Multi-Source Unsupervised Transfer Components Learning for Cross-Domain Speech Emotion Recognition

Shenjie Jiang [1] Peng Song [1] * Shaokai Li[1] Run Wang [1] Wenming Zheng [2]

[1] School of Computer and Control Engineering, Yantai University, Yantai 264005, China

[2] Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, Southeast University, Nanjing 210096, China

Introduction

As an important research direction in the field of speech signal processing, **cross-domain speech emotion recognition (SER)** has attracted extensive attention. In practice, it is challenging to collect enough labeled samples from single source domain to train robust classifiers. To this end, this paper presents a novel method named **multi-source unsupervised transfer components learning (MUTCL)** for cross-domain SER. In MUTCL, We conduct experiments on **five benchmark datasets**, and the results show that MUTCL achieves **excellent performance** compared with some state-of-the-art methods.

The Proposed Method

Multi-source principal component learning

We perform a PCA-like strategy in multiple source and target domains separately. This can preserve the domain-specific components in each domain, which can fully utilize different but valuable information in multiple source domains and contribute to the cross-domain knowledge transfer. The objective function is written as follows:

$$\begin{aligned} &\min_{\widehat{\Phi}} \sum_{v=1}^{m} ||X_s^{(v)} - Q_s^{(v)} {P_s^{(v)}}^T X_s^{(v)}||_F^2 + ||X_t - Q^* {P^*}^T X_t||_F^2 \\ &s.t.\ {P_s^{(v)}}^T P_s^{(v)} = I, {P^*}^T P^* = I, {Q_s^{(v)}}^T Q_s^{(v)} = I, {Q^*}^T Q^* = I \end{aligned} \tag{1}$$

where $X_s^v \in \mathbb{R}^{d \times n_s^v}$ represents the v-th source data matrix, $X_t \in \mathbb{R}^{d \times n_t}$ represents the target data matrix, $\widehat{\Phi} = \{P_s^{(v)}, P^*, Q_s^{(v)}, Q^*\}$ is a set of variables, $P_s = \{P_s^{(v)} \in \mathbb{R}^{d \times k}\}_{v=1}^m$ are the multi-source projections that preserve the domain-specific transfer components. $P^* \in \mathbb{R}^{d \times k}$ is the common projection that preserves the common components in the target domain. $Q_s = \{Q_s^{(v)} \in \mathbb{R}^{d \times k}\}_{v=1}^m$ and $Q^* \in \mathbb{R}^{d \times k}$ are the reconstruction matrices responsible for mapping the projected cross-domain data to individual domains.

Note that different source domains have different feature distributions, which means that they play different roles in the process of knowledge transfer. We impose an adaptive weight to each source domain to weigh the importance of them. The objective function is written as follows:

$$\begin{aligned} &\min_{\widehat{\Phi}, \alpha^{(v)}} \sum_{v=1}^{m} \alpha^{(v)} ||X_s^{(v)} - Q_s^{(v)} {P_s^{(v)}}^T X_s^{(v)}||_F^2 + ||X_t - Q^* {P^*}^T X_t||_F^2 + \sum_{v=1}^{m} \gamma (\alpha^{(v)})^2 \\ &s.t.\ {P_s^{(v)}}^T P_s^{(v)} = I, {P^*}^T P^* = I, {Q_s^{(v)}}^T Q_s^{(v)} = I, {Q^*}^T Q^* = I, \alpha^v \geq 0, \sum_{v=1}^{m} \alpha^v = 1 \end{aligned} \tag{2}$$

where α^v is the adaptive weight of the $v-$th source domain, and $\gamma \geq 0$ is a regularization parameter.

Multi-source domains alignment

From Eq. (1), we can get the following equation:

$$\left.\begin{aligned} &X_s^{(v)} \approx Q_s^{(v)} {P_s^{(v)}}^T X_s^{(v)}, \quad X_t \approx Q^* {P^*}^T X_t \\ &{Q_s^{(v)}}^T Q_s^{(v)} = I, \quad {Q^*}^T Q^* = I \end{aligned}\right\} \Rightarrow \left.\begin{aligned} &{Q_s^{(v)}}^T X_s^{(v)} \approx {P_s^{(v)}}^T X_s^{(v)} \\ &{Q^*}^T X_t \approx {P^*}^T X_t \end{aligned}\right\} \tag{3}$$

We hope that all cross-domain samples follow a similar geometric structure, which can significantly reduce the differences in feature distributions between multiple source and target domains, i.e., $\phi_s({P_s^{(v)}}^T X_s^{(v)}) \approx \phi_t({P^*}^T X_t)$. This problem can be solved by a simple strategy with the following formula:

$$\min_{Q_s^{(v)}, Q^*} \sum_{v=1}^{m} ||Q_s^{(v)} - Q^*||_F^2 \tag{4}$$

By minimizing Eq. (4), we have the following equation:

$$Q_s^{(v)} \approx Q^* \Rightarrow \phi_s({Q_s^{(v)}}^T X_s^{(v)}) \approx \phi_t({Q^*}^T X_t) \Rightarrow \phi_s({P_s^{(v)}}^T X_s^{(v)}) \approx \phi_t({P^*}^T X_t) \tag{5}$$

To make the multi-source domain and the target domain further be aligned, we minimize the following problem:

$$\min_{P_s^{(v)}, P^*} \sum_{v=1}^{m} ||P_s^{(v)} - P^*||_F^2 \tag{6}$$

By incorporating Eq. (2), Eq. (4) and Eq. (6), **the overall objective function of the proposed MUTCL** can be written as follows:

$$\begin{aligned} &\min_{\widehat{\Phi}, \alpha^{(v)}} \sum_{v=1}^{m} \alpha^{(v)} ||X_s^{(v)} - Q_s^{(v)} {P_s^{(v)}}^T X_s^{(v)}||_F^2 + ||X_t - Q^* {P^*}^T X_t||_F^2 \\ &\quad + \sum_{v=1}^{m} \beta ||P_s^{(v)} - P^*||_F^2 + \sum_{v=1}^{m} \lambda ||Q_s^{(v)} - Q^*||_F^2 + \sum_{v=1}^{m} \gamma (\alpha^{(v)})^2 \\ &s.t.\ {P_s^{(v)}}^T P_s^{(v)} = I, {P^*}^T P^* = I, {Q_s^{(v)}}^T Q_s^{(v)} = I, {Q^*}^T Q^* = I, \alpha^v \geq 0, \sum_{v=1}^{m} \alpha^v = 1 \end{aligned} \tag{7}$$

Experimental Settings

Dataset

Five emotional datasets: EMO-DB (B), CVE (C), EMOVO (E) , IEMOCAP (I) , and TESS (T).
Four common emotional categories: anger (AN), neutral (NE), happiness (HA), and sadness (SA).

Feature Extraction

Low-level feature: 1582-dimensional standard feature set used in INTERSPEECH 2010 Paralinguistic challenge. **Deep feature:** 2048-dimensional deep features extracted by ResNet-50 model on Mel spectrograms.

Emotional Evaluation

Training data: source database + random 7/10 of target database. **Testing data:** the rest 3/10 of target database. **Classifier:** linear SVM. **Evaluation metric:** the weighted average recall (WAR).

Results and Discussions

Results for Low-level Feature

Settings	Single-source methods (optimal source domain results)					Multi-source methods		MUTCL
	PCA	JDA	BDA	JTSLR	DLAD	MMFT	MDSA	
{C, E, I, T}→B	63.21±1.23	67.70±2.76	65.38±0.92	66.72±1.69	69.83±0.26	69.00±2.03	69.87±0.41	**75.73±1.98**
{B, E, I, T}→C	52.65±0.75	65.51±1.78	60.28±1.34	64.36±0.96	67.91±0.13	63.33±1.79	67.58±1.07	**69.68±0.71**
{B, C, I, T}→E	43.00±1.87	48.00±1.54	44.00±0.58	42.00±1.11	**51.00±0.89**	47.00±0.63	**51.00±0.74**	45.50±1.57
{B, C, E, T}→I	47.25±0.63	46.95±0.72	48.52±1.45	48.13±0.52	45.76±0.11	52.08±0.62	52.49±0.09	**57.03±0.21**
{B, C, E, I}→T	55.45±1.56	63.95±1.29	63.78±0.81	58.73±1.78	61.28±0.75	64.58±0.15	64.58±0.81	**65.60±1.02**
Average	52.31	58.42	56.39	55.99	59.15	59.19	61.10	**62.71**

Observations:

- The proposed MUTCL achieves the best results in all settings, which are 3.52% and 1.61% higher than that of the multi-source domain methods MMFT and MDSA, respectively.
- Most of unsupervised methods achieve better results than the regression method JTSLR.
- All multi-source transfer learning methods perform better than the single-source transfer learning methods.

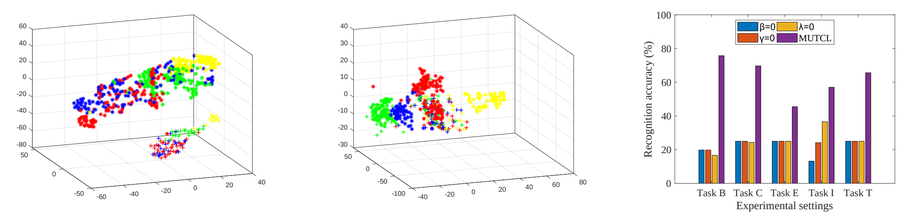
Results for Deep Feature

Settings	Single-source methods (optimal source domain results)					Multi-source methods		MUTCL
	JTSLR	DLAD	MRAN*	DAR*	DIFEX*	MMFT	MDSA	
{C, E, I, T}→B	63.92±1.02	70.88±0.16	70.68±1.65	70.83±1.12	67.42±0.60	71.87±2.14	71.87±0.65	**75.00±0.81**
{B, E, I, T}→C	49.36±1.29	**68.59±0.55**	45.49±0.71	60.76±1.43	59.58±1.22	64.74±2.27	66.66±0.74	60.13±1.56
{B, C, I, T}→E	49.00±0.69	49.00±1.41	53.56±1.40	53.84±1.93	51.75±0.83	51.00±2.31	**55.00±1.01**	54.00±1.40
{B, C, E, T}→I	**48.85±1.95**	44.50±0.61	48.24±0.87	42.49±0.85	46.20±1.03	48.66±1.17	44.79±1.02	44.28±0.21
{B, C, E, I}→T	69.71±0.54	65.94±0.12	58.47±1.50	68.54±1.57	63.22±0.98	68.25±0.87	68.75±0.68	**75.71±0.63**
Average	56.17	59.78	55.29	59.29	57.63	60.90	61.41	**61.82**

Observations:

- It can be noticed that MUTCL obtains the best recognition accuracy using deep features. This result is consistent with the performance using low-level features.
- The proposed MUTCL achieves much better performance than the deep learning methods. This demonstrates the effectiveness and superiority of our method.

t-SNE visualization and ablation study



(a) Original data (b) MUTCL (c) ablation study

Observations:

- Fig. (a) and (b) show the t-SNE visualization of data representation on the task {C, E, I, T}→B. From the figure, we can find that in our method, the samples from source and target domains are well intertwined, the cross-domain samples of the same category are closer.
- Fig. (c) shows the results of ablation study. From the figure, we can find that the recognition accuracy decreases regardless of which term is ignored, which indicates that all the modules contribute positively to MUTCL.

Conclusions

In this paper, we propose a new multi-source **cross-domain SER** method called **multi-source unsupervised transfer components learning (MUTCL)**. To be specific, MUTCL first performs a **PCA-like strategy** in each source domain separately to preserve **common and domain-peculiar information** in each **source domain** and the **target domain**. Then, it balances the contributions of multi-source domains by **assigning weights** to them and further **reduces the differences** between multi-source domains and target domains by aligning the domains. Finally we can obtain a **common projection subspace**. Extensive experiments are carried out on five popular datasets, and the results validate the effectiveness of the proposed MUTCL method for cross-domain SER.

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

pengsongseu@gmail.com