文章编号:2095-6134(2017)04-0431-08

基于局部特征优化的语音情感识别*

隋小芸[1,2],朱廷劭[1†],汪静莹[1,2]

(1 中国科学院心理研究所, 北京 100101; 2 中国科学院大学心理学系, 北京 100049)

(2016 年 3 月 29 日收稿; 2016 年 6 月 17 日收修改稿)

Sui X Y, Zhu T S, Wang J Y. Sample optimization based on local features in speech emotion recognition[J]. Journal of University of Chinese Academy of Sciences, 2017,34(4):431-438.

摘 要 情感识别在人机交互领域具有广阔前景。由于情感表达在时间上具有一定的持续性,统计特征更能体现不同情绪语音的差异和动态变化,大多数语音情感识别研究都使用全局特征(如最大值、最小值等),并没有充分挖掘局部特征(如单帧的短时能量、过零率等)中的信息。提出一种基于局部特征优化的方法,对每个情感语音样本做进一步提纯,通过聚类分析对情感特征相对不显著的帧进行过滤,在此基础上进行统计计算和分类,以提高预测的准确率。实验结果表明,基于优化后的样本进行情感分类,3 个语料库的平均准确率提高 5% ~17%。进一步的研究发现这种优化方法可能更适合于语音长度较长的情感识别场景。

关键词 语音情感识别; 局部特征; 全局特征; 聚类分析; 数据优化

中图分类号:TN391.4 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2017.04.004

Sample optimization based on local features in speech emotion recognition

SUI Xiaoyun[1,2], ZHU Tingshao[1], WANG Jingying[1,2]

(1 *Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*;
2 *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract Emotion recognition is one of the most prospective technics in human-machine interaction process. Most researches prefer statistical functional features because these features are more consistent with the speech variation as emotion changes. However, local features, i.e., short-term or temporal features extracted from single frame also contain useful information. In this work, a new approach is proposed to optimize samples via local features. To achieve this, a K-means cluster is employed to cluster each sample with 2 groups: frames with obvious emotion and frames with emotion which is not that obvious. It is hypothesized that the cluster with more frames should be emotionally obvious. It is observed in the results that the classification performs better on optimized samples than on original ones. The method was tested on 3 corpora and the classification accuracy increases by 5% -17%. It is also found the improvement increases as speech length grows, which implies the optimization approach may be more applicable to the longer speech recognition.

* 国家重点基础研究发展(973)计划(2014CB744600)资助

† 通信作者,E-mail:tszhu@psych.ac.cn

Keywords speech emotion recognition; local features; global features; cluster analysis; sample optimization

语音情感识别研究的开展距今已有 30 余年的历史，自动语音情感识别是计算机对人类上述情感感知和理解过程的模拟，它的任务就是从采集到的语音信号中提取表达情感的声学特征，并找出这些声学特征与人类情感的映射关系[1]。情感识别在人机交互领域具有广阔的前景，比如设计能够识别人类情感、具有社交功能的机器人；或者通过统计客户来电的情绪状态，分析并改善客服人员的服务质量[2]；在以电话方式进行的产品市场调查中，情绪数据也可以帮助调查人员对客户的满意度或购买倾向进行判断。

语音特征可以分为以统计函数计量的全局特征和以时间为自变量的局部特征[3]，全局特征是对整个语音文件的统计描述，如基频均值、方差等；而局部特征是对一个很短的时间（一般为一帧）内语音特征的测量，比如短时能量、短时过零率等。由于情感的表达在时间上具有一定的持续性，单帧的局部特征对于情感分析意义不大[4]，全局特征更能够体现情感的变化，大多数研究也证实基于全局特征的情感分类速度更快、准确率更高[3,5]。但这不代表局部特征对情感的识别没有用处，隐马尔可夫模型就常用在基于短时特征的语音识别[6]，以及根据频谱短时特征将原始数据按清音和浊音进行的分段建模[7]。

一般来说，情绪的酝酿是一个逐渐的过程，因此即便在同一句带有情感色彩的语音中，不同时间段的语音特征也存在差异。有研究发现中间位置的语音片段识别准确率最高[5]，也有研究发现句尾的单词或音节的情感区分度比其他位置的更高[8]。基于这种思路，本文提出每个情感语音样本中应该都存在少数情感特征不明显或者说与大多数帧特征不一致的帧（以下简称少数类），而根据局部特征进行聚类分析，可以把这些占少数的帧过滤掉，用剩下的情感特征更明显的帧（以下简称多数类）组成新的样本，再计算全局特征训练模型，即通过对局部特征的优化提高情感识别的准确率。

为了验证这种方法的适用性，本文采用 3 种不同语言、相互独立的语料库：EMO-DB[9]，RAVDESS[10]，以及在一次语音识别的预研究中采集的汉语普通话情感语音。其中前 2 个数据库自带的情感类别不完全相同，本文选取最典型的高兴、生气、悲伤、中性来简化实验设计。汉语的语音由于受预研究实验设计的限制，只有中性、负性 2 个类别，但语音文件长度较长，范围：70 ~ 157 s，因此本文进一步研究这种优化方法在不同语音长度时的效果。

1 语料库

EMO-DB 是由 10 位演员（5 位男性，5 位女性）阅读 10 个内容完全中性的句子，德语发音，7 种情感类别：中性、生气、恐惧、高兴、悲伤、厌恶、厌倦。数据经过情感感知实验的校验。本文选取 4 种情感类型（高兴、生气、悲伤、中性）的语音文件共 339 个。文件时长范围：1 ~ 8 s。

RAVDESS 是由 24 位演员（12 位男性，12 位女性）以不同情感阅读 2 个中性的句子，北美洲英语口音，包括 8 种情感：中性、冷静、高兴、悲伤、生气、害怕、惊讶、厌恶。每个句子以 2 种情感强度阅读 2 次：自然和强烈。并由 297 位参与者通过情感感知实验对该数据库的语音进行了校验。本文选取 4 种情感类型（高兴、生气、悲伤、中性）的语音文件共 384 个。文件时长范围：3 ~ 4 s。

汉语普通话语料库是在预研究中，通过让被试阅读有情感内涵的文章段落来诱发情绪，分为中性阅读和负性阅读 2 个任务。我们邀请 44 位年龄在 18 ~ 24 岁的在校大学生（男生 16 人，女生 28 人）参与录制。语音文件共 88 个。数据经过人为校验确认。中性阅读任务由被试阅读一段对交通堵塞的模型研究的科技文章段落，文件时长范围：73 ~ 125 s。负性阅读任务由被试阅读一段余华的小说《活着》中主人公孙子死亡时的场景段落，文件时长范围：70 ~ 157 s。

3 个语料库均为实验室安静环境下采集的高质量语音，目的是排除无关变量对研究结论的影响。

对 EMO-DB 和 RAVDESS 选择 4 种情绪类别主要是出于以下考虑：

1）情感分得越细，情感特征越加模糊，识别率将会大大地降低。所以现阶段的情感识别中，多采用 4 ~ 6 种情感分类[11]；

2）各种语料库对情绪的定义缺乏客观标

准[12]，EMO-DB 和 RAVDESS 中的情绪类别也不完全相同，不方便实验对照，因此选取最典型的高兴、生气、悲伤、中性来简化实验设计；

3）由于基频、能量或语速等特征的变化上存在着一定的相似性，在一些情绪之间存在着较高的混淆度，如高兴—惊讶，害怕—愤怒等[13]。解决方法通常是长短时特征融合[13-14]，而本文研究的重点是通过局部特征对样本数据的优化，所以没有选取更多的情绪类别。

2　研究方法

2.1　特征提取

本文使用 openSMILE[15] 工具对上述语音数据进行预处理和特征提取，该工具的特点是可定制化、灵活、高效。预处理包括以帧长 25 ms、帧移 10 ms 的汉明窗为参数进行加窗分帧，经过快速傅里叶变换，再对语音的高频部分进行加重，去除口唇辐射的影响，增加语音的高频分辨率，并辅以平滑处理以提高语音质量。

在特征的选取上，参考以往的研究[3,5,16-18]，并结合自身实践，确定 7 类语音特征：基频、均方根能量、前 3 个共振峰、过零率、谐噪比、浊音概率、MFCC（Mel frequency cepstral coefficient，梅尔频率倒谱系数）。基频和能量被诸多研究证实与情绪的激活强度相关[3]，MFCC 是语音识别领域使用最广泛的特征之一，由于其在高频区域对噪音更加敏感，而在低频区域具有很好的频率分辨率，因此大多数研究只使用低阶 MFCC 系数[19]，本文选用传统的 12 阶系数序列。

根据以上特征分类，对每一个特征计算其一阶导数（Δ），以便比较特征的动态变化。局部特征维度为 40。

全局特征则是在这 40 个维度上对同一个文件的所有帧的取值进行统计运算，包括最大值、最小值、均值、标准差、动态范围、峰度、偏度、斜率、截距、均方误差等 12 个统计函数，最终每一个样本文件对应一个 1 ×480 的特征向量，即构成分类算法的输入。

2.2　数据优化

数据的优化是针对每一个语音样本，以帧为单位，剔除情感特征不明显的帧，剩余帧组成新的样本。比如一个语音长度为 5 s 的样本文件，当帧长为 25 ms 时，共可以提取出 498 帧的局部特征数据，局部特征维度为 40。这样一个 498 ×40 的特征向量即原始样本，在 40 个局部特征中选取一个进行 K-means（K-平均算法）聚类，$K = 2$（即两类：情感特征明显和不明显），“不明显”类别的特征值所在的帧就会被剔除，具体过程见图 1。

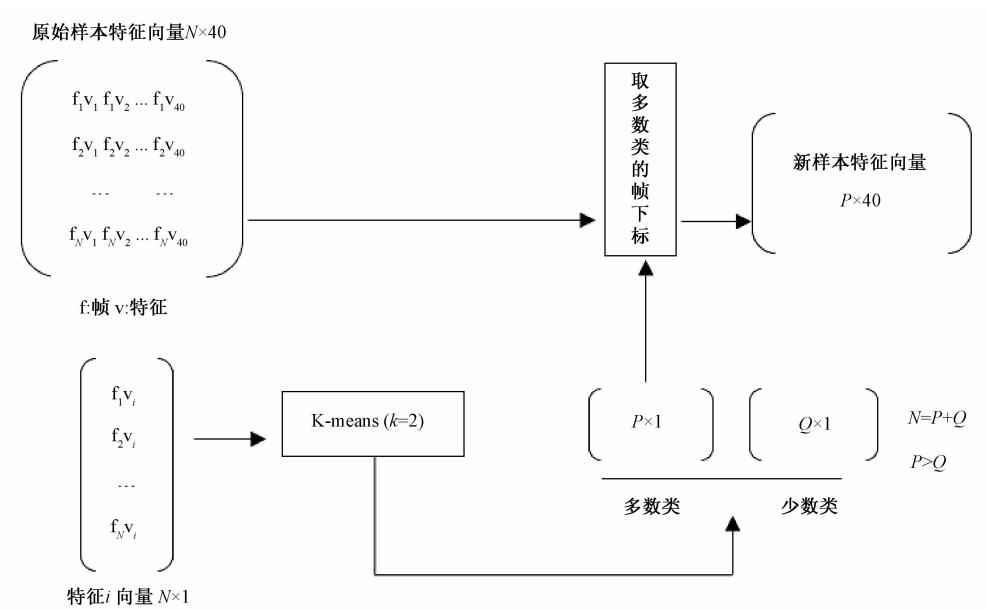

图 1　单样本聚类优化示意图

Fig. 1　Schematic of single sample optimization via cluster analysis

图 2 为 EMO-DB 训练集中某个悲伤语音样本的均方根能量频次分布，可以看出同一特征在不同帧的取值的确存在不平衡的现象。

可见大多数帧处于相对一致的能量值，而其

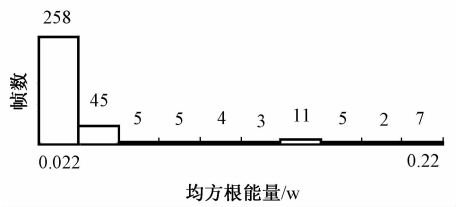

图 2 悲伤语音样本均方根能量频次分布

Fig. 2 Frequency distribution of RMS energy in one sample with sadness emotion

他帧的能量各异,它们可能是背景噪音或者情绪特征不明显的语音片段,因此是需要被剔除的部分。按照图 1 的优化方式实际操作该样本,K-means 算法在所有帧的均方根能量特征值中寻找两个中心点,围绕着两个中心点形成两类数据,每一类的数据到该类中心点位置最近。使用 Weka 聚类[20]后有 312 个帧被归为多数类构成这段悲伤语音的新样本,而剩下的 33 个帧被归为少数类应当被筛除掉。

2.3 实验流程

实验最初的设计是选择一个区分能力最好的(即优化后的数据识别准确率最高)特征进行聚类,剔除情感特征不明显的帧。在实验过程中发现,对于不同的语料库,最优的特征并不一致,EMO-DB 中 MFCC Δ 的区分能力最好,分类准确率为 80%;RAVDESS 中则为浊音概率 Δ:87.5%;而汉语语料库中的最优特征则是 ZCR(zero-crossing rate,过零率):100%。这也证实了以往研究的说法:由于说话人、语音内容、发音风格、语速等因素上的变异会直接影响提取的特征值,因此确定情感识别的最优特征充满挑战[3]。

为了保证在不同场景下自动情感识别的最佳效果,实验设计被修改为用 2.1 中列举的 7 类特征及它们的 Δ 特征分别优化数据,共 14 种优化方式,为每种优化方式建立一个 SVM(support vector machine,支持向量机)的分类器,再加上基于原始样本建立的分类器,共 15 个分类器:SVM N(N = 1,…,15),以各自分类结果的准确率为权重,进行投票,得到最终(Classifier16)的分类结果。流程如图 3。

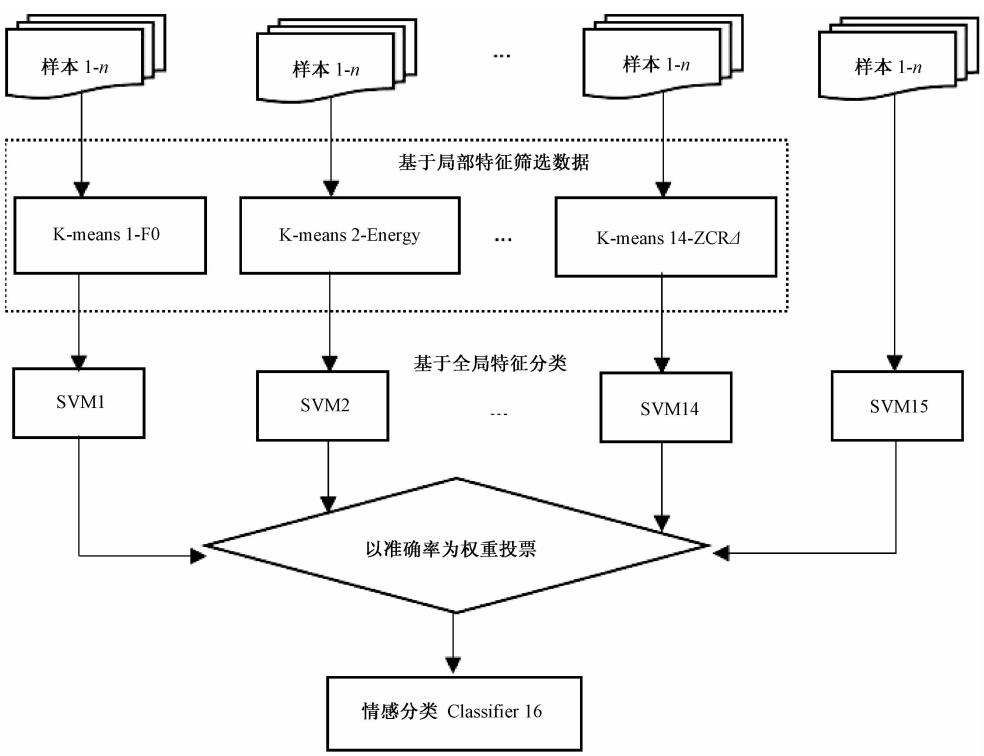

图 3 实验流程

Fig. 3 Flow chart of the experiment

3　结果与分析

3.1　聚类效果验证

本文的实验假设是每一个语音样本文件的所有帧都可以基于某种局部特征聚为两类：情感特征明显和不明显的类。并假设帧数较多的类（多数类）应该为情感特征更明显的类。多数类样本的分类效果应该比少数类样本更好，因此仅选择多数类数据产生新样本。

为了验证对两类的取舍是否合理，本文比较了 2.3 中的 14 种特征（优化方式）得到的多数类样本和少数类样本的分类效果，如图 4 所示。在 3 个语料库中，多数类样本的分类效果均有明显优势，说明两类的聚类原则是可行的。但是二者之间的差距并不如实验前预期的那么大，而且个别特征的少数类样本分类效果更佳，说明剔除的样本数据中仍然存在着可以利用的特征，在未来的研究中可以施加更加精细的聚类对其加以甄别。

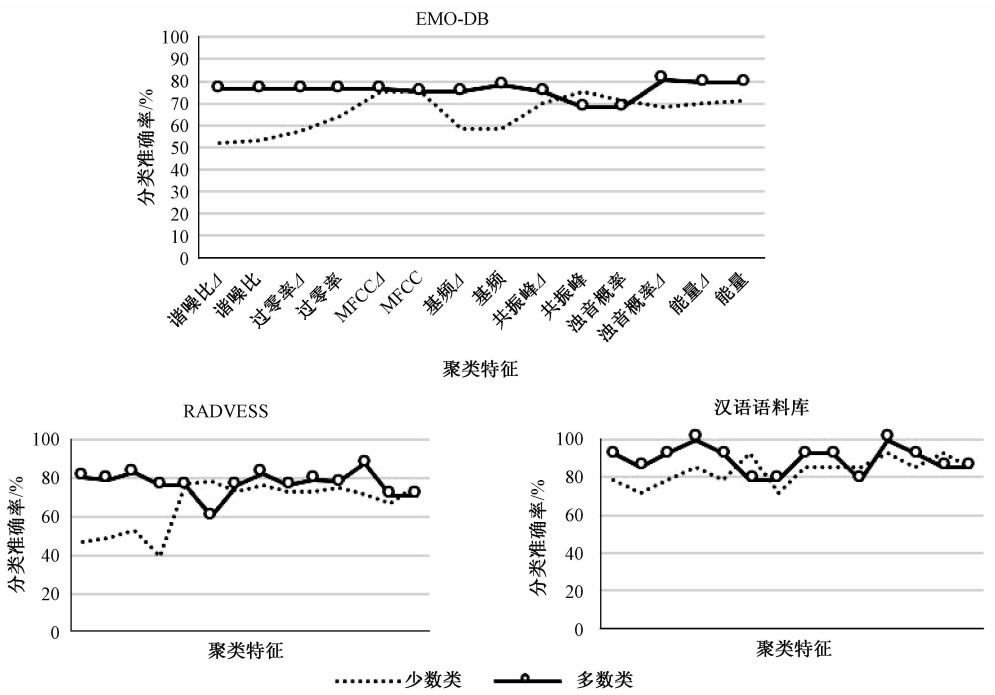

图 4　少数类与多数类样本分类效果对比

Fig. 4　Accuracy comparison between minority and majority samples

3.2　分类结果

分类算法采用支持向量机即 SVM 算法，这种算法在解决小样本、非线性、高维模式识别中表现出许多特有的优势，因此在语音情感识别领域被广泛应用。SVM 算法的目的是在多维（本文为 480 维）空间中寻找一个超平面，能够把各种类别区分开来，并且最大化这个超平面到各类数据的距离，以降低分类器的泛化误差。本文使用 LIBSVM[21] 软件进行分类，并使用网格搜索的交叉验证方法选取最优的分类。

3 个语料库的训练集和测试集的比例均为 80∶20。同时所有的数据是按照非特定说话人的方式分配，即同一个被试的语音不会同时出现在训练集和测试集中。这样确保了识别的效果不依赖于说话人的影响[17]，也更符合现实中语音识别的应用场景：训练数据和模型基于有限的样本建立，但可以用来识别范围更广泛的人群。

从表 1 可以看出，基于优化后的样本进行情感分类，准确率普遍高于未优化时，3 个语料库的平均准确率提高 5% ~17% 。

表 1　情感识别准确率百分比优化前后对比

Table 1　Accuracies before and after sample optimization　%

情绪	EMO-DB					RAVDESS					汉语语料库		
	高兴	生气	伤心	中性	平均	高兴	生气	伤心	中性	平均	负性	中性	平均
原始样本	23.5	96.3	100	68.8	73.9	83.3	83.3	66.7	79.2	78.1	100	71.4	85.7
优化后样本	23.5	100	100	81.3	78.1	91.7	91.7	75	83.3	85.4	100	100	100

需要注意的是 EMO-DB 中“高兴”的语音识别准确率在优化前后都非常低,这可能与样本本身的问题有关,因为当以相同的算法,以随机分配(同一个被试的语音在训练集和测试集中都可能出现)的方式分配训练集和测试集时,这一类别的准确率可以达到 78.6% ,说明这一类别的语音数据在不同说话人之间特征差异性较大,这种差异可能来源于不同演员演绎高兴情绪的表现偏差,因此基于这种数据建立的模型缺乏通用性。

3.3 结果对比

情感分类结果受算法、样本数量、特定说话人[17]、情绪类别个数[11]等因素影响,而且以往的大多数研究都没有注明是否特定说话人,或者只有平均值的报告,缺少各类情绪的详细结果。为了客观比较本文方法与其他研究的结果,本文选择几个基础算法相同(均为 SVM)、情绪分类相同(EMO-DB:4, RAVDESS:6)的研究进行对比,而本文也对 EMO-DB 和 RAVDESS 数据集重新执行了情绪识别,在测试中放弃了“非特定说话人”的严格设定,训练集和测试集随机分配产生。

从表 2 可以看出,使用随机分配的数据集,对比表 1 使用非特定说话人的数据集,准确率明显上升,验证了文献[17]中特定说话人对实验结果的影响。使用 EMO-DB 测试,本文方法比其他两篇研究的结果总体更好。

表 2 本文与其他研究在 EMO-DB 数据集上的准确率比较

Table 2 Accuracy comparison with other researches based on EMO-DB dataset %

	高兴	生气	伤心	中性	平均
本文方法	78.6	96	100	95.3	92.5↑
普通 SVM 方法[22]	88.89[1]	90.2	100	94.74	86.36[2]
基于多重分形的 SVM 方法[23]	90	70	93	76.6	82.4

注:1 为该研究中 10 次测试识别“高兴”情绪的最高准确率,其他 3 种情绪类别的结果同理;2 为该研究中 10 次测试的平均准确率。

表 3 中由于加入了“冷静”、“害怕”两个类别,与表 1 的结果相比,大多数指标都有下降,其中“高兴”和“害怕”尤其偏低。但总体均值仍然比文献[24]结果略高。

文献[22-24]均使用基于 SVM 算法的更复杂的模型,训练集和测试集的样本个数也与本文不完全相同,但从比较的结果来看,本文的方法至少不亚于其他优化方式。而且下一步改进的空间也很大,比如“高兴”类别的准确率始终偏低,可能需要对这种情绪的局部特征进一步分析,改进聚类算法。

表 3 本文与其他研究在 RAVDESS 数据集上的准确率比较

Table 3 Accuracy comparison with other researches based on RAVDESS dataset %

	高兴	生气	伤心	中性	冷静	害怕	平均
本文方法	71.9	87.5	91.7	79.2	83.3	66.7	80.1↑
Simple Model[24]	96	82	58	70	84	88	79.7

3.4 语音长度分析

汉语语料库的语音文件长度较长(70 ~ 157 s),因此本文进一步分析了数据优化在不同语音长度时的效果。将每个语音文件截取固定长度:5,10,20,30,40,50 s,对每一种长度,对图 3 中前 14 种分类器(基于优化过的样本,SVM N,N = 1,…,14)分别计算其与第 15 种分类器(基于原始样本,SVM15)的准确率差异,取均值,即 AVG(SVM N-SVM 15),结果以图 5 中的虚线表示。在语音较短时,只有少数特征有优化效果,随着语音长度增加,更多特征表现优异,因此平均差异由负值到正值,并逐渐增加。表明本文的优化方式在较长的样本上识别效果更好,在表 1 中汉语语料库优化后的分类准确率也的确是 3 个语料库中增幅最大的。图 5 的实线为使用本文方法分类(即图 3 中的 Classifier16)后的结果与 SVM15 的准确率差值,也可以看出本文方法在多数情况下分类结果优于未优化时的结果,并且随语音长度增加,优势更加明显。

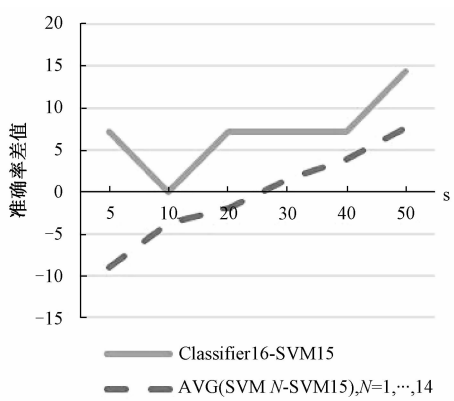

图 5 分类准确率差异均值与语音长度关系

Fig. 5 Relation between mean accuracy difference and utterance length

4 总结

本文在文献[5,8]的基础上,充分利用情感

语音原始样本中的局部特性，提出一种新的基于局部特征优化的方法，以 K-means 算法将原始数据的每一个样本聚为两类，剔除帧数较少的一类，再以新的样本数据，进行统计运算得到全局特征，运行 SVM 分类。在此过程中，也引入多个分类器以准确率为权重投票的机制，提高了分类方法的健壮性和适用性。

在 3 个语料库上的分类结果均比未优化时有明显提高，表明该方法的有效性。本文也分析了不同语音长度时优化效果的变化，发现随着语音长度的增加，各种特征优化数据的效果都得到了改善，最终的分类准确率也随之提高，说明本文的优化方法更适合于持续时间较长的语音情感识别。

该方法虽然简单易行，但由于使用全部的特征产生多个分类器进行投票，程序运行的时间成本也相应增加，下一步的研究中可以深入探讨最优特征的选择，通过减少分类器提高分类效率。优化数据时的聚类原则也比较笼统，在 3.1 的分析中可以看出，用少数类数据组成的样本，最终的分类效果虽然不如多数类样本，但差距并不是很大，这说明被剔除的少数类数据中依然存在着数量相当可以利用的情绪特征，在接下来的研究中可以尝试更精细的聚类分析和数据优化方式。对于易混淆的情绪类别如“高兴”、“害怕”等也需要进一步探索其局部特征的特殊之处来改进聚类算法，提高准确率。对于自然场景中的语音识别，即背景噪声的影响更大时，也有待进一步验证该方法的健壮性。

参考文献

[1] 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.

[2] Gupta P，Rajput N. Two-stream emotion recognition for call center monitoring [C]//Proceedings of International Conference on Spoken Language Processing(Interspeech)2007. Antwerp：International Speech Communication Association (ISCA)，2007：2 241-2 244.

[3] El Ayadi M，Kamel M S，Karray F. Survey on speech emotion recognition：features，classification schemes，and databases [J]. Pattern Recognition，2011，44 (3)：572-587.

[4] Vogt T，André E. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition [C] // Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2005). Amsterdam：IEEE，2005：474-477.

[5] Schuller B，Rigoll G. Timing levels in segment-based speech emotion recognition [C]//Proceedings of Interspeech 2006. Pittsburgh：ISCA，2006：1 818-1 822.

[6] 林奕琳，韦岗. 基于短时和长时特征的语音情感识别研究[J]. 科学技术与工程，2006，6(4)：450-454.

[7] Kim E H，Hyun K H，Kim S H，et al. Speech ermotion recognition separately from voiced and unvoiced sound for emotional interaction robot [C]//International Conference on Control，Automation and Systems 2008. Seoul：IEEE，2008：2 014-2 019.

[8] Rao K S，Koolagudi S G，Vempada R R. Emotion recognition from speech using global and local prosodic features [J]. International Journal of Speech Technology，2013，16(2)：143-160.

[9] Burkhardt F，Paeschke A，Rolfes M，et al. A database of German emotional speech [C]//Proceedings of Interspeech 2005. Lisbon：ISCA，2005：1 517-1 520.

[10] Livingstone S R，Peck K，Russo F A. Ravdess：the ryerson audio-visual database of emotional speech and song [C]// Proceedings of the 22nd Annual Meeting of the Canadian Society for Brain，Behaviour and Cognitive Science (CSBBCS). Kingston：CSBBCS，2012：71-72.

[11] 余伶俐，蔡自兴，陈明义，等. 语音信号的情感特征分析与识别研究综述[J]. 电路与系统学报，2007，12(4)：76-84.

[12] Eyben F，Batliner A，Schuller B，et al. Cross-Corpus classification of realistic emotions：some pilot experiments [C] // Proc. 3rd International Workshop on Emotion (satellite of LREC). Valletta：The Association for the Advancement of Affective Computing，2010：77-82.

[13] 蒋丹宁，蔡莲红. 基于语音声学特征的情感信息识别[J]. 清华大学学报(自然科学版)，2006，46(1)：86-89.

[14] 韩文静，李海峰，韩纪庆. 基于长短时特征融合的语音情感识别方法[J]. 清华大学学报(自然科学版)，2008，48(S1)：708-714.

[15] Eyben F，Weninger F，Gross F，et al. Recent developments in opensmile，the munich open-source multimedia feature extractor[C] // Proceedings of the 21st ACM international conference on Multimedia. Barcelona：Association for Computing Machinery (ACM)，2013：835-838.

[16] Petrushin V A. Emotion recognition in speech signal：

experimental study, development, and application [C]// Proceedings of Interspeech 2000. Beijing: ISCA, 2000:222-225.

[17] Bhaykar M, Yadav J, Rao K S. Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM [C]// Proceedings of National Conference on Communications (NCC) 2013. New Delhi: IEEE, 2013: 1-5.

[18] Kwon O W, Chan K, Hao J, et al. Emotion recognition by speech signals [C]// Proceedings of Interspeech 2003. Geneva: ISCA, 2003:125-128.

[19] 韩一,王国胤,杨勇. 基于 MFCC 的语音情感识别[J]. 重庆邮电大学学报(自然科学版),2008(5):597-602.

[20] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update [J]. ACM SIGKDD explorations newsletter, 2009, 11(1): 10-18.

[21] Chang C C, Lin C J. LIBSVM: a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 75-102.

[22] 朱菊霞,吴小培,吕钊,等. 基于 SVM 的语音情感识别算法[J]. 计算机系统应用,2011,20(5):87-91.

[23] 叶吉祥,张密霞,龚希龄,等. 基于 MF-DFA 的语音情感识别[J]. 长沙理工大学学报(自然科学版),2011,8(2):67-71.

[24] Zhang B, Essl G, Provost E M. Recognizing emotion from singing and speaking using shared models [C]// Proceedings of Affective Computing and Intelligent Interaction (ACII) 2015. London: IEEE, 2015: 139-145.