Uncorrelated Transferable Feature Extraction for Signal Classification in Brain-Computer Interfaces

Honglei Shi, Jinhua Xu, 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, jhxu@cs.ecnu.edu.cn, slsun@cs.ecnu.edu.cn

Abstract—**This paper presents a novel dimensionality reduction method, called uncorrelated transferable feature extraction (UTFE), for signal classification in brain-computer interfaces (BCIs). Considering the difference between the source and target distributions of signals from different subjects, we construct an optimization objective that finds a projection matrix to transform the original data in a high-dimensional space into a low-dimensional latent space and that guarantees both the discrimination of different classes and transferability between the source and target domains. In the low-dimensional latent space, the model constructed in the source domain can generalize well to the target domain. Additionally, the extracted features are statistically uncorrelated, which ensure the minimum informative redundancy in the latent space. In the experiments, we evaluate the method with data from nine BCI subjects, and compare with the state-of-the-art methods. The results demonstrate that our method has better performance and is suitable for signal classification in BCIs.**

I. Introduction

Brain-computer interfaces (BCIs) are direct communication and control pathways between the brain and external devices, which are often used at assisting, augmenting, or repairing human cognitive or sensory-motor functions [21]. For example, they can predict the movement intentions of subjects (or users), e.g., left or right hand movement, by analyzing electrophysiological signals of the brain and translating the signals into certain physical commands [26]. In this way, BCIs can help the patients who suffer from motor disabilities to interact with the environment. During the last decade, the research of BCIs has quickened greatly, to which many machine learning methods have been successfully applied [2], [25], [27]. Due to the diversity among persons, the patterns of the recorded signals may differ considerably among subjects, especially at the early training stage. However, the latent characteristics of data may not change drastically according to the assumption in [17]. Therefore, some common knowledge can be learned from some subjects of BCIs to accelerate the training procedures for other subjects. An appealing approach in BCI systems is to use the inter-subject transformation in which one subject helps another to train a classifier for later tasks [1], [8], [23], [24]. In this paper, we focus on the knowledge transfer in the feature extraction stage and propose a novel transferable feature extraction method for signal classification in BCIs to quicken the training session in the systems.

Feature extraction is important for many applications in machine learning and data mining. In the single domain scenario where the training and test data are from the same distribution, there have been many methods proposed, such as principal component analysis (PCA) [16], linear discriminant analysis (LDA) [11], locality preserving projection (LPP) [13] and Fisher discriminant analysis (FDA) [9]. LDA seeks an optimal linear transformation by maximizing the ratio of the between-class distance to the within-class distance of a given data set. A major disadvantage of LDA is that the scatter matrices must be nonsingular. But in many applications, such as pattern recognition [3], [11], face recognition [14], [22], and microarray analysis [7], the between-class and within-class scatter matrices are usually singular for undersampled data, which is known as the $singularity\ problem$.

Many extensions are proposed to address the singularity problem, such as regularized linear discriminant analysis [6], [10], orthogonal LDA (OLDA) [30], subspace LDA [22], etc. Another important extension of LDA is uncorrelated LDA (ULDA) [5], [14], [15], [29]–[31], which is motivated by extracting features with uncorrelated attributes. The feature vectors extracted via ULDA are shown to be statistically uncorrelated, which is greatly desirable for many applications, because they contain minimum redundancy. The proposed algorithm in [14] involves d generalized eigenvalue problems, if there exist d optimal discriminant vectors. It is computationally expensive for high-dimensional and large dataset. And it does not address the singularity problem either. In [29], they addressed the singularity problem in classical ULDA by introducing an optimization criterion that combines the key ingredients of ULDA (which is based on QR-decomposition) and regularized LDA. Then they employed generalized singular value decomposition (GSVD) tool [12] to solve the singularity problem directly, thus avoiding the information loss in the subspace. The effectiveness of ULDA has been demonstrated by many numerical experiments [5], [14], [29], [31].

In BCIs, the signals recorded for one subject may be largely different from the signals for another subject. As a result, former feature extraction methods may not be suitable in these applications. Thanks to the development of transfer learning methods [20], one can learn classification models on one subject (the source domain) and apply them to another subject (the target domain). The work in this paper aims to design a new inter-subject transferable dimensionality reduction method for BCIs by improving the previous framework of uncorrelated discriminative dimensionality reduction methods, which is called uncorrelated transferable feature extraction, UTFE for short. The purpose of the UTFE approach is to learn a low-dimensional space in which the source and target distributions can be close to each other and the discrimination can be preserved. In this way, both the discrimination and the transferability of the transformed space are considered. Furthermore, the extracted features are statistically uncorrelated, indicating the minimum informative redundancy. We firstly design mathematical terms that evaluate the discrimination in domain-merged training data and the transferability of a latent space to bridge the source and target domains. Then we merge these terms into the previous framework of uncorrelated discriminative dimensionality reduction to minimize the distance between distributions of the data in different domains and maximize the discrimination in merged data simultaneously. The proposed feature extraction method not only preserves the discrimination but also bridges the source and target domains.

The organization of this paper is as follows. Section II reviews some related work on transfer learning, the BCI application and ULDA. Section III introduces the proposed method in detail, which is called uncorrelated transferable feature extraction. The experiment results on data from nine BCI subjects are presented in Section IV. The conclusion is given in Section V.

II. Related work

In [8], Fazli et al. introduced an ensemble method for the BCI application which was built upon common spatial pattern filters (CSP) for spatial filtering. They utilized a large database of pairs of spatial filters and classifiers from 45 subjects to learn a sparse subset of these pairs which were predictive across subjects. The quadratic regression with l_1 norm penalty was used to guarantee the sparsity. Using a leave-one-subject-out cross-validation procedure, the authors then demonstrated that the sparse subset of spatial filters and classifiers could be applied to new subjects with only a moderate performance loss compared to subject-specific calibration.

In [1], a multitask learning framework to construct a BCI was proposed, which could be used without any subject-specific calibration process. Each subject in their framework was treated as one task. They designed a parametric probabilistic approach that uses shared priors. 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 $y_i^t = f_t(\mathbf{x}_i^t; \mathbf{w}_t) + \varepsilon_t$, they firstly trained off-line tasks to learn the model parameters and the shared prior parameters. Then an out-of-the-box BCI with these shared prior parameters was defined and used to adapt to new subjects in an online fashion.

In [23], Tu and Sun introduced a novel dimensionality reduction method for transfer learning in BCI applications. By minimizing the distance between domains and maximizing the distance between classes, they found a low-dimensional latent space that ensure the discrimination of merged training data and the transferability between the source domain and the target domain, improving the original discriminative dimensionality reduction method. The experimental results on a real BCI dataset with two subjects demonstrate the effectiveness of their work. However in their work the extracted features may contain some redundancy information, degrading the classification performance.

A recent work by [24] introduced a subject transfer framework for EEG classification, which could achieve positive subject transfer with improvement on both feature extraction and classification stages. At the feature extraction stage, two kinds of spatial filter banks, i.e., robust filter bank and adaptive filter bank, were constructed for each subject. Then for each training set projected by each bank one classifier was trained with some strategies. At the classification stage, an ensemble strategy was employed to combine the outcomes of the classifiers trained above into a single one. 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.

To address the singularity problem in LDA, Ye et al. [29] proposed a feature extraction method called ULDA/GSVD. The optimization problem in their work was $G = \arg\max_G trace((S_T^L + \mu I_l)^{-1} S_B^L)$, subject to $G^\top S_T G = I_l$. Here $S_T^L + \mu I_l$ is always nonsingular for $\mu > 0$. The key of ULDA/GSVD is that the optimal solution is independent of the perturbation μ, i.e., $G_{\mu_1} = G_{\mu_2}$, for any $\mu_1, \mu_2 > 0$. So ULDA/GSVD is easier to use without adjusting the parameter μ.

III. The proposed method

In this section, we describe the proposed dimensionality reduction approach, uncorrelated transferable feature extraction (UTFE) in detail. As an extension of the work of [23], this approach extracts uncorrelated features that contain minimum redundancy.

Suppose we have three kinds of data available, 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. Each column represents one data point in these matrices. The unlabeled target test data X_{te}^T are used for later tasks (e.g., classification or regression). For transfer learning problems, the difficulties are that the target domain distribution relates to but differs from the source domain distribution. The methods learning on the single distribution may perform poorly. Thus the classifiers trained on the source domain may generalize badly to the target domain. Therefore, the quality of a low-dimensional space should be considered both by its discriminability in the merged data and transferability from the source domain to the target domain simultaneously. In this section, we consider both the transferability and discriminability to construct an objective function for effective feature extraction. We'll see in subsection III-B that the difference between the objective function we propose and the one in [23] is that the constraint induced here makes the extracted features uncorrelated, indicating the minimum informative redundancy.

A. Domain-merged and between-domain scatter matrices

When given source domain training dataset and target domain training dataset as mentioned above, we can compute the within-class and between-class scatter measurements on the dataset merged by them which is called merged training dataset X_{tr}^M. In transfer learning we should consider the different importance of the target domain and the source domain to take more advantages of the target because training and test data points in the target are drawn from the same distribution.

Hence a weight W_{tr}^T can be added into the target to control the influence of its training samples. Since the reliability of the distribution estimation of the target training set is constrained by its sample size intuitively, the weight should relate to the number of the target training samples. We define the weight as $W_{tr}^T = 1 + N_T/N_S$, where N_T and N_S are the numbers of training data points from the target domain and the source domain, respectively. This weight attaches more importance to the target training data. Thus the merged training dataset is defined 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 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.

To measure the transferability between the source and the target domains, a between-domain scatter matrix related to the distance between the source and target distributions is defined. Here three different forms of the between-domain scatter matrix , i.e., $supervised$, $semi\text{-}supervised$ and $unsupervised$ between-domain scatter matrices, are considered.

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 distributions of the source and target domains should be minimized.

2) $Unsupervised$ between-domain scatter matrix. In many transfer learning problems, however, the cases are that no labeled target data are available. So we can not measure the class mean per class. One alternative way to obtain the between-domain scatter matrix 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. If there are both a few labeled target samples and a large number of unlabeled target samples available, a semi-supervised between-domain scatter matrix can be obtained by combining the supervised and unsupervised ones. Similar to the weight definition in the merged training set X_{tr}^M, we give the supervised and unsupervised between-domain scatter matrices different importances using 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. UTFE for transfer learning

As in classical LDA, in order to formulate the criterion for class separability and domain transferability, we need to convert these scatter matrices defined above to an objective function. Its value should be larger when the between-class scatter is larger or the within-class and between-domain scatters are smaller. 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, $trace(S_B^M)$ measures the distance between classes and $trace(S_W^M)$ measures the closeness of the data within the classes over all c classes. According to [11], 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 transfer learning problem, $trace(S^{ST})$ measures the closeness of the two domains, i.e., the source domain and the target domain. So we can define a generalized $\widetilde{S}_W$ as $\widetilde{S}_W = S_W^M + \alpha S^{ST}$ to obtain both class closeness and domain closeness.

The goal in our method is to find a projection matrix G to transform the original space in the high-dimensional space (R^M) into a low-dimensional latent space (R^L). The desired projection matrix $G = [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_L]$ should be with the following characteristics:

- $G^\top \in R^{L\times M}$, where $L \ll M$;
- $Z = G^\top X, X \in X_{tr}^M$, so that $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are uncorrelated, where $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are the ith and jth feature components of Z.

The second condition makes sure the features extracted uncorrelated, which means that the low-dimensional space obtained contains the minimum informative redundancy.

In the low-dimensional space, the scatter matrices can be written in the form:

$$S_B^{ML} = G^\top S_B^M G, \qquad S_W^{ML} = G^\top S_W^M G,$$

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

As mentioned above, one of the reasonable criteria to characterize the class separability and domain transferability is to minimize $trace(\widetilde{S}_W^L)$ where $\widetilde{S}_W^L = S_W^{ML} + \alpha S^{STL}$, and maximize $trace(S_B^{ML})$ simultaneously, resulting in the optimization problem of UTFE method 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 ensures that the extracted features are mutually uncorrelated. 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.

Let X be an original feature vector, Z be the transformed feature vector of X with $Z = G^\top X$. Let $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ be the ith and jth feature components of Z and $\boldsymbol{z}_i = \boldsymbol{g}_i X$. The covariance between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ can be easily computed as

$$\begin{aligned} Cov(\boldsymbol{z}_i, \boldsymbol{z}_j) &= \mathbb{E}(\boldsymbol{z}_i - \mathbb{E}\boldsymbol{z}_i)(\boldsymbol{z}_j - \mathbb{E}\boldsymbol{z}_j) \\ &= \boldsymbol{g}_i^\top \{\mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top\} \boldsymbol{g}_j \\ &= \boldsymbol{g}_i^\top S_T \boldsymbol{g}_j. \end{aligned}$$

The correlation coefficient is

$$Cor(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{g}_i^\top S_T \boldsymbol{g}_j}{\sqrt{\boldsymbol{g}_i^\top S_T \boldsymbol{g}_i}\sqrt{\boldsymbol{g}_j^\top S_T \boldsymbol{g}_j}}, \quad (8)$$

for $i \neq j$, $Cor(\boldsymbol{z}_i, \boldsymbol{z}_j) = 0$ iff $\boldsymbol{g}_i^\top S_T \boldsymbol{g}_j = 0$. Therefore, the condition $G^\top S_T^M G = I$ guarantees the extracted features are mutually uncorrelated, and the optimization problem also makes sure both the transferability between domains and the separability between classes at the same time.

C. Solution of the UTFE problem

Let $\widetilde{S}_W = S_W^{ML} + \alpha S^{STL}$, $\widetilde{S}_B = S_B^{ML}$. From linear algebra, there exists a nonsingular matrix Y such that

$$Y^\top \widetilde{S}_W Y = I_M, \quad Y^\top \widetilde{S}_B Y = \Lambda = diag\{\lambda_1, \ldots, \lambda_M\}, \quad (9)$$

where $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$.

It can be shown that the matrix consisting of the first q columns of Y solves the optimization problem in (7), where $q = rank(\widehat{S}_B)$ (for proof, refer to [28]).

Considering that the scatter matrices are usually singular in real applications, we employ the method based on the generalized singular value decomposition (GSVD) [29] to solve the optimization objective. The GSVD is a common way to diagonalize two matrices together, which is in our situation that $\widetilde{S}_W$ and $\widetilde{S}_B$ should be simultaneously diagonalized.

We decompose the two matrices $\widetilde{S}_W$ and $\widetilde{S}_B$ as

$$\widetilde{S}_W = \widetilde{H}_W \widetilde{H}_W^\top,$$

$$\widetilde{S}_B = \widetilde{H}_B \widetilde{H}_B^\top.$$

Let

$$\Gamma = \begin{bmatrix} \widetilde{H}_B^\top \\ \widetilde{H}_W^\top \end{bmatrix}, \quad (10)$$

which is an $(n + c) \times M$ matrix with n being the number of merged data points and M the number of dimensions.

According to the generalized singular value decomposition [19], there exist orthogonal matrices $U \in R^{c \times c}$, $V \in R^{n \times n}$, and a nonsingular matrix $E \in R^{M \times M}$, such that

$$\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^\top \Gamma E = \begin{bmatrix} \Sigma_1 & 0 \\ \Sigma_2 & 0 \end{bmatrix}, \quad (11)$$

where

$$\Sigma_1 = \begin{bmatrix} I_B & 0 & 0 \\ 0 & D_B & 0 \\ 0 & 0 & 0_B \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0_W & 0 & 0 \\ 0 & D_W & 0 \\ 0 & 0 & I_W \end{bmatrix}, \quad (12)$$

$\Sigma_1^\top \Sigma_1 = \text{diag } (\alpha_1^2, \ldots, \alpha_t^2)$, $\Sigma_2^\top \Sigma_2 = \text{diag } (\beta_1^2, \ldots, \beta_t^2)$, $1 \geq \alpha_1 \geq \cdots \geq \alpha_q > 0 = \alpha_{q+1} = \cdots = \alpha_t$, $0 \leq \beta_1 \leq \cdots \leq \beta_t \leq 1$, $\alpha_i^2 + \beta_i^2 = 1$, for $i = 1, 2, \ldots, t$, $D_B = \text{diag } (\alpha_1^2, \ldots, \alpha_t^2, 0, \ldots, 0)$, $D_W = \text{diag } (\beta_1^2, \ldots, \beta_t^2, 0, \ldots, 0)$, $t = rank(S_T)$.

From (11), we have

$$\widetilde{H}_B^\top E = U \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix},$$

$$\widetilde{H}_W^\top E = V \begin{bmatrix} \Sigma_2 & 0 \end{bmatrix},$$

and

$$E^\top \widetilde{H}_B \widetilde{H}_B^\top E = \begin{bmatrix} \Sigma_1^\top \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \equiv D_B \ ,$$

$$E^\top \widetilde{H}_W \widetilde{H}_W^\top E = \begin{bmatrix} \Sigma_2^\top \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} \equiv D_W \ .$$

Hence, after computing GSVD on the matrix pair $(\widetilde{H}_B^\top, \widetilde{H}_W^\top)$ and obtaining the matrix E as Eq.(11), we can choose the first q columns to form the desired projection matrix G^*, that is $G^* \leftarrow [E_1, \ldots, E_q]$. The optimization problem in (7) is solved.

IV. Experiments

The EEG data used in this study were provided by Dr. Allen Osman of University of Pennsylvania [18]. 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 3.75 s imagining which hand to move, then the second one appeared at 5.0s indicating that it was time to carry out the assigned response. For each subject, a

TABLE 1. The classification accuracies without adaptation when $k = 5$ in kNN and there are ten labeled target data available.

Target / Source	S1(%)	S2(%)	S3(%)	S4(%)	S5(%)	S6(%)	S7(%)	S8(%)	S9(%)
S1	60.0	53.7	60.0	67.5	47.5	53.7	68.7	50.0	56.2
S2	56.3	67.5	58.7	57.5	48.5	47.5	58.7	50.0	56.2
S3	60.0	58.7	65.0	71.2	47.5	50.0	66.2	50.0	56.2
S4	66.3	50.0	60.0	77.5	56.2	47.5	72.5	52.5	56.2
S5	63.8	53.7	62.5	61.2	67.5	47.5	63.7	50.0	56.2
S6	61.2	47.5	58.7	72.5	55.0	58.7	70.0	50.0	56.2
S7	66.2	46.2	67.5	77.5	53.7	47.5	68.7	50.0	56.2
S8	60.0	51.2	60.0	63.7	51.2	47.5	61.2	66.2	55.0
S9	55.0	48.7	58.7	60.0	50.0	47.5	63.7	51.2	58.7

TABLE 2. The classification accuracies when $k = 5$ in kNN and there are ten labeled target data available. Each column reports three accuracies, using UTFE, TDDR, SDA for classification, respectively.

Target / Source	S1(%)			S2(%)			S3(%)		
	UTFE	TDDR	SDA	UTFE	TDDR	SDA	UTFE	TDDR	SDA
S1		-		60.0	63.8	**65.0**	**75.3**	65.9	53.75
S2	**67.1**	67.0	61.2		-		**75.3**	65.0	61.2
S3	**83.5**	67.5	51.3	73.6	**80.0**	58.8		-	
S4	**71.3**	64.7	61.3	53.8	**76.7**	65.0	**76.5**	64.4	62.5
S5	**76.3**	70.0	57.5	**77.6**	63.8	55.0	**75.3**	66.7	65.0
S6	**71.3**	66.7	60.0	63.5	63.4	**67.5**	**75.6**	62.5	58.7
S7	**75.3**	62.5	65.0	70.0	**80.0**	65.0	**77.6**	65.6	65.0
S8	**86.3**	63.8	51.3	**78.8**	64.7	53.8	**77.5**	71.1	68.8
S9	**75.3**	64.4	65.0	**77.6**	68.2	68.8	**78.8**	67.5	58.8

TABLE 2 (continue). The classification accuracies when $k = 5$ in kNN and there are ten labeled target data available. Each column reports three accuracies, using UTFE, TDDR, SDA for classification, respectively.

Target / Source	S4(%)			S5(%)			S6(%)		
	UTFE	TDDR	SDA	UTFE	TDDR	SDA	UTFE	TDDR	SDA
S1	**68.8**	68.2	50.0	68.8	**70.0**	62.5	**70.0**	66.3	51.3
S2	54.1	**61.2**	57.5	62.5	**67.8**	53.8	55.0	**61.1**	48.8
S3	68.8	**71.3**	52.5	**70.0**	68.2	53.8	**75.0**	68.8	52.5
S4		-		71.8	**73.8**	52.5	**73.0**	68.8	50.0
S5	**77.5**	63.5	55.0		-		**76.5**	66.3	63.8
S6	**76.3**	63.3	58.7	**80.0**	70.6	57.5		-	
S7	**82.5**	65.0	47.5	**82.4**	66.3	53.8	**81.3**	65.0	56.3
S8	**71.8**	63.7	68.8	**72.9**	65.9	67.5	**74.4**	67.1	58.8
S9	**71.3**	65.5	53.8	**76.5**	67.8	55.0	**83.8**	62.4	56.3

TABLE 2 (continue). The classification accuracies when $k = 5$ in kNN and there are ten labeled target data available. Each column reports three accuracies, using UTFE, TDDR, SDA for classification, respectively.

Target / Source	S7(%)			S8(%)			S9(%)		
	UTFE	TDDR	SDA	UTFE	TDDR	SDA	UTFE	TDDR	SDA
S1	**78.9**	66.3	56.3	65.9	**71.3**	45.0	**72.9**	67.1	50.0
S2	68.8	67.1	**72.5**	60.0	70.6	**76.3**	66.3	**68.9**	57.5
S3	**84.4**	71.7	61.3	69.4	**70.6**	36.2	**73.7**	71.3	66.2
S4	**84.4**	65.9	57.5	**70.6**	66.3	41.3	**70.0**	66.7	53.8
S5	**91.3**	63.8	57.5	70.0	71.8	**75.0**	**82.5**	68.8	47.5
S6	**86.3**	73.8	66.3	**72.5**	66.7	68.7	**77.5**	69.4	46.2
S7		-		**73.8**	63.5	43.7	**77.5**	65.0	66.3
S8	**80.0**	62.4	63.7		-		**75.3**	66.3	50.0
S9	**76.3**	72.5	66, 2	**70.6**	70.0	67.5		-	

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. The original dimension number M of each data point is eight. After solving the UTFE problem, we follow the approach in [28] and reduce the dimension to two.

In our experiments, we design a one-source-vs-one-target transfer task for all the nine subjects, i.e., (S_i, S_j) with $i, j = \{1, 2, \ldots, 9\}$, and $i \neq j$. The pair (S_i, S_j) presents that subject S_i acts as the source domain and S_j acts as the target domain. When no labeled target data points are available in the training session, we have $S^{ST} = S^{ST}_U$ in the UTFE approach. Additionally, to simulate the real conditions that the target domain only has a few labeled data points for training, which is common in transfer learning, we also select some target training samples to help the classification. The number of labeled target samples n^T_{tr} is set to five or ten. In these settings, $S^{ST} = S^{ST}_U + (1 + n^T_{tr}/n^T_{te})S^{ST}_L$. 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 experiment settings totally with a combination of k and n^T_{tr}, where $n^T_{tr} = \{0, 5, 10\}$. In order to verify the effectiveness of the proposed method, we perform a naive classification approach without adaptation and two previous methods, i.e., semi-supervised discriminant analysis (SDA) [4] and transferable discriminative dimensionality reduction (TDDR) [23], in the same settings as comparisons. The naive classification approach used in the experiment simply trained a classifier using the source labeled data and applied it to the target unlabeled data without adaptation.

Firstly, we need to determine whether S_i can help S_j without adaptation. The classification task is done on the original datasets. We use the k-nearest-neighbor (kNN) classifier with $k = 5$ and offer ten labeled target samples for training. Table 1 reports the classification accuracies. Compared with Table 2, we can see in Table 1 that S_i can not help S_j for classification without adaptation methods due to the data differences between them. Therefore, a more effective method is needed which takes the differences between the source domain and the target domain into consideration. Domain adaptation method can learn information from the source and then use it for the target carefully. The method we proposed can deal with such problem effectively. In Table 2, we can observe the significant accuracy improvements when applying adaptation method under the same condition, especially for subjects S_5, S_7, S_8 and S_9. Significantly, the accuracy of our method is 91.3% when S_5 helps S_7.

Secondly, we compare the proposed UTFE with previous methods, TDDR and SDA, to demonstrate the advantages of uncorrelated feature extraction approach. We have in all nine settings due to the different combinations of k in kNN and n^T_{tr}, where $k = \{1, 3, 5\}$ and $n^T_{tr} = \{0, 5, 10\}$. Due to the limitation of space, taking $k = 5$ and $n^T_{tr} = 10$ for example, we report the classification accuracies in Table 2. We can see from the results that in general the UTFE method outperforms the other two methods, TDDR and SDA. All of the three methods seek a low-dimensional latent space that preserves the discriminative characteristic. Considering the differences between the source domain and the target domain in transfer learning, our method and TDDR take the transferability into account. What's more, in our method, uncorrelated features are extracted, resulting in the minimum redundancy compared with TDDR.

Beside Table 2, for all the experimental settings, we list the average classification accuracies that source S_i (as the source) helps all the left eight subjects S_j (as the target) in Fig. 1. Each subfigure in Fig. 1 reports the average classification accuracies that the nine sources help the targets under each

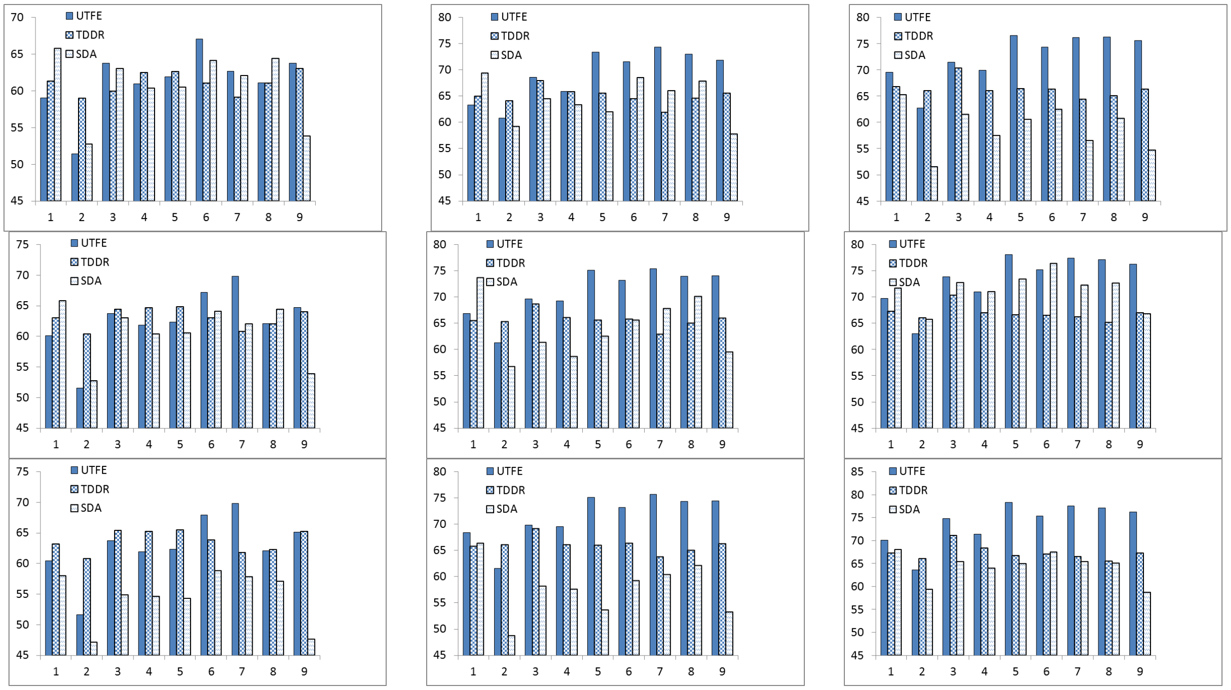

Fig. 1. The average classification results under nine experimental settings $\{(k, n_{tr}^T)\}$. The (k, n_{tr}^T) for each subfigure form top to down and left to right are (1,0), (1,5), (1,10), (3,0), (3,5), (3,10), (5,0), (5,5), (5,10). In each subfigure, the average classification accuracies of the case where each subject S_i (as the source) helps all the left eight subjects (as the targets) are reported, $i = \{1, 2, \ldots, 9\}$. The horizontal and vertical axes stand for the indexes of the source subjects, and the classification accuracy, respectively.

experimental setting, respectively. When there are no labeled target data available, our method is comparable with the other two. When there are some available labeled target data, the UTFE outperforms in a large scale. Due to the noise of the dataset, our method benefits more from the labeled target data, even there are only five, to obtain a more informative low-dimensional space for the classification.

In the end, a two-dimensional visualization of the data of subject S_7 helps by subject S_4 is shown in Fig. 2, where the samples are projected using the UTFE method and TDDR method. We can see from Fig. 2 that the data projected by UTFE has better class discrimination quality compared with TDDR. This is because, as mentioned above, the UTFE method seeks an uncorrelated feature space which contains the minimum informative redundancy and keeps the discrimination between classes, while the TDDR method contains informative redundancy.

V. Conclusions

In this paper, we present a new dimensionality reduction method for transfer 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 a low-dimensional space which obtains the maximum discrimination and transferability between the source and target domains. Meanwhile, the extracted features are statistically uncorrelated. By introducing the uncorrelated constraint the low-dimensional latent space reduces the informative redundancy, which improves the classification performance. The evaluations on real BCI data 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] D. Cai, X. He, and J. Han. *Semi-supervised discriminant analysis*. In Proceedings of the IEEE 11th International Conference on Computer Vision, pp. 1-7, 2007.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] J. Friedman. *Regularized discriminant analysis*. Journal of the American Statistical Association, vol. 84, pp. 165-175, 1989.

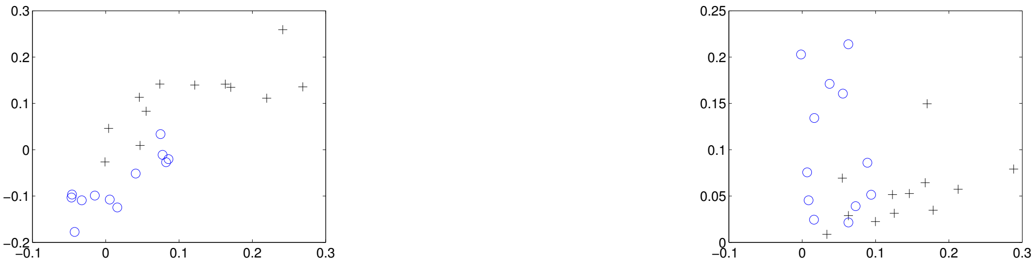

Fig. 2. 2D visualization of data of subject S_7 by the help of subject S_4. In the left subfigure, the samples are projected using the UTFE method, and in the right subfigure, the samples are projected onto the first two vectors by TDDR method.

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

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

[13] X. He, and P. Niyogi. *Locality preserving projections*. Advances in Neural Information Processing Systems, pp. 153-160, 2003.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] C. Paige, and M. Saunders. *Towards a generalized singular value decomposition*. SIAM Journal on Numerical Analysis, vol. 18, pp. 398-405, 1981.

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

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] J. Ye, T. Li, T. Xiong, and R. Janardan. *Using uncorrelated discriminant analysis for tissue classification with gene expression data*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 1, pp. 181-190, 2004.

[29] 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.

[30] 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.

[31] X. Zhang, and D. Chu. *Sparse uncorrelated linear discriminant analysis*. In Proceedings of the 30th International Conference on Machine Learning, pp. 45-52, 2013.