LETTER

Transfer Semi-Supervised Non-Negative Matrix Factorization for Speech Emotion Recognition

Peng SONG[†], ***Member***, **Shifeng OU**[††], **Xinran ZHANG**[†††], **Yun JIN**[†††], **Wenming ZHENG**[†††a)], **Jinglei LIU**[†], ***and*** **Yanwei YU**[†], ***Nonmembers***

SUMMARY In practice, emotional speech utterances are often collected from different devices or conditions, which will lead to discrepancy between the training and testing data, resulting in sharp decrease of recognition rates. To solve this problem, in this letter, a novel transfer semi-supervised non-negative matrix factorization (TSNMF) method is presented. A semi-supervised negative matrix factorization algorithm, utilizing both labeled source and unlabeled target data, is adopted to learn common feature representations. Meanwhile, the maximum mean discrepancy (MMD) as a similarity measurement is employed to reduce the distance between the feature distributions of two databases. Finally, the TSNMF algorithm, which optimizes the SNMF and MMD functions together, is proposed to obtain robust feature representations across databases. Extensive experiments demonstrate that in comparison to the state-of-the-art approaches, our proposed method can significantly improve the cross-corpus recognition rates.

key words: *speech emotion recognition, transfer learning, non-negative matrix factorization, semi-supervised learning*

1. Introduction

Speech emotion recognition is a hot research topic in speech signal processing areas. It has been found very useful in many real applications [1], such as helping doctors diagnose patients' psychological illness in the medical field, computer tutoring service in the education area, the human computer interaction (HCI) based entertainment industry.

In recent years, there have been many developments on speech emotion recognition techniques. The popular methods include support vector machine (SVM), Gaussian mixture model (GMM), artificial neural network (ANN) and deep neural network (DNN) [1], [2]. These approaches achieve satisfactory performance to some extent. However, they are all conducted on the assumption that the training and testing data sets are from the same corpus. In practice, these data are often collected in different scenarios, e.g., ages, genders, languages, noises, which will significantly degrade the recognition performance.

Manuscript received March 27, 2016.
Manuscript revised June 2, 2016.
Manuscript publicized July 1, 2016.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China.
[††]The author is with the School of Science and Technology for Opto-electronic Information, Yantai University, Yantai 264005, P.R. China.
[†††]The authors are with the Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, P.R. China.
a) E-mail: wenming_zheng@seu.edu.cn

DOI: 10.1587/transinf.2016EDL8067

In dealing with these mismatch problems, some efforts have been made in the past few years. Schuller et al. [3] investigate the performance of data agglomeration and decision-level fusion for cross-corpus speech emotion recognition. Deng et al. [4] present an autoencoder based unsupervised domain adaptation approach to reduce the discrepancy of the training and testing conditions. Abdelwahab et al. [5] explore a supervised model adaptation algorithm. These algorithms can solve the discrepancy problem to some extent. However, they are inapplicable to real-world tasks in that they do not take into account the difference between training and testing corpora. In practical situations, the discrepancy between feature distributions of these two corpora is often very large, which will lead to a sharp decrease in recognition accuracy.

Motivated by recent progress in non-negative matrix factorization (NMF) [6] and transfer learning [7] techniques, in this letter, a novel transfer semi-supervised NMF (TSNMF) approach is presented to investigate the robust feature representations between different corpora. In contrast to traditional dimension reduction based transfer learning algorithms [8], the STNMF approach optimizes the dimension reduction and similarity measurement function together, in which the semi-supervised (SNMF) algorithm is employed for dimension reduction, while the maximum mean discrepancy (MMD) is used for distance criterion between two data sets. Meanwhile, it is important to note that the similar idea has also been presented in our previous study [9], in which, a transfer sparse coding (TSC) approach is proposed. It is worthwhile to highlight the differences between TSC algorithm and our proposed TSNMF algorithm. First, in the proposed approach, the matrix factorization technique, called NMF, is employed to learn robust low dimensional feature representations, while the sparse coding algorithm is used to achieve sparse feature representations in the TSC method. Second, the TSC is an unsupervised transfer learning approach, while our proposed TSNMF method is a semi-supervised transfer learning approach, in which the label information is considered as an additional information. Thus, our approach can have more discriminating power than the TSC approach.

The remainder of this letter is organized as follows. In Sect. 2, the semi-supervised NMF algorithm is introduced. In Sect. 3, the TSNMF method is proposed. Experimental results are provided in Sect. 4. Finally, Sect. 5 summarizes the letter.

2. Semi-Supervised Non-Negative Matrix Factorization

2.1 Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) is a very popular machine learning method that can solve many real-world problems with non-negative data [6]. The goal of NMF is to find two matrices whose product can obtain a good approximation of the original data matrix. It has been found very useful in many real applications, e.g., gene expression, document clustering and text mining.

Given the non-negative data $X = [x_1, \ldots, x_N] \in R^{M\times N}$ and the reduced rank K, NMF aims to approximate X by a linear combination of dictionary $U = [u_{ik}] \in R^{M\times K}$ and codes $V = [v_{kj}] \in R^{K\times N}$:

$$\min_{U,V} \|X - UV\|_F^2 \tag{1}$$

with constraints $U,\ V \geq 0$, where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

2.2 Semi-Supervised NMF

The NMF algorithm is an unsupervised method. That is, it is inapplicable to many real-world problems, e.g., classification task. In reality, the labeled training data is often expensive and insufficient, but there exist a large number of unlabeled data. A possible solution, called semi-supervised learning has been presented, in which only the unlabeled data along with a small amount of labeled data can significantly improve the learning accuracy. Therefore, it is natural and reasonable to extend NMF to the semi-supervised case. A semi-supervised NMF (SNMF) algorithm, called constrained NMF is proposed in [10]. In this method, the label information is utilized as a hard constraint, and the data points with the same labels will be merged together in the new representations.

Let $X_s = [x_1, \cdots, x_{n_l}] \in R^{M\times n_l}$ and $X_t = [x_{n_l+1}, \cdots, x_N] \in R^{M\times n_u}$ denote the labeled source and unlabeled target feature matrices, respectively, and $L = [l_1, \cdots, l_{n_l}] \in R^{c\times n_l}$ is the corresponding label indicator matrix, where n_l, n_u, c denote the number of labeled source features, the number of unlabeled features, and the number of emotion categories, respectively, and $N = n_l + n_u$. For each feature vector, one class is labeled, and the indicator matrix satisfies

$$l_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is labeled with the } j\text{−th class} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

A label constraint matrix C can be defined with L as follows

$$C = \begin{pmatrix} L_{c\times n_l} & 0 \\ 0 & \mathbf{I}_{n_u} \end{pmatrix} \tag{3}$$

where $\mathbf{I}_{n_u} \in R^{n_u\times n_u}$ is an identity matrix. By introducing an auxiliary matrix D, the V can be expressed as

$$V = DC \tag{4}$$

From the above equation, it can be easily found that if x_i and x_j have the same emotion labels, then $v_i = v_j$. With the label constraint, the NMF can be extended to the semi-supervised NMF. By using the matrix product of U, D and C, the approximation function of the original NMF can be rewritten as follows

$$\min_{U,D} \|X - UDC\|_F^2 \tag{5}$$

3. Proposed Methodology

3.1 Minimizing the Distribution Divergence

By utilizing the SNMF algorithm, the low dimensional latent feature representations can be obtained for the labeled source and unlabeled target data sets. However, it should be noted that the divergence between the two feature distributions is often very large [4], [8], which will greatly affect the recognition performance. To address this shortcoming, in this letter, the distribution divergence is considered, and the empirical maximum mean discrepancy (MMD) [7] is employed to compare two distributions. The distance between the feature distributions of the labeled source and unlabeled target corpora using the K-dimensional embeddings is given as

$$\begin{aligned} Dist &= \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} v_i - \frac{1}{n_u}\sum_{j=n_l+1}^{N} v_j \right\|^2 \\ &= \sum_{i,j=1}^{N} v_i^T v_j m_{ij} \\ &= tr(VMV^T) \end{aligned} \tag{6}$$

where $V = [V_{src}, V_{tar}]$, in which $V_{src} = [v_1, \ldots, v_{n_l}] \in R^{K\times n_l}$ and $V_{tar} = [v_{n_l+1}, \ldots, v_N] \in R^{K\times n_u}$ are the coding matrices of labeled source and unlabled emotional features, respectively, $tr(\cdot)$ denotes the trace of a matrix, ^{T} is the transformation of a matrix, and $M = [m_{ij}]_{i,j=1}^N \geq 0$ is the MMD matrix, and is defined as

$$m_{ij} = \begin{cases} \frac{1}{n_l^2} & v_i, v_j \in V_{src} \\ \frac{1}{n_u^2} & v_i, v_j \in V_{tar} \\ \frac{-1}{n_l n_u} & \text{otherwise} \end{cases} \tag{7}$$

3.2 TSNMF Model

In speech emotion recognition, the labeled information is crucial to ensure the recognition accuracy. In practice, the labeled data are often very sparse, one may expect to leverage the abundant labeled emotional speech in source corpus for training the classifier in the target corpus. To solve this situation, the transfer learning technique is employed in this letter. Transfer learning has been widely studied over the last decade [7]. It has been proven very promising in many

fields, e.g., image classification, text categorization, sentiment analysis, video summarization and object recognition.

In this letter, to efficiently perform cross-corpus speech emotion recognition, a novel transfer learning algorithm, called transfer semi-supervised non-negative factorization (TSNMF) is put forward, in which the MMD is integrated with SNMF to achieve effective and robust feature representations. Therefore, by incorporating Eq. (6) into Eq. (5), the TSNMF optimization problem can be written as

$$\begin{aligned} &\min_{U,D} \|X - UDC\|_F^2 + \lambda tr(DCMC^T D^T) \\ &\text{s.t. } U,\ D \geq 0 \end{aligned} \tag{8}$$

where $\lambda \geq 0$ is a regularization parameter trading off the weight between the SNMF and distribution matching. As NMF, the above objective function is not convex in both U and D together. Therefore, it is unrealistic to design an optimization algorithm to obtain the global minima. In the following, an iterative algorithm is presented to obtain the local optima. The Eq. (8) can be rewritten as

$$\begin{aligned} &\min_{U,D} tr(XX^T) + tr(U(DC)(DC)^T U^T) \\ &\quad - 2tr(X(DC)^T U^T) + \lambda tr((DC)M(DC)^T) \\ &\text{s.t. } U,\ D \geq 0 \end{aligned} \tag{9}$$

where the matrix properties $tr(AB) = tr(BA)$ and $tr(A) = tr(A^T)$ is applied. Given the Lagrange multiplier matrices $\alpha = [\alpha_{ik}] \geq 0$ and $\beta = [\beta_{kj}] \geq 0$, the Lagrange function $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} =& tr(XX^T) + tr(UDC(DC)^T U^T) - 2tr(X(DC)^T U^T) \\ &+ tr(\lambda DCM(DC)^T) + tr(\alpha U) + tr(\beta DC) \end{aligned} \tag{10}$$

The derivatives of $\mathcal{L}$ with respect to U and D vanish, and given as

$$\frac{\partial \mathcal{L}}{\partial U} = 2UDCC^T D^T - 2XC^T D^T + \alpha = 0 \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial D} = 2U^T UDCC^T - 2U^T X + 2\lambda DCMC^T + \beta = 0 \tag{12}$$

Using the KKT conditions $\alpha_{ik}u_{ik} = 0$ and $\beta_{kj}d_{kj} = 0$, the following equations for u_{ik} and d_{kj} will be obtained

$$(UDCC^T D^T)_{ik} u_{ik} - (XDC)_{ik} u_{ik} = 0 \tag{13}$$

$$(U^T U)_{kj} d_{kj} + \lambda (DCMC)_{kj} d_{kj} - (U^T X)_{kj} d_{kj} = 0 \tag{14}$$

Then the following updating rules will be obtained as follows

$$u_{ik} \leftarrow u_{ik} \frac{(XDC)_{ik}}{(UDCC^T D^T)_{ik}} \tag{15}$$

$$d_{kj} \leftarrow d_{kj} \frac{(U^T X + V\lambda M^-)_{kj}}{(DCU^T U + DC\lambda M^+)_{kj}} \tag{16}$$

where M^+ and M^- are the positive and negative parts of M, respectively. Applying Eq. (15) and Eq. (16) iteratively until the convergence is reached, an optimal local minima will be obtained.

4. Experiments

4.1 Emotional Speech Databases

To evaluate the performance of our proposed method for cross-corpus speech emotion recognition, extensive experiments are conducted on two widely used data sets, i.e., Berlin database[†] and eNTERFACE database[††]. The Berlin database is a very public emotional speech data set, and consists of seven basic emotion categories, including neutral, happiness, anger, boredom, disgust, fear and sadness. Total 494 speech utterances in German are collected from 10 subjects, which are all utilized in our experiments. The eNTERFACE database is an audio-visual widely used data set. It includes six types of basic emotions, i.e., happiness, anger, disgust, fear, sadness and surprise. This database is recorded in English by 42 subjects from 14 countries. Finally total 1170 audio-visual samples are collected and used for our experiments.

4.2 Experimental Setup

To efficiently extract the acoustic features, the openSMILE toolkit[†††] is employed, and the 1582 dimensional standard feature set of INTERSPEECH 2010 Paralinguistic Challenge [11] is adopted.

To demonstrate how the recognition performance can be improved by our method, the following seven types of methods are compared. The traditional method (*Traditional*), in which the classifier trained in source database is directly applied for recognition in the target database, the baseline method (*Baseline*), in which the training and testing procedures are carried out on the same single database, the dimension reduction based transfer learning method (DR) [8], the transfer component analysis method (TCA) [7], the transfer sparse coding method (TSC) [9], the NMF method (NMF) [6], the semi-supervised NMF method (SNMF) [10], and our proposed TSNMF method (Ours). As classifier, the classic SVM algorithm is chosen in our experiments. The trade-off parameter λ is set by searching $\lambda \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, and finally is optimized as 0.1 for our evaluations.

4.3 Experimental Results

Two types of cross-corpus strategies are utilized to compare the performance of the aforementioned seven speech emotion recognition methods, i.e., *case1* and *case2*. Specifically, in *case1*, the labeled Berlin corpus is used for training,

[†] http://emodb.bilderbar.info/docu/
[††] http://enterface.net/enterface05/main.php?frame=emotion
[†††] http://sourceforge.net/projects/opensmile/

Table 1 The recognition performance in *case1* (anger: A, disgust: D, fear: F, happiness: H, sadness: S, average: Avg).

Methods	Recognition rates (%)					
	A	D	F	H	S	Avg
Traditional	37.24	19.22	17.97	27.17	28.43	28.90
DR	47.01	25.12	29.09	44.01	41.12	37.13
TCA	50.16	28.91	34.55	45.34	44.05	40.93
TSC	52.04	29.31	38.01	47.52	46.05	44.87
NMF	39.13	21.26	20.08	26.85	30.16	28.51
SNMF	41.23	22.47	20.68	29.01	32.15	30.25
Ours	**52.73**	**30.02**	**38.24**	**47.80**	**46.13**	**45.21**
Baseline	74.41	55.35	54.02	60.01	60.98	61.38

Table 2 The recognition performance in *case2* (anger: A, disgust: D, fear: F, happiness: H, sadness: S, average: Avg).

Methods	Recognition rates (%)					
	A	D	F	H	S	Avg
Traditional	31.52	53.05	16.44	20.02	47.21	34.64
DR	34.76	72.13	17.89	25.32	69.06	45.82
TCA	35.43	72.98	19.02	25.95	69.75	49.63
TSC	36.21	73.85	21.12	26.97	71.13	52.67
NMF	33.41	68.20	17.03	22.32	50.01	38.18
SNMF	34.52	69.08	19.05	24.02	51.68	39.81
Ours	**36.57**	**74.69**	**21.35**	**27.13**	**71.69**	**53.02**
Baseline	72.98	81.09	68.54	53.02	79.35	70.98

and the unlabeled eNTERFACE corpus is chosen for testing. Meanwhile, in *case2*, the labeled eNTERFACE databse is chosen for training, and the unlabeled Berlin database is used for testing. The two databases are partitioned into five parts with equal size, in each test, four of five subsets of each corpus are utilized for training, while the others are used for testing. The five common emotion categories of the two databases including happiness, anger, disgust, fear and sadness are employed for evaluation.

The experimental results are depicted in Table 1 and Table 2. In each table, the recognition rates of all emotion categories are shown, and the overall average recognition results are also shown. First, from the two tables, it can be easily seen that the DR, TCA, TSC and our proposed method achieve much better results than the other methods. It can be attributed to that the power of transfer learning, in which the distribution discrepancies between two data sets can be efficiently reduced. Second, it can be also found that compared to the traditional automatic recognition methods, the NMF, SNMF and our proposed method achieve better recognition. The reason is that the non-negative matrix factorization can obtain better robust feature representations. Finally, it can be observed that, either in Table 1 or Table 2, our proposed TSNMF approach obtains the best results among all the cross-corpus speech emotion recognition methods, and is competitive to the baseline method conducted on single corpus.

5. Conclusions

In this letter, a novel method called transfer semi-supervised non-negative factorization (TSNMF), which makes use of both semi-supervised NMF and transfer learning algorithms, has been presented for speech emotion recognition. The semi-supervised NMF can give the new representations of the features more discriminating power. Meanwhile, the transfer learning technique can efficiently reduce the differences between feature distributions of two data sets. The experimental results on cross-corpus speech emotion recognition evaluations show that our proposed TSNMF method can achieve better or competitive performance compared to other methods.

Acknowledgements

This work was supported by the Natural Science Foundation of Shandong Province under Grant ZR2014FQ016, the National Natural Science Foundation of China under Grants 61231002, 61403328 and 61572419, and the Fundamental Research Funds for the Southeast University under Grant CDLS-2015-04.

References

[1] E.A. Moataz, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," Pattern Recognition, vol.44, no.3, pp.572–587, 2011.
[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," Proc. Interspeech, pp.223–227, Singapore, 2014.
[3] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," Proc. Interspeech, pp.1553–1556, Florence, Italy, 2011.
[4] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," IEEE Signal Process. Lett., vol.21, no.9, pp.1068–1072, 2014.
[5] M. Abdelwahab and C. Busso, " Supervised domain adaptation for emotion recognition from speech," Proc. ICASSP, Brisbane, Australia, pp.5058–5062, 2015.
[6] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, no.6755, pp.788–791, 1999.
[7] S.J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol.22, no.10, pp.1345–1359, 2010.
[8] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," IEICE Trans. Inf. & Syst., 2014, vol.97, no.9, pp.2530–2532, 2014.
[9] P. Song, W. Zheng, and R. Liang, "Speech emotion recognition based on sparse transfer learning method," IEICE Trans. Inf. & Syst., 2015, vol.98, no.7, pp.1409–1412, 2015.
[10] H. Liu, Z. Wu, X. Li, D. Cai, and T.S. Huang, "Constrained nonnegative matrix factorization for image representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.7, pp.1299–1311, 2012.
[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C.A. Muller, and S.S. Narayanan, "The interspeech 2010 paralinguistic challenge," Proc. Interspeech, pp.2794–2797, Makuhari, Japan, 2010.