基于 AdaBoost 的音乐情绪分类

王 磊[①②] 杜利民[①] 王劲林[①]
[①](中国科学院声学研究所 北京 100080)
[②](中国科学院研究生院 北京 100039)

摘 要：随着流媒体应用的蓬勃兴起，音频信号的自动分类开始成为工程与学术关注的热点之一。根据音乐信号对乐曲表现的情绪进行分类，由于涉及音乐信号的社会属性和自然属性的综合表征与模糊分类，因此处理方法相应需要在各种传统表征与分类方法的基础上进行机制筛选与架构优化。该文探讨了在 AdaBoost 算法，K-L 变换和 GMM 模型的基础上构造弱分类器的方法，采用多层分类器结构，成功地实现了对音乐信号进行情绪分类。初步的实验对 163 首歌曲进行平静(Calm)，悲伤(Sad)，激动(Exciting)以及愉悦(Pleasant)4 种类别的分类，训练集和测试集的分类准确率分别达到 97.5%和 93.9%，展示了这种方法的可行性和进一步发展的潜在价值。

关键词： AdaBoost；音乐；情绪；音色；节奏；分类；K-L 变换；多层分类器
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2007)09-2067-06

Mood Classification of Music Using AdaBoost

Wang Lei[①②] Du Li-min[①] Wang Jin-lin[①]
[①](*Institute of Acoustics of the Chinese Academy of Sciences*, *Beijing* 100080, *China*)
[②](*Graduate School of the Chinese Academy of Sciences*, *Beijing* 100039, *China*)

Abstract: With fast development and boosting of stream media applications, automatic classification of audio signals becomes one of the hotspots on research and engineering. Since mood classification of music is involved with integrated representation and classification of social and natural properties of music, mechanism selection and architecture optimization should be implemented on the basis of different traditional music representations and classification methods. This paper discusses formation of weak classifiers in AdaBoost algorithm based on K-L transformation and GMM training and realizes mood classification of music with multi-layer classifier architecture. The experiments classify 163 songs into four mood classes: calm, sad, exciting and pleasant with 97.5% accuracy on training data and 93.9% accuracy on test data, which proves feasibility and potential value of this method.

Key words: AdaBoost; Music; Mood; Timbre; Tempo; Classification; K-L transformation; Multi-layer classifier

1 引言

上世纪 90 年代以来，随着计算机网络的不断发展和日益普及，人们可以方便快捷地获取日益丰富的音乐资源。因此，人们迫切需要新的技术对音乐资源进行有效的管理，实现对海量音乐资源的检索和访问。传统上对音乐的检索利用了文本检索，比如通过在 MP3 文件中写入歌曲名称、歌手姓名等方法，利用文本搜索技术实现对音乐的管理和检索。严格地说，这种技术并没有实现对音乐本身的检索，因此无法从内容上对音乐的管理和检索。

基于多媒体内容的管理和检索是当前学术研究的一个热点。对于音乐来说，音乐内容包括旋律、乐器和歌手等方面，而音乐内容作为情绪的主要载体和激励之一，研究如何对音乐的情绪分类和检索，具有实用和研究的重要意义。Marc Landy对音乐和心理的相互作用进行了深入的研究[1]，证实了不同的乐器、不同风格的曲调使听者的情绪变化也完全不同。虽然对于“音乐究竟是如何表达和激励情绪的？”等问题还需要音乐、美学、哲学等学科的进一步研究，但是通过机器对音乐情绪的自动分类的尝试，我们希望可以对基于音乐内容的检索及其实用做出有用探索。

在本文中，音乐被分为了 4 种不同的情绪：平静的、悲伤的、激动的以及愉悦的。在音乐与情绪的相互作用的研究中，一般使用两个要素，第 1 个是反映情绪正面或负面的指标(valence)，第 2 个是反映情绪强烈程度的指标(arousal)。在我们所使用的 4 种情绪中，愉悦的和悲伤的分别是典型的正面情绪(positive valence)和负面情绪(negative valence)，而平静的和激动的分别是强烈程度较低(low arousal)和较高(high arousal)的情绪。这 4 种情绪是音乐情绪相关研究中使用最广泛的 4 种情绪，因此可以作为典型的分类类型。

在机器学习和人工智能等领域中，Boosting的方法被广泛应用以改进分类的准确率。Freund和Schapire[2]提出的AdaBoost算法解决了早期一些Boosting算法存在的问题，通过自适应地调整不同数据的权重信息，循环筛选出若干最佳

2006-01-24 收到，2006-05-17 改回

的弱分类器，并把它们组合成最终的强分类器。作为AdaBoost的典型应用，Viola[3]在人脸识别的任务中利用AdaBoost进行特征选择，获得了成功。在音频方面的应用上，Guo[4]等人通过对 16 种声音进行分类，比较了AdaBoost算法和SVM算法的分类性能；Xiong[5]等人则利用AdaBoost算法实现了对语音和非语音信号的区分；Sourabh和David[6]把AdaBoost用作特征选择工具，从提取的 256 维特征中选择若干维，对 4 类音频信号进行分类，并比较了AdaBoost和PCA在特征选择方面的性能。

可见，AdaBoost 算法在模式匹配问题中有着极为广泛的应用。本文试将 AdaBoost 算法引入到音乐情绪分类的任务中。为了充分挖掘出 AdaBoost 算法的性能，本文在使用 AdaBoost 时，构造了不同的弱分类器组进行实验，各种不同的弱分类器将会在实验部分详细介绍。

2 特征提取

为了有效反应出音乐的情绪，我们提取了两类特征，第一类称为音色特征，第二类称为节奏特征。

2.1 音色特征(timbre feature)

音色通常被定义为“在相同的响度水平下，听者用来分辨不同声音和乐器的一种度量”[7]。不同的声音和乐器，其音色取决于发声器官振动源的物理特性。人们对音色的感知已经研究了很长时间，音频信号的短时频谱可以有效地反映出音频信号的音色。

音色特征主要由其频谱形状体现出来。对于一帧音频信号，其频谱形状是指其短时傅里叶变换的频谱分布特征。在本文中，把每段音乐表示成帧长为 20ms，帧移也为 20ms的互不交叠的帧序列，通过对音频帧的短时频谱分析，我们得到一个 26 维的特征矢量[8]。具体说明如下：

(1) 子带特征　把每一帧的短时频谱按照倍频程划分为 7 个子带，每一个子带包含若干频谱分量，每个子带的最大频谱分量、最小频谱分量以及平均值均作为一个特征，因此每个子带有 3 个特征，所有 7 个子带即有 21 维特征，有效地表示出了频谱能量在每一个子带的分布情况。

(2) 谱质心(spectral centroid) 即频谱各分量幅度值的加权平均，相当于频谱分布的“质心”。其计算公式如式(1)所示。其中，$S_t[n]$是第 t 帧的短时傅里叶变换的幅度值。通常来说，轻松愉快的歌曲中高频分量较多，因此谱质心也较高；而感觉哀伤的歌曲中低频分量较多，因此谱质心也较低。

$$C_t = \sum_{n=1}^{N} S_t[n]\cdot n \Big/ \sum_{n=1}^{N} S_t[n] \tag{1}$$

(3) bandwidth　即各频谱分量到谱质心距离的加权平均。

$$B_t = \sum_{n=1}^{N} S_t[n]\cdot |n - C_t| \Big/ \sum_{n=1}^{N} S_t[n] \tag{2}$$

其中 $S_t[n]$ 的意义同式(1)，C_t 是第 t 帧的谱质心。该特征体现了频谱分布较集中还是较分散。

(4) roll-off frequency　定义为频谱总能量的 95%处的截止频率。

$$R_t = \arg_{n_t}\left(\sum_{n=1}^{n_t} S_t[n] = 0.95\cdot \sum_{n=1}^{N} S_t[n]\right) \tag{3}$$

设 roll-off frequency 的值为 f_r，则低于 f_r 的所有频谱分量的能量和占频谱总能量的 95%；

(5) spectral flux　定义为前后相邻帧的频谱幅度差的二阶距离。反映了前后帧之间的频谱变化的大小。

$$SF_t = \sum_{n=1}^{N} (S_t[n] - S_{t-1}[n])^2 \tag{4}$$

(6) 短时能量(short time energy) 即每一帧的所有频谱能量，一般使用其对数值。

$$E_t = \log\left(\sum_{n=1}^{N} S_t^2[n]\right) \tag{5}$$

在以上 26 维特征中，1–21 维描述了各个子带的分布情形；22 维和 24 维反映了频率的高低；23 维反映了频谱分布的紧密情况；而 25 维则反映了前后帧之间的频谱变化；最后，26 维则给出了频谱的能量大小。这 26 个特征从各个不同的角度刻画了音乐的短时频谱形状，准确把握了短时频谱的特点，对于音乐的情绪分类起着重要的作用。

2.2 节奏特征(tempo feature)

在不同种类的音乐中，其节拍或者节奏也不相同。音乐的节奏有两个特征：快慢和强弱。在本文要划分的 4 个类别中，平静的和悲伤的音乐一般节奏较缓慢，而激动的和愉悦的音乐一般节奏较快；另一方面，平静的和愉悦的音乐的节奏感较弱，而悲伤的和激动的音乐来说较强。可见，音乐的节奏特征对于其情绪表达起着重要的作用。

一般的音乐中，节奏通常由乐器所体现出来。而节奏乐器又主要是一些低音乐器，如贝斯鼓、低音吉他等。由于 40–150Hz的频率范围基本上涵盖了常见的低音乐器的频带，所以，为了分析音乐的节奏，我们只需分析 40–150Hz频带的时域调制谱(modulation spectrum)即可。对于一段音乐的帧序列，音色特征的提取方法如下[9]：

在短时幅度谱的 40–150Hz 的范围内选取 3 个频率值，则对于整个帧序列，每一帧的这 3 个频率的幅度值在时域上构成了 3 个信号序列。对于其中每一个信号序列：

(1) 10Hz 低通滤波，并求其 1 阶差分序列，对差分序列做 FFT 变换，求其调制谱；

(2) 在调制谱上 0– 5Hz 的范围内(对应于每分钟 0 – 300 个节拍)，以 0.5Hz 为间隔，划分 10 个子带，每个子带的平均能量作为一个特征，则共有 10 个特征；

(3) 3 个信号序列共有 30 维特征，即为音乐的节奏特征。

不难看出，这样提取出的调制谱子带能量有效地表示出了音乐节奏的快慢和强弱。如果音乐的节奏较快，则频率较高的子带的平均能量值较大；另一方面，如果音乐的节奏感

较强，则子带的平均能量值较大。

3　AdaBoost 算法

Boosting算法通过对若干分类器的组合可以使分类性能得到有效的提高。AdaBoost算法由Freund和Schapire与 1995 年提出并发展完善。作为一种“自适应”的Boosting算法，AdaBoost算法在若干轮循环中选择出若干个弱分类器，并把它们组合起来形成最终的分类器。由于这个特性，AdaBoost也经常被使用为特征选择的工具[3,6]。在每一轮循环中，AdaBoost都会自动提高上一轮循环中训练错误的数据的权重，降低训练正确的数据的权重。所以，每一轮循环所选出的弱分类器都针对训练数据中训练错误的那一部分。

理论上说，AdaBoost 算法可以无限缩小训练错误(training error)，这是 AdaBoost 算法一个基本的理论特性[2]。Freund和Schapire在另一篇文章[10]里分析了AdaBoost算法的测试错误(generalization error)，指出AdaBoost的测试错误和训练集的大小、弱分类器的状态空间以及循环进行的次数存在着概率关系。

AdaBoost 算法的主要步骤如下。

输入：训练数据($\boldsymbol{x}_1, y_1$)，…，($\boldsymbol{x}_n, y_n$)和循环的次数 T，其中 $\boldsymbol{x}_i$ 表示一个训练数据， $y_i \in \{0,1\}$,表示 $\boldsymbol{x}_i$ 所属的类别。

初始化： $w_{1,i} = \begin{cases} 1/2m, \text{ when } y_i = 0 \\ 1/2n, \text{ when } y_i = 1 \end{cases}$ ， m 为类别 0 的训练数据个数， n 为类别 1 的训练数据个数。

循环训练：for ($t = 1, \cdots, T$)

{

(1) 归一化权重： $w_{t,i} = w_{t,i} / \sum_{j=1}^{n} w_{t,j}$ ；

(2) 训练弱分类器 $h_j, j \in [1,\cdots,M]$ ，计算 h_j 的错误： $\varepsilon_j = \sum_i w_{t,i} \left| h_j(\boldsymbol{x}_i) - y_i \right|$ ；

(3) 从 M 个弱分类器中选择 ε_j 最小的一个，令其为 h_t ，其所对应的错误值为 ε_t ；

(4) 更新权重： $w_{t+1,i} = w_t \beta_t^{1-e_i}$ ，其中 $\beta_t = \varepsilon_t/(1-\varepsilon_t)$ ，当 $\boldsymbol{x}_i$ 被正确分类时， $e_i = 0$ ；反之为 1；

}

输出：最终的分类器为： $h(\boldsymbol{x}) = \text{sgn}\left(\sum_{t=1}^{T} \alpha_t (h_t(\boldsymbol{x}) - 0.5)\right)$,其中 $\alpha_t = \log(1/\beta_t)$ 。

上面所述的 AdaBoost 算法应用于两个类别的分类问题。实际上，AdaBoost 算法也可以被扩展用于多类别分类问题。在上面算法的每一轮循环中，都要训练 M 个弱分类器，其中每个弱分类器的性能只要比随机分类稍好即可。在上文所引的各个应用中[3−6]，每一个弱分类器均建立在单维特征上。这种情况下，假设有 N 维特征，则共有 N 个弱分类器，每一轮都从这 N 个弱分类器中找出性能最好的那一个。在本文中，由于特征的维数较少，建立在单个特征上的弱分类器并不能带来令人满意的分类性能，所以必须构造建立在多维特征上的弱分类器组。

4　弱分类器的构造

4.1 基于单维特征的弱分类器的构造

在AdaBoost算法的每一轮循环中，都需要构造一组弱分类器，AdaBoost从这组弱分类器中选出性能最好的一个作为本轮的选择。在Viola的人脸识别的应用[3]中，针对每一维特征构造了一个弱分类器。而在Guo等人利用AdaBoost算法对声音分类的应用[4]中，弱分类器的定义如下：

$$h_i(\boldsymbol{x}) = \begin{cases} 0, & |x_i - \mu_{i0}| < |x_i - \mu_{i1}| \\ 1, & |x_i - \mu_{i0}| > |x_i - \mu_{i1}| \end{cases}, \quad 1 \le i \le D \tag{6}$$

在Xiong等人区分语音和非语音的工作[5]中，弱分类器定义如下：

$$h_i(\boldsymbol{x}) = \begin{cases} 0, & x_i > \theta_i \\ 1, & \text{其他} \end{cases}, \quad 1 \le i \le D \tag{7}$$

式(6)和式(7)中， $\boldsymbol{x}$为特征矢量， x_i 为第i维特征， μ_{i0} 、 μ_{i1} 分别为类别 0 和类别 1 的第 i 维特征的均值，D为特征矢量的维数， θ_i 为第 i 维特征上设定的阈值。由式(6)，式(7)可知，这两种弱分类器都建立在单一维特征的基础之上，其区别在于：式(6)使用NC(Nearest Center)原则进行分类，而式(7)使用阈值进行分类。另外，在Sourabh[6]的工作中，AdaBoost算法被用作特征选择，其使用的弱分类器同样建立在单一维特征基础之上。

4.2 利用 K-L 变换和 GMM 模型构造弱分类器组

本文为了获得更好的分类性能，提出了一种新的弱分类器构造方法。与上述工作不同的是，在整个特征矢量的基础上，使用 K-L 变换得到不同维数的新矢量，再用包含不同个数高斯分量的 GMM 分类器来构造整个弱分类器组。构造过程如图 1 所示。

如图 1 所示， D 维特征矢量经过 K-L 变换后，可以得到 1–D 维的特征矢量，分别对这些不同维数的数据通过 EM 训练得到 GMM 分类器，其中，GMM 分类器中的高斯分量数目为 3–9，共 7 个 GMM 分类器。由此，用这种方法构造弱分类器组，每组可以得到 7D 个 GMM 弱分类器，用这 7D 个弱分类器分别对训练数据进行分类，选择其中错误最少的一个作为 AdaBoost 该轮循环中所选定的弱分类器。

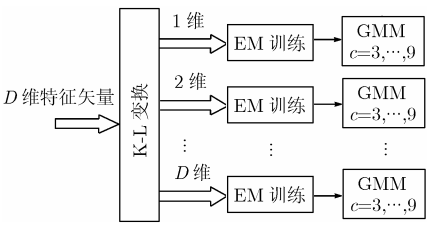

图 1 AdaBoost 中弱分类器的构造

4.3　对 K−L 变换和 EM 训练的修改

由AdaBoost算法可知，训练数据的权重在每一轮循环中都被归一化。因此可以把训练数据的权重看成是概率分布，并引入到K-L变换的算法以及训练GMM模型的EM算法中。假设我们有N个训练数据，其中 N_0 个属于类别 0，N_1 个属于类别 1，则 $N_0+N_1=N$；设每个训练数据的权重为 w_i，则有 $\sum_i w_i=1$。则对于K-L变换[11]，类别 0 和类别 1 的先验概率分别为

$$P_0=\sum_i w_i \ (i\in \text{ class 0}),\quad P_1=\sum_i w_i \ (i\in \text{ class 1}) \tag{8}$$

训练数据的类内分布为

$$w_i'=w_i/P_0 \ (i\in \text{ class 0}),\quad w_i'=w_i/P_1 \ (i\in \text{ class 1}) \tag{9}$$

类内散度阵 $\boldsymbol{S}_w$ 和类间散度阵 $\boldsymbol{S}_b$ 如式(10)所示，其中 c 为类别数，这里为 2。

$$\boldsymbol{S}_w=\sum_{i=1}^{c}P_i\boldsymbol{\Sigma}_i\ ,\qquad \boldsymbol{S}_b=\sum_{i=1}^{c}P_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^{\mathrm{T}} \tag{10}$$

$\boldsymbol{S}_w$ 和 $\boldsymbol{S}_b$ 中，有

$$\boldsymbol{\Sigma}_i=E[(\boldsymbol{X}-\boldsymbol{\mu}_i)(\boldsymbol{X}-\boldsymbol{\mu}_i)^{\mathrm{T}}]=\sum_{j=1}^{N_i}w_i'(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{\mu}_i=E(\boldsymbol{X})=\sum_{j=1}^{N_i}w_j'\boldsymbol{x}_j,\qquad \boldsymbol{\mu}=\sum_{j=1}^{N}w_j'\boldsymbol{x}_j=\sum_{i=1}^{c}P_i\boldsymbol{\mu}_i \tag{12}$$

对于 EM 训练[12]，则有如下公式，其中类别 0 时 $M=N_0$，类别 1 时 $M=N_1$。

$$\widehat{C}_k=\sum_{i=1}^{M}w_i'\gamma_k^i,\qquad \widehat{\boldsymbol{\mu}}_k=\sum_{i=1}^{M}w_i'\gamma_k^i\boldsymbol{x}_i\Big/\sum_{i=1}^{M}w_i'\gamma_k^i \tag{13}$$

$$\widehat{\boldsymbol{\Sigma}}_k=\sum_{i=1}^{M}w_k'\gamma_k^i(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^{\mathrm{T}}\Big/\sum_{i=1}^{M}w_k'\gamma_k^i \tag{14}$$

5 系统结构

5.1 双层结构及其训练过程

由于 AdaBoost 算法通常用来对两类问题进行分类，而本文要把音乐按照情绪分成 4 类，所以，我们针对每一种情绪训练了一个 AdaBoost 分类器，共计 4 个分类器。对于其中每一个，类别 0 的训练数据是该种情绪的歌曲，类别 1 是除了该种情绪的所有其它歌曲。另外，由于提取了音色特征和节奏特征，其中音色特征是短时特征，节奏特征是长时特征，为了充分利用这两种特征获得更好的分类性能，对于 4 个分类器中的每一个，我们都采取了两层结构。第 1 层根据音色特征进行归类，第 2 层根据节奏特征进行分类，两层分类器的结果综合到一起，得到最后的分类结果。其中，在训练第 2 层分类器时，类别 0 的训练集为第一层分类器误判为类别 1 的那些歌曲。下面以 Calm 为例分 3 个步骤详细描述训练过程。

对于第一层的 Calm 分类器，其类别 0 是训练集中平静的歌曲，类别 1 是其他 3 类歌曲。提取类别 0 和类别 1 中所有歌曲的音色特征，利用 AdaBoost 算法训练得到分类器“Calm $L1$”($L1$ 表示第 1 层)，用分类器“Calm $L1$”对训练集中类别 0 的所有歌曲进行测试；对于第 2 层的 Calm 分类器，其类别 0 是上一步中测试错误的歌曲，如果这类歌曲较少，则补充适当数目的歌曲，类别 1 仍然是所有其他 3 类歌曲。提取类别 0 和类别 1 中所有歌曲的节奏特征，利用 AdaBoost 算法得到分类器“Calm $L2$”($L2$ 表示第 2 层)。

5.2 歌曲的分类准则

图 2 给出了一首歌曲的判别过程。图中，$L1$ 为第 1 层分类器，$L2$ 为第 2 层分类器，Cij，Sij，Eij，Pij 分别表示第 i 层分类器归到类别 j 的帧数，其中，$i=1$ 表示第 1 层分类器；$i=2$ 表示第 2 层分类器；$j=0$ 表示归类到类别 0；$j=1$ 表示归类到类别 1。当要对一首音乐分类时，把音色特征矢量序列输入到 4 个第 1 层分类器，而节奏特征矢量序列则分别输入到 4 个第 2 层分类器。每 1 个 AdaBoost 分类器输出两个整数 Cij，Sij，Eij，Pij 等，分别表示归类到类别 0 的帧数和类别 1 的帧数。例如，“Calm $L1$”输出 C10 和 C11，其中，C10 表示归类到类别 0 的帧数，C11 表示归类到类别 1 的帧数。

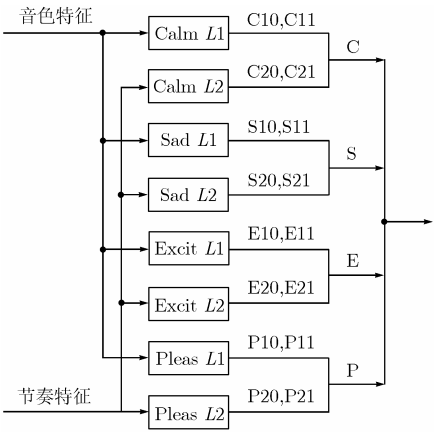

图 2 音乐按情绪分类过程

由于在训练时，第 2 层分类器的训练集中的类别 0 的数据是第 1 层分类器误判为类别 1 的歌曲，所以，在对两层分类器的分类结果进行综合时，我们对第 1 层分类器的分类结果分配了较大的权值，经过实验取为 0.7，而第 2 层分类器的分类结果的权值较小，取为 0.3。这样的话，根据上下两层共 8 个 AdaBoost 分类器的输出，我们可以进行如下计算：

$C=0.7\times$C10 / (C10 + C11)$+0.3\times$C20 / (C20 + C21)

$S=0.7\times$S10 / (S10 + S11)$+0.3\times$S20 / (S20 + S21)

$E=0.7\times$E10 / (E10 + E11) $+\ 0.3\times$E20 / (E20 + E21)

$P=0.7\times$P10 / (P10 + P11)$+0.3\times$P20 / (P20 + P21)

最终，对于输入的歌曲，如果 C 是 C，S，E，P 4 个值中最大的那个，则该歌曲被归类为平静的歌曲，如果 S 最大，则归类为悲伤的歌曲，以此类推。

6 实验结果

本文实验使用的数据包括 163 首歌曲，这些歌曲按照情绪被标记为 4 个类别，包括 33 首平静的(Calm)，30 首激动的(exciting)，50 首愉悦的(pleasant)和 50 首悲伤的(sad)。

平静的歌曲主要是一些经典音乐，包括小部分电影伴音和流行歌曲；激动的歌曲主要是以音量大、节奏强为特征的摇滚乐；愉悦的歌曲则是节奏较快的流行音乐或电子乐；悲伤的歌曲则是一些节奏较慢、音色颤抖的流行音乐。上述所有音乐的标注均由音乐方面的专业人员完成，但是按照情绪分类并不是绝对的，不同的人听完同一首歌曲在情绪上的反应会有所不同。例如，对于一首流行音乐，一部分人会认为很平静，而另一部分人却认为较悲伤。上述 163 首歌曲中，训练集和测试集之间的比例大约为 80%：20%。

本文进行了两项试验，并对比了它们的结果。第 1 项实验使用了文献[4]中的方法对音乐按情绪分类，对于任意两个类别都训练一个 AdaBoost 分类器，所以共有 6 个分类器，每 1 个分类器输出 1 个分类结果。如果一首歌曲被其中的 3 个分类器都归类为同一个类别，则该歌曲即被归类为该类别。在 AdaBoost 的训练中，弱分类器都是基于单维特征并使用 NC 准则进行判定。第 2 项实验采用了本文第 4 部分所介绍的方法构造弱分类器组，采用本文第 5 部分介绍的系统结构和训练方法进行音乐的情绪判别。表 1 和表 2 分别给出了两项实验的实验结果。

表 1 实验 1 结果(弱分类器：单一维上的 NC 准则分类器)

分类器	循环次数 T	错误歌曲	准确率
Calm v.s. Sad	5	4 Calm, 15 Sad	69.8%
Calm v.s. Exciting	5	0	100%
Calm v.s. Pleasant	5	0	100%
Sad.s.Exciting	5	1 Sad, 5 Exciting	92.5%
Sad v.s. Pleasant	5	2 Pleasant	98%
Exciting v.s.Pleasant	5	6Exciting,4Pleasant	87.5%
总共：163 个文件，37 个错误，总准确率：77.3%			

表 1 中，用来区分平静歌曲和悲伤歌曲的性能最差，其准确率只有 69.8%，其原因在于：平静类别和悲伤类别各包含一部分流行歌曲，而这些流行歌曲即使对于不同的人来说，也很可能得到不同的分类结果，对于计算机来说，区分这种有相当重叠区域的两个类别当然也是一个很困难的任务；另一方面，在这项实验中，我们所选用的 NC 准则弱分类器组实际上性能很差，这样的话，每一轮循环中所得到的 ε_t 也就较大，经过有限的几次循环后，ε_t 就已经逼近了其极限 0.5(ε_t 的实际意义是分类的错误概率)，因此，β_t 的值也就约等于 1，训练数据的权重基本上停止了更新，继续循环下去已经没有了意义。这一点同样也可以从表 1 中看出，所有的 6 个分类器在进行了仅仅 5 轮循环后就已无法再继续循环更新下去了。这样的话，最终的分类器也就是 5 个弱分类器的线性组合，自然没有多循环几次的效果好。

除了区分平静和悲伤这两个类别的分类器之外，其余的 5 个分类器的分类性能良好，这充分说明了所使用的 26 维音色特征能够准确表现出音乐的情绪。

表 2　实验 2 结果(两层分类器，弱分类器组：KL + GMM)

分类器	经过第 1 层后的结果			经过两层后的最终结果	
	循环次数 T	错误歌曲数	准确率	错误歌曲数	准确率
Calm	30	27/163	83.4%	2/163	98.8%
Sad	30	25/163	84.7%	0/163	100%
Exciting	30	9/163	94.5%	1/163	99.4%
Pleasant	30	15/163	90.8%	1/163	99.4%
总计	经过两层分类器后，全部 163 首歌曲中(训练集：130，测试集：33)，错了 4 首(训练集 2 首，测试集 2 首)，准确率为 97.5%。其中，测试集准确率为 93.9%				

在实验 2 中，使用了本文提出的两层分类器的结构，在构建 AdaBoost 循环中的弱分类器组时使用了经过 K-L 变换后的 GMM 分类器。第 2 层分类器利用节奏特征修正了第 1 层分类器的分类结果，由表 2 可见，第 2 层分类器可以大幅度地提高分类的准确率，一方面是由于第 2 层的分类器的训练负样本采用的是第 1 层分类错误的那些歌曲，另一方面也说明了节奏特征在音乐情绪的分类任务中具有良好的区分性。

7　结束语

作为基于音乐内容的检索的有益探索，本文在音乐的音色特征和节奏特征的基础上，使用 AdaBoost 算法实现了把音乐按情绪分成 4 个类别。本文深入讨论了 AdaBoost 算法中弱分类器组设计对其性能的影响，提出了一种能够充分发挥 AdaBoost 算法性能的弱分类器组的设计方法，设计了双层分类器的结构，并在实验中得到了良好的效果，最终的分类准确率达到 97.5%。本文所论述的方法对 AdaBoost 算法在音频方面的应用具有一定的意义。

参 考 文 献

[1] Landy M. Emotions and music, how does music convey emotion? From learning to performing. *Canadian Music Educator*, 2004, 45(4): 28–33.

[2] Schapire R E. A brief introduction to boosting, proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, 1999: 1401–1406.

[3] Viola P and Jones M J. Robust real-time face detection. *International Journal of Computer Vision*, 2004, 57(2): 137–154.

[4] Guo G D, Zhang H J, and Li S Z. Boosting for content-based audio classification and retrieval: An evaluation. International Conference on Multimedia and Expo, Tokyo, 2001: 253–256.

[5] Xiong Z Y and Huang T S. Boosting speech / non-speech classification using averaged mel-frequency cepstrum coefficients features. IEEE Pacific Rim Conference on Multimedia, Hsinchu, 2002: 573–580.

[6] Ravindran S and Anderson D. Boosting as a dimensionality reduction tool for audio classification, Proceedings of IEEE International Symposium on Circuits and Systems, Vancouver, 2004: III–465–8.

[7] Zhang T and Kuo J. Hierarchical system for content-based audio classification and retrieval. Proceedings of SPIE's Conference on Multimedia Storage and Archiving Systems III, Boston, 1998: 398–409.

[8] Liu D, Lu L, and Zhang H J. Automatic mood detection from acoustic music data. Proceedings of International Symposium on Music Information Retrieval, Baltimore, 2003: 81–87.

[9] Scheirer E D. Tempo and beat analysis of acoustic musical signals. *Journal of Acoustic Society of America*, 1998, 103(1): 588–601.

[10] Freund Y and Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997, 55(1): 119–139.

[11] 边肇祺，张学工. 模式识别. 第二版.北京：清华大学出版社，2002: 第 9 章.

[12] Huang X D, Acero A, and Hon H W. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. 1st Edition. Upper Saddle River, NJ, Prentice Hall PTR, 2001, Chapter 5.

王 磊： 男，1978 年生，博士生，研究方向为音频信号处理和流媒体检索.

杜利民： 男，1957 年生，教授，博士生导师，主要研究方向为信号处理、语音交互技术研究.

王劲林： 男，1964 年生，教授，博士生导师，主要研究方向为数字视音频广播、宽带网络通信.