doi:10.12068/j.issn.1005－3026.2024.06.003

音乐多模态数据情感识别方法的研究

韩东红[1]，孔彦茹[2]，展艺萌[1]，刘 源[1]
（1.东北大学 计算机科学与工程学院，辽宁 沈阳 110169；2.国网电力科学研究院 南瑞集团有限公司，江苏 南京 211000）

摘 要：音乐情感识别研究在音乐智能推荐和音乐可视化等领域有着广阔的应用前景.针对该研究中存在的仅利用低层音频特征进行情感识别时效果有限且可解释性差的问题,首先,构建能够学习音符语义信息的基于乐器数字接口(MIDI)数据的情感识别模型ERMSLM(emotion recognition model based on skip-gram and LSTM using MIDI data),该模型的特征是由基于跳字模型(skip-gram)和长短期记忆(LSTM)网络提取的旋律特征,利用预训练的多层感知机(MLP)提取的调性特征以及手动构建的特征3部分连接而成;其次,构建融合歌词和社交标签的基于文本数据的情感识别模型ERMBT(emotion recognition model based on BERT using text data),其中歌词特征是由基于BERT(bidirectional encoder representations from trans formers)提取的情感特征、利用英文单词情感标准(ANEW)列表所构建的情感词典特征以及歌词的词频—逆文本频率(TF-IDF)特征所组成;最后,围绕MIDI和文本两种数据构建特征级融合和决策级融合两种多模态融合模型.实验结果表明,ERMSLM和ERMBT模型分别可达到56.93%,72.62%的准确率,决策级多模态融合模型效果更优.

关 键 词：音乐情感识别;深度学习;多模态;长短期记忆
中图分类号：TP 391.1 **文献标志码**：A **文章编号**：1005-3026(2024)06-0776-11

Research on Emotion Recognition Method of Music Multimodal Data

HAN Dong-hong[1], *KONG Yan-ru*[2], *ZHAN Yi-meng*[1], *LIU Yuan*[1]
(1. School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China; 2. NARI Group Corporation, State Grid Electric Power Research Institute, Nanjing 211000, China. Corresponding author: KONG Yan-ru, Email: kong19960103@163.com)

Abstract: The research of music emotion recognition has broad application prospects in the fields of music intelligent recommendation and music visualization. Aiming at the problem that only using low-level audio features for emotion recognition has limited effectiveness and poor interpretability. Firstly, an emotion recognition model ERMSLM based on MIDI (musical instrument digital interface) data is constructed, which can learn the semantic information of notes. The features of this model are composed of melodic features extracted with skip-gram and LSTM (long short-term memory), tonal features extracted by pre-trained MLP and manually constructed features. Secondly, an emotion recognition model ERMBT based on text data that integrates lyrics and social tags is constructed. The lyrics features are composed of emotional features extracted with BERT, emotional dictionary features constructed by using ANEW lists and TF-IDF features of lyrics. Finally, two multimodal fusion models of feature-level fusion and decision-level fusion are constructed based on MIDI and text data. The experimental results show that the ERMSLM and ERMBT models can achieve accuracies of 56.93% and 72.62% respectively. And the decision-level multimodal fusion model is more effective.

Key words: music emotion recognition; deep learning; multimodal; LSTM

音乐是人类情感的载体,也是情感表达的重要途径.音乐情感识别(music emotion recognition,

收稿日期：2023-02-09
基金项目：国家自然科学基金资助项目(61672144);国家重点研发计划项目(2019YFB1405302).
作者简介：韩东红(1968-),女,河北平山人,东北大学教授.

MER)是指利用计算机技术提取和分析音乐特征,形成音乐特征与情感空间之间的映射关系并识别音乐所表达情感的过程[1],属于音乐心理学、音频信号处理和自然语言处理等多学科的交叉研究领域,在智能搜索与推荐[2]、音乐可视化[3]、自动编曲[4]和音乐治疗[5]等领域均有广泛应用,该研究符合现今海量音乐作品发布和管理的实际需求.

随着人工智能技术的快速发展,基于深度学习的音乐情感识别研究已引发更多学者的关注.尽管现有研究已获得不错的性能,但仍然存在一些待研究的问题:

1)在音乐特征提取方面,大多已有研究直接从波形音频文件中提取低层音频特征.但有研究表明,低层音频特征是为了其他音频任务而设计的,与音乐的语义和情感之间没有直接联系[6-7],因此效果有限且可解释性差.音乐的旋律、调性等中高层概念更接近人的认知,利用音乐中高层概念进行特征抽取可提高情感识别的准确率[7-9].

2)受到版权影响,现有公开歌词数据集仅有The musiXmatch dataset,其仅有词干化的BOW(bags-of-words)特征,即一首歌的歌词文件只包含其中出现的高频词及词频计数.数据集匮乏不利于文本特征提取和模型设计,使得基于歌词的情感识别模型性能较差.

3)利用音频或歌词等单一模态数据进行音乐情感识别已经触及研究领域的"天花板"[10-11],融合多种模态数据使得不同模态信息互补会提升音乐情感识别任务的性能.

为解决上述问题,围绕MIDI数据、歌词和社交标签等文本数据展开多模态数据音乐情感识别研究.首先,利用MIDI数据提取与人的情感感知更接近的速度、旋律等中高层语义信息,构建ERMSLM模型;其次,从有限的歌词数据中提取情感信息,并加入社交标签文本进一步挖掘情感,构建ERMBT模型;最后采用特征级融合和决策级融合两种方式构建多模态音乐情感分类模型,并验证了所提方法的有效性.

1　相关工作

近年来深度学习技术在人工智能的诸多领域均取得了显著成效,越来越多的MER研究直接从原始的数据中自动学习最佳特征[12].现从基于深度学习的歌曲级音乐情感分类、歌曲级音乐情感回归和连续音乐情感变化检测(music emotion variation detection,MEVD)3个研究方向总结现有研究成果.

在歌曲级音乐情感分类方向,文献[13]将经过短时傅里叶变换(short time Fourier transformation,STFT)计算得到的语谱图作为输入,语谱图经过卷积层、池化层和隐藏层的计算,最终经过softmax模块得到情感标签.除感知特征外,文献[14]利用脑电图(electroencephalogram,EEG)和其他生理特征等唤起特征来预测情感标签.文献[15]尝试了基于VGGNet的模型,并对中级特征进行了探索,该文献独特之处是构建了人类可理解的模型.

在歌曲级音乐情感回归方向,Ma等[16]提出了一种基于多尺度上下文的注意力模型(multi-scale context based attention model,MCA),其创新之处在于利用注意力机制集成不同的时间尺度以动态学习音乐的时序和层次信息.Liu等[17]使用BiLSTM(bidirectional LSTM)模型提取音频特征,同时提取速度、能量等特征并与LSTM提取的音频特征集成后进行情感识别.文献[18]提出多任务学习模型,该模型的主任务是情感回归,辅任务是流派分类.该文献指出唤起情感跟音乐流派、年龄以及性别等因素相关,添加流派分类这一辅任务来学习音乐情感和流派的关系有益于情感识别过程.

在连续MEVD方向,文献[19]分别利用机器学习和深度学习模型完成静态和动态的音乐情感识别任务,实验结果表明,在静态情感识别任务中,高斯过程回归模型和支持向量回归模型的性能胜过了多元线性回归模型,而使用大量特征和循环神经网络的组合对于动态任务能达到最佳性能.Li等[20]认为音乐中某一点的情感不仅与该点之前的内容有关,也与该点之后的内容有关,提出了用DBLSTM(deep bidirectional LSTM)提取两个方向的信息.Chaki等[21]提出Attentive LSTM模型,此模型是一种结合了改进注意力机制的LSTM,其认为音乐在某一时刻情感只取决于该点之前的音乐内容,当计算特定时刻上下文向量时,仅考虑之前的隐藏状态.

2　模　　型

2.1　基于MIDI数据的情感识别模型ERMSLM

2.1.1　问题定义

利用音乐的MIDI数据识别音乐情感,情感识别的过程如下:从第i首音乐的MIDI数据中提取音符音高集$P_i=\{p_1,p_2,\cdots,p_n\}$,音符音强集$I_i=$

$\{i_1, i_2, \cdots, i_n\}$，音符音长集 $D_i=\{d_1, d_2, \cdots, d_n\}$，音乐速度 BPM_i 和调性 $\mathbf{KEY}_i$，用训练数据集学习得到情感识别模型 ERMSLM，以预测未标注数据集中第 i 首音乐的情感标签 y_i'.

ERMSLM 框架图如图 1 所示. 该模型分为特征提取模块和情感识别模块两部分，特征提取模块从输入数据中提取带有情感信息的特征，即从音乐的 MIDI 数据提取音乐的旋律、调性和速度等特征，情感识别模块利用提取的特征进行情感分类.

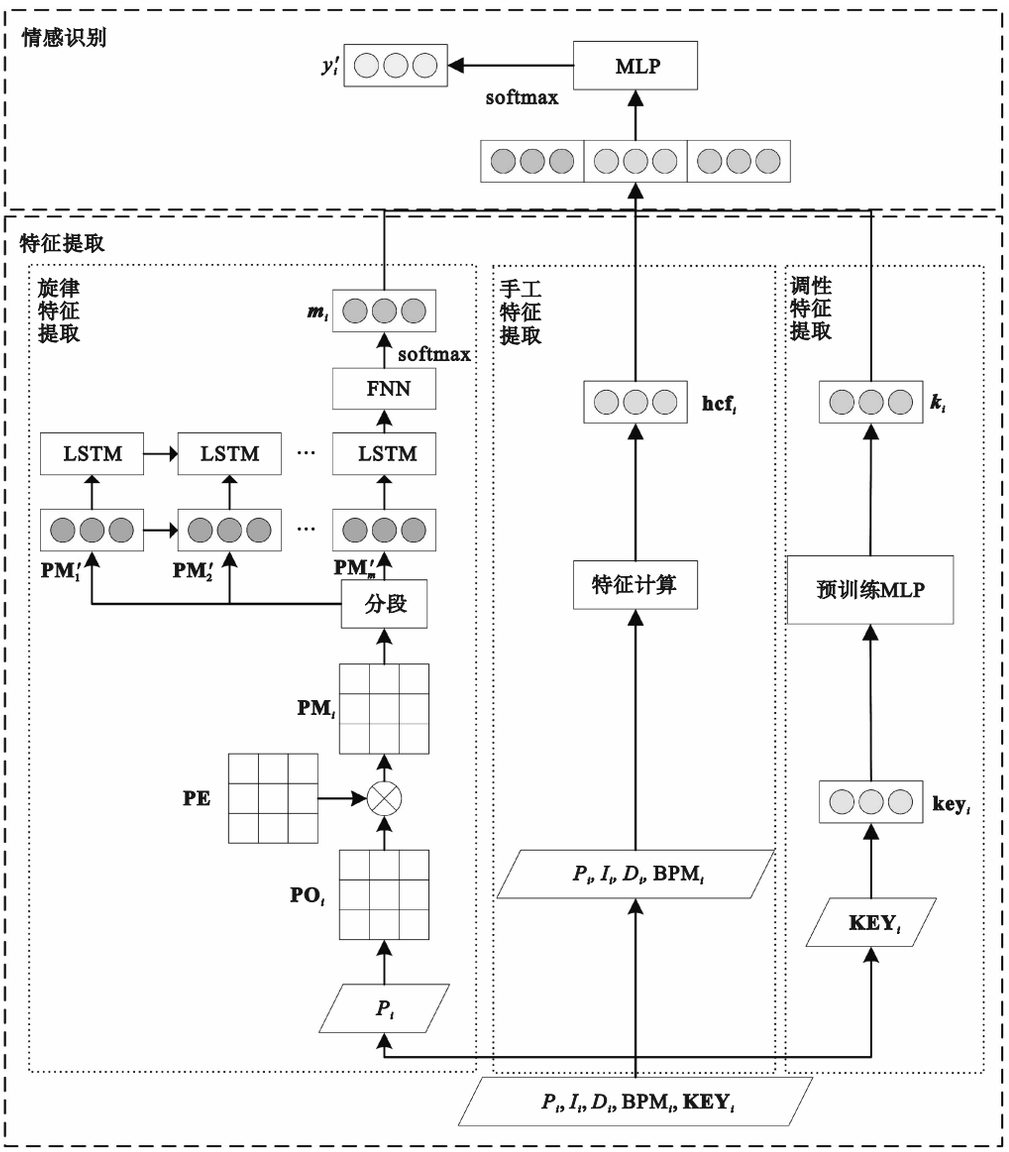

图 1 ERMSLM 框架图

Fig. 1 Diagram of ERMSLM frame

2.1.2 特征提取

1）旋律特征提取. 如图 1 所示，旋律特征提取是从第 i 首音乐的音符音高集 P_i 中提取旋律特征向量 $\boldsymbol{m}_i$. 首先通过 skip-gram 模型训练得到音符嵌入 **PE** 矩阵，再利用 **PE** 矩阵将音符音高集 P_i 对应的音符独热矩阵 $\mathbf{PO}_i$ 转化为包含音符语义信息的矩阵 $\mathbf{PM}_i$，之后对 $\mathbf{PM}_i$ 分段成 $\mathbf{PM}'_i$ 并利用 LSTM 模型进行上下文信息提取，最终获得旋律特征 $\boldsymbol{m}_i$.

2）调性特征提取. 调性特征提取需要从音乐的调性数据中提取调性特征 $\boldsymbol{k}_i$. 即从音乐中获取调性信息 $\mathbf{KEY}_i$，并将 $\mathbf{KEY}_i$ 转换为独热编码 $\mathbf{key}_i \in \mathbf{R}^{1\times 24}$，之后将调性独热向量 $\mathbf{key}_i$ 输入一个预训练过的包含 3 层全连接层的 MLP，获得的输出结果为调性特征 $\boldsymbol{k}_i$. 调性是调的主音和调式类别的总称，共有 24 种调性，故其独热编码的维度为 24. 预训练 MLP 的目的是让 MLP 的参数学习到调性与情感的联系，MLP 的输入为调性独热向量 $\mathbf{key}_i$，输出为预测的情感标签 y_i'，然后用真正的标签 y_i 来计算损失并反向调参. 预训练的过程如图 2 所示.

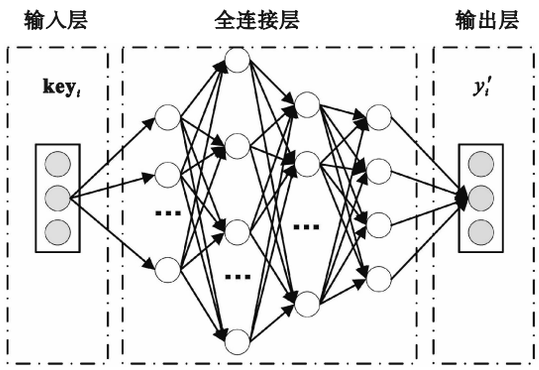

图2　预训练MLP

Fig. 2　Pretrained MLP

图2中,输出层y_i'用式(1)表示.式(2)与式(3)中的h^{k}为各层隐藏层的结果,W^{k}和b^{k}为各层全连接网络的参数,relu为激活函数.

$$y_i'=\mathrm{softmax}(h_2^{\mathrm{k}}\times W_3^{\mathrm{k}}+b_3^{\mathrm{k}}),\tag{1}$$

$$h_2^{\mathrm{k}}=\mathrm{relu}(h_1^{\mathrm{k}}\times W_2^{\mathrm{k}}+b_2^{\mathrm{k}}),\tag{2}$$

$$h_1^{\mathrm{k}}=\mathrm{relu}(\mathbf{key}_i\times W_1^{\mathrm{k}}+b_1^{\mathrm{k}}).\tag{3}$$

在反向调参中用到的损失函数为交叉熵损失函数,使用式(4)计算,N为样本数量.

$$\mathrm{LOSS}=-\sum_{i=1}^{N}y_i\times\lg y_i'.\tag{4}$$

对MLP进行预训练后,若将其表示为Pre-trained MLP,则调性特征的计算如式(5)所示.

$$\boldsymbol{k}_i=\mathrm{Pre-trainedMLP}(\mathbf{key}_i)\tag{5}$$

3) 手工特征提取.手工特征提取需要从音乐的主音轨中提取出音符的音高、力度、时值和音乐的速度4种信息来构建手工特征$\mathbf{hcf}_i$,此处的特征提取方式参考了文献[9,22].提取特征时用到了音符音高集、音符音强集、音符音长集和音乐速度.手工特征$\mathbf{hcf}_i$的计算方式如下:

① 对于音符音高集P_i,计算均值、标准差以及旋律走势.均值和标准差的计算公式分别为式(6)和式(7).式(6)中的n为音符音高集P_i的长度,pitch_{ij}表示第j个音符的音高;式(7)中的$\overline{\mathrm{pitch}_i}$表示音高集$P_i$的平均值.旋律走势根据音高变化进行标记,若相邻两个音的音程差为负即为上行旋律,标记为0.音程差为正则为下行旋律,记为1.用a和b统计0和1的个数,a和b连接即为式(8)的旋律走势特征hc_3.

$$\mathrm{hc}_1=\frac{1}{n}\sum_{j=1}^{n}\mathrm{pitch}_{ij},\tag{6}$$

$$\mathrm{hc}_2=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\mathrm{pitch}_{ij}-\overline{\mathrm{pitch}_i})^2},\tag{7}$$

$$\mathrm{hc}_3=[a{:}b].\tag{8}$$

② 音符音强集I_i和音长集D_i,计算音强平均值、音强标准差、音长平均值和音强差的时长加权平均值,使用式(9)至式(12)计算.式(9)中的n为音强集的长度,$\mathrm{intensity}_{ij}$表示第j个音符的音强.式(10)中的$\overline{\mathrm{intensity}_i}$表示音强集$I_i$的平均值.式(11)中的$\mathrm{duration}_{ij}$表示是第$j$个音符的音长.式(12)计算了时间差的加权平均值.

$$\mathrm{hc}_4=\frac{1}{n}\sum_{j=1}^{n}\mathrm{intensity}_{ij},\tag{9}$$

$$\mathrm{hc}_5=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\mathrm{intensity}_{ij}-\overline{\mathrm{intensity}_i})^2},\tag{10}$$

$$\mathrm{hc}_6=\frac{1}{n}\sum_{j=1}^{n}\mathrm{duration}_{ij},\tag{11}$$

$$\mathrm{hc}_7=\frac{\sum_{j=1}^{n}[(\mathrm{intensity}_j-\mathrm{intensity}_{j-1})\times\mathrm{duration}_j]}{\sum_{j=1}^{n}\mathrm{duration}_j}.\tag{12}$$

③ 音乐的速度BPM_i为浮点数,无需再进行计算.将上述计算结果连接,构成手工特征$\mathbf{hcf}_i$,如式(13).

$$\mathbf{hcf}_i=[\mathrm{hc}_1{:}\mathrm{hc}_2{:}\mathrm{hc}_3{:}\mathrm{hc}_4{:}\mathrm{hc}_5{:}\mathrm{hc}_6{:}\mathrm{hc}_7{:}\mathrm{BPM}_i].\tag{13}$$

2.1.3　情感识别

如图1所示,情感识别阶段先将上述提取的3类特征进行连接,然后输入MLP和softmax得到预测的情感标签y_i'.特征连接如式(14)所示,情感标签y_i'利用式(15)计算得到.式(16)至式(17)中,h^{s}为各层的隐层输出,W^{s}和b^{s}为各层的参数.反向调参过程同样使用交叉熵损失函数,使用式(4)计算.

$$\boldsymbol{f}_i=[\boldsymbol{m}_i{:}\mathbf{hcf}_i{:}k_i],\tag{14}$$

$$y_i'=\mathrm{softmax}(h_2^{\mathrm{s}}\times W_3^{\mathrm{s}}+b_3^{\mathrm{s}}),\tag{15}$$

$$h_2^{\mathrm{s}}=\mathrm{relu}(h_1^{\mathrm{s}}\times W_2^{\mathrm{s}}+b_2^{\mathrm{s}}),\tag{16}$$

$$h_1^{\mathrm{s}}=\mathrm{relu}(f_i\times W_1^{\mathrm{s}}+b_1^{\mathrm{s}}).\tag{17}$$

2.2　基于文本的情感识别模型ERMBT

2.2.1　问题定义

基于音乐的文本数据识别音乐情感的任务定义如下:从第i首音乐的歌词数据中获取该歌词的单词集合$W_i=\{w_1,w_2,\cdots,w_n\}$以及对应的词频集合$\mathrm{WC}_i=\{\mathrm{wc}_1,\mathrm{wc}_2,\cdots,\mathrm{wc}_n\}$,从第$i$首音乐的社交标签数据中获得社交标签集$\mathrm{tag}_i=\{\mathrm{ta}_1,\mathrm{ta}_2,\cdots,\mathrm{ta}_n\}$以及标签对应的标签得分集$\mathrm{t_score}_i=\{\mathrm{ts}_1,\mathrm{ts}_2,\cdots,\mathrm{ts}_n\}$,用训练数据集学习得到情感识别模型ERMBT,以预测未标注数据集中第i首音乐的情感标签y_i'.基于文本数据的音乐情感识别模型ERMBT的框架图如图3所示.

2.2.2　歌词特征提取

1) 基于BERT模型的情感特征构建.利用预

训练好的BERT模型获得词向量,该词向量一定程度上反映词汇的情感极性,如表1所示.

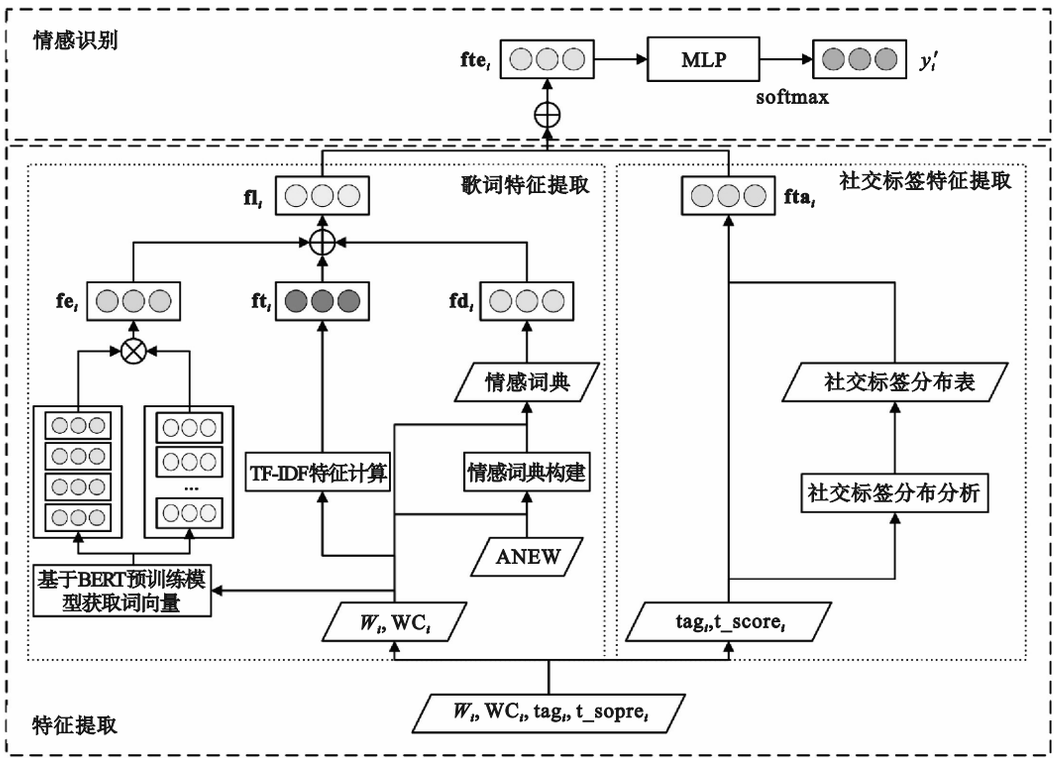

图3 ERMBT框架图

Fig. 3 Diagram of ERMBT frame

表1 词汇相似度

Table 1 Lexical similarity

词汇	Happy	Anxious	Sad	Relaxed
Happy	—	0. 368	0. 326	0. 329
Anxious	0. 368	—	0. 416	0. 276
Sad	0. 326	0. 416	—	0. 341
Relaxed	0. 329	0. 276	0. 341	—

表1中的数值是对应两个词汇的相似度(similarity),利用式(18)计算其归一化后的向量点积,式中的$\boldsymbol{A}$和$\boldsymbol{B}$表示两个向量.

$$\text{similarity}=\frac{\sum_{i=1}^{n}\boldsymbol{A}_i\times\boldsymbol{B}_i}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{A}_i)^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{B}_i)^2}}. \tag{18}$$

基于预训练的BERT模型来构建歌词的情感特征.提取出的情感特征对应图3中的$\mathbf{fe}_i=[\boldsymbol{h}_1,\cdots,\boldsymbol{h}_c]$,其表示第$i$首音乐的基于BERT的情感特征,$c$为情感类别数.式(19)表示$\mathbf{fe}_i$第$j$个维度的计算过程,$\mathbf{ew}_{ik}$为第$i$首音乐的第$k$个单词使用BERT预训练模型获得的词向量,$\mathbf{em}_j$为BERT预训练模型获得的VA情感空间中第$j$个象限情感单词的词向量,$n$为第$i$首音乐的单词数.

$$h_j=\sum_{k=1}^{n}\mathbf{ew}_{ik}\times\mathbf{em}_j. \tag{19}$$

2) 基于词典的情感特征构建. 词典的构建需要完成ANEW列表词干,The musiXmatch dataset提供了未词干化及词干化形式的对照表,依据该对照表对ANEW列表进行了词干化.之后对词干化后的ANEW列表与数据集中的高频词求取交集以找出重合部分,并保存重合词汇的效价度(Valence)和唤起度(Arousal)得分;找出重合部分后,则利用词汇的Valence和Arousal值(VA)进行词性标注.标注的方法如下:首先计算某个单词在VA情感空间中的对应点与4个情感标签单词在VA空间中对应点的欧氏距离,使用式(20)计算.

$$\rho=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}. \tag{20}$$

然后选取4个距离中最短的情感标签作为该单词的情感标注,标注后的词汇分布如表2所示.

表2 情感词典词汇分布

Table 2 Distribution of sentiment dictionary vocabulary

情感类别	V^+A^+	V^-A^+	V^-A^-	V^+A^-
单词数	199	133	48	119

注:V^+A^+(happy),V^-A^+(anxious),V^-A^-(sad),V^+A^-(relaxed).

构建好情感词典后,使用该情感词典构建词典情感特征,构建好的特征对应图3中的$\mathbf{fd}_i=[d_1,d_2,d_3,d_4]$,其中每一个维度对应第$i$首音乐拥有的对应情感类别的词汇数,例如$d_1$代表的是第$i$首音乐歌词拥有的第一象限的词汇数.

3) TF-IDF特征.特征构建目的是使用TF-IDF计算一个单词对于特定情感类别的重要性[23-25].数据集歌词中每个单词的TF-IDF值将作为权重以表示该单词与情感类别的相关性.先将整体数据集中属于同一情感类别e_j的歌词汇总在一个文档doc_j中,构成一个文档集合D.对于一个单词w_k,$\mathrm{tf}_j(w_k)$表示单词w_k对于情感类别e_j的重要性,$\mathrm{idf}(w_k)$表示单词w_k区别不同情感的能力,$\mathrm{tf}_j(w_k)$和$\mathrm{idf}(w_k)$使用式(21)和式(22)计算.式(21)中,$n_{k,j}$表示w_k在文档doc_j中的计数.式(22)中的分母表示包含w_k的文档数,分子表示文档的总数.

$$\mathrm{tf}_j(w_k)=\frac{n_{k,j}}{\sum_z n_{z,j}},\tag{21}$$

$$\mathrm{idf}(w_k)=\frac{|D|}{|\{\mathrm{doc}_j:w_k\in\mathrm{doc}_j\}|}.\tag{22}$$

第i首音乐的TF-IDF特征向量表示为$\mathbf{ft}_i=[\mathrm{ti}_1,\cdots,\mathrm{ti}_c]$,$c$为情感类别数,$\mathbf{lrc}_i$为第$i$首音乐的歌词数据.其中$\mathbf{ft}_i$的第$j$个维度的计算使用式(23).

$$\mathrm{ti}_j=\sum_{\{k|w_k\in\mathbf{lrc}_i\}}\mathrm{tf}_j(w_k)\times\mathrm{idf}(w_k).\tag{23}$$

BERT情感特征、词典情感特征和TF-IDF特征计算完成后,将3个特征归一化后相加即可得到歌词特征$\mathbf{fl}_i=[\mathbf{ly}_1,\cdots,\mathbf{ly}_c]$,$c$为情感类别数.使用的归一化方法是最小最大值标准化,使用式(24)计算.式(24)中,x指的是待处理的样本数据,$\min(x)$指的是样本数据中的最小值,$\max(x)$指的是样本数据中的最大值,x_i指的是当前样本取值,y_i指的是归一化后的结果.

$$y_i=\frac{x_i-\min(x)}{\max(x)-\min(x)}.\tag{24}$$

2.2.3　社交标签特征提取

社交标签特征提取需要进行数据预处理,提出一种社交标签分布分析算法来分析社交标签与情感类别的关联,得到社交标签分布表,最后根据分布表、社交标签及标签得分构建社交标签特征.

1) 数据预处理.社交标签集Tag由j个子集组成,每个子集Tag_j由数据集中属于同一情感类别e_j的音乐数据所包含的社交标签汇总组成,每一个子集按照社交标签出现在对应情感数据集中的频率由高到低排列.社交标签集Tag共有4个子集,对照$\mathrm{V^+A^+}$,$\mathrm{V^-A^+}$,$\mathrm{V^-A^-}$和$\mathrm{V^+A^-}$4种情感.标签汇总集T由数据集中出现的所有社交标签汇总而成,并按出现的频率由高到低进行排序.数据预处理后得到社交标签集Tag和标签汇总集T.

2) 社交标签分布分析算法.社交标签分布分析算法的伪代码描述见表3.表3中第2)~3)行为数据读取,第4)~9)行为当某一社交标签出现在Tag集的2个、3个或4个子集时的处理方式,第10)~13)行为判断社交标签的保留位置,第14)~15)行为获得社交标签分布表的过程.

表3　社交标签分布分析算法

Table 3　Social tag distribution analysis algorithm

输入:经过整理后的社交标签集Tag,标签汇总集T
输出:社交标签分布表
1) **Begin**
2)　由社交标签集Tag生成初始社交标签分布表
3)　读取标签汇总集T中的每个社交标签t_i
4)　**For each** t_i **in** T
5)　　若t_i在Tag的3个或4个子集中出现:
6)　　从初始社交标签分布表中删除所有t_i
7)　　若t_i在社交标签集Tag的两个子集中出现:
8)　　将t_i加入临时集合Temp
9)　**End for**
10)　读取临时集合Temp中的每个社交标签ta_i
11)　**For each** ta_i **in** Temp
12)　　判断ta_i出现的位置,将其从靠后位置对应的初始社交标签分布表中删除
13)　**End for**
14)　获得社交标签分布表
15) **End**

利用社交标签分布分析算法获得社交标签分布表的具体流程如下:首先将社交标签集Tag转换成列表形式,并将每列按照社交标签出现在该类情感数据集中的频率由高到低进行排列,完成上述处理后的列表作为社交标签分布表的初始列表;读取标签汇总集T,按照社交标签出现的频率由高到低依次判断该社交标签出现在Tag子集中的次数,若出现了3次及以上,则将其从初始社交标签分布表中全部删除,若出现2次,则暂时保留,并将该社交标签保存在1个临时列表Temp中,若该社交标签仅出现1次,则不进行操作;依次遍历临时列表Temp,判断每1个社交标签在2个子集中出现的位置,位置靠前表示该社交标签在该子集对应情感的音乐数据中出现频次高,位置靠后则表示在该子集对应情感的音乐数据中出现频率低,因此在初始社交标签分布表中,将该社交标签从靠后位置的对应列删除;完成上述操作后就得到了社交标签分布表.利用上述算法得到的社交标签分布表(Top 5)如表4所示.

3) 社交标签特征提取.第i首音乐社交标签特征的提取思路如下:对于第i首音乐的社交标

签集 $\mathrm{tag}_i=\{\mathrm{ta}_1,\cdots,\mathrm{ta}_n\}$，依此判断社交标签 ta_k 是否出现在社交标签分布表中，若不出现则跳过；若出现在社交分布表中，判断其出现在社交标签分布表中的哪 1 列，该列对应 1 个情感类别，将该社交标签的得分 ts_k 加到该情感类别在社交标签特征 $\mathbf{fta}_i$ 的对应维度上. 社交标签特征 $\mathbf{fta}_i$ 中维度 t_j 利用式(25)计算，式中的 n 表示第 i 首音乐所拥有的社交标签数，stat_j 代表与情感类别 e_j 相关的重要性排名靠前的社交标签集合. 由式(25)可知，维度 t_j 的数值为属于情感 e_j 的社交标签得分的累加和. 每个维度计算后归一化即可，得到社交标签特征 $\mathbf{fta}_i$.

$$t_j=\sum_{k=1}^{n}\mathrm{ts}_k,\ \mathrm{ta}_k\in\mathrm{stdt}_j. \tag{25}$$

表 4 社交标签分布

Table 4 Social tag distribution

V^+A^+	V^-A^+	V^-A^-	V^+A^-
happy	heartbreak	sad	chillout
upbeat	angry	soft	soul
fun	epic	acoustic	smooth
party	heartache	emotional	relax
catchy	aggressive	dark	relaxing

2.2.4 情感识别

情感识别阶段的任务是将歌词和社交标签特征进行相加，然后输入 MLP 和 softmax 得到情感标签 y_i'. 特征相加用式(26)计算，公式中的 β 为表示社交标签特征权重的常量. y_i' 使用式(27)计算，在式(28)至式(29)中，h^{t} 为各层隐层输出，W^{t} 和 b^{t} 为各层参数. 反向调参过程使用交叉熵损失函数公式(4)计算.

$$\mathbf{fte}_i=\mathbf{fl}_i+\beta\mathbf{ft}_i, \tag{26}$$

$$y_i'=\mathrm{softmax}(h_2^{\mathrm{t}}\times W_3^{\mathrm{t}}+b_3^{\mathrm{t}}), \tag{27}$$

$$h_2^{\mathrm{t}}=\mathrm{relu}(h_1^{\mathrm{t}}\times W_2^{\mathrm{t}}+b_2^{\mathrm{t}}), \tag{28}$$

$$h_1^{\mathrm{t}}=\mathrm{relu}(\mathbf{fte}_i\times W_1^{\mathrm{t}}+b_1^{\mathrm{t}}). \tag{29}$$

2.3 基于多模态融合的音乐情感识别模型

利用特征级融合方法和决策级融合两种方法，融合多种模态数据进行音乐情感识别方法的研究.

2.3.1 特征级融合模型

提出的特征级融合情感识别模型 FF-ERM (feature fusion emotion recognition model)处理框架如图 4 所示. 首先分别利用数据集中的 MIDI 数据和文本数据，依据之前提出的方法分别提取 MIDI 特征 $\boldsymbol{f}_i$ 和文本特征 $\mathbf{fte}_i$，然后连接两个特征，得到融合后的特征 $\mathbf{fu}_i$，最后输入 MLP 和 softmax 层得到情感结果 y_i'.

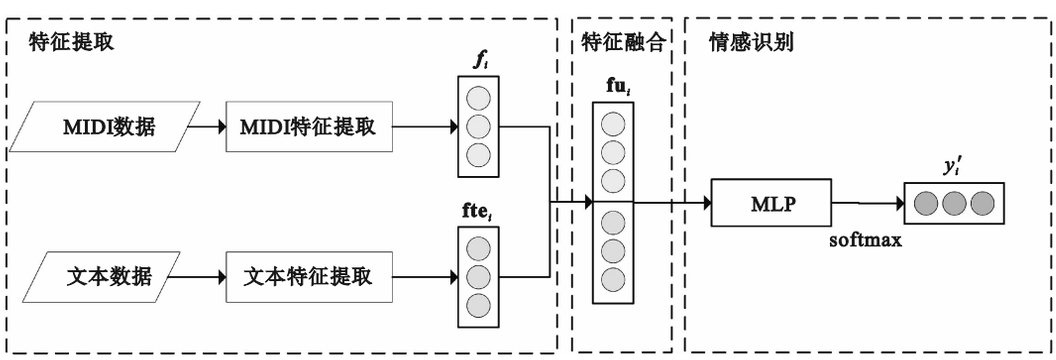

图 4 FF-ERM 框架图

Fig. 4 Diagram of FF-ERM frame

情感的预测结果 y_i' 使用式(30)计算，在式(31)至式(32)中，h^{ef} 为各层的隐藏层输出，W^{ef} 和 b^{ef} 为各层的参数.

$$y_i'=\mathrm{softmax}(h_2^{\mathrm{ef}}\times W_3^{\mathrm{ef}}+b_3^{\mathrm{ef}}), \tag{30}$$

$$h_2^{\mathrm{ef}}=\mathrm{relu}(h_1^{\mathrm{ef}}\times W_2^{\mathrm{ef}}+b_2^{\mathrm{ef}}), \tag{31}$$

$$h_1^{\mathrm{ef}}=\mathrm{relu}(\mathbf{fu}_i\times W_1^{\mathrm{ef}}+b_1^{\mathrm{ef}}). \tag{32}$$

2.3.2 决策级融合模型

提出的决策级融合情感识别模型(decision fusion emotion recognition model, DF-ERM)处理框架如图 5 所示. 首先，分别利用数据集中的 MIDI 和文本数据，依据之前提出的方法分别提取 MIDI 和文本特征 $\boldsymbol{f}_i$ 和 $\mathbf{fte}_i$；之后分别将两种特征输入各自的 MLP 和 softmax 层进行情感分类训练，预测结果分别为 $\boldsymbol{y}_{\mathrm{m}}=[x_1^{\mathrm{m}},x_2^{\mathrm{m}},x_3^{\mathrm{m}},x_4^{\mathrm{m}}]$ 和 $\boldsymbol{y}_{\mathrm{t}}=[x_1^{\mathrm{t}},x_2^{\mathrm{t}},x_3^{\mathrm{t}},x_4^{\mathrm{t}}]$，其中 x_j^{m} 和 x_j^{t} 分别表示 MIDI 模态和文本模态在第 j 类情感上的概率预测值. $\boldsymbol{y}_{\mathrm{m}}$ 利用式(33)得到，式(34)至式(35)中，h^{lfm} 分别为各层的隐藏层输出，W^{lfm} 和 b^{lfm} 为各层的参数. $\boldsymbol{y}_i$ 的计算使用了式(36)，式(37)至式(38)中，h^{lft} 分别为各层的隐藏层输出，W^{lft} 和 b^{lft} 为各层的参数. 之后将 $\boldsymbol{y}_{\mathrm{m}}$ 和 $\boldsymbol{y}_{\mathrm{t}}$ 进行线性加权求和得到融合结果 $\mathbf{rf}_{\mathrm{i}}=$

$[\mathrm{rf}_1,\mathrm{rf}_2,\mathrm{rf}_3,\mathrm{rf}_4]$,其中 $\mathbf{rf}_j$ 使用式(39)计算,式(39)中参数 ∂ 表示 MIDI 模态分类结果所占的比重.最后将 $\mathbf{rf}_j$ 经过 softmax 层,得到多模态融合结果 y'_i.

$$\boldsymbol{y}_{\mathrm{m}}=\mathrm{softmax}(h_2^{\mathrm{lfm}}\times W_3^{\mathrm{lfm}}+b_3^{\mathrm{lfm}}), \tag{33}$$

$$h_2^{\mathrm{lfm}}=\mathrm{relu}(h_1^{\mathrm{lfm}}\times W_2^{\mathrm{lfm}}+b_2^{\mathrm{lfm}}), \tag{34}$$

$$h_1^{\mathrm{lfm}}=\mathrm{relu}(\boldsymbol{f}_i\times W_1^{\mathrm{lfm}}+b_1^{\mathrm{lfm}}), \tag{35}$$

$$\boldsymbol{y}_{\mathrm{t}}=\mathrm{softmax}(h_2^{\mathrm{lft}}\times W_3^{\mathrm{lft}}+b_3^{\mathrm{lft}}), \tag{36}$$

$$h_2^{\mathrm{lft}}=\mathrm{relu}(h_1^{\mathrm{lft}}\times W_2^{\mathrm{lft}}+b_2^{\mathrm{lft}}) \tag{37}$$

$$h_1^{\mathrm{lft}}=\mathrm{relu}(\mathbf{fte}_i\times W_1^{\mathrm{lft}}+b_1^{\mathrm{lft}}) \tag{38}$$

$$\mathbf{rf}_j=\partial\times x_j^{\mathrm{m}}+(1-\partial)\times x_j^{\mathrm{t}}. \tag{39}$$

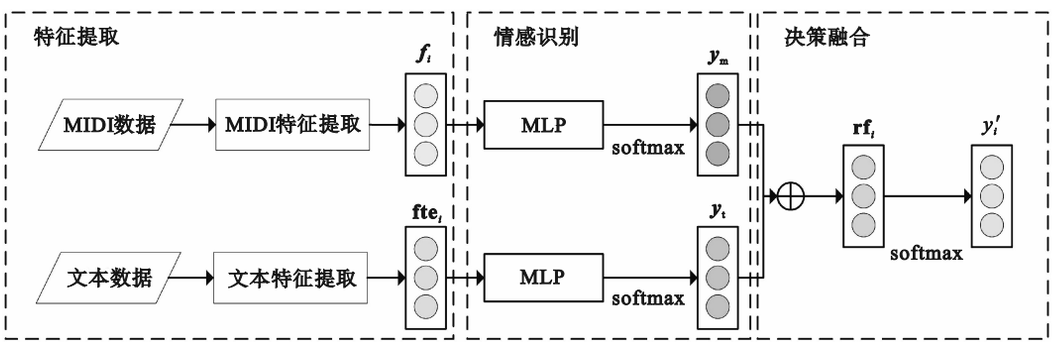

图5　DF-ERM 框架图

Fig. 5　Diagram of DF-ERM frame

3　实验与分析

3.1　数据集

使用了 MER 领域内数据量最大且模态信息最多的公开数据集 MSD,它整合了包括 The musiXmatch Datasets, The Last. fm Dataset 及 Lakh MIDI Dataset 等在内的多个权威音乐社区的数据,其中 The musiXmatch Datasets 提供歌词数据,The Last.fm Dataset 提供与音乐流派等信息相关的社交标签数据,Lakh MIDI Dataset 提供 MIDI 文件.现将音乐情感识别视为分类问题,将连续标签转化成对应的 4 类情感,即 V^+A^+(happy),V^-A^+(anxious),V^-A^-(sad)和 V^+A^-(relaxed).

3.2　参数设置

ERMSLM 模型设置音符嵌入的维度(e_{p})的取值为 50,段落数(m)的取值为 200,LSTM 网络隐藏层的向量维度(ls)的取值为 300.

ERMBT 模型设置社交标签分布表的行数(α)的取值为 10,歌词特征与社交标签特征相加时社交标签的权重(β)的取值为 3.

决策级融合模型(DF-ERM)设置 MIDI 模态分类结果所占的比重 ∂ 的取值为 0.3.

此次各模型的学习率(l)均设置为 0.000 1.学习率根据经验设定,而其余上述 6 个参数均通过大量实验确定.

3.3　对比与消融实验

为评估模型性能,选取领域内有代表性的方法进行复现,与 ERMSLM,ERMBT 进行性能对比,同时进行各模型的消融实验.

3.3.1　ERMSLM 模型

1)对比实验.① 手工特征+神经网络:该方法是文献[26]使用的模型,其从 MIDI 文件中手动提取了 9 种特征,并将提取的特征输入三层神经网络进行情感识别;② MFCC+SVM:其为音乐情感识别领域常用基线模型.复现时使用了 MSD 提供的 MFCC(mel-frequency cepstral coefficient),对于 SVM(support vector machine)的内核使用 sigmoid 和 RBF(radial basis function)两种核函数;③ MFCC+DBM:Huang 等[27]使用 MSD 提供的 MFCC 特征,将其输入 3 层 DBM(deep boltzmann machine)模型预测情感标签.

2)消融实验.① ERMSLM($\boldsymbol{m}_i$):仅使用旋律特征预测情感;② ERMSLM($\boldsymbol{m}_i$+$\boldsymbol{k}_i$):使用旋律和调性特征预测;③ ERMSLM(full):完整使用 MIDI 数据预测情感.

3.3.2　ERMBT 模型

1)对比实验.① BOW+DBM:该模型是文献[27]提出的模型.复现 DBM 模型的参数设置如下:可见层参数为 5 000,两层隐藏层的参数分别为 2 048 和 1 024;② TF-IDF+KNN:在提取 TF-IDF 特征时借鉴了文献[23]的特征提取过程.将提取的 TF-IDF 特征输入 KNN(k-nearest neighbor)进行情感识别,参数设置为 k=1.

2)消融实验.① ERMBT($\mathbf{fl}_i$):仅利用歌词特征进行情感预测;② ERMBT($\mathbf{fta}_i$):仅利用社交标签特征预测情感;③ ERMBT(full):利用完整

文本模态识别情感.

3.4 实验结果分析

实验时将数据集随机划分成85%和15%,其中85%用作训练集,15%用作测试集.评价指标是准确率(Accuracy)和Marco-F1值.

3.4.1 ERMSLM模型

ERMSLM模型的对比和消融实验结果见表5.

表5 对比和消融实验结果

Table 5 Contrast and ablation experiment results

类别	模型	Accuracy	Marco-F1
对比实验	手工特征+神经网络	0.383 2	0.388 4
	MFCC+SVM(RBF)	0.551 1	0.538 6
	MFCC+SVM(sigmoid)	0.554 7	0.519 4
	MFCC+DBM	0.551 1	0.594 4
消融实验	ERMSLM($\boldsymbol{m}_i$)	0.463 5	0.489 5
	ERMSLM($\boldsymbol{m}_i$+$\boldsymbol{k}_i$)	0.547 4	0.597 2
	ERMSLM	**0.569 3**	**0.599 9**

对比实验结果表明,手工特征+神经网络模型的准确率是所有模型中最低的,可能是由于该模型仅考虑了音高、音强和音长信息,并没有加入旋律、调式等其他音乐的中高层特征.MFCC+SVM模型以及MFCC+DBM模型的性能与手工特征+神经网络模型相比有了大幅提升,证明了MFCC特征在音频情感识别任务上的良好性能.MFCC+SVM的两个实验结果也反映出不同的核函数会对模型效果产生影响.MFCC+DBM模型的准确率与原文相比有所降低,这是因为数据量的减少导致的,说明数据量的大小对于深度学习模型而言十分重要.

通过分析消融实验结果发现,当仅使用旋律特征时,模型性能较低,可能是因为所提取的旋律特征仅包含音高和旋律走势信息,特征量不足.加入调性特征后,准确率有所上升,说明调性对于音乐情感的表达是有作用的,也说明调性信息提取模块是有效的.完整的ERMSLM模型准确率最高,达到了56.93%,不仅证明了ERMSLM模型的有效性,也说明旋律、调性和速度等音乐的中高层特征对于情感识别任务而言是十分重要的.

3.4.2 EMRBT模型

EMRBT模型的对比和消融实验结果见表6.对比实验结果表明,TFIDF+KNN模型准确率最低,BOW+DBM模型的准确率比TFIDF+KNN模型提高了1.69%,与原文相比有所降低,可能源于数据量的减少.

表6 对比和消融实验结果

Table 6 Contrast and ablation experiment results

类别	模型	Accuracy	Marco-F1
对比实验	TFIDF+KNN	0.534 2	0.462 6
	BOW+DBM	0.551 1	0.594 4
消融实验	ERMBT($\mathbf{fl}_i$)	0.551 1	0.594 4
	ERMBT($\mathbf{fta}_i$)	0.715 3	0.743 4
	ERMBT	**0.726 2**	**0.794 7**

通过分析消融实验可得出,利用抽取的歌词特征进行情感分类准确率达55.11%,与对比实验中性能最优的BOW+DBM模型相当;仅利用社交标签特征进行情感分类,准确率为71.53%,说明社交标签能在很大程度上反映音乐的情感信息,由于数据集的社交标签与音乐流派相关,也能间接说明音乐流派与音乐情感可能存在相关性;将歌词和社交标签进行融合之后准确率为72.62%,比单纯使用社交标签提高了1.09%,证明了对两种文本数据进行融合能进一步提升情感识别的准确率.

3.4.3 特征级融合模型和决策级融合模型

从整体准确率以及在各个情感类别上的准确率来对比4个模型的性能,4个模型分别为仅使用MIDI模态模型(ERMSLM)、仅使用文本模态模型(ERMBT)、特征级融合模型(FF-ERM)和决策级融合模型(DF-ERM).对比结果见表7,为了直观展示数据,绘制了图6.

表7 每个情感类别上的准确率

Table 7 Accuracy on each sentiment category

模型名称	V^+A^+	V^-A^+	V^-A^-	V^+A^-	四类均值
ERMSLM	0.538 0	0.55	0.636 4	0.5	0.569 3
ERMBT	0.722 6	0.75	0.848 5	0.6	0.726 2
FF-ERM	0.711 5	0.70	0.787 9	0.6	0.708 0
DF-ERM	0.737 2	0.75	0.878 8	0.7	**0.737 2**

通过分析图6可得:①在使用两种单模态数据进行情感识别时,文本模态能够取得更好的情感识别效果,其在四类上的准确率比MIDI模态要高15.69%,并且文本模态在每种情感类别上的分类准确率均高于MIDI模态;②在使用多模态数据进行情感识别时,决策级融合的效果比特征级融合的效果要好,其在四类上的准确率比特征级融合高2.92%,而特征级融合的效果比仅使用文本模态还要低1.82%;③四种分类模型均在V^-A^-情感类别上取得了最高的情感识别准确率,这可能是由于该情感类别的音乐在调式、速度、歌

词用词等方面存在独特性;④四种分类模型均在V^+A^-情感类别上取得了最低的情感识别准确率,这可能是由于属于该情感类别的音乐数量较少.

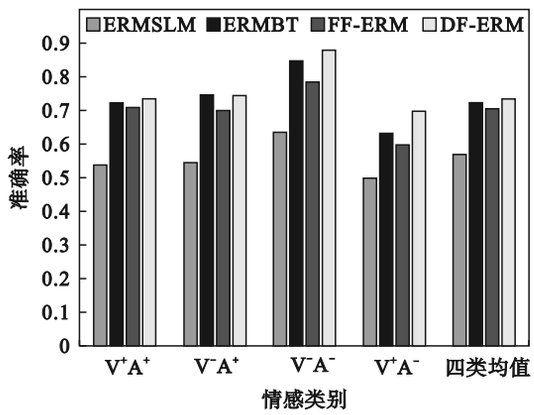

图6　实验结果图

Fig. 6　Histogram of experiment results

4　结　　语

MER在智能推荐,音乐可视化和自动编曲等领域都有着广泛的应用和理论研究价值.利用音乐的MIDI、歌词和社交标签数据进行基于多模态数据的MER研究,提出了基于音乐MIDI数据的情感识别模型ERMLSM以及基于音乐文本数据的情感识别模型ERMBT.同时采用了特征级融合和决策级融合两种融合方式,展开了基于音乐的MIDI数据和文本数据的多模态音乐情感识别研究,证实了多模态融合的有效性.

参考文献:

[1] Han D H, Kong Y R, Han J Y, et al. A survey of music emotion recognition [J]. *Frontiers of Computer Science*, 2022, 16(6): 166335.

[2] Jazi S Y, Kaedi M, Fatemi A. An emotion-aware music recommender system: bridging the user's interaction and music recommendation [J]. *Multimedia Tools and Application*, 2021, 80(9): 13559-13574.

[3] Dharmapriya J, Dayarathne L, Diasena T, et al. Music emotion visualization through colour [C]//2021 International Conference on Electronics, Information, and Communication (ICEIC). Jeju, 2021: 1-6.

[4] Novelli N, Proksch S. Am I (deep) blue? music-making AI and emotional awareness [J]. *Frontiers in Neurorobotics*, 2022, 16: 897110.

[5] Shukuroglou M, Roseman L, Wall M, et al. Changes in music-evoked emotion and ventral striatal functional connectivity after psilocybin therapy for depression [J]. *Journal of Psychopharmacology*, 2023, 37(1): 70-79.

[6] 陈晓鸥,杨德顺.音乐情感识别研究进展[J].复旦学报(自然科学版),2017,56(2):136-148.
(Chen Xiao-ou, Yang De-shun. Research progresses in music emotion recognition [J]. *Journal of Fudan University (Natural Science)*, 2017, 56(2): 136-148.)

[7] Panda R, Malheiro R, Paiva R P. Novel audio features for music emotion recognition [J]. *IEEE Transactions on Affective Computing*, 2020, 11(4): 614-626.

[8] Singh Y, Biswas A. Robustness of musical features on deep learning models for music genre classification [J]. *Expert Systems with Applications*, 2022, 199: 116879.

[9] 邓永莉,吕愿愿,刘明亮,等.基于中高层特征的音乐情感识别模型[J].计算机工程与设计,2017,38(4):1029-1034.
(Deng Yong-li, Lyu Yuan-yuan, Liu Ming-liang, et al. Music emotion recognition based on middle and high level features [J]. *Computer Engineering and Design*, 2017, 38(4): 1029-1034.)

[10] Qiu L, Zhong Y, Xie Q, et al. Multi-modal integration of EEG-fNIRS for characterization of brain activity evoked by preferred music [J]. *Frontiers in Neurorobotics*, 2022, 16: 823435.

[11] Delbouys R, Hennequin R, Piccoli F, et al. Music mood detection based on audio and lyrics with deep neural net [C]//Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR). Paris, 2018: 370-375.

[12] Jia X S. A music emotion classification model based on the improved convolutional neural network [J]. *Computational Intelligence and Neuroscience*, 2022, 2022: 6749622.

[13] Liu X, Chen Q, Wu X, et al. CNN based music emotion classification [J]. *arXiv prePrint arXiv*, 2017: 1704.05665.

[14] Keelawat P, Thammasan N, Kijsirikul B, et al. Subject-independent emotion recognition during music listening based on EEG using deep convolutional neural networks [C]//2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA). Penang, 2019: 21-26.

[15] Chowdhury S, Vall A, Haunscmid V, et al. Towards explainable music emotion recognition: the route via mid-level features [C]//Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR). Delft, 2019: 237-243.

[16] Ma Y, Li X X, Xu M X, et al. Multi-scale context based attention for dynamic music emotion prediction [C]// Proceedings of the 25th ACM international conference on Multimedia. Mountain View, 2017: 1443-1450.

[17] Liu H, Fang Y, Huang Q. Music emotion recognition using a variant of recurrent neural network [C]//Proceedings of the International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA). Chengdu, 2018: 15-18.

[18] Chang W H, Li J L, Lin Y S, et al. A genre-affect relationship network with task-specific uncertainty weighting for recognizing induced emotion in music [C]// 2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego, 2018: 1-6.

[19] Soleymani M, Aljanaki A, Yang Y, et al. Emotional analysis of music: a comparison of methods [C]//Proceedings of the ACM Conference on Multimedia (MM). Orlando: ACM, 2014: 1161-1164.

[20] Li X X, Tian J S, Xu M X, et al. DBLSTM-based multi-scale fusion for dynamic emotion prediction in music [C]//2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, 2016: 1-6.

[21] Chaki S, Doshi P, Patnaik P, et al. Attentive RNNs for continuous-time emotion prediction in music clips [C]// Proceedings of the 3rd Workshop in Affective Content Analysis co-located with Thirty-Fourth AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 36-46.

(下转第 792 页)