基于熵和支持向量机的音乐分类方法①

高林杰, 张 明

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 音乐分类研究已经持续多年, 但目前检索效率并不理想. 提出了一种基于熵和支持向量机的音乐分类方法. 利用滤波器把音乐片段分解成不同的频率通道, 然后通过离散傅里叶变换转换为频谱图后计算信息熵, 并使用支持向量机在四个类别的音乐集上进行训练和测试. 同时, 比较了三种不同的滤波器, 其中 Bark 滤波取得了80%的识别率, 实验结果表明其比使用 MFCC 特征分类效果要好.

关键词: 音乐分类; 滤波器; 频谱图; 信息熵; 支持向量机

Music Classification Method Based on Entropy and Support Vector Machine

GAO Lin-Jie, ZHANG Ming

(Department of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Research on music classification has been processing years, but the performance of each method is not very well. This paper proposes a new method based on entropy and support vector machine for music classification. It uses bank of filters to decompose the music clip into different channels. Then the filters turns it into spectrum through discrete Fourier transform and compute the information entropy and uses support vector machine training and testing on a dataset containing four categories of music. The experiment compares three different kinds of filters, among which the Bark filter achieves an accuracy of 80%. The result shows that the proposed feature vector is better than MFCC.

Key words: music classification; filters; spectrum; entropy of information; support vector machine

1 引言

互联网的发展促进了多媒体的推广与传播, 而音乐作为主要的多媒体形式之一, 其数量得到了迅猛的增加. 如何有效的管理数量繁多的音乐, 从而为用户提供更好的服务, 是许多唱片公司和音乐网站考虑的主要问题之一. 分类是管理数量庞大的音乐数据的必要策略, 如果能按某种标签将各种音乐分门别类, 不但可以促进大型音乐数据库的管理, 而且可以让用户在检索某一类音乐时方便很多.

音乐分类是多媒体信息检索(Multimedia Information Retrieval, MIR)领域的一个热门方向, 一方面研究者不断提出新的音频特征, 另一方面分类算法也在持续的发展. 目前大多数音乐数据库都是根据音乐的某一属性来组织, 例如标题、日期、专辑、艺术家等等, 这些信息都是歌曲的元数据, 并不能很好的反映歌曲的内容和风格. 近些年国外开始研究音乐的流派分类, 比如从 2004 年起每年举办的 MIREX 大赛就是以音乐分类为主要内容的. 对于音乐分类的标准研究也比较多, 目前主要有基于流派分类(Genre Classi?cation)、基于情感分类(Mood Classi?cation)、基于艺术家分类(Artist Identi?cation)、基于乐器分类(Instrument Recognition)、基于音乐注释(Music Annotation)等. 基于流派的音乐分类是使用最广泛的分类方式. G. Tzanetakis and P. Cook(2002) [1]率先对音乐进行流派划分, 根据三种音乐特征(音色、旋律、高音)将音乐分为 10 种流派: 古典(classical)、乡村(country)、迪斯科(disco)、嘻哈(hiphop)、爵士(jazz)、摇滚(rock)、布鲁斯(blues)、雷鬼(reggae)、流行(pop)、金属(metal). Costa Y M G, Oliveira L S, Koerich A L, et al(2012)[2]提出了一种对音乐进行流派划分的方法, 通

① 收稿时间:2013-09-08;收到修改稿时间:2013-10-28

过将音频信号转换为频谱图并抽取文本特征进行流派划分，实验证明取得了很好的效果. Poria S, Gelbukh A, Hussain A, et al(2013)[3]提出了一个音乐流派分类器，通过乐器、高音、旋律、和声等特征对音乐分类，对随机采集到的音乐取得了 97.1%的分类精度.

然而对国内众多用户而言，“摇滚”“爵士”这些描述词语并不常用. 通常在查询音乐的时候检索关键字多为“伤感音乐”“轻音乐”“民族乐器”等大众化的描述词语，而这些词语都是人工标注上去的，效率有限，主观性较强，且耗时耗力. 如若能按照常用标签进行自动音乐分类，不仅能极大地减少分类操作人员的工作量，由于主观因素的减少还能提高用户的音乐体验. 张燕，唐振民，李燕萍等人(2008)[4]采用基于 Mel 倒谱系数特征的隐马尔可夫模型对音乐进行分类，实验表明该方法具有更好的抗干扰能力和正确率.

本文在综合分析国外现存优秀音乐分类方法的基础上，主要考虑本国的音乐特点进行分类. 目前国内网络音乐的格式以 MP3 为主，但是研究却很少. 胡景凯，吴磊，高阳(2007)[5]提出一种基于学习分类器的 MP3 音乐分类方法，张家发，胡景凯，高阳等人(2007)[6]提出一种基于 MDCT 域特征的 MP3 音乐分类. 本文的研究对象选择具有代表性的 MP3 格式音乐. 本文首先利用滤波器把分割后的 MP3 音乐片段分解成不同的通道，然后通过离散傅里叶变换(DFT)将各通道转换成频谱图，对各个频谱图计算信息熵，得到最终的特征向量. 实验阶段使用支持向量机(Support Vector Machine, SVM)进行分类，支持向量机作为统计学习理论的一项重大研究成果，其解决非线性和小样本的模式识别问题时有很多的优势，在分类、回归等方面得到了很好的应用，因此本文使用 SVM 进行音乐分类. 实验结果证明本文的分类方法能取得很好的分类效果.

2　音乐数据的特征提取

进行分类前先对 MP3 文件进行特征提取. 本文将 MP3 音乐数据集分为四个类别: 中国风、伤感音乐、笛子和古筝，这四个标签是国内普通大众检索音乐的常见词语，具有代表性和典型性. 每个类别都选取了 50 首歌曲进行特征提取.

考虑到选取的歌曲是整首?MP3 音乐，并且 MP3 的采样率一般是 44.1KHz，如直接处理不但计算量大，也可能会影响最后的分类结果. 因此本文首先对每首歌曲进行了简单的预处理，代表性的截取了歌曲的开始、中间和结尾三个部分各?30 秒歌曲片段，分别可以作为一首歌的序曲、高潮和结束曲，而这三个部分最能体现一首歌的曲风，并且每个片段的采样率都降低到 22050Hz. 见图 1.

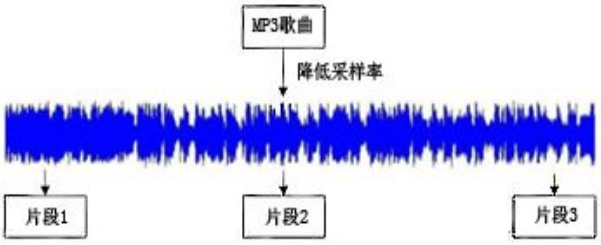


图 1　歌曲预处理

每个音频片段都是一段数字信号，近似可看成是由无限个正弦波叠加而成，根据信号处理理论，本文首先使用带通滤波器，只允许一定频率范围的信号通过，将音乐片段分割成不同频率范围的通道，如图 2. 这样的处理是考虑到在一首歌曲中，不同的乐器、人声，其频率都各不相同，反映在听觉上就是音调的高低.

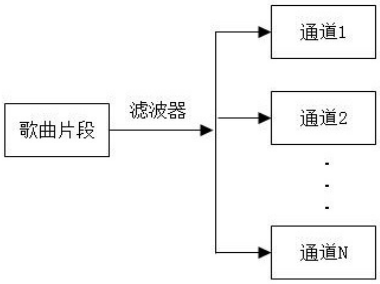
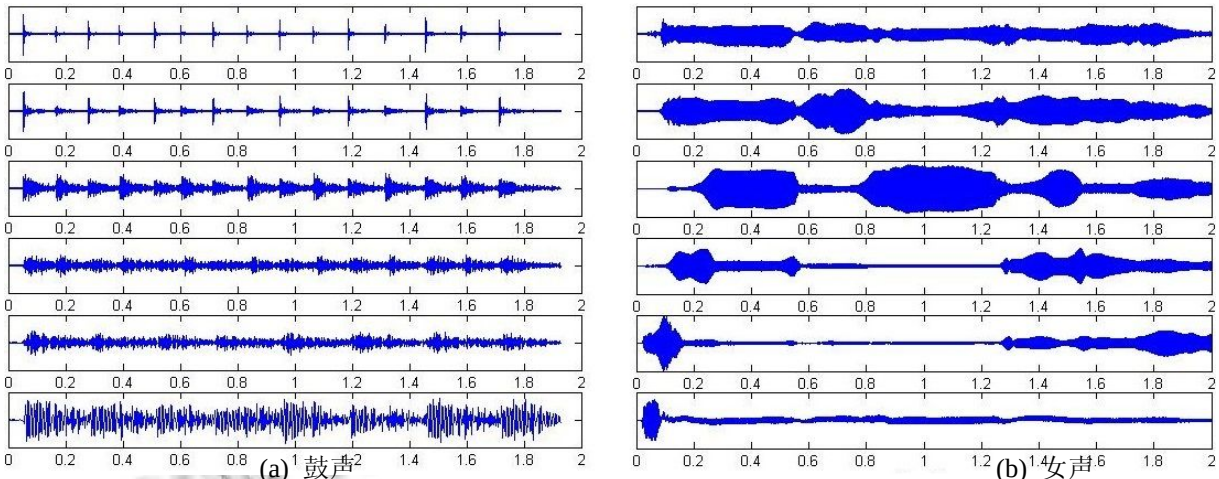


图 2　滤波处理

为了直观的反映这一特征，本文选取了两种高低频率的音乐: 鼓声和女声. 为了比较不同的滤波器对最后的分类精度的影响，本文选择了三种滤波器处理音频片段，分别是线性滤波器、Bark 滤波器和 Scheirer 滤波器. 线性滤波器易于实现且应用很普遍，Bark 滤波器和 Scheirer 滤波器能较好的模拟人耳感知声音的特性，在音频处理方面有广泛的应用. 图 3 是经过 Scheirer 滤波后的两种声音的波形，从下到上频率逐渐升高. 鼓声的音调偏低，声音低沉，其在分解后主要集中的低频部分；而女声的音调偏高，声音尖锐，其在分解后主要集在高频部分. 这样通过滤波器的分

解，可将不同音调的声音分割在不同的通道里，便于分开研究. 三种滤波器的频率带区间如表 1-表 3 所示，其中线性滤波、Bark 滤波和 Scheirer 滤波的通道数分别为 10、23、6. 由于采样率为 22050Hz，所以每个滤波器的最后一个区间的上限为 11025Hz.



(a) 鼓声 (b) 女声

图 3 滤波后的波形

表 1 线性滤波

通道	1	2	3	4	5	6
区间	0-1100	1100-2200	2200-3300	3300-4400	4400-5500	5500-6600
7	8	9	10			
6600-7700	7700-8800	8800-9900	9900-11050			

表 2 Bark 滤波[7]

通道	1	2	3	4	5	6	7
区间	0-100	100-200	200-300	300-400	400-510	510-630	630-770
8	9	10	11	12	13	14	15
770-920	920-1080	1080-1270	1270-1480	1480-1720	1720-2000	2000-2320	2320-2700
16	17	18	19	20	21	22	23
2700-3150	3150-3700	3700-4400	4400-5300	5300-6400	6400-7700	7700-9500	9500-11025

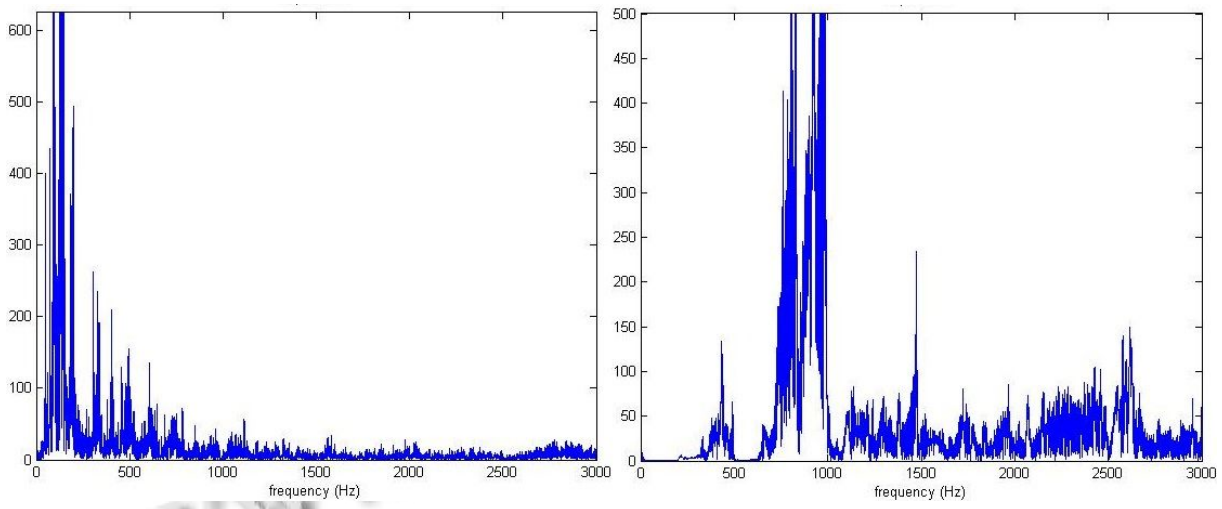
表 3 Scheirer 滤波[8]

通道	1	2	3	4	5	6
区间	0-200	200-400	400-800	800-1600	1600-3200	3200-11025

然后对每个通道按照公式(1)进行离散傅里叶变换，计算每个通道的频谱图，将时域信号转化为频率上能量的变化. 离散傅里叶变换是多媒体处理的常用方法，是连续傅里叶变换在时域和频域上都离散的形式，将时域信号的采样变换为在离散时间傅里叶变换频域的采样. (1)式中 $x(n)$ 是波形序列，N 是序列长度，j 是虚数单位.

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N}\quad k=0,\ldots,N-1 \qquad (1)$$

对上述两种未经滤波的音乐实行 DFT 变换，结果如图 4. 可以看出，鼓声的能量主要集中在 0-500Hz 之间，而女声的能量主要集中在 500-1000Hz 之间，分布在不同的通道中.



(a) 鼓声　　(b) 女声

图 4　频谱图

最后根据公式(2)计算各通道频谱图对应的信息熵. 信息熵是由 Shannon 从热力学中引入来的[9]，用来衡量一个随机变量出现的期望值. 为了排除序列长度的影响，得到的信息熵还要除以序列的长度，作为最终的特征向量.

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \qquad (2)$$

本文使用 MIRtoolbox[10]来帮助实现特征提取过程. 通过以上的处理，我们从每首歌曲中可以得到 3N(N 是通道的个数)维的特征向量，用于训练和测试支持向量机. 此外还提取了 13 维的 MFCC 特征，然后测试数据集作为对比.

2　支持向量机的使用

支持向量机在解决非线性和小样本的模式识别问题时有很多的优势，是由 Cortes 和 Vapnik[11]提出来的，其基本思想如图 5 所示. H 是两类样本(实心点和空心点表示)的分界线，H_1、H_2 分别是过各类中离分类线最近的样本且平行于分类线的直线. 支持向量机的目的就是找出这样的最优分界线，不但能准确的分开两类样本，而且使 H_1、H_2 之间的距离最大. 前者保证经验风险最小(如使训练误差为 0)，后者是使推广性的界中的置信范围最小，从而使真实风险最小. 在高维空间，最优分界线就变成了最优分类曲面.

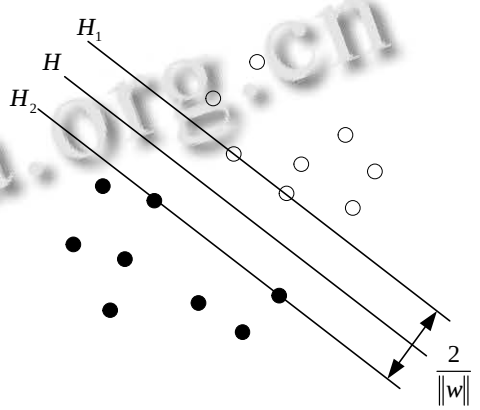


图 5　最优分类线

LIBSVM[12]是台湾大学林智仁博士开发设计的一个易于使用的模式识别软件包，本文使用该软件包进行了音乐分类的相关实验. 为了使分类结果更有说服性，本文采用了分类方法中比较常用的 k 次交叉验证，这样可避免数据集的分布不均导致结果的过度拟合. 具体过程如下:

(1) 数据预处理，把提取出的数据转换为LIBSVM要求的格式:

<label> <index1>:<value1> <index2>:<value2> ...

其中 label 表示样本的类别标签，用 1, 2, …等标注即可; index 是特征向量的维数序号; value 是具体的特征向量值. 图 6 是一部分经过预处理的数据:

```
1   1:6.940669e-001 2:7.108526e-001 3:7.518428e-001 4:8.107229e-001 5:8.559350e-001 6:9.635586e-001 7:6.924620e-001
1   1:7.203817e-001 2:7.275576e-001 3:7.831755e-001 4:8.024088e-001 5:8.445212e-001 6:9.671610e-001 7:6.860338e-001
1   1:9.112664e-001 2:8.534828e-001 3:7.367006e-001 4:7.897271e-001 5:8.372955e-001 6:9.606146e-001 7:7.095373e-001
2   1:6.816589e-001 2:6.776122e-001 3:7.170629e-001 4:7.954388e-001 5:8.463550e-001 6:9.525864e-001 7:7.150956e-001
2   1:7.171628e-001 2:6.882027e-001 3:7.220489e-001 4:7.850114e-001 5:8.550313e-001 6:9.677700e-001 7:6.920255e-001
2   1:6.935315e-001 2:6.856607e-001 3:7.425599e-001 4:7.995660e-001 5:8.513662e-001 6:9.621069e-001 7:6.916545e-001
```

图 6 部分预处理后的数据

(2) 采用交叉验证，将数据集平均分成 5 份，轮流将其中的 1 份作为测试集，而剩下的 4 份作为训练集，并使用网格搜索寻找最优的惩罚因子和核参数，其中 C 和 γ 的搜索区间分别为 $2^{-5}-2^{15}$、$2^{-15}-2^{3}$.

(3) 使用径向基函数(RBF)作为核函数，使用上一步得到的 C 和 γ 训练支持向量机. RBF 函数具有良好的泛化能力，并有很快的学习收敛速度，所以经常用来做核函数.

(4) 得到支持向量机模型后，用第 1 章中提取的测试集在上一步训练好的模型上进行测试，观察结果.

3 实验结果与分析

表 4 显示了在不同滤波器下，最终的 5 次交叉验证识别率. 可以看出，线性的滤波的识别率最差，而 Bark 最好，达到了 80%.

表 4 不同滤波下分类器识别率

滤波器	线性	Bark	Scheirer
识别率	69.5%	80%	73%

表 5 不同滤波下不同类别歌曲的识别率

	线性	Bark	Scheirer
中国风	82%	82%	78%
伤感音乐	66%	70%	72%
笛子	70%	90%	76%
古筝	60%	78%	66%

另外，表 5 比较了不同滤波器对不同类别的歌曲识别率的影响. 其中 Bark 滤波使笛子乐曲的识别率达到了 90%. 对同一滤波器而言，中国风和笛子的识别率都比伤感音乐要高，其原因是这两个类别的曲目中都有类似笛子等曲调婉转的传统乐器的演奏，音调分明，有别于现代伤感音乐中数目繁多的电子乐器. 对同一类别而言，Bark 滤波总体比线性和 Scherier 滤波要好.

根据 MFCC 分类是目前音乐分类的常用方法[1][4]，本文也通过对测试集提取 MFCC 进行了实验. 表 6 是提取 MFCC 特征分类后的分类精度. 对比 Bark 滤波后提出的信息熵特征，可以看到不管从整体识别率还是单一类别的识别率，本文提出的方法得出的结果都比 MFCC 特征要好. 不过 MFCC 特征对笛子乐曲有较高的识别率，这也为进一步的研究工作提供了一定启示.

表 6 MFCC 特征的识别率

中国风	伤感音乐	笛子	古筝	平均
60%	68%	92%	60%	71%

4 结论

本文探讨了一种基于熵和支持向量机的音乐分类方法，首先利用滤波器把音乐片段分解成不同的频率通道，然后通过离散傅里叶变换转换为频谱图后，计算信息熵，再使用支持向量机在四个类别的音乐集上进行训练和测试. 同时，对三种不同的滤波器进行了比较. 实验结果表明本文所提出的方法优于 MFCC 特征分类方法.

参考文献

1 Tzanetakis G, Cook P. Musical genre classi cation of audio signals. IEEE Trans. on Speech Audio Process., 2002, 10(5): 293–302.

2 Costa YMG, Oliveira LS, Koerich AL, et al. Music genre classification using LBP textural features. Signal Processing, 2012, 92(11): 2723–2737.

3 Poria S, Gelbukh A, Hussain A, et al. Music genre classification: A semi-supervised approach. Pattern Recognition.

Springer Berlin Heidelberg. 2013: 254−263.

4 张燕,唐振民,李燕萍,邹益.基于 MFCC 和 HMM 的音乐分类方法研究.南京师范大学学报(工程技术版),2008,(4): 112−114.

5 胡景凯,吴磊,高阳.基于学习分类器(LCS)的 MP3 音乐分类方法.重庆邮电大学学报(自然科学版),2007,(4):417−421.

6 张家发,胡景凯,高阳,黄建东.基于 MDCT 域特征的 MP3 音乐分类.江南大学学报(自然科学版),2007,(6):769−773.

7 Zwicker E. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). The Journal of the Acoustical Society of America, 1961, 33: 248.

8 Scheirer ED. Tempo and beat analysis of acoustic musical signals. The Journal of the Acoustical Society of America, 1998, 103: 588.

9 Shannon CE, Weaver W. A Mathematical Theory of Communication. 1948.

10 Lartillot O, Toivianen P. A Matlab toolbox for musical feature extraction from audio. International Conference on Digital Audio Effects. 2007. 237−244.

11 Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995,20(3):273−297.

12 Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.