基于 APR-SVM 的音频分类方法

王晓峰,蒋先涛
(上海海事大学 信息工程学院,上海 201306)

摘 要:音频分类在多媒体应用中十分广泛,主要有时域分析和频域分析方法。文中提出了一种基于自适应间距比(APR)算法和支持向量机(SVM)算法的音频分类方法,先用 APR 算法区分语音与非语音;对于非语音,再通过 SVM 进行音频分类。APR 算法是比较 PR 参数和阈值来区分语音和非语音,它和信噪比密切相关;而将非语音分成四组:音乐,汽车,会议,雨声,提取特征因子。实验结果表明:文中设计的分类器的精度达到 93.75% 以上,能很好地把各类型音频分开。

关键词:音频分类;特征提取;支持向量机;自适应间距比;信噪比
中图分类号:TP315 **文献标识码**:A **文章编号**:1673-629X(2012)10-0059-03

Audio Classification Based on APR-SVM

WANG Xiao-feng, JIANG Xian-tao
(Information Engineering College, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Audio classification is widely applied in multimedia applications, which mainly has time domain analysis and frequency domain analysis methods. In this paper, an audio classification method based on APR algorithm and SVM algorithm is proposed, first use the APR algorithm to distinguish between voice and non voice, for non-voice take audio classification by SVM. APR algorithm is to compare the PR parameters and thresholds to distinguish between voice and non voice, is closely related to SNR, and non-voice is divided into four groups: music, cars, meeting, rain, extract the feature factor. The experimental results show that: the accuracy of the classifier designed in this paper is to reach over 93.75%, good separation of various types of audio.

Key words: audio classification; feature extraction; SVM; adaptive pitch ratio; SNR

0 引 言

音频分类在理论研究和实际应用中都很重要,特别是在多媒体、视频监控等方面。音频具有无结构的特点,所以在语义内容和内容的基础上,从二进制流中提取音频的结构特征。一般来说,音频分类是一个模式识别问题,由两部分组成:特征提取和基于特征的分类[1]。

目前国内外主要的音频分类方法有:

1)基于支持向量机(SVM)的分类[2~4];

2)基于小波变换和 SVM 的分类;

3)基于 HMM 和 SVM 的音频分类;

4)基于听觉模型的分类方法。

由上面的研究方法可见,虽然有多种分类,基于支持向量机(SVM)分类的性能在几种分类算法中最佳。由于 SVM 是最初开发的二元分类,其多类扩展仍然是一个重要的研究课题。因此,支持向量机算法用于分类不同的音频的文件,已经对算法做了进一步的改进和研究。

文中的主要思路是:首先区分语音与非语音,对于非语音,再通过 SVM 进行音频分类。区分语音与非语音,用自适应间距比(APR)算法[5]。

1 音频特征分析

音频特征能反应音频频域和时域的特性[6],在提取特征前,需对音频数据做预处理。

1)基于帧的音频特征。

频域能量(frequency energy):频域能量也是区分音乐和静音的有效特征。定义如公式(1):

$$E = \log\left(\int_0^{w_0} | F(w) |^2 \mathrm{d}w\right) \tag{1}$$

式中 $F(w)$ 是该帧的 FFT 的变换系数[7],w_0 是采样频率的一半。

频率中心:频率中心是度量音频亮度的指标。定义如公式(2):

收稿日期:2012-01-31;**修回日期**:2012-05-03
基金项目:上海市科技计划重点项目(08240510800)
作者简介:王晓峰(1958-),男,工学博士,教授,博士生导师,研究方向为人工智能及其在交通信息与控制工程中的应用、数据挖掘与知识发现;蒋先涛(1983-),男,湖北孝感人,硕士研究生,研究方向为港行和物流信息管理、数据挖掘、嵌入式系统。

$$FC = \frac{\int_0^{w_0} w \mid F(w) \mid^2 \mathrm{d}w}{\int_0^{w_0} \mid F(w) \mid^2 \mathrm{d}w} \tag{2}$$

带宽(band width):带宽是衡量音频频域范围的指标。定义如公式(3):

$$BW = \sqrt{\frac{\int_0^{w_0} (wFC)^2 \mid F(w) \mid^2 \mathrm{d}w}{\int_0^{w_0} \mid F(w) \mid^2 \mathrm{d}w}} \tag{3}$$

过零率:可以用来表示信号频率量的度量。如公式(4) 所示:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \mid \mathrm{sgn}[x(m+1)] - \mathrm{sgn}[x(m)] \mid \tag{4}$$

$x(m)$ 为离散音频信号。

MFCC 和ΔMFCC:其中 MFCC 是频率的倒谱系数,ΔMFCC 为 MFCC 的差分系数[8]。

2) 基于片段的音频特征。

子带能量比(sub - band energy ratio):频域为 4 个子带区间 $[0, w/16]$, $[w/16, w/8]$, $[w/8, w/4]$ 和 $[w/4, w]$,其子带能量 $SW_i = \int_{w \in sb_i} \mid F(w) \mid^2 \mathrm{d}w$,则子带能量比为 $SWR_i = SW_i / E$。

频谱流量:频谱流量定义为一个片段中相邻两帧之间频谱变化量的均值。

和谐度:某一 clip 中基音频率不等于 0 的帧数所占的比例。

2 APR 算法

APR 算法把音频段分成帧[9],统计语音在每帧中的间距,得到一个 *PR* 参数。如果 *PR* 参数大于确定的阈值,则是语音频段,否则是非语音频段。

输入音频片段正常,低通滤波截止频率为 900Hz,每个片段分成帧,设帧的大小 *FS* 定义为:

$20\mathrm{ms} \leqslant FS \leqslant SD$

其中 *SD* 为片段的持续时间。

人的音高范围一般在 70Hz ~ 280Hz 之间。用计数器去统计在每个片段中语音的帧数,则 *PR* 可以定义为片段中语音帧的数目与总帧数的比率。

$$PR = \frac{NP}{NF} (0 \leqslant PR \leqslant 1)$$

其中 *NP* 是含有语音帧的数目,*NF* 是帧的总数。

总的帧数可以通过下面的公式计算:

$$NF = \left(\frac{SD - FS}{1 - OR}\right) / FS + 1$$

其中 *SD* 是片段的持续时间,*FS* 是帧的大小,*OR* 是重叠比率。

为了辨别非语音片段,阈值用 *PR* 参数。大于 *PR* 参数的则为语音,低于 *PR* 参数的为非语音。当信噪比大于 20dB 的片段,*PR* 阈值是 *FS*,*OR* 和 *SD* 的函数。当信噪比小于 20dB 时,*PR* 是 SNR 的函数。

A. 信噪比 SNR。

信噪比是正常声音信号强度与噪声信号强度的比值。

假设噪音具有高斯分布,其噪音的峰度等于 0,则音频能量值 $E_{\mathrm{NoisySpeech}}$ 可以求方差得到,噪音的能量值 E_{Noise} 通过下面公式计算得到:

$$E_{\mathrm{Noise}} = E_{\mathrm{NoisySpeech}} - E_{\mathrm{Speech}}$$

$$\mathrm{SNR} = \frac{E_{\mathrm{Speech}}}{E_{\mathrm{Noise}}}$$

B. ARP 算法流程。

1. 先根据输入音频片段计算信噪比 SNR,确定阈值;

2. 分解片段成帧,计算总共的帧数;

3. 对于每帧,先判断是否是基音检测,若是进行下一步,否则该帧语音 pitch 记为 0;

4. 再判断语音 pitch 的范围是否在 70 ~ 280Hz 之间,若是帧语音 pitch 记为 1,否则记为 0;

5. 最后计算 *PR* 参数;

6. 比较 *PR* 参数和阈值大小,若 *PR* 大于阈值则为语音,否则为非语音。

3 基于 SVM 算法

支持向量机远远超过其他传统的非参数的有效分类[10](例如,RBF 神经网络,近邻(NN),最近中心(NC)分类)的分类精度、计算时间、稳定的条件参数设置。SVM 使用已知的内核函数定义超平面以分成两个预定义的已知点类。

A. SVM 算法原理[6,11]。

假设 $x_i \in X \subseteq R^n$, $y_i \in Y = \{-1, 1\}$ 作为输入向量和目标变量,集合 $S = \{(x_1, y_1), \cdots, (x_l, y_l)\}_{i=1}^{l} \subseteq (X \times Y)^l$ 和核函数 $K(x_i, y_j) = < \varnothing(x_i), \varnothing(y_j) >$ 给定,它将输入空间 X 映射到另一个高维特征空间 F 上,对于给定的 $\varnothing$,非线性问题 S 转化为线性可分问题 F。

许多超平面可以实现上述分离的目的,必须找到正样本与负样本到超平面的最小距离。超平面记为 $(w, b) \in R^n \times R$,解空间满足 $< w, x > + b = 0$。问题可归结为二次型问题,模型为:

$$\text{Minimize} \quad \varnothing(w, b) = \frac{1}{2} \| w \|^2$$

$$y_i(w * x + b) - 1 \geqslant 0, i = 1, 2, \cdots$$

可以用拉格朗日函数解决支持向量机的问题,常

数 C 是它的上界,拉格朗日函数 α_i,则上面的公式可转化为:

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j < x_i, y_j >$$

其中 $\sum_{i=1}^{l} \alpha_i y_i = 0$ 和 $0 \leqslant \alpha_i \leqslant C$。

支持向量机通过核函数把数据从低维空间映射到高维空间中,在高维空间为低维空间数据构造线性可分的超平面。一般而言主要有三种核函数可以选择:

1) ERBF 核函数: $K(x, \bar{x}) = \exp\{ - \frac{| x - \bar{x} |}{2\sigma^2} \}$;

2) Gaussion 核函数: $K(x, \bar{x}) = \exp(- \frac{| x - \bar{x} |^2}{2\sigma^2})$;

3) polynomial 核函数:

$K(x, \bar{x}) = \exp (< x, \bar{x} > + 1)^d$ 。

B. 基于 SVM 的分类器设计。

需要将环境音为汽车声音、音乐、机械设备声音、下雨声四种类型。

首先将音频信号分帧处理,再对每一帧分别提取其特征:将基音频率,频域能量,频率中心,带宽,过零率,静音比例,子带能量比,高 ZCR 比率,频谱流量,和谐度,6 维音频特征及 12 维的 MFCC+ΔMFCC 音频特征构建特征向量集作为输入。

接下来分别训练 SVM 对音频进行分类:

SVM1 区分汽车声音,音乐,SVM2 区分音乐,机械设备声音,SVM3 区分机械设备声音,下雨声[12]。

4　实验数据及结果分析

音频数据可以通过实际场景采集得到,音频数据的采样频率为 22KHz,精度为 16 位,音频数据集中 2/3 用于训练,1/3 用于测试,每个片段 200ms,每个片段分成 20ms 的帧。数据总量为 256MB,总共 300min,其中语音 75min,汽车声 45min,音乐 60min,会议 50min,下雨声 70min。

首先区分语音和非语音[13],再把训练音频分割成音乐、汽车、会议和雨声 4 部分,最后使用 4 类音频训练 SVM 分类器。在测试阶段记录正确分类的片段和错误的分类片段,并计算分类的精确度。

分类的准确率采用分类精度来衡量:

分类精度=分类正确样本片段/样本总片段

基于 APR 和 SVM 的音频分类结果如表 1。

从分类的结果看:语音准确率为 93.75%,汽车的准确率为 97.20%,音乐的准确率为 93.45%,会议模式的准确率为 74.35%,下雨声的准确率为 90.90%。相对文献[13]采用 HMM 和 SVM 的分类器的结果:语音 90.01%,音乐 96.41%,说明文中采用的 APR 和 SVM 分类器能很好地实现音频的分类。

表 1　基于 APR 和 SVM 的音频分类结果

音频类型	测试片段	测试片段	正确分类片段	误分的片段	分类准确率/%
语音	255	128	120	8	93.75
音乐	268	143	139	4	97.20
汽车	327	168	157	11	93.45
会议	152	78	58	20	74.35
雨声	89	55	50	5	90.90

5　结束语

音频分类技术随着技术的发展会越来越多地受到人们的重视,它是提取音频信息和内容语义的基础,在生活中各个领域都有应用。

文中先用 APR 算法区分语音与非语音;对于非语音,再通过 SVM 进行音频分类,有效地提高了分类的精度。但非语音的分组还不够多,以后可以继续添加分组。

参考文献:

[1] Lin C C, Chen S H, Truong T K. Audio Classification and Categorization Based on Wavelets and Support Vector Machine [J]. IEEE Transaction on Speech and Audio Processing, 2005, 13(5): 644-651.

[2] Tran H D, Li Haizhou. Jump Function Kolmogorov for Audio Classification in Noise-Mismatch Conditions[J]. IEEE transactions on signal processing, 2009, 57(8): 2908-2918.

[3] Wu Chung-Hsien, Hsieh Chia-Hsin. Multiple change point audio segmentation and classification using an MDL based Gaussian model[J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(2): 647-657.

[4] Ghaemmaghami S. Audio segmentation and classification based on a selective analysis scheme [C]//IEEE Multimedia Modeling Conference. [s. l.]: [s. n.], 2004: 42-48.

[5] Ghoraani B, Krishnan S. Time-frequency Matrix Feature Extraction and Classification of Environmental Audio Signals [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2197-2209.

[6] Kiranyaz S, Qureshi A F, Gabbouj M. A generic audio classification and segmentation approach for multimedia indexing and retrieval[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(3): 1062-1081.

[7] 白　亮, 老松杨, 陈剑赟, 等. 音频自动分类中的特征分析和抽取[J]. 小型微型计算机系统, 2005, 26(11): 2028-2034.

[8] 史东承, 韩玲艳. 基于 HMM/SVM 的音频自动分类[J]. 长春工业大学学报, 2008, 29(2): 177-182.

(下转第 65 页)

Memory:2GB

2.4.2　实验结果

测试平台和测试用例来自于 2010RoboCup 中国公开赛(鄂尔多斯)的服务机器人仿真项目的比赛题库。给出 5 个家庭环境场景描述,每一个场景 8 个复合任务。要求在 5 秒内规划出服务机器人的动作序列,并根据比赛规则对动作序列评分。实验分两个部分,RobotA 没有对场景信息进行预处理,RobotB 对场景信息进行了预处理。

表 1 中,从任务完成情况来看,RobotB 比 RobotA 多完成一个,但都接近全部完成,在可接受响应时间上还是可以接受的。在时间上看,RobotB 规划每一个任务所用的平均时间比 RobotA 少了 1.238 秒,在可接受响应时间为 5 秒的情况下,这还是提高了较大的效率。最后的规划结果评分,RobotB 也比 RobotA 多。因此可以得出,对场景信息的预处理能较大地提高服务机器人的规划效率。

表 1　测试结果

	完成任务数(个)	完成任务的平均时间(S/个)	得分
RobotA	38	3.327	963
RobotB	39	2.089	987

3　结束语

实验证明,ASP 能规划出服务机器人的最优行动序列。其优点是方式灵活、方法可靠、程序可扩展性好。当场景中物品描述形式或机器人原子动作有所变化,只用相应地修改 ASP 规则即可。其目前最大的瓶颈是效率问题,试验中,在可接受响应时间内会有规划不出动作序列的情况发生。ASP 的常例化过程相当于是穷举,也就是相当于求解所有行动序列的可能组合,ASP 求解器通常需要很长时间才能求解出大规模或者需要行动步数太多的问题。因此,如何提高 ASP 的求解效率将是一个未来的热点研究。

参考文献:

[1] 田国会,李晓磊,赵守鹏,等. 家庭服务机器人智能空间技术研究与进展[J]. 山东大学学报(工学版),2007,37(5):53-59.

[2] 田国会. 家庭服务机器人研究前景广阔[J]. 国际学术动态,2007(1):28-29.

[3] 吉建民. 提高 ASP 效率的若干途径及服务机器人上应用[D]. 合肥:中国科学技术大学,2010.

[4] 周北海. 模态逻辑导论[M]. 北京:北京大学出版社,1996.

[5] 蔡自兴,徐光祐. 人工智能及其应用[M]. 第 3 版. 北京:清华大学出版社,2004:72-78.

[6] Gelfond M, Lifschitz V. The stable model semantics for logic programming[C]//Proceedings of the 5th International Conference on Logic Programming (ICLP-88). Seattle, Washington: MIT Press, 1988:1070-1080.

[7] Simons P, Niemela I, Soininen T. Extending and implementing the stable model semantics[J]. Artif. Intell., 2002, 138(1-2):181-234.

[8] Gelfond M. Answer sets[M]//Handbook of Knowledge Representation. [s. l.]: Elsevier, 2007:285-316.

[9] Gebser M, Kaminski R, Kaufmann B, et al. Engineering an Incremental ASP Solver[M]. [s. l.]: [s. n.], 2008.

[10] Gebser M, Kaminski R, Kaufmann B, et al. A User's guide to gringo, clasp, clingo and iclingo[M]. [s. l.]: [s. n.], 2008.

[11] 吕勇全,陈　寅,邬家炜,等. 基于回答集程序的排课系统设计与实现[J]. 计算机技术与发展,2010,20(6):228-232.

[12] 赵岭忠,王雪松,钱俊彦,等. 从经典逻辑知识构建 ASP 知识库的新方法[J]. 计算机应用,2010,30(11):2932-2936.

(上接第 61 页)

[9] Briggs F, Raich R, Fern X Z. Audio Classification of Bird Species: A Statistical Manifold Approach[C]//IEEE International Conference on Data Mining. [s. l.]: [s. n.], 2009:51-60.

[10] Zhang J X, Brooks S. Audio classification based on adaptive partitioning[C]//IEEE International Conference on Multimedia and Expo.. [s. l.]: [s. n.], 2009:490-493.

[11] Chu Wei, Champagne B. A Noise-robust FFT-based Auditory Spectrum with Application in Audio Classification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(1):137-150.

[12] Shiazi J, Ghaemmaghami S. Audio classification based on sinusoidal model: a newfeature[C]//TENCON 2008. [s. l.]: [s. n.], 2008:1-5.

[13] Chen Lei, Gunduz S, Ozsu M T. Mixed type audio classification with support vector machine[C]//IEEE international conference on multimedia and expo. [s. l.]: [s. n.], 2006:781-784.

基于APR-SVM的音频分类方法

作者： 王晓峰， 蒋先涛
作者单位： 上海海事大学 信息工程学院,上海 201306
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(10)