基于SVM的音频分类系统设计及实现

孙文静[1] 李士强[2]

(南京审计学院 南京 211815)[1] (南京信息工程大学 南京 210044)[2]

摘 要 分析音频时域特征及提取方法,研究基于支持向量机的语音分类系统流程、分类系统架构以及SVM语音分类器的设计,并进行了相关实验。结果表明,设计的基于SVM的音频分类系统能够有效地对音频进行分类,平均识别准确率达到90%以上。

关键词 音频分类,SVM,MFCC,系统

Design and Implementation of a Audio Classification System Based on SVM

SUN Wen-jing[1] LI Shi-qiang[2]

(Nanjing Audit University, Nanjing 211815, China)[1]

(Nanjing University of Information Science and Technology, Nanjing 210044, China)[2]

Abstract The time-domain feature of audio and its extraction method were analyzed. The research of the process and architecture of a audio classification system based on SVM was made, and the SVM audio classifier was designed. The results of experiments show that the audio classification system based on SVM designed in the paper can classify audio effectively, and the average identification accuracy reaches more than 90%.

Keywords Audio classification, SVM, MFCC, System

基于内容的音频分类与识别技术的研究始于20世纪末,它在远程教学、数字图书馆、新闻节目检索等众多领域具有极大的应用价值。

南京大学卢坚等人在文献[1]中提出了一种基于隐马尔可夫模型的音频分类方法,用于语音、音乐以及它们的混合声音的分类,最优分类精度达到90.28%。

浙江大学赵雪雁等人在文献[2]中提出了基于非监督机制的音频分类检索方法,其直接从压缩域提取音频特征,用基于时空约束的模糊聚类进行特征降维,以加快检索速度,同时使用相关反馈机制提高分类检索的准确率。

Chih-Chieh Cheng等人在文献[3]中采用ellipsoid距离方法对乐器声、男声、女声、环境音等音频类型进行分类,使用的特征有短时能量、过零率、频率质心和频谱带宽等,分别计算各声音类型在这些特征上的均值和标准方差,在特征的选取方法上提出采用优化的对称矩阵来衡量特征的可用度,取得了良好的实验结果,区分环境音准确率达到100%,但对于男声和女声的分类不是很理想,准确率分别只有63%和77%。

Li,S. Z. 等人[4]以美尔倒谱系数MFCC(Mel Frequency Cepstral Coefficients)为特征,建立特征向量,设计实现了基于支持向量机的音频多级分类器,为多级音频分类技术进行了有益的探索。

Erlin Wold等人详细分析了音频的区别性特征,包括响度、音调(pitch)和谐度(harmonicity)等,并且根据最近邻准则(Nearest Neighbor,NN)设计音频的分类器[5],所用的数据集包括笑声、铃声、电话声等16类样本数据。

文献[6]采用最近特征线方法设计分类器对铃声、笑声和水声等进行分类。文献[7]采用相位补偿滤波器组提取音频特征,并用于音频的分割、音乐内容的分析检测等方面。

本文选择短平均过零率、短时能量、频谱质心、子带能量分布、MFCC系数等特征参数,基于SVM(Support Vector Machines)设计音频分类系统。实验表明,其具有令人鼓舞的音频分类能力。

1 音频时域特征提取

(1)短时平均过零率,指单位时间内信号通过零值的次数。对于离散音频信号而言,其实质就是信号采样点符号变化的次数。短时平均过零率可以在一定程度上反映信号的频谱性质,可以通过短时平均过零率获得信号谱特性的一种粗略估计。

短时平均过零率的计算公式如下:

$$Z_n=\frac{1}{2}\sum_{k=-\infty}^{\infty}|\mathrm{sgn}[s(k)]-\mathrm{sgn}[s(k-1)]|w(n-k)$$

$$=\frac{1}{2}\sum_{k=n}^{n+N-1}|\mathrm{sgn}[s_w(k)]-\mathrm{sgn}[s_w(k-1)]| \quad (1)$$

式中,$w(n)$为窗函数,$s_w(k)$表示sgn[·]经过加窗处理后的信号,窗函数的长度为N,sgn[·]是符号函数。

由式(1)可见,短时平均过零率容易受到低频的干扰,如果音频中有反复穿越坐标轴的随机噪声,会使得过零率检测结果虚高。为了提高算法的鲁棒性,此处应设立一个阈值T,将过零率的含义修改为跨过正负阈值的次数:

$$Z_n=\frac{1}{2}\sum_{k=-\infty}^{\infty}\{|\mathrm{sgn}[s(k)-T]-\mathrm{sgn}[s(k-1)-T]|+$$

到稿日期:2010-01-21 返修日期:2010-04-05 本文受南京审计学院资助课题成果资助。

孙文静(1979-),女,博士生,讲师,主要研究方向为信息安全。

$$|\mathrm{sgn}[s(k)+T]-\mathrm{sgn}[s(k-1)+T]|\}w(n-k) \quad (2)$$

这样，即使存在小的随机噪声，其震荡幅度只要不超过正负阈值所构成的区域，就不会产生虚假过零率。

(2) 短时能量。对于音频信号$\{s(n)\}$，短时能量的定义如下：

$$E_n=\sum_{k=-\infty}^{\infty}[s(k)w(n-k)]^2=\sum_{k=-\infty}^{\infty}s^2(k)h(n-k)$$
$$=s^2(n)*h(n) \quad (3)$$

或

$$E_n=\sum_{k=n}^{n+N-1}s_w^2(k) \quad (4)$$

式中，$h(n)=w^2(n)$，$s_w(k)$表示$s(k)$经过加窗处理后的信号，窗函数的长度为N。短时能量可以用来衡量音频信号的强度[8]，并可用于进行有声/无声的判定，该特征对音频信号的检测非常重要。

(3)频谱质心，指一个音频帧的频谱能量分布的平均点，反映了音频信号频率分布的中心，是度量音频亮度(brightness)的指标[8]：

$$SC=\frac{\int_0^{\omega}\omega\,|F(\omega)|^2\mathrm{d}\omega}{E} \quad (5)$$

式中，$\omega=\omega_k$，ω_k为中心频率，E为能量，$|F(\omega)|^2$为功率谱。

(4)子带能量比是用来衡量不同子带的能量占整个频带能量的比例。音乐的子带能量分布较为均匀，而语音的频谱能量主要集中在第一个子带。

各子带能量分布计算如下：

$$D=\frac{1}{E}\int_{I_j}^{H_j}|F(w)^2|\mathrm{d}\omega \quad (6)$$

(5) 美尔倒谱系数(MFCC)，是根据人耳听觉机理导出的声学特征[8]。研究表明，人类对1000Hz以下的声音频率范围的感知遵循近似线性关系，对1000Hz以上的声音频率范围的声音的感知则不遵循线性关系，而遵循对数频率坐标上的近似线性关系。MFCC则是在Mel标度频率域提取出的倒谱参数。

MFCC系数的具体计算过程如下：

① 首先确定每一帧语音采样序列的点数，本文取$N=256$点。对每帧序列$s(n)$进行预加重处理后再经过离散FFT变换，取模的平方得到离散功率谱$S(n)$。

② 计算通过M个$H_m(n)$后所得的功率值$S(n)$，即计算$S(n)$和$H_m(n)$在各离散频率点上的乘积之和，得到M个参数P_m，$m=0,1,\cdots,M-1$。

③ 计算P_m的自然对数，得到L_m，$m=0,1,\cdots,M-1$。

④ 对$L_0,L_1,\cdots,L_{m-1}$计算其离散余弦变换，得到D_m，$m=0,1,\cdots,M-1$。

⑤ 舍去代表直流成分的$L_0,L_1,\cdots,L_{m-1}$，取$L_0,L_1,\cdots,L_{m-1}$作为MFCC参数。此处$K=12$。

2 基于SVM的语音分类系统设计

2.1 分类流程

基于支持向量机的语音分类系统流程如图1所示。首先对原始音频库中的音频数据进行预处理得到音频帧信号，然后对音频帧信号提取包括短时过零率、短时能量、频谱质心、子带能量分布、美尔倒谱系数等帧层次特征，并计算部分帧特征值的统计量，如均值、方差及高过零率比例、低短时能量比率、静音帧比率、平滑基因帧比率等作为音频片段特征，得到完整的特征向量集。将训练样本及测试样本的特征向量集送入支持向量集进行训练和测试，得到满足分类要求的语音分类器。将剩余的样本作为未知样本，进行分类实验。

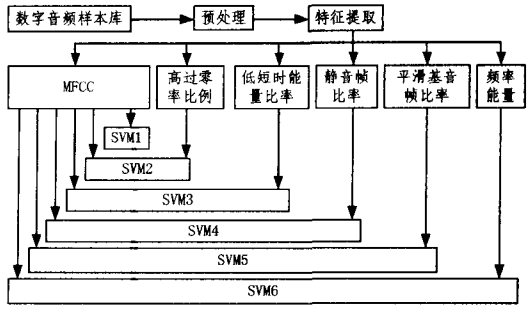

图1 基于支持向量机的语音分类流程图

2.2 分类系统架构

基于支持向量机的分类系统通常由两部分组成，分别为支持向量机分类器训练子系统和分类子系统。其结构如图2、图3所示。

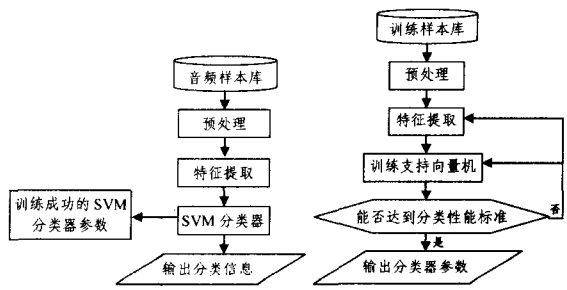

图2 SVM训练子系统结构框图

图3 SVM分类子系统结构框图

训练子系统首先读取音频样本数据，进行预处理和特征提取后，将训练样本特征数据输入支持向量机进行训练，得到一组支持向量机参数。然后运用测试样本对训练好的支持向量机进行测试，利用评价标准进行评价，如果效果达不到预期值，则调整特征向量集的结果和支持向量机参数。重新训练支持向量机，得到新的参数并进行测试和评价，如此循环，直到达到预期要求。

分类子系统首先将未知类别的音频信号进行加窗、分帧、分片段等预处理工作，然后按照训练子系统分类效果较好的特征向量图取待分类样本集的特征，将其送入分类器进行分类，最后将得到的分类结果进行标注并分析结果。

$$\text{分类精度}=\frac{\text{正确分类的音频片段数}}{\text{音频样本中片段总数}} \quad (7)$$

2.3 SVM分类器

SVM分类器工作流程如图4所示。

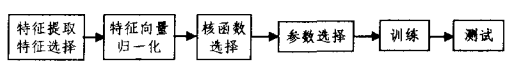

图4 SVM分类器工作流程

(1)特征提取和特征选择模块首先提取出基于音频帧和片段的各类特征，构建相应的特征集。本文选用MFCC的均值和方差作为基本特征集，并且分别在基本特征集的基础上加上部分基于片段的特征组成新的特征向量集，并依次进行训练、测试，寻找最优特征向量集合。

(2)归一化模块对特征数据进行归一化。归一化后的特征数据可以避免向量集中数值过大的某些特征在分类过程中起到决定性作用，还可以减小分类过程中的计算复杂度。

(3)核函数选择模块首选RBF核函数。RBF核函数具有

(下转第223页)

参 考 文 献

[1] Yao Y Y. Interval Sets and Interval-Set Algebras[C]// The 8th IEEE International Conference on cognitive informatics. Hong Kong: IEEE Computer Society, 2009: 307-314

[2] 俞峰,杨成梧. 直觉区间值模糊集的熵、距离测试和相似测试[J]. 计算机科学,2008,35(6):199-201,205

[3] Pomykala J, Pomykala J A. The Stone Algebra of Rough Sets [J]. Bull. Polidh Acad. Sci. Math. ,1988,36:495-508

[4] 薛占熬,何华灿. 粗糙蕴涵[J]. 计算机科学,2003,30(11):18-20

[5] 张小红. 模糊逻辑及其代数分析[M]. 北京:科学出版社,2008,7:122-168

[6] 徐扬. 格蕴涵代数[J]. 西南交通大学学报,1993,28(1):20-26

[7] 朱怡权. 泛蕴涵代数与 FI-代数[J]. 数学杂志,2004,4(24):411-415

[8] 王国俊. MV-代数、BL-代数、R_0-代数与多值逻辑[J]. 模糊系统与数学,2002,16(2):1-15

[9] 刘春辉,徐罗山. 赋范 Fuzzy 蕴涵代数[J]. 扬州大学学报:自然科学报,2009,12(3):1-5

[10] 朱怡权. 格蕴涵代数与 Lukasiewicz 逻辑系统[J]. 内蒙古大学学报,2004,35(2):121-123

[11] 代建云,吴洪博. 分配的 Fuzzy 蕴涵代数[J]. 模糊系统与数学,2008,22(1):26-32

(上接第 210 页)

很多优点:相对于线性核函数而言,RBF 核函数可以有效解决非线性问题;相对于多项式核函数,RBF 核函数的参数较少,因此计算的复杂性较低;另外,与多项式核函数及 Sigmoid 核函数相比,RBF 核函数计算方便。

(4)参数选择子模块主要用来选择合适的聚类中心数 C 和 σ,以提高分类精度。

(5)训练模块将特征向量集输入 SVM 分类器,根据不同的 C 和 σ 分别训练 SVM 分类器,最终得到满足目标的分类器。

(6)测试模块利用训练好的 SVM 分类器测试未分类的音频信号,得到分类结果,并计算正确率。

本文设计的支持向量机分类系统用来区分语音和音乐两类,选择美尔倒谱系数(MFCC)作为特征参数,主要是因为其具有良好的识别性能和抗噪能力。T. Foote[12],Zhang 和 Kuo 等人[14,15]在研究语音与音乐的分类中均采用了美尔倒谱系数,且取得了不错的分类效果。

3 系统实验及分析

实验选择 800 个语音 clip 和 800 个音乐 clip 作为训练集对 SVM 进行训练,剩下的 300 个语音 clip 及 400 个音乐 clip 作为测试集Ⅰ,得到的结果如表 1 和表 3 所列。由于测试集中数目有限,本文又在训练集中随机抽取了 400 个语音 clip 和 400 个音乐 clip 加入到测试集中,形成测试集Ⅱ,得到的实验结果如表 2 和表 4 所列。

表 1 美尔倒谱系数,测试集Ⅰ实验结果

样本		分类结果		正确率(%)
类别	数量	语音	音乐	
语音	300	274	26	91.33%
音乐	400	41	359	89.75%
平均识别率				90.43%

表 2 美尔倒谱系数,测试集Ⅱ实验结果

样本		分类结果		正确率(%)
类别	数量	语音	音乐	
语音	700	657	43	93.86%
音乐	800	54	746	93.25%
平均识别率				93.53%

表 3 美尔倒谱系数+各类其他特征,测试集Ⅰ实验结果

特征	MFCC	MFCC+高过零率比例	MFCC+低短时能量比率
正确率(%)	90.43%	91.29%(+0.86%)	91.86%(+1.43%)
特征	MFCC+静音帧比率	MFCC+平滑基音帧比率	MFCC+频率能量
正确率(%)	89.71%(−0.72%)	89.71%(−0.72%)	92.14%(+1.71%)

表 4 美尔倒谱系数+各类其他特征,测试集Ⅱ实验结果

特征	MFCC	MFCC+高过零率比例	MFCC+低短时能量比率
正确率(%)	93.53%	94.20%(+0.67%)	94.47%(+0.94%)
特征	MFCC+静音帧比率	MFCC+平滑基音帧比率	MFCC+频率能量
正确率(%)	92.87%(−0.66%)	92.53%(−1.00%)	94.47%(+0.94%)

实验结果表明,随着分类样本的增加,分类的正确率有所上升,无论哪种特征的组合,在测试集Ⅱ上的测试正确率都比该类特征组合在测试集Ⅰ上的高,平均高出了 3.155%。在各特征组合实验中,低短时能量及频率能量对结果的提升较为明显。在测试集Ⅰ上,加上低短时能量比率特征后,结果高出了 1.43%,加上频率能量特征后,结果高出了 1.71%。在测试集Ⅱ上,低短时能量比率和频率能量特征都对结果有近 1%的提升。

参 考 文 献

[1] 卢坚,陈毅松,孙正兴. 基于隐马尔可夫模型的音频自动分类[J]. 软件学报,2002,13(8):1593-1597

[2] 赵雪雁,吴飞,刘骏伟. 基于模糊聚类表征的音频例子检索及相关反馈[J]. 浙江大学学报,2003,37(3):264-268

[3] Cheng Chih-chieh, Hsu Chiou-ting. Content-Based Audio Classification with Generalized Ellipsoid Distance[C]// Proc. PCM. Berlin Heidelberg: Springer-Verlag, Dec. 2002: 328-335

[4] Li S Z, Guo Guo-dong. Content-Based audio classification and retrieval using SVM learning[C]// Proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia. Sydney, Australia, 2000: 156-163

[5] Wold E, Blum T, Keislar D, et al. Content-Based classification, search and retrieval of audio[J]. IEEE Multimedia Magazine, 1996, 3(3): 27-36

[6] Tsekeridou S, Pitas I. Content-based video parsing and indexing based on audio-visual interaction[J]. IEEE Transactions on Circuits and Systems For Video Technology, 2001, 11(4): 522-535

[7] Li S Z. Content-Based classification and retrieval of audio using the nearest feature line method[J]. IEEE Transactions on Speech and Audio Processing, 2000, 8(5): 619-625

[8] 韩纪庆,等. 音频信息处理技术[M]. 北京:清华大学出版社,2007