A Generalized Subspace Distribution Adaptation Framework for Cross-Corpus Speech Emotion Recognition

Shaokai Li [1] Peng Song [1 *] Liang Ji [1 4] Yun Jin [2] Wenming Zheng [3]

[1] Yantai University [2] Jiangsu Normal University

[3] Southeast University Key Laboratory of Child Development and Learning Science The State Key Laboratory of Tibetan Intelligent Information Processing and Application

Introduction

The exists transfer learning methods can effectively reduce the discrepancy between different corpora. However, they do not efficiently exploit the structural information between the samples from different corpora, which would affect the recognition performance to some extent. Different from most existing approaches, in this paper, we propose a generalized subspace distribution adaptation (GSDA) framework for cross-corpus SER. Specifically, we put forward a novel distance metric, which simultaneously explores the similarity and dissimilarity relationships between cross-domain samples. To verify the effectiveness of the proposed framework, we apply it to two popular subspace learning algorithms, i.e., PCA and LDA. It is worth mentioning that the proposed framework can be easily extended to combine with other types of subspace learning algorithms.

The Proposed Method

Problem Formulation

We aim to learn a common projection subspace P by aligning the source and target distributions, where the corpus discrepancy would be well reduced. The objective function of the proposed GSDA can be formulated as follows:

$$\min_{P} \ \mathcal{F}(P,X)+\mathcal{G}(P,X)+\gamma\,\mathcal{S}(P)$$
$$\text{s.t. } P^TP=I$$

The first item $\mathcal{F}(P,X)$ is a generalized subspace learning method, in which the original feature space is projected into a low-dimensional common subspace. The second item $\mathcal{G}(P,X)$ is the distance metric learning strategy. The third item $\mathcal{S}(P)$ is a sparse regularization term.

Distance Metric

To measure the distance across different corpora, we develop a novel distance metric strategy. It simultaneously takes into account the similarity and dissimilarity information during the process of knowledge transfer. Meanwhile, it can enhance the discriminative ability of the learned common subspace P. Thus, we minimize the following objective function:

$$\min_{P}\sum_{i,j}^{n}\|P^T(x_i-x_j)\|^2S_{ij}-\sum_{i,j}^{n}\|P^T(x_i-x_j)\|^2D_{ij}$$

where S_{ij} is a similarity weighted value, and D_{ij} is a dissimilarity weighted value.

According to the above equation, the term $\mathcal{G}(P,X)$ can be expressed as:

$$\mathcal{G}(P,X)=\alpha\|V\odot S\|_1-\beta\|V\odot D\|_1$$

Examples of GSDA

1) GSDA-PCA: By setting $\mathcal{F}(P,X)=\min_P -\text{Tr}(P^TXXP)$ and $\mathcal{S}(P)=\|P\|_{2,1}$, we get the objective function as follows:

$$\min_{P} -\text{Tr}(P^TXX^TP)+\alpha\|V\odot S\|_1-\beta\|V\odot D\|_1+\gamma\|P\|_{2,1}$$
$$\text{s.t. } P^TP=I$$

2) GSDA-LDA: By setting $\mathcal{F}(P,X)=\min_P \text{Tr}(P^T(S_w-\mu S_b)P)$ and $\mathcal{S}(P)=\|P\|_{2,1}$, we get the objective function as follows:

$$\min_{P} \text{Tr}(P^T(S_w-\mu S_b)P)+\alpha\|V\odot S\|_1-\beta\|V\odot D\|_1+\gamma\|P\|_{2,1}$$
$$\text{s.t. } P^TP=I$$

Experimental Setup

Dataset

Three emotional datasets: Berlin (B), IEMOCAP (I), CVE (C).

Four common emotional categories: anger (AN), neutral (NE), happiness (HA), and sadness (SA).

Feature Extraction

Low-level feature: we use the openSMILE toolkit to extract 1,582-dimensional low-level features, which is the standard emotional feature set of the INTERSPEECH 2010 Paralinguistic challenge.

Descriptors	Number of features
MFCC [0–14]	630
LSP frequency [0–7]	336
Log mel freq band [0–7]	336
Voicing prob	42
Loudness	42
F0 envelope	42
F0	38
Shimmer	38
Jitter	38
Jitter consecutive frame pairs	38
F0 number of onesets	1
Turn duration	1

Deep feature: we extract the Mel spectrograms to learn a 2,048-dimensional deep feature by ResNet50. Specifically, given a cross-corpus task, we fine-tune a pre-trained ResNet-50 model on the source corpus, and extract 2048-dimensional deep features of the target corpus using the fine-tuned model.

Emotional Evaluation

Training data: source database + random 7/10 target database.

Testing data: the remainder 3/10 target database.

Classifier: linear SVM.

Evaluation metric: the weighted average recall (WAR).

Results and Discussion

Results for Low-level Feature and Deep Feature

Tasks	Traditional methods		Transfer learning methods						GSDA -PCA	GSDA -LDA
	PCA	LDA	TCA	JDA	DaLSR	JGSA	LPJT	TSDSL		
B→I	44.21	40.11	46.54	46.06	49.19	45.24	45.96	49.25	**50.07**	50.35
B→C	45.74	39.35	49.83	48.16	48.22	49.67	46.87	52.12	**52.74**	51.32
I→B	30.31	40.62	55.85	53.66	49.37	59.70	**60.20**	59.79	57.10	59.18
I→C	35.16	30.32	43.38	47.12	51.61	46.32	53.21	51.19	**53.41**	48.38
C→B	56.54	56.33	59.81	62.62	49.47	58.14	59.18	63.70	63.95	**66.22**
C→I	44.62	32.39	47.13	48.11	49.94	47.78	46.69	46.51	**50.65**	50.45
Average	42.76	39.85	50.42	50.95	49.63	51.14	52.01	53.76	**54.65**	54.31

Important observations:

- The transfer learning methods can efficiently reduce the feature distribution discrepancy.
- The proposed transfer metric strategy can handle the data distribution discrepancy problem between corpora, which is neglected in traditional subspace learning methods.
- The feature distribution discrepancy between domains can be effectively reduced by considering the structural information of cross-domain samples.

Results for Deep Feature

Tasks	Traditional methods		Transfer learning methods								GSDA -PCA	GSDA -LDA
	PCA	LDA	TCA	JDA	DaLSR	JGSA	LPJT	TSDSL	DAN*	DSAN*	DAR*		
B→I	40.19	37.74	43.61	44.28	43.03	43.74	44.09	44.87	46.96	**47.63**	47.23	44.57	44.94
B→C	41.93	42.87	50.96	49.67	45.16	50.61	53.54	55.06	50.14	46.06	54.93	54.19	**55.80**
I→B	42.70	43.75	59.37	59.53	59.67	63.20	66.66	65.62	55.70	62.96	62.50	**66.75**	65.54
I→C	32.25	26.45	45.80	48.38	49.03	46.80	48.18	49.16	45.74	47.02	46.42	49.09	**49.23**
C→B	61.35	59.37	64.58	65.62	62.38	63.75	**69.79**	67.71	62.16	63.58	62.50	**69.79**	67.70
C→I	35.06	32.42	44.42	43.02	39.37	43.98	43.11	44.02	43.51	40.77	42.19	**45.39**	44.56
Average	42.24	40.43	51.45	51.75	49.77	52.01	54.22	54.40	50.70	51.33	52.62	**54.79**	54.62

Important observations:

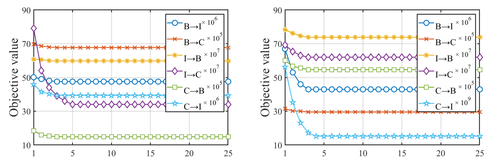
- These results are consistent with those of low-level features, which further validates the efficacy of the proposed framework.
- It would be interesting to integrate the proposed strategy into the deep learning framework, which might further improve the recognition performance.

Confusion Matrices

(a) B→I (b) B→C (c) I→B

(d) I→C (e) C→B (f) C→I

Convergence Analysis



(g) GSDA-PCA (h) GSDA-LDA

Conclusions

In this paper, we propose a novel transfer learning framework, called generalized subspace alignment adaptation (GSDA), for cross-corpus SER. The proposed GSDA utilizes a novel distance metric learning strategy to reduce the discrepancy between different corpora. In the future, we will investigate to develop the deep transfer learning methods using the the proposed strategy to solve the cross-corpus SER problem.

* The corresponding author. Peng Song's Space (https://drpengsong.github.io/)

48^{th} ICASSP, Rhodes Island, Greece

pengsong@ytu.edu.c