latent space: 1). The distance between the source and target distributions can be minimized; 2). The discrimination among classes can be preserved; 3). The extracted features are statistically uncorrelated, indicating the minimum information redundancy; 4). A sparse LDA transformation is computed, simultaneously. In this way, we consider the discrimination and transferability of the transformed space as well as the sparsity of the solution. To evaluate the discrimination in domain-merged training data and the transferability of the latent space to bridge the source and target domains, we firstly design mathematical terms to reveal these two characteristics using the scatter matrices of the data sets. Then we characterize all the solutions of the generalized ULDA by solving the optimization problem proposed in [22]. By finding the minimum l_1-norm solution from all the solutions with minimum dimension, we compute the sparse solution of ULDA to find a sparse uncorrelated solution for SUFE. In this paper, we employ the accelerated linearized Bregman method [4], [14] to solve the l_1-minimization problem.

The rest of this paper is organized as follows. In Section II we review some related work on transfer learning, the BCI application, ULDA and sparse LDA. The proposed sparse uncorrelated cross-domain feature extraction method (SUFE) is introduced in detail in Section III. We display in Section IV the experimental results on data from multiple signal datasets. Finally, the conclusion is given in Section V.

II. Related work

In [1], Alamgir et al. utilized the framework of multitask learning to construct a BCI system to learn feature characteristics that are consistent across subjects, which could be used without any subject-specific calibration process. In their framework, each subject was regarded as one task, and shared priors were employed in a parametric probabilistic approach. Specifically, they firstly trained off-line tasks to learn the model parameters and the shared prior parameters, by inferring K linear functions $f_t(\mathbf{x}; \mathbf{w}_t) = \langle \mathbf{w}_t, \mathbf{x} \rangle$ associated to each task such that $y_i^t = f_t(\mathbf{x}_i^t; \mathbf{w}_t) + \varepsilon_t$. Then for a new subject, the subject-specific parameters were inferred with respect to these shared prior parameters in an online fashion and an out-of-the-box BCI was defined to adapt to the new subject.

In [10], Fazli et al. introduced a subject-independent zero-training procedure for BCI applications. They constructed an ensemble method which was built upon common spatial pattern filters (CSP) for spatial filtering. Using a large database of pairs of spatial filters and classifiers from 45 subjects, they proposed to construct an ensemble of classifiers and learned a sparse subset of these pairs which were predictive across subjects. Then the quadratic regression with l_1 norm penalty is used to make sure the sparsity. The final classifier generalized well in a target subject. Finally, the authors demonstrated through a leave-one-subject-out cross-validation procedure that the sparse subset of spatial filters and classifiers could be applied to new subjects with only a moderate performance loss.

In [22], Shi et al. introduced a novel uncorrelated dimensionality reduction method for transfer learning in BCI applications. By maximizing the distance between classes and minimizing the distance between domains, their proposed method found a low-dimensional space that ensures the discrimination of merged training data and the transferability between the source domain and the target domain. By introducing a constraint $G^\top S_T^M G = I$ where S_T^M is the total scatter matrix of the merged data for the source domain and target domain, their proposed method extracted features that are mutually uncorrelated indicating the minimum information redundancy. The experimental results on a real BCI dataset with nine subjects demonstrated the effectiveness of their work. However, the authors ignored the sparsity in the solutions.

A recent work by Tu and Sun [27] introduced a two-level ensemble subject transfer learning framework for EEG classification. By dynamically and locally combining the outcomes of a robust classifier and an adaptive classifier to give the final classification results, their framework achieved positive subject transfer with improvements on both feature extraction and classification stages. Despite the encouraging results achieved by the proposed framework, the extracted features in their framework may not be uncorrelated, which may damage the classification performance.

III. The proposed method

Given three kinds of data matrix, i.e., a large number of source training data $X_{tr}^S \in R^{M \times N_S}$ with their labels $L_i^S \in \{1, 2, \ldots, c\}$ from a source domain S, a very small number of target training data $X_{tr}^T \in R^{M \times N_T}$ with their labels $L_i^T \in \{1, 2, \ldots, c\}$ from a target domain T, and a

large number of unlabeled target test data X_{te}^T. The unlabeled target test data X_{te}^T are used for later tasks (e.g., classification or regression). In these matrices, each column represents one data point. For cross-domain learning, both the differences and the relations between the target domain and the source domain should be considered. Due to the differences, the models learned on the single distribution may perform poorly on the target domain. On the other hand, we can expect the performance improvement by avoiding the differences and making full use of the relations. Therefore, the quality of a low-dimensional space should be taken into consideration both by its transferability from the source domain to the target domain and discriminability in the merged data simultaneously. To get a good computational acceleration for large scale datasets, the sparsity of the solution should also be considered. In this section, we firstly consider the transferability and the discriminability to construct an objective function for uncorrelated feature extraction. Then by finding the minimum l_1-norm solution we can get a sparse transformation for uncorrelated feature extraction to propose our sparse uncorrelated feature extraction method.

A. Domain-merged and between-domain scatter matrices

When given a source domain training dataset and a target domain training dataset as mentioned above, the within-class and between-class scatter measurements can be computed on the dataset merged by them which is called merged training dataset X_{tr}^M as in [22], [26]. As in cross-domain learning problems, the training and test data in the target domain are sampled from the same distribution and share the same characteristics. We should consider the different importances of the target domain and the source domain to take more advantages of the target. To this end we can add a weight W_{tr}^T into the target data points to control the influence of its training samples. Since the reliability of the distribution estimation of target training set is constrained by the sample size intuitively, the weight is related to the number of the target training samples. So we can add $W_{tr}^T = 1 + N_T/N_S$ to the target training samples to attach more importance to them, where N_T and N_S are the numbers of training data points from the target domain and the source domain, respectively. Thus we can define the merged training dataset as $X_{tr}^M = \{X_{tr}^S; X_{wtr}^T\}$, where X_{wtr}^T represents the weighted target training data using the weight defined above.

On the merged dataset, the between-class scatter matrix S_B^M, the within-class scatter matrix S_W^M and the total scatter matrix S_T^M are defined as follows:

$$S_B^M = \frac{1}{n}\sum_{i=1}^{c} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top, \tag{1}$$

$$S_W^M = \frac{1}{n}\sum_{i=1}^{c}\sum_{\boldsymbol{x}_j \in A_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top, \tag{2}$$

$$S_T^M = S_W^M + S_B^M, \tag{3}$$

where c is the class number, n is the sample number of the merged dataset, n_i is the number of samples belonging to the ith class, A_i is the set of ith class dataset, $\boldsymbol{\mu}_i$ is the class mean of ith class in the merged dataset, and $\boldsymbol{\mu}$ is the class mean of the merged dataset.

In cross-domain learning, only a few (even zero) labeled target samples can be used for later classification. Later classification models are vastly based on the source samples. So the low-dimensional latent space should bridge the source and target distributions to make sure the models trained on source samples can generalized well to the target domain. To measure the transferability between the source and target domains, a between-domain scatter matrix which bridges the source domain and the target domain is defined. The between-domain scatter matrix is related to the distance between the source and target distributions. Here as in our previous work we consider three different forms of the between-domain scatter matrix, i.e., $supervised$, $semi\text{-}supervised$ and $unsupervised$ between-domain scatter matrices.

1) $Supervised$ between-domain scatter matrix. In the supervised case, the between-domain scatter matrix S_L^{ST} is defined as follows:

$$S_L^{ST} = \sum_{i=1}^{c} \left(\boldsymbol{\mu}_i^S - \boldsymbol{\mu}_i^T\right)\left(\boldsymbol{\mu}_i^S - \boldsymbol{\mu}_i^T\right)^\top, \tag{4}$$

where $\boldsymbol{\mu}_i^S$ and $\boldsymbol{\mu}_i^T$ are the ith class means of the source training and target training datasets, respectively. The term $\boldsymbol{\mu}_i^S - \boldsymbol{\mu}_i^T$ reflects the scatter of class i between the source and target domains. To improve the transferability of the low-dimensional space, the distance between the distributions of the source and target domains should be minimized, thus making sure that the distributions of these two distributions can be as close as possible.

2) $Unsupervised$ between-domain scatter matrix. In many cross-domain learning problems, however, no labeled target data samples can be used in the training session. So the class mean can not be measured. One alternative way under such a setting is using the means of the whole data, i.e.,

$$S_U^{ST} = (\boldsymbol{\mu}^S - \boldsymbol{\mu}^T)(\boldsymbol{\mu}^S - \boldsymbol{\mu}^T)^\top, \tag{5}$$

where $\boldsymbol{\mu}^S$ and $\boldsymbol{\mu}^T$ are means of the source training and target training datasets, respectively. Similarly, the distance between the distributions of the source and target domains in the low-dimensional space should be minimized.

3) $Semi\text{-}supervised$ between-domain scatter matrix. When both a few labeled target samples and a large number of unlabeled target samples are available, we can combine the supervised and unsupervised between-domain scatter matrices to define a semi-supervised one. Similar with the weight definition in the merged training set X_{tr}^M, the different importances of the supervised and unsupervised between-domain scatter matrices should be considered with respect to the sample numbers of training (labeled) and test (unlabeled) samples from target domain as follows:

$$S^{ST} = S_U^{ST} + (1 + n_{tr}^T/n_{te}^T)S_L^{ST}, \tag{6}$$

where n_{tr}^T and n_{te}^T are the sample numbers of labeled and unlabeled target datasets. We attach more importance to the target data with label information.

B. SUFE

Once the scatter matrices are defined, the trace of the scatter matrices can be viewed as a measurement of the quality of the class structure and the domain characteristic. In particular, we can use $trace(S_B^M)$ to measure the distance between classes and use $trace(S_W^M)$ to measure the closeness of the data within the classes over all c classes. According to [12], there are some typical criteria, one of which is $J = trace(S_2^{-1}S_1)$, where S_1 and S_2 are combinations of S_W, S_B, and S_T. In a cross-domain learning problem, the closeness between the source domain and the target domain can be measured by $trace(S^{ST})$. So a generalized $\widetilde{S}_W$ can be defined as $\widetilde{S}_W = S_W^M + \alpha S^{ST}$ to measure both class closeness and domain closeness.

In the low-dimensional space mapped by the transformation $G^\top \in R^{L\times M}$, the between-class, within-class, between-domain and total scatter matrices can be written in the following form:

$$S_B^{ML} = G^\top S_B^M G, \qquad S_W^{ML} = G^\top S_W^M G,$$
$$S^{STL} = G^\top S^{ST} G, \qquad S_T^{ML} = G^\top S_T^M G.$$

To accomplish uncorrelated feature extraction, in [22] the optimization problem is defined as follows:

$$G^* = \arg\max_{G^\top S_T^M G = I} trace((S_W^{ML} + \alpha S^{STL})^{-1} S_B^{ML}). \quad (7)$$

The constraint $G^\top S_T^M G = I$ used here makes sure that the extracted features are mutually uncorrelated. So the low-dimensional space obtained contains the minimum information redundancy. Since the rank of the between-class scatter matrix is bounded by $c-1$, there are at most $c-1$ discriminant vectors in the solution. The uncorrelated feature extraction problem can be computed based on the generalized singular value decomposition (GSVD) [22], [33],

Next, let us characterize the solutions of the optimization problem (7). Inspired by [35], we employ the singular value decomposition (SVD) [13] to study the characteristics of all the solutions. Let $\widetilde{S}_W = S_W^M + \alpha S^{ST}$, $\widetilde{S}_B = S_B^M$ and $\widetilde{S}_T = S_T^M$. We decompose the three matrices $\widetilde{S}_W$, $\widetilde{S}_B$ and $\widetilde{S}_T$ as

$$\widetilde{S}_W = \widetilde{H}_W \widetilde{H}_W^\top, \quad \widetilde{S}_B = \widetilde{H}_B \widetilde{H}_B^\top, \quad \widetilde{S}_T = \widetilde{H}_T \widetilde{H}_T^\top.$$

Let the reduced SVD of $\widetilde{H}_T$ be

$$\widetilde{H}_T = U_1 \Sigma_T V_1^\top, \quad (8)$$

where $U_1 \in R^{*\times r}$ and $V_1 \in R^{*\times r}$ are column orthogonal, and $\Sigma_T \in R^{r\times r}$ is diagonal and nonsingular with $r = rank(\widetilde{H}_T)$. And let the reduced SVD of $\Sigma^{-1}U_1^\top \widetilde{H}_B$ be

$$\Sigma^{-1}U_1^\top \widetilde{H}_B = P_1 \Sigma_B Q_1^\top, \quad (9)$$

where $P_1 \in R^{*\times q}$ and $V_1 \in R^{*\times q}$ are column orthogonal, and $\Sigma_B \in R^{q\times q}$ is diagonal and nonsingular with $q = rank(\widetilde{H}_B)$. Then G is a solution of the optimization problem (7) if and only if $q \le L \le r$ and

$$G = (U_1 \Sigma_T^{-1}[P_1 M_1] + M_2)F, \quad (10)$$

where $M_1 \in R^{r\times(L-q)}$ is column orthogonal satisfying $M_1^\top P_1 = 0$, $M_2 \in R^{m\times L}$ is an arbitrary matrix satisfying $M_2^T U_1 = 0$, and $F \in R^{L\times L}$ is orthogonal.

For data dimensionality reduction, the goal is to find an optimal linear transformation G^* which transforms the original space into a low dimensional space. On the other hand, the dimension of the low space should be as small as possible. We propose a sparse solution to find the sparsest transformation of uncorrelated feature extraction from all G satisfying problem (7). A natural way to do this is to find a matrix G that minimizes the l_0-norm. However, l_0-norm is non-convex and NP-hard. So we use the convex relaxation of l_0-norm, say, l_1-norm to compute the sparsest solution of problem (7), resulting the following optimization problem, which is the main optimization problem proposed in this paper:

$$\begin{aligned} G^* &= \arg\min_G \|G\|_1, \\ &s.t. \quad U_1^\top G = \Sigma_T^{-1} P_1 F, F^\top F = I, \end{aligned} \quad (11)$$

where $\|G\|_1 = \sum_{i=1}^M \sum_{j=1}^q \|G_{ij}\|$.

When $q = 1$, the l_1-norm minimization problem in (11) can be reduced to the following problem

$$x_1^* = \arg\min \|x\|_1 : x \in R^n, Ax = b. \quad (12)$$

The numerical methods to solve problem (12) can be extended to solve problem (11) without any change. In this paper, we employed the accelerated linearized Bregman method [14] to solve problem (12). According to [14], the accelerated linearized Bregman method for solving (12) is :

$$\begin{cases} x^{i+1} = \beta \mathcal{S}_\xi(\tilde{y}^i), \\ y^{i+1} = \tilde{y}^i - \tau \mathcal{A}^\top(\mathcal{A}x^{i+1} - b), i \ge 0, \\ \tilde{y}^{i+1} = \gamma_i y^{i+1} + (1-\gamma_i) y^i, \end{cases} \quad (13)$$

where β, ξ and τ are positive parameters,$\gamma_i = \frac{2i+3}{i+3}$, $\tilde{y}^0 = y^0 = \tau \mathcal{A}^\top b$, and $\mathcal{S}_\xi(\cdot)$ is the component-wise soft-thresholding operator $\mathcal{S}_\xi(x) = sign(x) = \max\{\|x\| - \xi, 0\}$.

Now we can extend the accelerated linearized Bregman method (13) to the proposed optimization problem (11) as follows:

$$\begin{cases} G^{i+1} = \beta \mathcal{S}_\xi(\tilde{Y}^i), \\ Y^{i+1} = \tilde{Y}^i - \tau U_1(U_1^\top G^{i+1} - \Sigma_t^{-1} P_1 Z), \\ \tilde{Y}^{i+1} = \gamma_i Y^{i+1} + (1-\gamma_i) Y^i, \end{cases} \quad (14)$$

where $\tilde{Y}^0 = Y^0 = \tau U_1 \Sigma_t^{-1} P_1 Z$. As to the convergence of the above accelerated linear Bregman method, researchers can refer to the results in [14].

IV. Experiments

A. Datasets

In this work, we used three datasets to evaluate the proposed sparse uncorrelated cross-domain feature extraction method: the EEG dataset and two sEMG basic hand movements datasets.

The EEG dataset used in this study was a BCI dataset and provided by Dr. Allen Osman of University of Pennsylvania [19]. There were a total of nine subjects denoted as $S_1, S_2, S_3, \ldots, S_9$, respectively. Each subject was required to

imagine moving either the left or right index finger in response to a highly predictable visual cur. EEG data were recorded from 59 channels mounted according to the international 10/20 system. The sampling rate was 100 HZ. Each movement lasted for six seconds with two cues. The first cue turned up at 3.75s imagining which hand to move, and the second one appeared at 5.0s indicating that it was time to carry out the assigned response. For each subject, a total of 180 movements were recorded, with 90 trials labeled as left and the rest as right. Ninety movements with half labeled as right and half as left were used to training, while the other 90 for test in the experiments. We preprocessed the dataset as in [26].

The sEMG basic hand movements datasets were collected at a sampling rate of 500 Hz, using as a programming kernel the National Instruments (NI) Labview. Although they are not BCI datasets, their characteristics make them ideal to test the cross-domain feature extraction algorithm. The signals were band-pass filtered using a Butterworth Band Pass filter with low and high cutoff at 15Hz and 500Hz respectively and a notch filter at 50Hz to eliminate line interference artifacts. There are two different databases including: 1) data obtained from 5 healthy subjects (two males and three females, denoted as M_1, M_2 and F_1, F_2, F_3) of the same age approximately [20]. We regarded it as sEMG-1 dataset. 2) data obtained from 1 healthy subject during three days (denoted as D_1, D_2, D_3) [21]. We regarded it as sEMG-2 dataset. The subjects were asked to perform repeatedly the following six movements, which can be considered as daily hand grasps: spherical, tip, palmar, lateral, cylindrical, and hook. The five subjects in sEMG-1 were asked to conducted the six grasps for 30 times each and each grasp held for 6 seconds. The subject in sEMG-2 was asked to conducted the six grasps for 100 times each for 3 consecutive days and each grasp held for 5 seconds. For both datasets, we randomly select half of each movement for training and the rest data for test in the experiments.

B. Results and discussions

In our experiments, we design a one-source-vs-one-target transfer task for the three real datasets. The one-source-vs-one-target transfer task selects one domain (one subject in EEG dataset, one person in sEMG-1 dataset, or one day in sEMG-2 dataset) to act as the source domain, and selects another domain in the corresponding dataset as the target domain. In order to verify the effectiveness of the proposed method, we perform two previous methods, i.e., uncorrelated transferable feature extraction (UTFE) [22] and transferable discriminative dimensionality reduction (TDDR) [26], in the same setting as comparisons. When no labeled target data points are available in the training session, we have $S^{ST} = S_U^{ST}$ in the SUFE and UTFE approaches. Additionally, to simulate the real conditions that the target domain has only a few labeled data points for training, which is common in transfer learning, we also select some target training samples to help the classifications. The number of labeled target samples n_{tr}^T is set to five or ten. In these settings, $S^{ST} = S_U^{ST} + (1 + n_{tr}^T/n_{te}^T)S_L^{ST}$. The parameter α in our objective function Eq. (7) is selected from [0.1, 0.15, 0.2, 0.25, ..., 1] using 10-fold cross-validation technology. We employ the k-nearest-neighbor (kNN) classifier with $k = \{1, 3, 5\}$ to perform classifications. Therefore, there are nine experimental settings in total for each dataset with a combination of k and n_{tr}^T, where $n_{tr}^T = \{0, 5, 10\}$.

TABLE 1. The classification accuracies (%) for EEG dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$S1$	$S2$	$S3$
	UTFE TDDR SUFE	UTFE TDDR SUFE	UTFE TDDR SUFE
$S1$	-	60.0 63.8 **80.0**	75.3 65.9 **85.0**
$S2$	67.1 67.0 **76.3**	-	75.3 65.0 **87.5**
$S3$	83.5 67.5 **83.8**	73.6 80.0 **86.3**	-
$S4$	71.3 64.7 **80.0**	53.8 76.7 **77.5**	76.5 64.4 **80.0**
$S5$	76.3 70.0 **80.0**	77.6 63.8 **81.3**	75.3 66.7 **87.5**
$S6$	71.3 66.7 **77.5**	63.5 63.4 **77.8**	75.6 62.5 **78.8**
$S7$	75.3 62.5 **76.3**	70.0 80.0 **82.5**	77.6 65.6 **82.5**
$S8$	**86.3** 63.8 83.8	78.8 64.7 **83.8**	77.5 71.1 **86.3**
$S9$	75.3 64.4 **77.5**	77.6 68.2 **81.3**	78.8 67.5 **86.3**

TABLE 1 (continued). The classification accuracies (%) for EEG dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$S4$	$S5$	$S6$
	UTFE TDDR SUFE	UTFE TDDR SUFE	UTFE TDDR SUFE
$S1$	68.8 68.2 **77.5**	68.8 70.0 **78.8**	70.0 66.3 **81.3**
$S2$	54.1 61.2 **77.5**	62.5 67.8 **72.5**	55.0 61.1 **71.3**
$S3$	68.8 71.3 **80.0**	70.0 68.2 **78.8**	75.0 68.8 **78.8**
$S4$	-	71.8 73.8 **82.5**	73.0 68.8 **80.0**
$S5$	77.5 63.5 **81.2**	-	76.5 66.3 **80.0**
$S6$	76.3 63.3 **83.7**	80.0 70.6 **85.0**	-
$S7$	82.5 65.0 **83.7**	82.4 66.3 **83.8**	81.3 65.0 **82.5**
$S8$	71.8 63.7 **77.5**	72.9 65.9 **77.5**	74.4 67.1 **81.3**
$S9$	71.3 65.5 **80.0**	**76.5** 67.8 73.8	83.8 62.4 **86.2**

TABLE 1 (continued). The classification accuracies (%) for EEG dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$S7$	$S8$	$S9$
	UTFE TDDR SUFE	UTFE TDDR SUFE	UTFE TDDR SUFE
$S1$	78.9 66.3 **87.5**	65.9 71.3 **82.5**	72.9 67.1 **73.8**
$S2$	68.8 67.1 **86.3**	60.0 70.6 **86.3**	66.3 68.9 **78.8**
$S3$	84.4 71.7 **85.0**	69.4 70.6 **78.8**	73.7 71.3 **83.7**
$S4$	84.4 65.9 **91.5**	70.6 66.3 **80.0**	70.0 66.7 **81.3**
$S5$	**91.3** 63.8 90.0	70.0 71.8 **78.8**	**82.5** 68.8 78.7
$S6$	86.3 73.8 **88.8**	72.5 66.7 **86.3**	77.5 69.4 **81.3**
$S7$	-	73.8 63.5 **82.5**	77.5 65.0 **78.8**
$S8$	80.0 62.4 **85.0**	-	75.3 66.3 **78.8**
$S9$	76.3 72.5 **87.5**	70.6 70.0 **83.8**	-

TABLE 2. The classification accuracies (%) for sEMG-1 dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$M1$	$M2$	$F1$
	UTFE TDDR SUFE	UTFE TDDR SUFE	UTFE TDDR SUFE
$M1$	-	31.2 31.8 **43.5**	23.6 28.2 **30.0**
$M2$	36.5 27.6 **40.0**	-	21.2 22.4 **32.4**
$F1$	25.3 29.4 **34.1**	30.0 26.7 **40.0**	-
$F2$	22.4 18.8 **32.9**	26.5 22.9 **31.2**	25.3 40.0 **35.9**
$F3$	18.2 20.6 **27.6**	14.7 22.4 **24.7**	30.0 20.6 **37.1**

TABLE 2 (continued). The classification accuracies (%) for sEMG-1 dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$F2$	$F3$
	UTFE TDDR SUFE	UTFE TDDR SUFE
$M1$	15.9 **31.8** 29.4	15.3 24.1 **25.9**
$M2$	17.1 25.3 **30.6**	13.5 25.3 **29.4**
$F1$	20.6 33.5 **34.1**	22.9 22.9 **27.1**
$F2$	-	15.8 17.7 **27.7**
$F3$	18.2 21.8 **24.1**	-

TABLE 3. The classification accuracies (%) for sEMG-2 dataset when $k = 5$ in kNN and there are ten labeled target data points available. Each column reports three accuracies, using UTFE, TDDR, SUFE for classification, respectively.

S \ T	$D1$	$D2$	$D3$
	UTFE TDDR SUFE	UTFE TDDR SUFE	UTFE TDDR SUFE
$D1$	-	16.6 18.9 **19.8**	16.9 18.8 **19.3**
$D2$	16.6 **19.3** 19.2	-	16.6 19.5 **20.7**
$D3$	18.9 20.5 **20.9**	14.7 17.3 **23.2**	-

In Table 1∼3 we report the classification accuracies for the EEG dataset, the sEMG-1 dataset, and the sEMG-2 dataset, respectively, when there are ten labeled target data points available with $k = 5$ in the kNN classifier. Because we observed the similar comparison results and due to limitation of space, we don't report in this paper the results on the other eight settings. We can see from the results that the proposed sparse uncorrelated cross-domain feature extraction method (SUFE) outperforms the other two methods, TDDR and UTFE. For the EEG dataset, our method achieves a significant improvement in the classification accuracies. For the sEMG-1 dataset and the sEMG-2 dataset, to the best of our knowledge, we make the first attempt for cross-domain classification on these two datasets. To sum up, we offer a new baseline for the researchers who are interested in domain adaptation.

V. Conclusion

In this paper, we present a new sparse uncorrelated dimensionality reduction method for cross-domain learning in brain-computer interface systems. By maximizing the trace of the between-class scatter matrix and minimizing the trace of the within-class scatter matrix and the trace of the between-domain scatter matrix, the new method seeks an uncorrelated low-dimensional space which obtains the maximum discrimination and transferability between the source and target domains. What's more, by adding the l_1-norm penalty to the objective function, our method obtains a sparse LDA transformation, simultaneously. The evaluations on three real datasets demonstrate that our method outperforms the previous methods.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Project 61370175, and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

[1] M. Alamgir, M. Grosse-Wentrup, and Y. Altun. *Multitask learning for brain-computer interfaces*. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 17-24, 2010.

[2] B. Blankertz, G. Dornhege, M. Krauledat, K. R. Müller, and G. Curio. *The non-invasive berlin brain-computer interface: fast acquisition of effective performance in untrained subjects*. NeuroImage, vol. 37, pp. 539-550, 2007.

[3] C. M. Bishop, and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] J. Cai, S. Osher, and Z. Shen. *Convergence of the linearized Bregman iteration for l_1-norm minimization*. Mathematics of Computation, vol. 78, pp. 2127-2136, 2009.

[5] L. Clemmensen, T. Hastie, D. Wiiten, and B. Ersbøll. *Sparse discriminant analysis*. Technometrics, vol. 53, pp. 406-413, 2011.

[6] D. Chu, S. Goh, and Y. S. Hung. *Characterization of all solutions for undersampled uncorrelated linear discriminant analysis problems*. SIAM Journal on Matrix Analysis and Applications, vol. 32, pp. 820-844, 2011.

[7] D. Dai, and P. Yuen. *Regularized discriminant analysis and its application to face recognition*. Pattern Recognition, vol. 36, pp. 845-847, 2003.

[8] S. Dudoit, J. Fridlyand, and T. P. Speed. *Comparison of discrimination methods for the classification of tumors using gene expression data*. Journal of the American Statistical Association, vol. 87, pp. 77-87, 2002.

[9] M. M. Dundar, G. Fung, J. Bi, S. Sathyakama, and B. Rao. *Sparse Fisher discriminant analysis for computer aided detection*. In Proceedings of the SIAM International Conference on Data Mining, pp. 476-480, 2005.

[10] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K. R. Müller, and C. Grozea. *Subject-independent mental state classification in single trials*. Neural Networks, vol. 22, pp. 1305-1312, 2009.

[11] R. A. Fisher. *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, vol. 7, pp. 179-188, 1936.

[12] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[13] G. H. Golub, and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996.

[14] B. Huang, S. Ma, and D. Goldfarb. *Accelerated linearized Bregman method*. Journal of Scientific Computing, vol. 54, pp. 428-453, 2013.

[15] Z. Jin, J. Yang, Z. Hu, and Z. Lou. *Face recognition based on the uncorrelated discriminant transformation*. Pattern Recognition, vol. 34, pp. 1405-1416, 2001.

[16] Z. Jin, J. Yang, Z. Hu, and Z. Lou. *A theorem on the uncorrelated optimal discriminant vectors*. Pattern Recognition, vol. 34, pp. 2041-2047, 2001.

[17] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[18] S. G. Mason, A. Bashashati, M. Fatourechi, K. F. Navarro, and G. E. Birch. *A compenhensive survey of brain interface technology designs*. Annals of Biomedical Engineering, vol. 35, pp. 137-169, 2007.

[19] A. Osman, and A. Robert. *Time-course of cortical activation during overt and imagined movements*. In Proceedings of the Cognitive Neuroscientists Annual Meetings, pp. 1842-1852, 2001.

[20] C. Sapsanis, G. Georgoulas, A. Tzes, D. Lymberopoulos. *Improving EMG based classification of basic hand movements using EMD*. In Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5754-5757, 2013.

[21] C. Sapsanis. *Recognition of basic hand movements using electromyography*. 2013.

[22] H. Shi, J. Xu, and S. Sun. *Uncorrelated transferable feature extraction for signal classification in brain-computer interfaces*. In Proceedings of the International Joint Conference on Neural Networks, pp. 1-7, 2015.

[23] S. Sun, H. Shi, and Y. Wu. *A survey of multi-source domain adaptation*. Information Fusion, vol. 24, pp. 84-92, 2015.

[24] S. Sun, and J. Zhou. *A review of adaptive feature extraction and classification methods for EEG-based brain-computer interfaces*. In Proceedings of the International Joint Conference on Neural Networks, pp. 1746-1753, 2014.

[25] D. L. Swets, and J. Weng. *Using discriminant eigenfeatures for image retrieval*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, pp. 831-836, 1996.

[26] W. Tu, and S. Sun. *Transferable discriminative dimensionality reduction*. In Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence, pp. 865-868, 2011.

[27] W. Tu, and S. Sun. *A subject transfer framework for EEG classification*. Neurocomputing, vol. 82, pp. 109-116, 2012.

[28] R. Tibshirani. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, vol. 58, pp. 15149-15154, 1996.

[29] C. Vidaurre, A. Schlöegl, B. Blankertz, M. Kawanabe, and K. R. Müller. *Unsupervised adaptation of the LDA classifier for brain-computer interfaces*. In Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course, pp. 122-127, 2008.

[30] D. M. Witten, and R. Tibshirani. *Penalized classification using Fisher's*

linear discriminant. Journal of the Royal Statistical Society, vol. 73, pp. 753-772, 2011.

[31] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. *Brain-computer interfaces for communication and control*. Clinical Neurophysiology, vol. 113, pp. 767-791, 2002.

[32] J. R. Wolpaw, and D. J. McFarland. *Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans*. In Proceedings of the National Academy of Sciences of the United States of America, pp. 17849-17854, 2004.

[33] J. Ye, R. Janardan, Q. Li, and H. Park. *Feature extraction via generalized uncorrelated linear discriminant analysis*. In Proceedings of the 21st International Conference on Machine Learning, pp. 113-120, 2004.

[34] J. Ye, *Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems*. Journal of Machine Learning Research, vol. 6, pp. 483-502, 2005.

[35] X. Zhang, and D. Chu. *Sparse uncorrelated linear discriminant analysis*. Journal of Machine Learning Research, vol. 28, pp. 45-52, 2013.