第37卷 第2期 华侨大学学报(自然科学版) Vol. 37 No. 2
2016年3月 Journal of Huaqiao University (Natural Science) Mar. 2016

文章编号:1000-5013(2016)02-0196-05 **doi**:10.11830/ISSN.1000-5013.2016.02.0196

采用潜在概率语义模型和K近邻分类器的音频分类算法

辛欣[1,3], 陈曙东[2,3], 仝明磊[4], 胡文皓[1,3], 刘陈伟[1,3], 葛浩栋[3]

(1. 中国科学院大学 电子电气与通信工程学院, 北京 100049;
2. 中国科学院微电子研究所, 北京 100029;
3. 中国物联网研究发展中心, 江苏 无锡 214135;
4. 上海电力学院 电子与信息工程学院, 上海 200090)

摘要: 提出一种基于潜在概率语义(PLSA)模型和K近邻分类器的音频分类算法.首先,将信号特征向量送入潜在概率语义模型中训练,获得声音主题词袋模型;然后,使用K近邻分类器(KNN)进行分类.实验结果表明:与传统的K近邻分类算法相比,提出的算法在分类效果上有较明显的改善.

关键词: 梅尔频率倒谱系数; 词-频共现矩阵; 声音主题词袋模型; 潜在概率语义模型; K近邻分类器

中图分类号: TP 391 **文献标志码**: A

音频分类是异常声音检测系统中的重要模块,可以依据音频特征区分不同的音频信号.在音频分类研究中,Radhakrishnan等[1]给出在电梯里检测罪案的音频分类系统框架.该系统提取梅尔倒谱系数(Mel frequency cepstral coefficients,MFCC)特征,用高斯混合模型(GMM)进行分类和识别报警声、撞墙声等声音事件.Atrey等[2]采用过零率ZCR和线性倒谱系数(linear frequency cepstral coefficient,LFCC)等特征参数,使用GMM模型进行分类.结合词袋模型,Aucouturier[3]提出了对应的帧袋模型,将音频中的一些连续的音频帧作为一个整体提取相应的音频特征.冀中等[4]提出了一种与HMM相结合的分层音频分类算法.Zeng等[5]采用一些常用的音频特征,如梅尔倒谱系数、声音功率等进行聚类和分类,利用潜在概率语义模型(PLSA)对音乐片段进行分类.容宝华[6]提出了一种基于MFCC的简化的特征,选取最近邻分类器和K近邻分类器,对音频进行分类.目前,音频分类算法在异常声音监测系统中的效果仍然不够理想.为了改善分类准确率,本文在传统的K近邻分类算法基础上进行改进,在提取音频信号特征以后不立即进行分类,而是先送入PLSA概率语义模型获取声音主题词袋模型,降低语音信号特征矩阵的维数,再使用分类器进行分类.

1 音频分类算法

算法分为3个模块:特征提取与处理模块、概率潜在语义模型模块和分类器模块,分别完成特征矩阵提取、获取声音主题词袋模型和分类功能.

1.1 特征提取与处理

由于梅尔频率倒谱系数含有语义信息,并结合人耳感知特性与语音信号的产生机制,具有良好的识别能力和抗噪性能.而差分倒谱系数能表现出2个音频帧之间的关联,体现帧与帧之间的信息量.因此,选择提取前12维的MFCC系数和12维的MFCC一阶差分系数.

在特征提取之前,需要对每一段音频进行预处理.首先,对原始音频信号进行预加重处理,以提升高

收稿日期: 2015-09-18

通信作者: 陈曙东(1977-),女,研究员,博士,主要从事大数据管理的研究.E-mail:chenshudong@ciotc.org.

基金项目: 江苏省基础研究计划(自然科学基金)面上项目(BK20141116)

频部分的能量，减少尖锐噪声的影响；然后，分割音频为 1 s 大小的片段，相邻片段间无重叠部分；再对每个片段加汉明窗形成帧，帧长约 23 ms，相邻帧之间有 50%的重叠部分. 特征提取流程，如图 1 所示.

对得到的语音信号进行滤波、去噪，获取 24 维 MFCC 特征向量. 类似于文本文件，每一段语音信号相当于一篇文本. 文本中的词对应语音信号中的 MFCC 帧向量. 对于已知类别的语音信号训练集，由于每段语音信号长度不一样，假如第 i 个语音信号 d_i 包含 M 帧，每一帧是 N 维，那么每一个语音信号表示为 $d_i=\{w_1,w_2,w_3,\cdots,w_M\}\in\mathbf{R}^{M\times N}$，$i=1,2,3,\cdots,n$. 全部音频文件构成一个 $V\times N$ 的矩阵，其中，$V=M\times n$. 这个矩阵是一个由帧向量构成的全部特征集合，进而可以得到一个 $V\times n$ 的词-频共现矩阵 $\boldsymbol{T}_{i,j}=\mathrm{num}(w_j,d_i)$，它由音频文件在这个特征集合中出现的频率构成. 对于测试音频文件(语音信号)集 $d_{\mathrm{test},i}$，$i=1,2,3,\cdots,n_{\mathrm{test}}$，同样会得到一个 $V\times n_{\mathrm{test}}$ 的词-频共现矩阵 $\boldsymbol{T}_{\mathrm{test}}=\mathrm{num}(w_j,d_{\mathrm{test}})$.

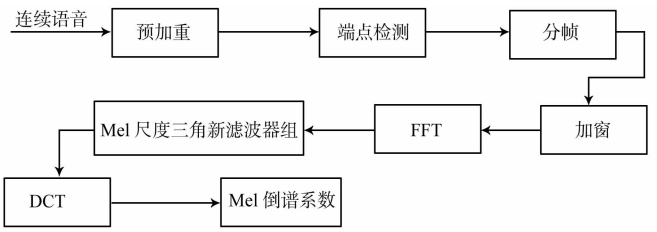


图 1　特征提取流程图

Fig. 1　Feature extraction flow chart

1.2　概率潜在语义分析模型

概率潜在语义分析模型(probabilistic latent semantic analysis，PLSA)通过[7]奇异值分解，将高维词汇-文档共现矩阵降到低维的潜在语义空间，使表面上不相关的词汇能体现深层次的语义联系，从而能解决部分文档语义分析中同义词与异义词的问题.

在 PLSA 的方法中，将隐含变量 $z\in Z=\{z_1,z_2,z_3,\cdots,z_k\}$ 与词汇 $w\in W=\{w_1,w_2,\cdots,w_M\}$ 在文档 $d\in D=\{d_1,d_2,\cdots,d_N\}$ 中出现的频率(词频共现矩阵)联系起来，该统计模型表示为[8]词汇与文档的联合概率.

$P(d_i)$是每段语音信号出现的概率，$P(z_k\mid d_i)$表示潜在主题类别 z_k 的条件概率，$P(w_i\mid z_k)$表示在潜在主题 z_k 下产生音频帧 w_i 的概率，即

$$P(d_i,w_j)=P(d_i)P(w_j\mid d_i),\tag{1}$$

$$P(w_j\mid d_i)=\sum_{k=1}^{K}P(w_j\mid z_k)P(z_k\mid d_i).\tag{2}$$

PLSA 模型假设词-文档对之间相互独立，潜在语义词 z_k 在词和文档上的分布独立. 模型给出目标函数为

$$L=\sum_{i=1}^{N}\sum_{j=1}^{M}f(d_i,w_j)\log P(d_i,w_j).$$

上式中：$f(d_i,w_j)$为词汇 w_j 在文档 d_i 中的概率.

采用 EM 算法求解参数，有如下 2 个步骤.

步骤 E

$$P^{(r)}=(z_k\mid d_i,w_j)=\frac{P^{(r-1)}(w_j\mid z_k)P^{(r-1)}(z_k\mid d_i)}{\sum\limits_{k=K}^{k=1}P^{(r)}(w_j\mid z_k)P^{(r-1)}(z_k\mid d_i)},\tag{3}$$

步骤 M

$$P^{(r)}(w_j\mid z_k)=\frac{\sum\limits_{i=1}^{i=N}f(d_i,w_j)P^{(r)}(z_k\mid d_i,w_j)}{\sum\limits_{j=1}^{j=M}\sum\limits_{i=1}^{i=N}f(d_i,w_j)P^{(r)}(z_k\mid d_i,w_j)},\tag{4}$$

$$P^{(r)}(z_k\mid d_i)=\frac{\sum\limits_{j=1}^{j=M}f(d_i,w_j)P^{(r)}(z_k\mid d_i,w_j)}{\sum\limits_{k=1}^{k=K}\sum\limits_{j=1}^{j=M}f(d_i,w_j)P^{(r)}(z_k\mid d_i,w_j)}.\tag{5}$$

通过不停地迭代步骤 E 和步骤 M，直到收敛，使得 L 最大化. 训练结束后，得到的 $P(z_k\mid d_i)$即为声

音主题词袋模型.而对于测试语音,则得到 $P(z_k|d_{\text{test}})$. $P(z_k|d_i)$和 $P(z_k|d_{\text{test}})$作为分类器的输入进行分类.由于文中算法是在K近邻算法上的改进,所以选择K近邻(K-nearest neighbor,KNN)分类器.

1.3 分类器

K 近邻是一种基本分类与回归方法[9].设特征空间 χ 是 n 维实数向量空间 $\mathbf{R}^n$, $\boldsymbol{x}_i, \boldsymbol{x}_j \in \chi$, $\boldsymbol{x}_i=(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^{\mathrm{T}}$, $\boldsymbol{x}_j=(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)})^{\mathrm{T}}$, $\boldsymbol{x}_i, \boldsymbol{x}_j$ 的 L_p 距离为

$$L_p(\boldsymbol{x}_i, \boldsymbol{x}_j)=\left(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p\right)^{1/p}, \qquad p \geqslant 1. \tag{6}$$

当 $p=2$ 时, L_p 称为欧氏距离(Euclidean distance).文中算法的KNN采用欧氏距离.

K近邻算法具体如下.

输入:训练数据集

$$T=\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}.$$

其中: $\boldsymbol{x}_i \in \chi \subseteq \mathbf{R}^n$ 为训练实例的特征向量; $y_i \in \gamma=\{c_1, c_2, \cdots, c_k\}$ 为实例的类别, $i=1,2,\cdots,N$.

特征向量 $\boldsymbol{x}$ 为 $P(z_k|d_i)$, y 通过人工方法标注为已知的.测试特征向量 $\boldsymbol{x}_{\text{test}}$ 为 $P(z_k|d_{\text{test}})$.

输出:实例 $\boldsymbol{x}_{\text{test}}$ 所属的类 y.

1) 根据距离度量,在训练集 T 中找出与 $\boldsymbol{x}_{\text{test}}$ 最近邻的 k 个点,涵盖 k 个点的 $\boldsymbol{x}$ 的邻域记为 $N_k(\boldsymbol{x})$.

2) 在 $N_k(\boldsymbol{x})$ 中,根据分类决策准则(如多数表决规则)决定 $\boldsymbol{x}$ 的类别 y,即

$$y=\arg\max_{c_j}\sum_{x_i \in N_k(x)} I(y_i=c_j), \qquad i=1,2,\cdots,N; \quad j=1,2,\cdots,k. \tag{7}$$

式(7)中: I 为指示函数,即当 $y_i=c_j$ 时, I 为1;否则, I 为0.

2 实验与结果分析

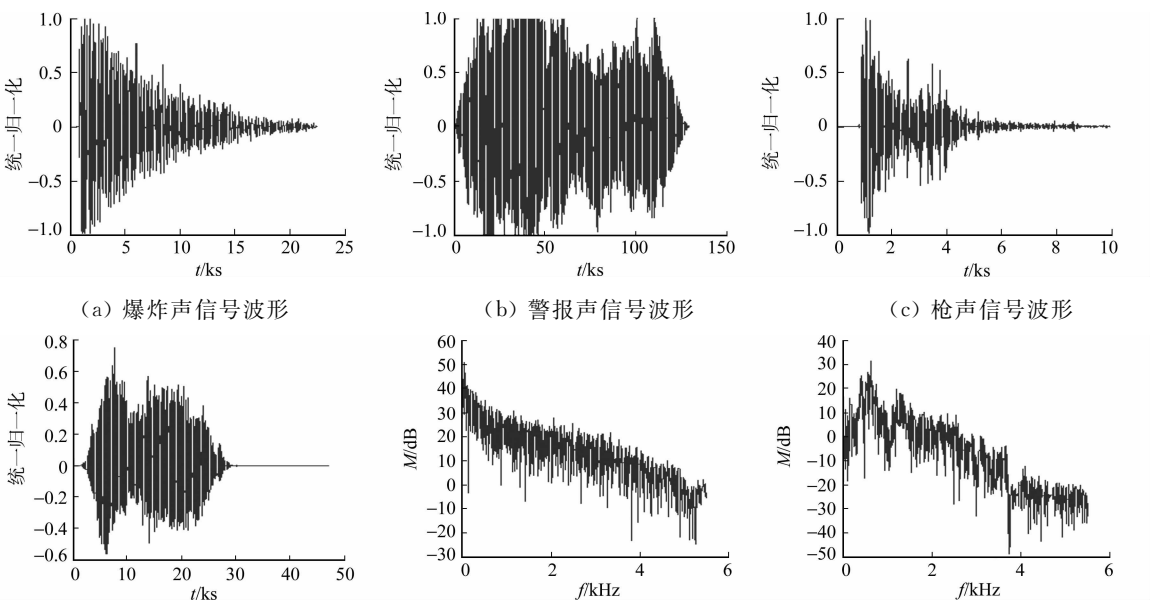
2.1 训练样本

为了验证文中算法的有效性,采用已获得标记后的音频样本集,在单机实验环境中验证,并与传统的KNN算法作对比.

1) 测试环境搭建:4 GB内存的PC机;Windows 7操作系统;Matlab R2010b.

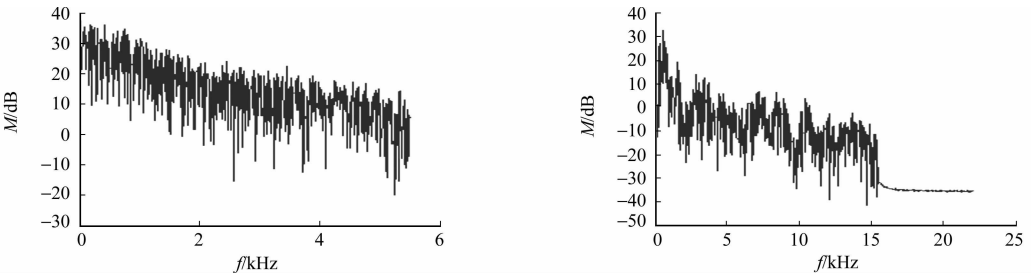
2) 测试数据集:4类音频文件,分别是爆炸声、枪声、警报声、人声(呼救声).其中,训练声音文件192个,测试音频文件90个.声音是未经过去噪的原始语音.

这4类声音是异常声音检测系统中最重要的检测对象,能够检测出这几种声音有很重要的实际价值.因此,实验中以此为测试数据集.训练语音的各类别样本,如图2所示.图2中: t 为时间; f 为频率;



(a) 爆炸声信号波形 (b) 警报声信号波形 (c) 枪声信号波形

(d) 人声信号波形 (e) 爆炸声信号频谱 (f) 警报声信号频谱



（g）枪声信号频谱　　（h）人声信号频谱

图 2　训练语音波形图和频谱图

Fig. 2　Training voice waveform and spectrum

M 为幅度.

2.2　实验验证

分别提取训练声音信号和测试声音信号的 MFCC 和一阶差分 MFCC 特征后，构建词-频共现矩阵，送入 PLSA 模型进行训练获得 $P(z_k|d_i)$和 $P(z_k|d_{\text{test}})$，再送入 KNN 分类器中，得到分类结果.

使用 K 近邻分类，当 k 取值不同时，分类的准确率是不同的. 当 k 取值不同时，使用 PLSA＋KNN 的分类方法与直接使用 KNN 分类方法的结果对比，如图 3 所示. 图 3 中：η 为分类准确率；k 为最近邻的数目.

由图 3 可知：当 $k=11$ 时，分类准确率最高. 取 $k=11$，对比其他分类方法的准确率(η)，如表 1 所示.

表 1　$k=11$ 时不同算法分类结果对比

Tab. 1　Comparisons of classification results when k is taken as 11

算法	PLSA＋KNN	KNN	PLSA＋SVM
η/%	38.89	35.56	37.78

由于文中实验数据处理及环境、参数设置的问题，实验的结果与原始论文中的算法的结果之间可能有所差别.

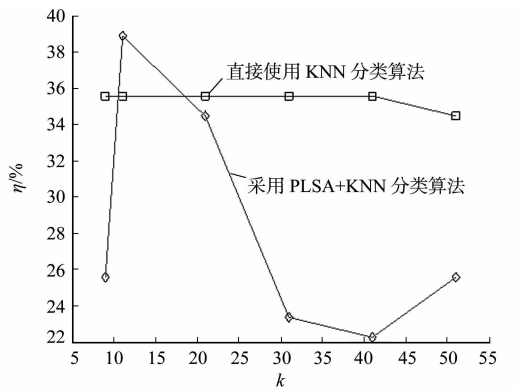


图 3　k 取不同值时分类结果

Fig. 3　Classification results of different values of k

2.3　结果分析

在实验中，对于 4 类音频文件组成的训练集，先构建词-频共现矩阵，使用 PLSA 模型计算隐含主题类别的条件概率，获得声音主题词袋模型，再送入 KNN 分类器的方法比直接用 KNN 分类器处理词频矩阵的分类方法有更高的准确率，准确率提高了大约 3.3%. 对于 KNN 分类器，k 值的选取也会影响到算法的分类精度. 文中算法对比了 PLSA＋KNN 算法与传统 KNN 算法分别在 k 取最优值时的分类效果. 此外，如果使用 SVM[10] 分类器(PLSA＋SVM)，相较于 KNN 也有约 2.2%的提高. 而同样是先送入 PLSA 模型中，选择了不同的分类器时，PLSA＋KNN 算法比 PLSA＋SVM 算法有更好的性能表现，提高了约 1.1%. 由于音频数据样本量少、特征提取方法等原因，整体准确率偏低. 但是，不同分类方法之间的区别还是较为明显的. 实验结果表明：文中算法相较于传统算法，在分类效果上有较明显的改善.

3　结束语

音频分类模块属于异常声音监测系统的重要模块. 目前，采用的传统分类算法在分类效果方面不能很好地满足实际的需求. 提出一种基于潜在概率语义模型和 K 近邻算法的音频分类算法. 该算法首先对音频信号进行信号处理获取特征矩阵，然后送入潜在概率语义模型，再使用 K 近邻分类器进行分类. 实验中，将文中算法应用于 4 类音频文件组成的数据集，并与传统的 KNN 分类算法进行了分类效果对比. 结果表明：文中算法在分类效果方面有较明显的改善. 同时，验证了使用其他分类器(PLSA＋SVM)时的分类效果. 由于验证用的实验对象数量和品种不够，算法本身也有一些局限，整个研究还有进一步提升的空间，未来将继续改进.

参考文献:

[1] RADHAKRISHNAN R, DIVAKARAN A. Systematic acquisition of audio classes for elevator surveillance[C]// SPIE Image and Video Communications and Processing. San Jose:[s. n.],2005:64- 71.

[2] ATREY P K, MADDAGE N C, KANKANHALLI M S. Audio based event detection for multimedia surveillance [C]// International Conference on Acoustics, Speech and Signal Processing. Toulouse:IEEE Press,2006:3-5.

[3] AUCOUTURIER J J. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban sound scapes but not for polyphonic music [J]. Journal of Acoustical Society of America,2007,122(2):881-891.

[4] 冀中. 面向新闻视频内容分析的音频分层分类算法[J]. 计算机应用研究,2009,26(5):1673-1675.

[5] ZENG Zhi,ZHANG Shuwu. A novel approach to musical genre classification using probabilistic latent semantic analysis model[C]// International Conference on Multimedia and Expo. New York:IEEE Press,2009:486- 489.

[6] 容宝华. 基于最小距离的音频分类方法的研究[J]. 电声技术,2012,36(11):46-51,65.

[7] 张宝印. 面向情感的电影背景音乐分类分类方法研究[D]. 武汉:华中科技大学,2011:26.

[8] 石晶,戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展,2007,44(2):242-248.

[9] 李航. 统计学习方法[M]. 北京:清华大学出版社,2013:37-40.

[10] 郭金玲,王文剑. 一种基于数据分布的 SVM 核选择方法统计学习方法[J]. 华侨大学学报(自然科学版),2013,34(5):525-528.

Audio Classification Algorithm Using Probabilistic Latent Semantic Models and K Nearest Neighbor Classifier

XIN Xin[1,3], CHEN Shudong[2,3], TONG Minglei[4], HU Wenhao[1,3], LIU Chenwei[1,3], GE Haodong[3]

(1. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China;
2. Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China;
3. China R&D Center for Internet of Things, Wuxi 214135, China;
4. School of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: The paper proposed an audio classification algorithm based on probabilistic latent semantic analysis model (PLSA) and K-nearest neighbor classifiers (KNN). The algorithm first feed the audio signal feature vector into the PLSA model training to get a bag of sound frames models, then classify with the KNN classifier. Experimental results showed that the proposed classification algorithm has better classification effect compared with the traditional KNN algorithm.

Keywords: Mel frequency cepstral coefficients; word-frequency of co-occurrence matrix; bag of sound frames models; probabilistic latent semantic analysis model; K-nearest neighbor classifiers

(**责任编辑**: 黄晓楠 **英文审校**: 吴逢铁)