ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 111－120
DOI:10. 16337/j. 1004-9037. 2023. 01. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于图卷积深浅特征融合的跨语料库情感识别

杨子秀[1]，金　赟[1,2]，马　勇[2,3]，戴妍妍[1]，俞佳佳[1]，顾　煜[1]

(1. 江苏师范大学物理与电子工程学院，徐州 221116；2. 江苏师范大学科文学院，徐州 221116；3. 江苏师范大学语言科学与艺术学院，徐州 221116)

摘　要：语音情感识别任务的训练数据和测试数据往往来源于不同的数据库，二者特征空间存在明显差异，导致识别率很低。针对该问题，本文提出新的构图方法表示源和目标数据库之间的拓扑结构，利用图卷积神经网络进行跨语料库的情感识别。针对单一情感特征识别率不高的问题，提出一种新的特征融合方法。首先利用 OpenSMILE提取浅层声学特征，然后利用图卷积神经网络提取深层特征。随着卷积层的不断深入，节点的特征信息被传递给其他节点，使得深层特征包含更明确的节点特征信息和更详细的语义信息，然后将浅层特征和深层特征进行特征融合。采用两组实验进行验证，第1组用 eNTERFACE 库训练测试 Berlin 库，识别率为 59.4%；第2组用 Berlin 库训练测试 eNTERFACE 库，识别率为 36.1%。实验结果高于基线系统和文献中最优的研究成果，证明本文提出方法的有效性。

关键词：图卷积神经网络；跨语料库；语音情感识别；构图；深层和浅层特征融合

中图分类号：TN912.34　　**文献标志码**：A

Deep and Shallow Feature Fusion Based on Graph Convolution for Cross-Corpus Emotion Recognition

YANG Zixiu[1], JIN Yun[1,2], MA Yong[2,3], DAI Yanyan[1], YU Jiajia[1], GU Yu[1]

(1. School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China;2. Kewen College, Jiangsu Normal University, Xuzhou 221116, China;3. School of Linguistics Sciences and Arts, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: The traning and testing data for speech emotion recognition often come from different corpora. In this case, the mode recognition performance decreases greatly due to the domain mismatch. To address this problem, we present a new composition method using graph convolutional network to represent the topological structure between the source and target databases for cross corpus speech emotion recognition. Besides, aiming at the problem of low accuracy of single feature in emotion recognition, a novel feature fusion method is proposed. Firstly, we extract the acoustic features by OpenSMILE, then extract deep features by graph convolutional neural network. With the proceeding of convolutional layers, nodes transmit the feature information to another nodes, making the deep features contain clearer feature information and more detailed semantic information. Finally, we fusion the shallow and deep features. Two classification experiments are carried out. eNTERFACE corpus is for training and Berlin corpus is for testing, and the recognition rate is 59.375%. Berlin corpus is for training and eNTERFACE corpus is for testing, and the

基金项目：江苏省高校自然科学基金(18KJB510013)。
收稿日期：2022-01-07；**修订日期**：2022-04-19

recognition rate is 36.111%. The experimental results are higher than the best research results in the baseline system and references, which proves the effectiveness of the method proposed in this paper.

Key words: graph convolutional network; cross-corpus; speech emotion recognition; composition; deep and shallow feature fusion

引　言

语音情感识别是指计算机从人类的语音信号中识别出情感，如愤怒、高兴、悲伤和恐惧等，它已经成为人机和谐交互的一个指标。传统的语音情感识别（Speech emotion recognition，SER）任务训练样本和测试样本来自同一个语料库。但在实际应用中，训练样本和测试样本可能属于不同的语料库，这给语音情感识别增加了难度。许多研究者针对这一具有挑战性的问题，提出了许多有效的方法。例如，Jin等[1]提出了一种半监督判别分析的跨语料库情感识别的方法。Song等[2]提出了一种基于迁移非负矩阵分解的跨语料库情感识别方法。Zhang等[3]进行了多任务学习跨语料库情感识别。Zhang等[4]将稀疏判别子空间学习用于跨语料库情感识别，提出了一种迁移稀疏判别子空间学习方法（Transfer sparse discriminant subspace learning，TSDSL）。近年来，图卷积网络（Graph convolutional network，GCN）以图或网络的形式应用于许多实词数据集，例如对合成孔径雷达（Synthetic aperture radar，SAR）图像中的目标进行分类[5]以及用于阿尔茨海默病分类[6]。在这些文献中，GCN因能传递节点之间的信息，取得了出色的结果。

情感特征是语音情感识别中的重要参数，如何提取具有显著性差异的情感特征，减少语音信号和情感间的差距是语音情感识别中的一个研究重点。在对大量的训练语料进行学习的过程中，挖掘由各种声学特征通往对应情感状态的映射通路，从而实现测试语料库情感状态的正确判断已经取得一定的情感识别效果[7]，但是用单一特征表达情感具有一定的局限性，因此研究人员提出了各种特征融合的方法。Jin等[8]提出了一种基于多核学习的特征选择与特征融合方法，提高了Berlin数据库7类情感的识别率；Bandela等[9]提出Teager能量算子（Teager energy operator，TEO）和Mel倒谱系数（Mel frequency cepstral coefficient，MFCC）的特征融合技术Teager-MFCC（T-MFCC），对语音信号中的应激情绪进行识别，达到了更好的效果；Liu等[10]提出了一种基于深度学习的特征融合方法，将基于谱的特征和基于音高的超韵律特征相结合，提高了语音情感识别系统的性能。而多层神经网络是一个功能强大的特征提取器，更高层次的特征放大了输入中对区分类别很重要的方面[11]；Lee等[12]提出了一种递归神经网络模型可以提取深层特征，并且这些方法已经得到了证明[13]。迁移学习和特征融合的跨语料库情感识别方法已经取得了初步进展，但是如何将处于不同数据库的情感特征投影到共同的情感子空间，并在数据库间传递更多信息，提高语音情感识别率是当前亟待解决的问题。针对该问题，本文提出基于图卷积神经网络的深层与浅层特征融合的方法。

1　图卷积神经网络

图卷积神经网络主要包括两个步骤：图构造（将数据集转换为图）和图卷积（用所构造的图预测未标记数据的标签）。设$\left\{(v_i,y_i)_{i=1}^{l},(n_j)_{j=l+1}^{l+u}\right\}$表示一个数据集，其中$(v_i,y_i)_{i=1}^{l}$表示有标签样本以及它的标签，$(v_j)_{j=l+1}^{l=u}$表示无标签样本，$l$表示有标签样本的个数，$u$表示无标签样本的个数。图构造是将数据集转换为图，表示为$G(V,E)$，其中$V=\left\{v_{ij}\right\}$表示节点集，$E=\left\{E_{ij}\right\}$表示边集，$i,j\in\{1,2,\cdots,l+u\}$。图卷积是标签传播的过程。

1.1 图构造

基于图算法的准确性很大程度上依赖输入的图结构,图构造包括节点选择和边的连接。图上的节点是数据集中的样本点v_i,其中$i\in\{1,l+u\}$。根据节点的特征添加边连接节点,得到特征之间的邻接矩阵$\boldsymbol{A}$:$\boldsymbol{A}\in\{0,1\}^{n\times n}$,$\boldsymbol{A}$的元素组成为

$$A_{ij}=\begin{cases}1 & i\neq j\text{并且}v_i,v_j\text{有连接}\\ 0 & i=j\text{或者}v_i,v_j\text{没有连接}\end{cases} \tag{1}$$

边的连接方法通常使用K近邻(K-nearest neighbor,KNN),每个节点与它K个最近邻点连接,这种方法具有较强的鲁棒性和简单性,在选择合适的K值时可以取得很好的效果。

1.2 图卷积

GCN是将传统的卷积神经网络推广到图域,是一种有效的图表示模型,它可以在学习过程中自然地将结构信息和节点特征结合起来,通过聚集来自其他邻居(包括自身)的特征向量来表示一个节点[14],本实验采用频谱卷积方法。GCN的传播规律可以概括为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{2}$$

式中:$\boldsymbol{H}^{(l)}$为第l层卷积运算后的矩阵;$\boldsymbol{H}^{(0)}=\boldsymbol{X}$;$\boldsymbol{D}$为对角矩阵,$\tilde{\boldsymbol{D}}$为$\boldsymbol{D}$的度矩阵;$\tilde{D}_{ii}=\sum_j\tilde{A}_{ij}$;$\boldsymbol{W}^{(l)}$为权矩阵;$\sigma$为激活函数。

图1是两层GCN框架,ReLU为整流线性单元,无标签节点更新它们的特征,将最终获得的无标签节点的特征输入Softmax分类器。本文Softmax分类器的输出是无标签节点的标签(情感)。因此基于该方法可以获得所有无标签数据的标签。两层GCN模型的整个过程可以表示为

$$\boldsymbol{Z}=f(\boldsymbol{X},\boldsymbol{A})=\text{softmax}(\hat{\boldsymbol{A}}\text{ReLU}\left(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}^{(0)}\right)\boldsymbol{W}^{(1)}) \tag{3}$$

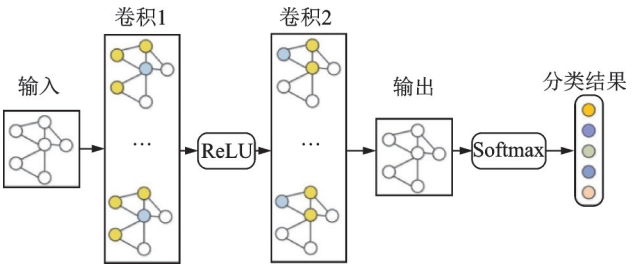


图1 图卷积神经网络

Fig.1 Graph convolutional neural network framework

式中:$\boldsymbol{W}^{(0)}$为输入层到隐藏层的权值;$\boldsymbol{W}^{(1)}$为隐藏层到输出层的权值;$\hat{\boldsymbol{A}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$,其中$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$,$\boldsymbol{A}$为邻接矩阵,$\boldsymbol{I}_N$为$N$维的单位矩阵。

损失函数定义为所有标签节点的交叉熵误差为

$$L=\sum_{i\in Y_D}\sum_{f=1}^{F}Y_{if}\ln Z_{if} \tag{4}$$

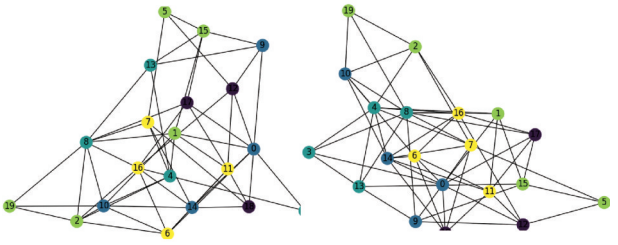
式中:Y_D为标签节点的索引集;F等于类的数量;Z_{if}为输出的标签集。

1.3 跨语料库的图构造

本文研究的跨语料库语音情感识别所使用的语料库都已被标记标签,但在实验中不使用目标语料库标签,即源语料库数据是有标签的,目标语料库数据是无标签的。该类问题一直是语音情感识别中的难题。原因在于源语料库和目标语料库的特征空间完全不一致,二者的分类面也不同,直接用源语料库数据训练得到的分类面去识别目标语料库,识别率很低,通常在30%~40%左右。因此如何利用源语料库特征向量的拓扑结构并传递给目标语料库,是本文的一个关键点。根据前述的构图方

法，对于源数据库和目标数据库分别建立各自的图。目标数据库中，每个类别随机选取一个点（有类别信息）和源数据库进行连接。这样就可以把原本两个独立的图（源数据库图和目标数据库图）建立关联。

具体步骤如下所述。先对所有数据进行归一化，统一样本的统计分布。对源语料库数据进行线性判别分析（Linear discriminant analysis，LDA）降维，降至4维，对目标语料库数据进行主成分分析（Principal component analysis，PCA）降维，降至4维，和源语料库维度保持一致。然后分别计算欧式距离，用KNN寻找最近邻点，从而得到相应的邻接矩阵 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$，空间中距离越近、邻接矩阵值越大的点表示情感越相似，最后构成图 G_1 和 G_2。本文 $\boldsymbol{X}_1$、$\boldsymbol{A}_1$ 分别表示Berlin库的特征向量和邻接矩阵，$\boldsymbol{X}_2$、$\boldsymbol{A}_2$ 分别表示eNTERFACE库的特征向量和邻接矩阵，邻接矩阵是由每两条语音特征的欧式距离构成。如图2所示，5种不同的颜色分别代表5种不同的情感，图2（a）是源语料库构成的图，图2（b）是目标语料库构成的图。为了解决两张独立的图无法相互间传递信息的问题，从目标语料库中选出5个不同类别情感（有监督）的点，并与源语料库中选出的5个不同类别情感的点一一对应连接，构成一张跨语料库图 G。通过这5个点将两个不同的情感语料库联系起来，从而源语料库有标签节点的信息可以传递到目标语料库，达到获得较好情感识别结果的目的。如图3所示，使用此构图方法，一方面考虑了不同数据库样本空间分布不同，缩小源域和目标域特征空间差异；另一方面引入图模型，构造跨库图建立不同数据库间的联系。



(a) Source corpus graph (b) Target corpus graph

图2 源语料库和目标语料库图

Fig.2 Source corpus graph and target corpus graph

图3 两个独立语料库图之间的连接

Fig.3 Connection between two independent-corpus graphs

遵循Kipf等[15]的方法，采用多层GCN识别语音情感。在构建GCN的图 G 之后，根据邻接矩阵 $\boldsymbol{A}$ 对图节点特征进行加权并融合到相邻节点，使得相邻节点获得相似的特征。

2 深、浅特征融合

2.1 浅层声学特征

本文采用Interspeech2010语音情感识别竞赛中使用的特征集[16]，共1 582维特征，包含34个低级描述符（Low level descriptor，LLD）和34个相应的delta作为68个LLD轮廓值，在此基础上应用21个函数得到1 482个特征，另外，对4个基于音高的LLD及其4个delta系数应用了19个函数得到152个特征，最后附加音高（伪音节）的数量和总数输入的持续时间（2个特征）。本文所用的38个LLD如表1所示。

表1 实验采用的低级描述符

Table 1 LLDs used in the experiment

LLD	数量
响度	1
梅尔频率倒谱系数[0～14]	15
Log梅尔带宽[0～7]	8
线谱对[0～7]	8
基音频率 F_0	1
F_0 包络	1
浊音频率	1
局部抖动	1
连续抖动帧对	1
局部微扰	1

2.2 图卷积的深层特征

图卷积则是在图网络中,把节点和边输进一个函数σ,每个节点就会得到相连点的信息,如图4所示。点1在输入网络后则会聚集点2、点3和点4的特征,这是1个前向传播的过程,在此过程中点1可以包含更多点的语义信息,且经过多层图卷积之后有标签的点将标签信息传递到未标记的点上,更有利于之后的情感信息传递,提高情感识别效果。浅层的特征可以提供全局信息,但是语义特征不明显[17]。单用浅层特征表达情感是不够的,1个好的网络结构要同时考虑到浅层和深层特征。

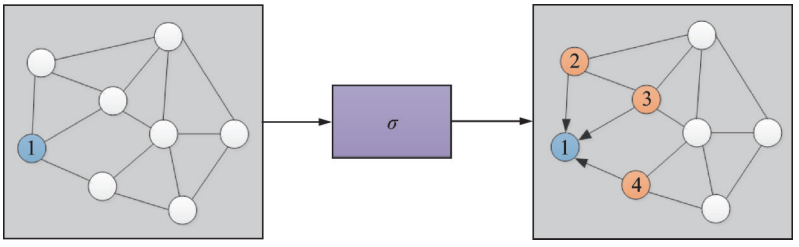


图4 图卷积

Fig.4 Graph convolution

2.3 特征融合

基于以上思想,本文将提取出的图卷积神经网络隐藏层的输出特征(深层特征)与声学特征(浅层特征)进行特征级融合。融合后的特征作为最终输入特征,可以很好地考虑到语义信息并传递更多标签信息。特征融合框架如图5所示。将隐藏层特征输出与1 582维声学特征融合,添加语音的深层信息,改变了输入$\boldsymbol{X}$,邻接矩阵$\boldsymbol{A}$不变。考虑到相邻层之间的相关性比较高,融合特征有局限性,且每层提取出来的特征对最后的识别结果贡献不一样。因此在实验中通过比较不同层特征对跨语料库情感识别率的贡献来提取最佳的特征向量。

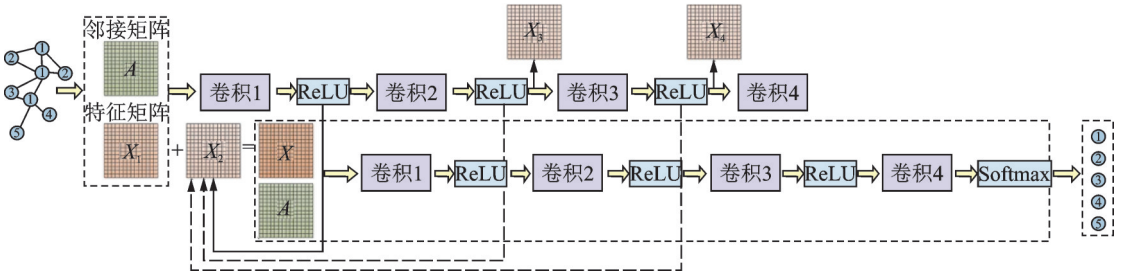


图5 特征融合框架

Fig.5 Feature fusion frame

本实验建立了2~4层的GCN,称之为2GCN、3GCN、4GCN,分别输出nGCN的前$n-1$层隐藏层特征,每层输出特征与1 582维声学特征融合,再输入相对应的GCN网络,观察实验结果。图5为4GCN框架示例图,每输出一个隐藏层特征$X_\alpha(\alpha\in\{1,\cdots,n-1\})$就与特征$\boldsymbol{X}_1$融合得融合特征$\boldsymbol{X}$,与邻接矩阵$\boldsymbol{A}$再一同输入4GCN网络,得到1个实验结果;2GCN和3GCN重复相同实验步骤,分别得到不同的结果,观察不同层数GCN和每层GCN输出特征对实验结果的影响。

3 实验验证

为了验证本文提出方法的有效性,选择Berlin数据库和eNTERFACE数据库,二者分别作为源语料库和目的语料库进行实验。首先设计3种不同层数的图卷积神经网络,确定最优的网络层数,再提取

不同深度的特征向量和浅层特征向量进行融合，确定最优特征的深度。最后利用上述构图方法对源数据库和目的数据库进行构图并连接，用GCN进行识别。

3.1 数据准备

Berlin数据库是1个德国情感语音数据库，有7种情绪：中立、恐惧、愤怒、高兴、悲伤、厌恶和无聊，录音由10名演员(5男5女)在专业录音室完成，经过20名参与者的录音测试，保留了535句日常生活语句。eNTERFACE数据库是1个英语情感语音数据库，有6种情绪：恐惧、愤怒、悲伤、高兴、厌恶和惊讶，录音工作由来自14个国家的42名受试者完成，其中81%为男性，19%为女性，最后通过2名专业人士判断该反应清楚的表达了情绪。如果情绪表达清楚，就将样本添加到数据库中。最终保留了1 170条语句。最后在Berlin数据库和eNTERFACE数据库中选取5种相同的情绪：恐惧、悲伤、高兴、厌恶和愤怒进行训练和测试，共1 395句，每种情绪的数量如表2所示。

表2 Berlin数据库和eNTERFACE数据库中相同情绪的语句数量

Table 2 The number of sentences for same emotions in Berlin and eNTERFACE databases

Database	Angry	Sad	Disgust	Happy	Fear
Berlin	70	62	46	71	69
eNTERFACE	216	216	216	213	216

3.2 实验方案

为了验证本文提出方法的普遍性，采用两种方案。方案1(e-Be)：eNTERFACE库是源语料库，Berlin库是目标语料库；方案2(Be-e)：Berlin库为源语料库，eNTERFACE库为目标语料库。每个数据集随机平均分成10份，每次测试对所有源语料库的数据(全部有标签)和8份目标语料库(无标签)的数据进行训练，然后识别剩余2份目标语料库的样本。

3.3 基于不同层数GCN的性能对比实验

跨语料库语音情感识别率与GCN网络层数的关系如表3所示。本实验采用的基线系统，用源语料库训练，直接测试目标语料库，之间没有任何的连接。e-Be是eNTERFACE库训练，Berlin库直接测试，基线的识别率为37.690%；Be-e是Berlin库训练，eNTERFACE库直接测试，基线的识别率为22.950%。两种方案的识别率都很低，说明二者的特征空间分布差别很大。从表3可以看出，2GCN的两个方案的识别率分别为50%和30.374%，3GCN的两个方案的识别率分别为45.833%和26.636%，4GCN的两个方案的识别率分别为48.611%和29.439%。2GCN的2个方案的识别率都要高于3GCN和4GCN，说明给出的方法较为有效。随着层数的增多，反复使用拉普拉斯平滑，使得两个距离很远的节点相似，很难区分类别信息，从而影响分类精度。同时，2GCN的两个方案比基线分别提高了12.31%和7.424%，也说明了本文提出的构图方法和图连接方法的有效性。

表3 不同模型的实验结果

Table 3 Experimental results of different models %

方案	基线	2GCN	3GCN	4GCN
e-Be	37.690	50.000	45.833	48.611
Be-e	22.950	30.374	26.636	29.439

3.4 基于不同深度的特征融合性能对比实验

本实验除了基线系统外，把文献[18]提出的联合迁移子空间学习方法(Joint transfer subspace learning and regression，JTSLR)的实验结果作为比较。实验框架如图6所示。由2个卷积层和1个Softmax层组成。该网络的输入是由1 582维特征通过归一化、降维、计算欧式距离、选择最近邻点和连接5个相同情感的点构成的图。第1个隐藏层输出特征的大小为16×1 395。ReLU作为激活函数[19]，提高了训练过程的效率。不同深度的特征融合的实验结果如表4所示。表4中的方案nGCN-m，n表示GCN的层数，m表示提取第m层的特征作为深度特征，然后与原始声学特征融合。从表中可以看出，2GCN提

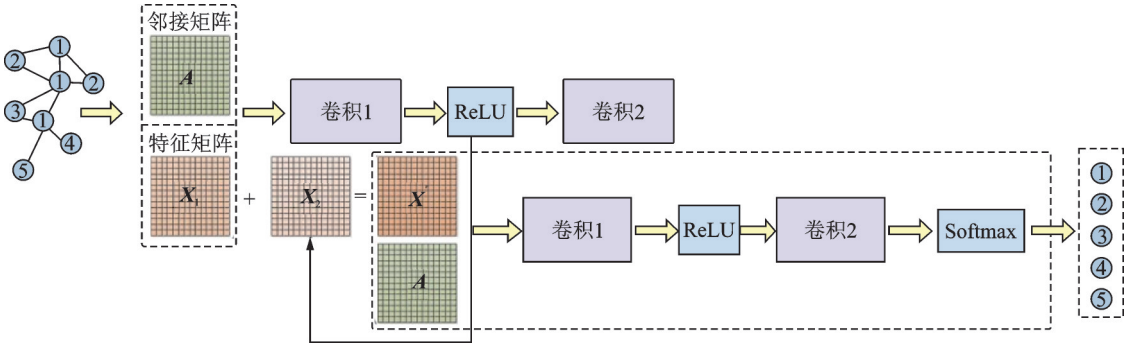


图6 实验框架

Fig.6 Experimental frame

表4 不同方案特征融合的实验结果

Table 4 Feature fusion experiment results of different methods %

方案	e-Be	Be-e	平均值
基线	37.690	22.950	30.32
JTSLR	44.230	38.140	41.19
2GCN-1	59.375	36.111	47.74
3GCN-1	51.389	27.103	39.25
3GCN-2	47.222	33.178	40.20
3GCN-1+2	48.611	30.374	39.49
4GCN-1	44.444	26.168	35.31
4GCN-2	48.611	28.037	38.32
4GCN-3	45.833	30.374	38.10
4GCN-1+2	55.556	24.299	39.93
4GCN-1+2+3	41.667	31.308	36.49

取的特征并进行融合后,两种方案得到的识别率分别为59.375%和36.111%,高于3GCN和4GCN的方案,说明2GCN提取的深度特征比3GCN和4GCN包含更多的情感信息。此外,2GCN的识别结果比基线分别提高了21.685%和13.161%。和JTSLR方法相比,e-Be方案的识别率提高了15.145%,Be-e方案比其低2.029%,总的平均值为47.74%,比JTSLR方法41.19%的平均值提高了6.55%。Be-e的识别率较低,是由于eNTERFACE库的语音质量相比Berlin库要差,其本身的识别率也要比Berlin库低很多,所以eNTERFACE库分类边界的不清晰导致不同类别节点之间有很多连接,类别标签进行传递时发生了错误,识别率较低。Be-e的识别率始终低于e-Be,因为Berlin数据集的数据量远远少于eNTERFACE数据集,导致使用Berlin数据集训练的模型不能很好地匹配eNTERFACE数据集的情感特征的分布,这些结果与相关文献一致[20-21]。本文所使用的图卷积神经网络减少了卷积核参数个数,参数复杂度降低;矩阵变换后无需做特征分解,直接使用拉普拉斯矩阵进行变换,这两步都可以有效降低识别速度。

本实验的混淆矩阵如图7所示,在e-Be中对悲伤、厌恶和愤怒情绪的识别率都在60%以上。而对恐惧情绪的预测能力低于30%。恐惧和愤怒、厌恶混淆,高兴和恐惧混淆。在Be-e中对高兴、恐惧和愤怒情绪的识别率都在40%以上,而对厌恶和悲伤情绪的预测能力低于20%。厌恶和高兴混淆,生气、悲伤和恐惧混淆,说明本文提出的模型虽然可以改善两个数据库在空间中的特征向量分布,但只能改善

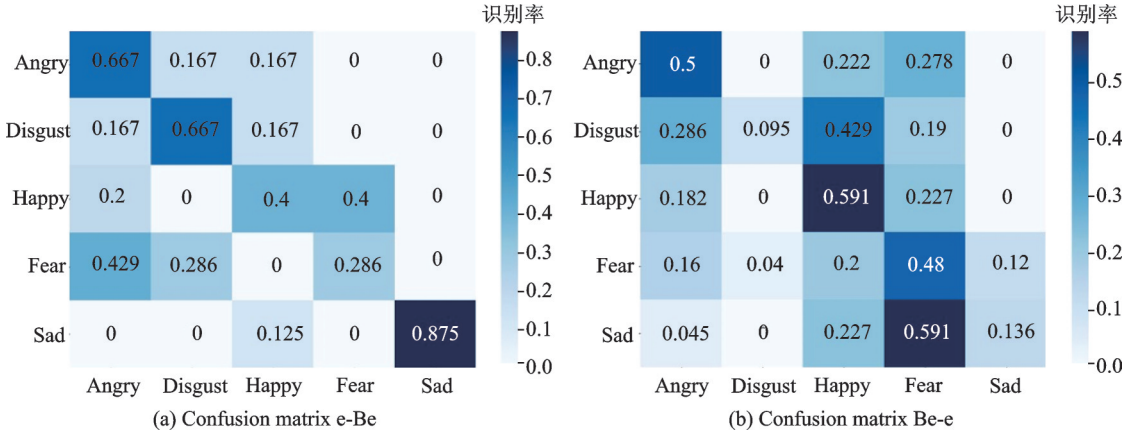


图7 混淆矩阵

Fig.7 Confusion matrix

部分情感的特征空间分布,还不能做到对5种情感的特征空间改进。这是由于在训练过程中,只是利用了目标语料库的数据样本和其中5个点的标签信息,在之后的实验中继续提高每种情感的识别率。

对特征融合前后的数据进行了可视化分析,用以直观地比较特征向量在空间分布的变化。图8与图9分别为特征融合前后数据空间分布情况,其中每个点代表网络中的1个节点,每个节点的颜色代表该节点的类标签,不同的颜色代表不同的类标签。为了更清楚地显示情感可视化,任意选择其中2种情感(害怕和悲伤)生成可视图,发现特征融合后情绪分类比特征融合前更明显,说明本文提出的方法是有效的。

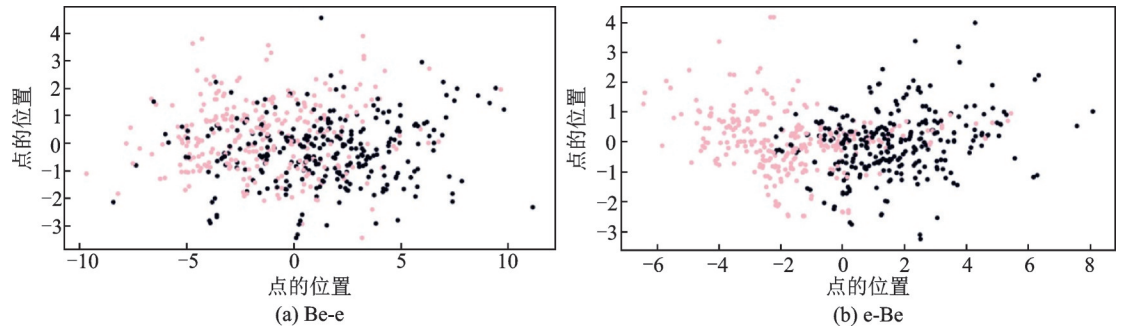


图8 特征融合前数据空间分布

Fig.8 Spatial distribution of data before feature fusion

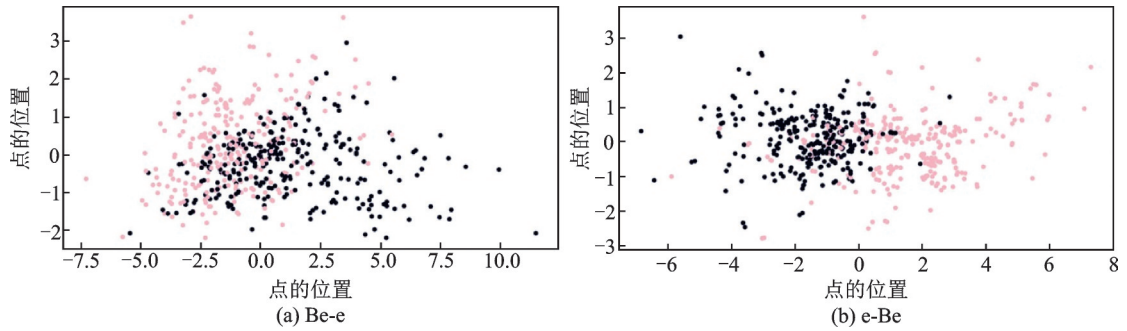


图9 特征融合后数据空间分布

Fig.9 Spatial distribution of data after feature fusion

4 结束语

本文把GCN引入跨库的语音情感识别,提出新的构图方法以增加两个独立情感语音数据库之间的关联。设计了一种利用GCN提取深度特征的方法,并与浅层声学特征进行融合,提高跨库语音情感识别率。通过比较2GCN、3GCN和4GCN,确定GCN在跨库语音情感识别中的最优深度是2层。通过比较提取不同深度的情感特征的识别率,获得了最优的深度特征。实验结果表明,本文提出的基于GCN的跨语料库语音情感识别方法优于基线系统,与参考文献中提出的JTSLR方法相比,也有了较大的提升,说明该方法的有效性。文中源数据库与目标数据库分别构图后,利用了目标数据库5个有标签的样本作为两个库的连接,其实是降低了跨语料库情感识别的难度。完全不利用任何目标语料库的类别信息,从而真正实现跨语料库的语音情感识别,还需要作更深入的研究。

参考文献:

[1] 金赟,宋鹏,郑文明,等. 半监督判别分析的跨库语音情感识别[J]. 声学学报,2015,40(1): 20-27.
JIN Yun, SONG Peng, ZHENG Wenming, et al. Cross corpus speech emotion recognition using semi-supervised discriminant analysis[J]. Acta Acustica, 2015, 40(1): 20-27.

[2] SONG Peng, OU Shifeng , ZHENG Wenming, et al. Speech emotion recognition using transfer non-negative matrix factorization[C]// Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: [s.n.], 2016: 5180-5184.

[3] ZHANG Biqiao, PROVOST E M, ESSL G. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences[J]. IEEE Transactions on Affective Computing, 2019, 10(1): 85-99.

[4] ZHANG Weijian, SONG Peng. Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition[J]. IEEE-ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 307-318.

[5] MA Fei, GAO Fei, SUN Jinping, et al. Attention graph convolution network for image segmentation in big SAR imagery data [J]. Remote Sensing, 2019, 11(21): 2586.

[6] SONG T A, CHOWDHURY S R, YANG Fan, et al. Graph convolutional neural networks for alzheimer's disease classification[C]// Proceedings of the Institute of Electrical and Electronics Engineers 16th International Symposium on Biomedical Imaging. Venice, Italy: [s.n.], 2019: 414-417.

[7] 韩文静,李海峰,阮华斌,等. 语音情感识别研究进展综述[J]. 软件学报,2014,25(1): 37-50.
HAN Wenjing, LI Haifeng, RUAN Huabin, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, 25(1): 37-50.

[8] JIN Yun, SONG Peng, ZHENG Wenming, et al. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition[C]// Proceedings of the Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: [s.n.], 2014: 4808-4812.

[9] BANDELA S R, KUMAR T K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC [C]// Proceedings of the 8th International Conference on Computing, Communication and Networking Technologies. Delhi, India: [s.n.], 2017: 1-5.

[10] LIU Gang, HE Wei, JIN Bicheng. Feature fusion of speech emotion recognition based on deep learning[C]// Proceedings of the International Conference on Network Infrastructure and Digital Content. Guiyang, China: [s.n.], 2018: 193-197.

[11] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature. 2015, 521(7553): 436-444.

[12] LEE J, TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C]// Proceedings of the International Speech Communition Association. Dresden, Germany: [s.n.], 2015: 1537-1540.

[13] CAO Junhong, WEI Zhuobin, HUANG Tao, et al. Analysis of feature extraction criterion function maximum in nonlinear multi-layer feedforward neural networks for pattern recognition[C]// Proceedings of the International Conference on Intelligent Computation Technology and Automation. Changsha, China: [s.n.], 2010: 655-658.

[14] ZHU Ruifeng, DORNAIKA F, RUICHEK Y. Joint graph based embedding and feature weighting for image classification[J].

Pattern Recogn, 2019,93: 458-469.

[15] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the International Conference on Learning Representations. San Juan, Puerto Rico: [s.n.], 2017.

[16] SCHULLER B, STEIDL S, BATLINER A, et al. The interspeech 2010 paralinguistic challenge[C]//Proceedings of the International Speech Communition Association. Makuhari, Chiba, Japan: [s.n.], 2010: 2794-2797.

[17] SUN Linhui, CHEN Jia, XIE Keli, et al. Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition[J]. International Journal of Speech Technology, 2018,21(4): 931-940.

[18] ZHANG Weijian, SONG Peng, CHEN Dongliang, et al. Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021,99: 1-1.

[19] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: [s.n.], 2010: 807-814.

[20] BUSSO C, BULUT M, LEE C C, et al. Iemocap: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008,42(4): 335.

[21] ZHENG Wenming, XIN Minghai, WANG Xiaolan, et al. A novel speech emotion recognition method via incomplete sparse least square regression[J]. IEEE Signal Processing Letters, 2014, 21(5): 569-572.

作者简介:

杨子秀(1997-),女,硕士研究生,研究方向:语音情感识别,E-mail:1921140990@qq.com。

金赟(1979-),**通信作者**,男,副教授,硕士生导师,研究方向:语音信号处理、人工智能、机器学习等,E-mail:jiny@jsnu.edu.cn。

马勇(1977-),男,讲师,研究方向:语音与音频信号处理、模式识别等,E-mail:may@jsnu.edu.cn。

戴妍妍(1994-),女,硕士研究生,研究方向:语音情感识别,E-mail:1329309625@qq.com。

俞佳佳(1997-),女,硕士研究生,研究方向:多模态语音情感识别,E-mail:1137218386@qq.com。

顾煜(1997-),男,硕士研究生,研究方向:语音信号处理,E-mail:guyuluck666@163.com。

(编辑:刘彦东)