第六届数学中国数学建模网络挑战赛
地址：内蒙古数学学会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2013@tzmcm.cn

第六届“认证杯”数学中国

数学建模网络挑战赛

承　诺　书

我们仔细阅读了第六届“认证杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

我们的参赛队号为：#2854

参赛队员　(签名)　：

队员 1：陈嘉鑫

队员 2：余相君

队员 3：涂丽

参赛队教练员 (签名)：

参赛队伍组别：本科组

比赛阶段：第一阶段

第六届数学中国数学建模网络挑战赛

地址：内蒙古数学学会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2013@tzmcm.cn

第六届“认证杯”数学中国

数学建模网络挑战赛
编 号 专 用 页

参赛队伍的参赛队号：（请各个参赛队提前填写好）：

#2854

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

第六届数学中国数学建模网络挑战赛
地址：内蒙古数学学会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2013@tzmcm.cn

2013 年第六届“认证杯”数学中国数学建模网络挑战赛

题　　目　多模态分析在流行音乐风格分类中的应用

关 键 词　多模态 标签 语义 LDA SVMs 分形维数 流行音乐 分类

摘　　要：

由于社会文化的发展，流行音乐的风格日趋多元化，风格之间相互融合发展，造成类别混乱，难以划分，存在分类不当等诸多不足。本文采用循序渐进的方法对流行音乐的风格给出多模型优势互补的分类方法。

模型一是基于标签的流行音乐风格分类模型。随着 Web 2.0 日益健全完善，标签数据日益丰富且趋于稳定，标签所表达的语义信息要比描述文档中的关键词更接近被描述事物的特性，标签资源在音乐风格分类领域有着广阔的应用。模型通过 LDA 分析器对抽取的标签资源进行语料库建模，达到音乐风格分类的目的。

模型二是基于语义的流行音乐风格分类模型。互联网上拥有海量的文本资源，当我们在网络上抓取数据的时候，如果一段流行音乐总与某个特定的流行音乐风格出现在一起，那么我们就认为此流行音乐与这个风格有着非常紧密的联系，从而我们就可以推断其风格，同时也能弥补某些类别标签不够准确、没有得到公认的缺点。

模型三是基于 LDA 和多类 SVM[1]的流行音乐风格分类模型。模型一和模型二都是在音乐文件具有大量相关数据的前提下具有较高分类准确率的，对于没有标签，数据较少的原创歌曲，体现了其局限性，此时需要提取底层的声学特征进行分类。这也是当前使用广泛且成熟的音频分类方法，本文通过对主流的特征提取及分类的算法进行改进，不但能取得最佳的分类精确率，而且也能实现最好的时间复杂度。

模型四是基于分形维数[2]的流行音乐风格分类模型。传统的基于声学特征的分类方法算法复杂，数据庞大，开销高，不具有深度推广的潜质，如何在准确率和开销之间做出权衡十分必要。基于分形维数的优势在于只用一维特征就能区分音乐的不同类型即分形刻画了音乐的内在特征——部分与整体的相似性。该方法具有应用简单，分类准确度较高，速度快等优点。

没有一种模型能应对所有情况，只有扩大其优点，针对性解决，优势互补才能达到理想的分类效果，本文在模型的优化中将多种模型综合，提出了多模态音乐风格分类方法，使分类更加自然、合理、准确。

参赛队号　#2854

所选题目　B

第六届数学中国数学建模网络挑战赛

地址：内蒙古数学学会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2013@tzmcm.cn

英文摘要（选填）

（此摘要非论文必须部分，选填可加分，加分不超过论文总分的 5%）

The development of social culture, pop music style diversification, integration development style, cause category clutter, is hard to classify, inappropriate classification of the deficiencies of existing. In this paper, using the step-by-step, complementary advantages of the model classification method of layer by layer analysis method is given for the popular style.

Model one is pop music style classification model based on tags. With the Web 2 is perfect, the tag data increasingly rich and tends to be stable, semantic information label expression to get closer to nature of things described than description document Keywords tag resources, has broad application in the field of music style classification. Model of corpus modeling through LDA analyzer on the selected tag resources .To achieve the purpose of music style classification.

Model two is the pop music style classification model based on semantic. The Internet has a mass text resources, when we crawl in the network data, if a pop music always appear together with a particular style of popular music, so we think the pop music has a very close relationship with this style, thus we can infer its style, also can make up for some of the class label is not accurate enough, not recognized.

Model three is the pop music style classification model based on LDA and SVM. Model one and two are with high classification accuracy premise with a large number of related data in the music file, not for the label, less data of original songs, reflects its limitation, this time need to extract acoustic features of the underlying classification. This is also the current audio classification method is widely used and mature, was improved through the feature extraction and classification of the mainstream algorithm, not only can achieve the best classification accuracy, but also can achieve the best time complexity.

Model four is the pop music style classification model based on fractal dimension. The traditional classification methods based on acoustic features complexity, huge data, high cost, does not have the depth expansion potential, how to make the necessary trade-offs between accuracy and overhead. The fractal dimension of the advantage of using only a one-dimensional feature can distinguish different types of fractal characterization of music is the inherent characteristics of music -- Based on the similarity between the part and the whole. The method has simple application, classification accuracy, the advantages of faster.

No model can deal with all cases, only the expansion of its advantages, solve, complementary effect of classification to achieve the ideal, in this paper, a variety of optimization model is put forward comprehensive, multimode music style classification method, the classification is more natural, reasonable, accurate.

参赛队号#2854

目录

一、问题的重述 2
1.1 问题背景 2
1.2 问题重述 2
二、问题分析 2
2.1 基于标签数据的分类方法 2
2.2 基于语义的分类方法 3
2.3 基于声学特征的分类方法 3
2.4 基于分形的分类方法 3
三、符号说明和问题假设 3
四、基于标签的流行音乐风格分类模型 5
4.1 抽取测试音乐的标题和音乐家的姓名 5
4.2 获取测试音乐的标签 5
4.3 基于 LDA 框架对标签进行分类 7
4.4 数据分析和结论 7
五、基于互联网语义的流行音乐风格分类模型 8
5.1 原始数据采集 9
5.2 互联网语义关系特征向量的表示 11
5.3 多个音乐风格同时存在的判断 12
5.4 数据分析和结论 12
六、基于 LDA 和多类 SVM 的流行音乐风格分类模型 12
6.1 音乐特征分析 13
6.2 语音信号预处理 16
6.3 音乐特征提取 17
6.4 使用 LDA 对特征向量进行降维 18
6.5 使用多类 SVM 对特征向量进行分类 21
6.6 数据分析和结论 22
七、基于分形维数的流行音乐风格分类模型 25
7.1 音乐与 1/f 噪声 25
7.2 关于 1/f 噪声的分形性质 27
7.3 音乐的分形维数计算 28
7.4 基于分形维数的音乐分类方法 29
7.5 数据分析和结论 31
八、模型评价及优化 31
九、模型的推广 33
十、参考文献 33

参赛队号#2854

一、问题的重述

1.1 问题背景

随着互联网的发展，流行音乐的主要传播媒介从传统的电台和唱片逐渐过渡到网络下载和网络电台等。网络电台需要根据收听者的已知喜好，自动推荐并播放其它音乐。由于每个人喜好的音乐可能横跨若干种风格，区别甚大，需要分别对待。这就需要探讨如何区分音乐风格的问题。

传统的基于关键字的分类技术需要获得音频数据的版权信息， 而当数据的版权信息不明确时， 就有必要采用基于内容的方法来实现自动分类的需求。面对海量的多媒体信息，如何在浩如烟海的信息中快速、容易地获得不同风格的音乐亦成为当前必须要解决的问题。

1.2 问题重述

由于社会文化的发展，流行音乐的风格日趋多元化，风格之间相互融合发展，造成类别混乱，难以划分，存在分类不当等诸多不足。请你建立合理的数学模型，对流行音乐的风格给出一个自然、合理的分类方法，以便给网络电台的推荐功能和其它可能的用途提供支持。

二、问题分析

伴随着数字技术的飞速发展，越来越多的音乐被上传到互联网上，正是这种海量的且不断增长的音乐资源使得用来处理音乐数据库的音乐信息检索(MIR)系统受到了越来越多的关注。网络用户更希望可以利用和音乐内容有关的信息来检索音乐。音乐的风格，例如蓝调(Blues)、摇滚(Rock)等就是经常被用户使用的检索词。目前，很多音乐网站例如 Last.fm，mp3.com 等都相继推出了基于风格的音乐检索系统。因此，音乐风格的准确分类对于现代音乐信息检索系统来说是至关重要的。

音乐风格分类历经了人工化和自动化两个阶段。早期，绝大多数的音乐网站都是对音乐的风格进行人工标注，其中最著名的就是潘多拉网站(pandora.com)聘请音乐专家所进行的“音乐染色体工程(music genome project)”。虽然人工地对音乐进行风格标注取得了一定的成功，但是这样做消耗了大量的人力成本、时间成本和资金成本；而更为严重的问题是，人工标注的速度显然已经不能满足网络中音乐资源飞速增长的需求。

2.1 基于标签数据的分类方法

Web 2.0 时代的到来恰恰提供了这样的一个机遇。Web 2.0 技术允许网络用户使用标签对其感兴趣的资源进行个性化标注，这其中自然也包括音乐资源，文献[3]通过实验证明了互联网中的标签数据的分布服从无标度网络的特征，在若干时间间隔后标签会趋于稳定，而且标签所表达的语义信息要比描述文档中的关键词更接近被描述事物的特性；文献[4]通过大量的实验数据验证了用户对歌曲进行标注时使用的标签与音乐专家对音乐的评价具有高度的一致性。因此，标签资源在音乐风格的自动分类领域同样有着广阔的应用前景。

在此提出了基于标签的音乐风格分类方法，使用 LDA 方法对由音乐标签组成的语料库进行建模。

2.2 基于语义的分类方法

模型一中存在标签信息的不明确，有的标签没有得到公认等问题，使其不具有普遍适用性。

在此提出了基于语义音乐风格分类模型，使用音乐名称和艺术家姓名这些与音乐有关的语义信息，通过搜索网络资源，计算音乐与不同音乐风格之间的联系紧密度并以此为依据进行音乐的风格分类。

2.3 基于声学特征的分类方法

模型一和模型二都是在音乐文件具有大量相关数据的前提下具有较高分类准确率的，当音乐文件最初发布时，其相关数据相对而言较少，基于标签和语义的方法不再适用。此时需要提取音乐自身的声学特征，然后根据这些特征对音乐进行风格分类。

在此提出了基于线性判别分析和支持向量机的音乐分类模型。

2.4 基于分形的分类方法

目前绝大多数音频分类算法集中在两方面——音频的特征提取以及根据音频特征进行分类。现有的音频特征算法有：短时过零率、时域的短时能量、谱质心分析、频域带宽等，还有基于听觉感受的 MFCC(Mel-frequency cepstral coefficients)梅尔倒频谱系数等。另一方面，分类算法可利用模式识别和模式分类中已知算法，如 CMM(Gaussian mixture model) 高斯混合模型、NN(Neural Network) 神经网络、HMM(Hidden Markov Model) 隐马尔可夫模型等。这些方法都存在着算法复杂，数据庞大，高精度带来的高开销等问题。

在此提出了基于分形维数的音乐分类模型，通过对不同风格音乐的分形维数的计算与比较，确定音乐分类的范围指标，然后利用此指标作为依据对音乐进行自动分类。由于维数为 1，所以此方法使用简单，速度较快，同时也具有较高的分类精度。

三、符号说明和问题假设

模型一：

X:歌曲集合

x_i：编号为 i 的歌曲名

$T(x_i)$：x_i的标签集合

C：流行音乐风格的集合

$Freq(t_j, x_i)$：标签t_j被用来标注音乐x_i的次数

t：音乐x_i包含的标签的数目

$pr(c_k|t_j)$：每个标签属于某个风格的概率

c_k：表示音乐x_i通过模型一求得的风格

参赛队号#2854

模型二：

P：两个词语同时出现在一个网页中的概率
M(a,b)：词语 a 和词语 b 同时出现的网页的个数
C(a)：只有 a 出现的网页的个数
T：流行音乐的互联网语义关系特征向量

模型三：

SC：频谱质心
SB：频谱带宽
SR：频谱滚降度
SF：频谱通量
P：Mel 倒谱系数
ω_c：谱质心
B：信号频谱成分与谱质心之差以能量进行加权的均值
S_T：总体离散度矩阵
S_B：类间离散度矩阵
S_W：类内离散度矩阵

模型四：

f：噪声的频率
F：盒维数
$\overline{\mathrm{DIM}_B(F)}$：上盒维数
$\underline{\mathrm{DIM}_B(F)}$：下盒维数
e：网格大小
2n：网格的最大边长
D:通过斜率表示的盒维数

当代流行音乐进入了高度细分化的时代，同时尚在继续细分更微观的类型。流行音乐电台细分市场的类型化方向正是来自流行音乐的细分化。对于流行音乐的分类目前尚无统一标准，我们使用目前认同度较高的分类标准，对流行音乐类型进行筛选合并后将流行音乐分为如下类型。

01. Pop 流行(Dream-Pop, Classical Pop, Britpop, Synth Pop)
02. R&B 节奏布鲁斯(Soul)
03. Hip-Hop 饶舌(Trip-Hop, Brit-Hop)
04. Rap 说唱(Gangsta Rap)
05. Jazz 爵士(Bossa Nova)
06. Rock 摇滚(Reggae，Punk)
07. Electro 电子（Techno，House，Disco,Chill Out）
08. Country 乡村
09. Blue 蓝调
10. Newage 新世纪

参赛队号#2854

四、基于标签的流行音乐风格分类模型

流行音乐风格的自动分类是一项非常具有挑战性的工作，虽然以之前传统的声学特征分类方法使得音乐风格分类的准确率有了一定程度的提高，但是仅仅使用底层声学特征对音乐进行音乐风格分类往往不能得到令人满意的结果，同时也不能满足实际音乐信息检索系统的需要。为此，需要找到研究方法来提高流行音乐风格分类的准确率。

基于互联网的应用正变得越来越普及，在这个过程中，有更多的站点将自身的资源开放给开发者来调用。对外提供的 API 调用使得站点之间的内容关联性更强，同时这些开放的平台也为用户、开发者和中小网站带来了更大的价值。当然也包括允许网络用户使用标签对其感兴趣的资源进行个性化标注，这其中自然也包括本次模型建立所要求的流行音乐资源，因此，标签资源在流行音乐风格的自动分类领域同样有着广阔的应用前景。

本模型提出于基于标签的流行音乐风格分类的概念，基于标签的流行音乐风格分类的框架如图 4.1 所示。

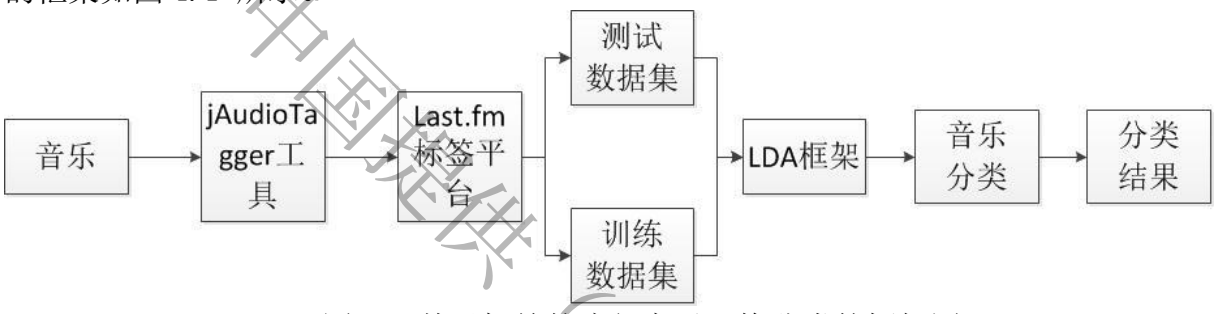

图 4.1 基于标签的流行音乐风格分类的框架图

4.1 抽取测试音乐的标题和音乐家的姓名

本模型的研究更关注于实际应用，所以选取整首流行音乐作为分类对象。数据集共包括10个流行音乐风格，分别是蓝调音乐(Blues)、乡村音乐(Country)、爵士乐(Jazz)、流行音乐(Pop)摇滚乐(Rock)、节奏布鲁斯(R&B)、嘻哈音乐(HIP-HOP)、灵魂音乐(Soul)、新世纪音乐(New Age)、说唱音乐(Rap)。其中每个风格由 10 首 mp3 格式的音乐组成。这些音乐是根据网站 Last.fm 列出的每个风格的经典专辑和经典曲目列表在 Google、百度和 Last.fm 上免费下载获得的。为了从音乐网站 Last.fm 上获取与上述音乐对应的标签,首先使用 jAudioTagger Library 从音乐文件包含的 ID3 标签中抽取音乐的标题和音乐家的姓名。

4.2 获取测试音乐的标签

根据所获得的音乐标题和音乐家姓名以此作为输入利用网站 Last.fm 提供的 API 接口(Track.GetTopTags) 下载和音乐对应的标签。Last.fm 是一个广受欢迎的音乐网站，它邀请用户对所听的音乐进行标注并通过用户添加的标签来来发掘用户的收听兴趣，以及向用户推荐音乐，当规模巨大的音乐资源、音乐家专辑和单曲被用户标注后，我们就能通过计算不同对象间公共标签的数量来分析对象之间的相似性，并能根据这些相似性来进行音乐风格分类。

参赛队号#2854

```
<toptags artist=" B.B.King " track="Three O' clock Blues">
  <tag>
    <name>blues</name>
    <count>100</count>
    <url>www.last.fm/tag/blues </url>
  </tag>
  <tag>
    <name>blues rock</name>
    <count>20</count>
    <url>www.last.fm/tag/blues rouck</url>
  </tag>
  ...
</toptags>
```

图 4.2：通过 API 接口（ Track.GetTopTags ）获得的音乐 Three O' clock Blues 的标签文件

Last.fm 网站规定每首歌曲的标签数量范围在到 100 之间， 而且除了标签以外还可以得到每个标签被使用的频率。图 4.2 展示了通过 API 接口（ Track.GetTopTags ）上获得的标签数据的原始数据。图 4.3 是展示了歌曲“Three O' clock Blues” 的标签。以序号为 1 的歌曲为例，歌曲的标题和艺术家的姓名分别是 Three O'clock Blues 和 B.B.King。历史上（时间 2013 年 4 月 13 日止）共有 25 个标签被用来标注这首歌曲，其中使用次数最多的标签就是 blues， 其频率为 100 次。

我们通过 API 接口（Track.GetTopTags）提取的标签是用户根据自己的理解对于音乐进行的标注，具有很大的随意性，而且其中的很多标签是与音乐风格信息无关的，因此我们需要对标签特征进行去噪、分离等处理。还是以歌曲 Three O' clock Blues 为例，我们需要将与风格无关的标签 bb king 等去掉；对于标签 blues rock 我们需要从中分离出音乐风格标签 Blues 和 rock 并分别计数；而对于标签 blues guitar 我们需要将 blues 提取出来并累积相加。图 4.3 中的标签特征经过上述操作以后就会得到如图 4.4 的结果：

```
1
artist:B.B. King
title:Three O'clock Blues
tags:    25
100      blues
20       blues rock
20       Classic Blues
13       jazz-blues
10       bb king
3        tioramon
3        elegidas
3        1951
3        whiskey drinking music
3        Diana Krall
3        blues guitar
3        electric blues
3        Awesome Guitar Jams
3        male vocalist
3        evil
3        Chicago Blues
3        melancholy
3        oldies
3        american
3        sad
3        swing
3        jazz
3        guitar
-1       endtags
```

图 4.3：描述歌曲 “Three O' clock Blues” 的标签

Tags:	4
152	blues
20	rock
20	classical
16	jazz

图 4.4：歌曲 Three O' clock Blues 预处理后的标签特征

4.3 基于 LDA 框架对标签进行分类

LDA 是一种生成模型(generative model)，也就是说，与直接根据观察到的文档来进行预测不同，LDA 首先假设了产生文档的一个过程，然后根据观察到文档，来预测背后的产生过程是怎样的。LDA 假设所有的文档存在 K 个主题（主题其实就是词的分布），要生成一篇文档，首先生成该文档的一个主题分布，然后再生成词的集合；要生成一个词，需要根据文档的主题分布随机选择一个主题，然后根据主题中词的分布随机选择一个词。在基于标签的音乐风格分类方法中， 将用到 LDA 方法对由音乐标签组成的语料库进行建模。根据文献[5],在本文的方法中将标签类比于文档中的词,而将每首歌曲对应的标签向量类比于语料库中的文档。

假设 k 维向量 α 是主题的先验分布的参数，K×V 的矩阵β是主题中词的分布的参数（V 为词的总数），即 $\beta_{ij} = p(w_j|z_i)$就是第 i 个主题中出现词w_j的概率，那么生成一个文档的主题分布、再生成 N 个主题、进而得到这篇文档的 N 个词的概率可以表示为：

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)$$

其中θ是文档的主题分布向量，z 是 N 维的主题向量，w 是 N 个标签组成的向量。由于θ和 z 是训练数据中观察不到的潜在变量，求边缘分布将其从左边消掉：

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta$$

对含有 M 篇文档的语料库 D 来说，$p(D|\alpha, \beta) = \prod_{d=1\cdots m} p(w_d|\alpha, \beta)$，所以：

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d$$

LDA 的训练过程就是求使得$p(D|\alpha, \beta)$最大的参数α 和 β 的值。求得 α 和 β，我们就可以对一篇文档的主题分布，以及每个词所属的主题进行预测，即求：

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

而使用 LDA 对由标签所组成的语料库进行建模恰恰是 LDA 生成文档的逆过程， 也就是说训练集中的标签数据实际上就构成了由多个文档所组成的语料库， 通过 Gibbs sampling 算法的不断迭代,可以求得矩阵 β,也就可以得到每个标签属于某个风格的概率,用$pr(c_k|t_j)$来表示。此时,定义基于标签的音乐风格分类方法如下：

定义集合$X = \{x_1, x_2, \cdots\cdots, x_n\}$为歌曲的集合， 对于每首歌曲$x_i$而言均有与之对应的标签集合 $T(x_i)$。定义集合$C = \{c_1, c_2, \cdots\cdots, c_n\}$为音乐的风格， 使得对$\forall x_i \in X$都有$C(x_i) \in C$。基于标签的音乐风格分类就是要求得函数 F 从而建立由 $X \to C$的映射。函数 F 可由下面的公式定义：

$$c_k = \arg\max_{c_k}\sum_{j=1}^{t}[Freq(t_j, x_i) \times pr(c_k|t_j)]$$

其中$Freq(t_j, x_i)$表示标签t_j被用来标注音乐x_i的次数， t 表示音乐x_i包含的标签的数目。

4.4 数据分析和结论

参赛队号#2854

为了能更好的与该模型提出的方法进行比较，我们使用同样的音乐数据进行测。试测试集由 10 个流行音乐风格和各 100 个艺术家组成（时间有限，理论上 200 个左右最佳），每个风格 10 个艺术家。对每一个艺术家，我们通过 API 接口（Track.GetTopTags）获取与之相关的最热门的100个标签，并用这些标签进行风格分类实验和分类结果分析，计算每个风格的分类准确率。实验采用 k 最近邻分类分类方法，且将 k 值设为 6。每个风格的 10 个艺术家分为两部分，随机选 5 个艺术家作为训练数据，另 5 个作为检验数据。为了尽可能减少实验的随机性，我们进行了次 10 重复实验（时间有限，理论上 100 次最佳）。

实验结果图 4.5 表明，每次实验的平均准确率是上下波动的，但是 10 次实验之后总的平均准确率趋近一个固定的值，大概为 84%。

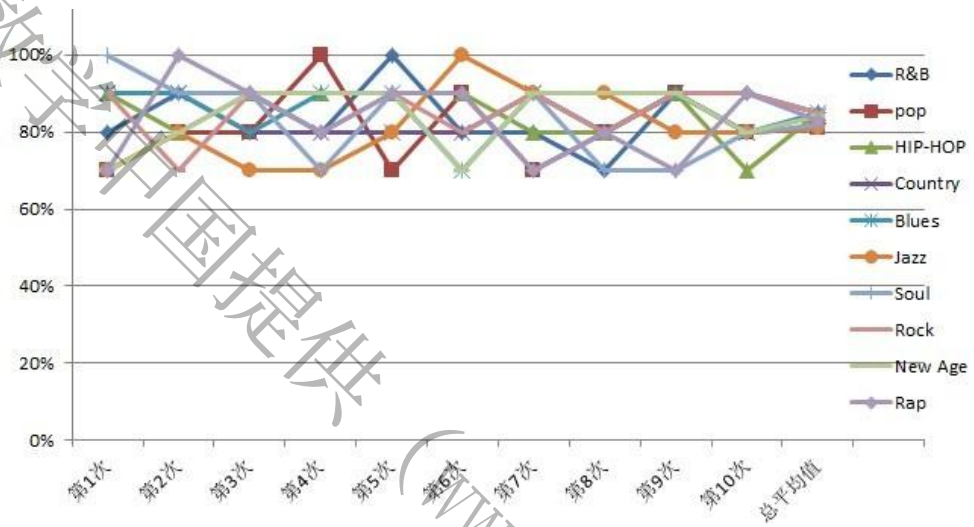

图 4.5：各个风格每次实验的平均准确率及总平均准确率

基于标签的方法对流行音乐风格进行分类的模型是一种比较好的分类方法，自动的音乐风格分类仅仅使用底层声学特征而进行的风格分类往往很难得到好的结果，由此对于网络资源在音乐风格分类的贡献不可忽视。但是，基于标签的方法对流行音乐风格进行分类也有它的局限性，例如对于没有标签的原创歌曲，只能用传统的底层声学特征来进行风格分类。

五、基于互联网语义的流行音乐风格分类模型

我们提出的基于互联网语义关系来对流行音乐风格进行分类是借鉴了C.McKay等人在文章[6]中提出的音乐的文化特征和甑超等人在文章[7]中提出的语义特征这一系列概念。这类的关系包含了相当松散的语义关联性，但是它对于本次模型的建立仍然是有用的。我们可以使用语义网络来呈现逻辑上的描述，例如：对于《东风破》这首歌曲必然在互联网上存在一系列的音乐风格与之相联系，当然也包括各种流行音乐风格。互联网上拥有海量的文本资源，例如某个用户对于音乐的评价、网站对音乐的介绍、音乐专家对音乐的评论文章等等。当我们在网络上抓取数据的时候，如果一段流行音乐总与某个特定的流行音乐风格出现在一起，那么我们就认为此流行音乐与这个风格有着非常紧密的联系，从而我们就可以推断该流行音乐就是属于这个风格的。

参赛队号#2854

5.1 原始数据采集

a) 关系紧密度

在我们的实验中对于每一首歌曲我们都要通过搜索网络资源计算歌曲名称与各个流行音乐风格的联系紧密度以及音乐家与各个流行音乐的联系紧密度。在这里，联系紧密度是我们通过 Google 搜索引擎抓取互联网上相关资源，然后计算两个词语同时出现在一个网页中的概率而得到，计算的公式是：

$$P = M(a,b) / C(a) \qquad \text{公式 5.1}$$

其中，M(a,b)代表 a 和 b 同时出现的网页的个数，C(a)代表只有 a 出现的网页的个数。

Google 搜索表达式分别是”a” AND “b”和”a”,例如：

M(东风破,R&B)的搜索表达式是：

”周杰伦” AND （”R&B” OR “节奏布鲁斯”）

C(周杰伦)的表达式是：

“周杰伦”

并且采用 Google 搜索工具的精确搜索以提高准确性。

b) 训练集数据采集

下面举例说明,我们从流行音乐训练集当中随机抽取的 6 个不同音乐家的流行音乐，其中音乐与音乐家的对应关系如表 5.1 所示。

表 5.1：音乐名称与对应的音乐家

音乐名称	音乐家
东风破	周杰伦
Rolling In The Deep	Adele
逝去的温柔	司徒骏文
Call Me Maybe	Carly Rae Jepsen
我的歌声里	曲婉婷
Baby	Justin Bieber

根据公式 5.1 定义 a=音乐名称，b=音乐风格，分别求出 M(a,b)和 C(a)。如表 5.2 所示。

其中

RB =（”R&B” OR ”节奏布鲁斯”）；

Pop =（”Pop”）;

HH =（”嘻哈” OR “HIP-HOP”）；

CM =（”乡村音乐” OR “Country”）；

BL =（”蓝调” OR “Blues”）；

JA =（”爵士” OR “Jazz”）；

SO =（”索尔” OR “Soul” ）；

RO =（”摇滚” OR “Rock” ）；

NA =（”New Age” OR “新世纪”）；

RP = （”Rap” OR “说唱”）。

参赛队号#2854

表 5.2：不同音乐与不同风格同时出现和单独出现的网页条数

	东风破	Rolling In The Deep	逝去的温柔	Call Me Maybe	我的歌声里	Baby
RB	817000	33,000,000	353,000	359,000,000	7,150,000	1,230,000,000
pop	644000	27,000,000	616,000	86,100,000	9,630,000	437,000,000
HH	1050000	8,950,000	1,030,000	147,000,000	10,100,000	165,000,000
CM	22000	30,700,000	36,000	178,000,000	524,000	254,000,000
BL	527000	13,100,000	1,240,000	24,000,000	6,680,000	277,000,000
JA	854000	11,900,000	1,490,000	22,600,000	1,740,000	226,000,000
SO	892000	24,900,000	1,350,000	63,700,000	7,670,000	657,000,000
RO	900000	45,400,000	3,030,000	111,000,000	3,770,000	1,350,000,000
NA	65200	16,000,000	88,100	108,000,000	1,600,000	312,000,000
RP	1410000	12,800,000	1,270,000	30,100,000	1,510,000	308,000,000
单独	3,020,000	532,000,000	3,500,000	1,190,000,000	25,400,000	1,910,000,000

根据公式 5.1 定义 a=音乐名称，b=音乐风格，以及表 2 中的数据，求出 P = M(a,b) / C(a)。如表 5.3 所示

表 5.3：参数 a 和 b 分别为音乐名和风格时概率 P 的值

	东风破	Rolling In The Deep	逝去的温柔	Call Me Maybe	我的歌声里	Baby
RB	0.27053	0.062030075	0.100857143	0.301680672	0.28149606	0.643979
pop	0.213245	0.05075188	0.176	0.072352941	0.37913386	0.228796
HH	0.347682	0.016823308	0.294285714	0.123529412	0.3976378	0.086387
CM	0.007285	0.057706767	0.010285714	0.149579832	0.02062992	0.132984
BL	0.174503	0.02462406	0.354285714	0.020168067	0.26299213	0.145026
JA	0.282781	0.022368421	0.425714286	0.018991597	0.06850394	0.118325
SO	0.295364	0.046804511	0.385714286	0.053529412	0.3019685	0.343979
RO	0.298013	0.085338346	0.865714286	0.093277311	0.1484252	0.706806
NA	0.021589	0.030075188	0.025171429	0.090756303	0.06299213	0.163351
RP	0.466887	0.02406015	0.362857143	0.025294118	0.05944882	0.161257

根据公式 5.1 的公式定义 a=音乐家，b=音乐风格，分别求出 M(a,b)和 C(a)。如表 5.4 所示。

参赛队号#2854

表 5.4：不同音乐家与不同风格同时出现和单独出现的网页条数

	周杰伦	Adele	司徒骏文	Carly Rae Jepsen	曲婉婷	Justin Bieber
RB	10,200,000	64,400,000	113,000	14,800,000	13,100,000	101,000,000
pop	131,000,000	318,000,000	56,600	106,000,000	14,000,000	624,000,000
HH	14,400,000	125,000,000	104,000	35,500,000	18,800,000	275,000,000
CM	70,500,000	502,000,000	8,180	71,300,000	5,890,000	743,000,000
BL	4,570,000	162,000,000	45,100	27,700,000	685,000	139,000,000
JA	192,000,000	136,000,000	35,400	29,400,000	23,100,000	181,000,000
SO	89,200,000	281,000,000	40,700	53,900,000	10,100,000	365,000,000
RO	126,000,000	672,000,000	218,000	128,000,000	44,100,000	883,000,000
NA	122,000,000	686,000,000	2,410	102,000,000	8,520,000	19,100,000
RP	185,000,000	242,000,000	350,000	56,500,000	23,600,000	3,700,000,000
单独	864,000,000	2,840,000,000	388,000	694,000,000	81,100,000	6,520,000,000

根据公式 5.1 中的公式定义 a=音乐家，b=音乐风格，以及表 2 中的数据，求出 P = M(a,b) / C(a)。如表 5.5 所示

表 5.5：参数 a 和 b 分别为音乐家名和风格时概率 P 的值

	周杰伦	Adele	司徒骏文	Carly Rae Jepsen	曲婉婷	Justin Bieber
RB	0.011806	0.022676	0.291237	0.021325648	0.161529	0.015490798
pop	0.15162	0.111972	0.145876	0.152737752	0.172626	0.095705521
HH	0.016667	0.044014	0.268041	0.051152738	0.231813	0.042177914
CM	0.081597	0.176761	0.021082	0.102737752	0.072626	0.113957055
BL	0.005289	0.057042	0.116237	0.039913545	0.008446	0.021319018
JA	0.222222	0.047887	0.091237	0.042363112	0.284834	0.027760736
SO	0.103241	0.098944	0.104897	0.077665706	0.124538	0.055981595
RO	0.145833	0.23662	0.561856	0.18443804	0.543773	0.135429448
NA	0.141204	0.241549	0.006211	0.146974063	0.105055	0.002929448
RP	0.21412	0.085211	0.902062	0.081412104	0.290999	0.567484663

5.2 互联网语义关系特征向量的表示

当计算出流行音乐名称、音乐家和流行音乐风格的联系紧密度之后，我们就可以定义流行音乐的语义特征了。例如，对于某段音乐的名称与流行音乐风格的联系紧密度用向量 X 表示，其音乐家与流行音乐风格的联系紧密度用向量 Y 表示。那么流行音乐的互联网语义关系特征向量将根据下面的公式计算得到：

$T = m \times X + n \times Y$（其中 m+n=1）

表 5.6 为训练数据集中的互联网语义关系特征向量表。

参赛队号#2854

表 5.6：各个流行音乐的互联网语义关系特征向量表

	东风破	Rolling In The Deep	逝去的温柔	Call Me Maybe	我的歌声里	Baby
RB	0.06821	0.023015704	0.219878328	0.077097454	0.14010365	0.161413
pop	0.125592	0.084931948	0.117923446	0.115972609	0.160456635	0.091782
HH	0.087815	0.033277414	0.214070524	0.049249961	0.200273166	0.038303
CM	0.061225	0.133353089	0.016019583	0.085648095	0.054713414	0.091706
BL	0.043806	0.043222327	0.124277403	0.030356806	0.066052502	0.039626
JA	0.18104	0.036348236	0.126528295	0.032125127	0.214310536	0.036174
SO	0.106995	0.075124603	0.124450231	0.059766763	0.120096711	0.095697
RO	0.132339	0.178742597	0.473721791	0.140280338	0.409514414	0.203814
NA	0.10604	0.181317932	0.007829542	0.112541405	0.080349979	0.040897
RP	0.198528	0.064190893	0.682600983	0.06138565	0.218754529	0.041241

5.3 多个音乐风格同时存在的判断

当然，有时候一首流行音乐同属 2 种或 3 种风格，此时，我们提出的解决方案是：

假定 T(i)是某首流行音乐与各个风格之间互联网语义关系特征向量绝对值经过排序之后的集合，T(i)>T(i+1)。伪码如下：

```
printf ("s% ", b[1]);          //输出互联网语义关系特征最明显的流行音乐风格
for (i = 1; i <= 2; i++)       //最多输出3种流行音乐风格
{
    float x = T(i+1)/T(i)*100;
    if (x >= 95)
        printf("%s ",b[i+1]);  //如果T(i+1)和T(i)的值非常接近，输出T(i+1)所对应的音乐风格
    else
        break;                 //否则，停止判定
}
```

图 5.1 判断音乐是否存在共存风格的伪码

5.4 数据分析和结论

经过几十次的测试，我们得出的实验数据表明，当 m=0.25，n=0.75 时对流行音乐的风格分类的准确性最高，可以达到 83.2%。说明音乐家和流行音乐风格的紧密度 Y 的权重更大，互联网语义关系特征越明显，而流行音乐名称与流行音乐风格之间的互联网语义关系特征稍弱。

基于互联网语义关系来对流行音乐风格进行分类是非常好的一个研究方向，由于时间有限，本论文此次只是提出这个模型，实验结果虽然较好，但是不能以一概全，需要用大量数据进行佐证。本模型可研究的潜力非常之大，它不仅与语义网络紧密相连，同时又和数据挖掘技术密切相关，这无疑为我们在以后的研究中提供了一个新的方向。

六、基于 LDA 和多类 SVM 的流行音乐风格分类模型

传统的基于声学特征的音乐风格分类方法最早由 George Tzanetakis 等人在 2002 年提出[8]，其过程大致可分为三个阶段：

- 基于短时音频帧的特征提取过程：在这个过程中一些描述音乐音色、节奏和音高的底层声学特征被计算出来；
- 特征选择过程：使用特征选择算法降低特征向量的维数，同时去除无关和冗余特征；
- 分类过程：使用模式识别及分类算法对特征向量进行处理，从而对音乐进行自动的风格分类。

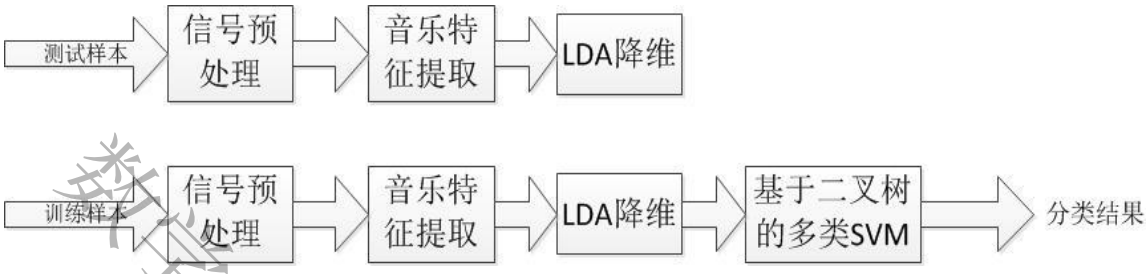

图 6.1 基于声学特征的音乐风格分类方法

如图 6.1 所示，首先随机地选定训练样本集合测试样本集，其中的样本都依次通过信号预处理、音乐特征提取和 LDA 降维这 3 个步骤，然后使用训练样本集中的特征向量对分类器进行训练，最后使用训练好的分类器对测试样本集中的测试特征向量进行分类实验，得到最终整个改进算法的实验性能结果。

6.1 音乐特征分析

有效提取并以正确的方式使用特征是进行模式识别和机器学习的前提。基于音频信号时频域分析的声学特征是音频信号分析和处理的基础，而处于声学特征与高层内容属性之间的乐理特征是联系两者的纽带。本章所提出的乐理层特征是指依据乐理知识提取的一系列特征和特征模式，反映了音乐基本元素在音频信号中的体现方式，是从音乐角度对声音信号的进一步抽象。

无论是音乐的聚类还是分类，基本思路都是将待分析的问题分解成可以被计算机量化处理的感知要素或音乐要素，然后对这些量化的要素进行建模以找到他们与类别之间的对应关系。因此，使用不同抽象层次的特征可以将复杂问题细化为若干复杂度较低的子问题，有助于在求解子问题时排除不必要的干扰因素，充分的利用先验知识建立更加细致的数学模型，以降低求解算法的复杂度并提高求解精度。

a) 音色特征

声音信号的声学特征反映了其在时间和能量上的基本物理特性。音乐音频的音色（Timbre）、节奏（Rhythm）和小波变换系数（Walet coefficients）直接影响人耳的对声音的感知体验。

1. 频谱质心（Spectral Centroid）

频谱质心又称为频谱亮度（Brightness），其定义为 FFT 能量谱中经能量加权的频率均值

参赛队号#2854

$$SC = \frac{\sum_{k=1}^{K}(E_k \times f_k)}{\sum_{k=1}^{K} E_k}$$

其中f_k和E_k表示 FFT 能量谱中第 k 个分量所对应的频率值及其能量，K 为由 FFT 变换长度和采样率决定的 FFT 谱中频率分量的数目。频谱质心可以看做频谱能量分布的均值（一阶矩），反映了声音的明亮程度。频谱质心不仅仅受演奏音符的基频影响还和乐器共振腔的特性密切相关。例如，作为低音乐器的长号由于铜管的共振效应高次谐波的能量得到加强，与相同音域的其他乐器相比声音显得明亮有力。

2. 频谱带宽（Spectral Bandwidth）

频谱带宽的定义为 FFT 谱中所有频率分量与频谱质心之间的距离的平方经能量加权后的平均数的平方根

$$SB = \sqrt{\frac{\sum_{k=1}^{K}[E_k \times (f_k - SC)^2]}{\sum_{k=1}^{K} E_k}}$$

频谱带宽描述了频谱能量分布的标准差（二阶矩的平方根），反映了频谱能量的集中程度，取决于乐器的共振峰的位置的形状与演奏时的音高的相关性不大。

3. 频谱滚降度（Spectral Rolloff）

频谱滚降度的定义为使得低频累积能量达到总量的一定百分比时的最小频率边界

$$SR = \left\{ f_k \,|\, \min(k \,|\, \frac{\sum_{i=1}^{k} E_i}{\sum_{j=1}^{K} E_j} > \gamma\%) \right\}$$

其中$\gamma\%$为低频累积能量占总能量的比例，对于音乐信号$\gamma\%$的取值可以设置在 85%左右。频谱滚降度反映了信号的频谱能量分布的高频衰减速度（又称倾斜度，steepness）。

4. 频谱通量（Spectral Flux）

频谱通量定义为一段时间内相邻两帧之间的频谱能量变化量的均值，本文使用 2 范数距离（欧氏距离）表示频谱向量间的变化量

$$SF = \frac{1}{N-1}\sum_{i=1}^{N-1}\sqrt{\sum_{k=1}^{K}(E_k(i) - E_k(i+1))^2}$$

其中 $E_k(i)$表示第 i 帧频谱向量中第 k 个频率分量的能量值，N 为总帧数。以 FFT 频谱能量分布的一阶差分为定义的频谱通量反映了音乐信号的平稳程度。

频谱形状反映了信号 FFT 能量谱分布的总体特性，频谱对比特征则进一步将频谱划分为若干子带，对每一个子带中的能量的分布进行了更为细致的统计分析。由前文的分析知，当外界激励经过乐器共振腔时，受到共振效应的影响不同频率能量重新分配，

有的频段能量得到强化，有的频段能量受到抑制，因此子带分析方法能够刻画乐器共振曲线的差别，有助于全面描述声音信号的物理特性。对于音乐音频一般采用基于对数比例的八度音程标准划分频域，以便于音乐理论保持一致。

设第 s 子带所包含的频率在 FFT 频谱向量中对应的分量为$\{s_1, s_2, \cdots\cdots, s_M\}$，其中$s_1 < s_2 < \cdots s_M$，$M_s$为该子带中频率分量的个数，它们对应的能量为$\{E_{s_1}, E_{s_2}, \cdots\cdots, E_{s_M}\}$，则子带能量定义为该自带内所有频率分量能量的总和

$$SubEng_s = \sum_{i=1}^{M_s} E_{s_i}$$

将频谱分量的能量按降序排列得到能量有序序列$\{E_{s_1'}, E_{s_2'}, \cdots\cdots, E_{s_M'}\}$，其中$E_{s_1'} > E_{s_2'} > \cdots > E_{s_M'}$，则子带能量峰值定义为能量有序序列中具有最大能量的前η（依据经验η设置为 20%左右）的频率分量的平均能量，即：

$$SubPeak_s = \frac{1}{\eta\% \times M_s} \sum_{i=1}^{\eta\% \times M_s} E_{s_i'}$$

同理子带能量谷值的定义为

$$SubValley_s = \frac{1}{(1-\eta\%) \times M_s} \sum_{i=(1-\eta\%) \times M_s}^{M_s} E_{s_i'}$$

为了描述子带内峰值和谷值之间的差异定义子带对比度

$$SubCtr_s = SubPeak_s / SubValley_s$$

子带对比度放映了子带内部能量分布的离散情况，当能量集中分布在少数频率分量上时对比度较高，反之对比度较低。

b) 节奏特征

前文讨论了在单音音乐中如何结合参考乐谱提取音符的起始点并进而分析出音乐的节奏。但是市场上出售的唱片往往不提供乐谱，同时唱片进过混音后各原始音轨的信号相互叠加进一步加大了提取节拍点的难度。本小节所述的节奏提取方法将不再试图精确定位节拍点的位置，而是以长度为 5-10 秒的时间段为基本单位，将音乐信号的波形转化为节奏强度曲线，从中提取反映声音节奏特性的可量化计算的参量作为节奏强度、规律性和速度的度量。

敲击强度曲线（Onset curve）的提取过程如图 6.2 所示。首先以八度音程为间隔将频谱划分为带宽呈指数增加的子带，构造带相应的通滤波器组将音乐信号划分成若干子带。在每一个子带内进行以下四步操作

- 为了去信号波形中存在的毛刺，利用升余弦窗（ Raised Cosine Window）的低通特性将子带信号与余弦窗卷积以提取波形的包络；
- 参考图像处理中经典的边缘提取算法，将振幅包络曲线与高斯内核的 Canny 算子进行卷积，得到振幅包络的差分曲线；
- 差分曲线经过半波整流得到整流差分曲线
- 对所有子带的整流差分曲线求和得到最终的敲击强度曲线

参赛队号#2854

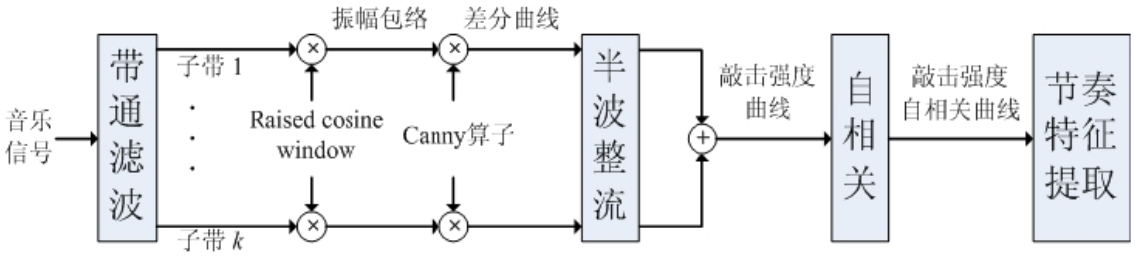

图 6.2 敲击强度曲线（Onset curve）的提取过程

敲击强度曲线$OC(i)$（其中$1 \le i \le L$，L 为敲击强度曲线的长度）中的峰值指示信号中能量急剧增加的位置。参考音高提取的自相关算法，计算敲击强度曲线的自相关函数$AOC(i)$。如果音频中周期性的出现敲击点，则在敲击点周期及其整数倍的位置自相关函数将出现峰值点，峰值点的强度可以作为敲击周期性强弱程度的度量。

c） 小波变换系数直方图（Wavelet Coefficients Histogram）

在前文所述的音色特征基于信号短时分析计算的局部帧级特征。由于各帧之间独立分析丧失了时间信息，无法对某一时间段内的频域特性或某一频率段内的时间特征进行建模，无法分析信号的时变特征（如特征向量的突变、偏移、趋势等）。而节奏特征则是对一段信号进行了时域上的动态关联性分析（差分、自相关）得到的长期性的全局特征，但是当音乐的节奏发生变化时用统计方法计算全局特征屏蔽了不同时间段内节奏的差别，不利于全面的刻画信号的特性。

与傅里叶变换通过平移固定宽度的滑动窗口分析得到固定时域和频域分辨率的思路相比，小波变换保持窗口面积不变窗口形状随分析频率的高低改变，能够保证低频段有较高的频率分辨率和较低的时间分辨率，在高频段有较高的时间分辨率较低的频率分辨率。通过小波变换能够将信号分解成不同层次的逼近分量和细节分量，因此基于小波变换提取的特征能够同时兼顾信号全局信息和局部信息的时频域表征能力。实践证明，在特定分析任务中小波变换系数能够很好的刻画同类音乐音频之间的共性同时也保证了不同类别音乐之间的可区分性。

从信号中提取小波变换相关特征的步骤如下：

- 将输入音频切分为 5-10 秒的片段；
- 使用 8 阶多波西小波（8-order Daubechies Wavelet, Db8）滤波器组对信号进行 7 级分解得到 8 个子带的小波变换系数。
- 对每一个子带的小波变换系数求直方图，并依据直方图中所有分量总和对各分量进行归一化
- 子带内系数分布用直方图的 1 阶矩（均值）、2 阶矩（方差）和 3 阶矩（偏度）描述。

6.2 语音信号预处理

文中所有音乐文件为 wave 格式, 采样率 11.025kHz,16 bit/sample。在特征提取之前, 要对每一段音乐进行预处理。首先对信号进行预加重(参数为 0.96), 以提升高频部分的能量; 然后对每一首音乐进行分帧, 帧长 256 点(约 23 ms),相邻帧之间有 128 点(50%)的重叠,每一帧都用汉明窗进行处理; 最后计算每一帧的能量,若该帧能量

低于一阈值(10，经验值)时，该帧被判为静音帧而被抛弃。

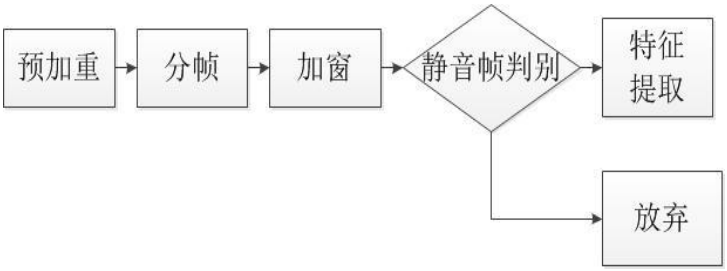

图 6.3 语音信号预处理过程

6.3 音乐特征提取

文中提取的特征包括:

- 感知特征(Perc)，包括帧能量、子带能量、谱质心、带宽和基音频率;
- Mel 倒谱系数。它们的定义如下，其中 F(ω) 是每一帧信号的离散傅里叶变换。

帧能量：该特征用每帧信号总能量的对数来表示,

$$P=\lg\left[\int_0^{\omega_o}|F(\omega)|^2\right]d\omega$$

其中，$|F(\omega)|^2$表示频率ω处的谱密度，$\omega_o = 5.5125\text{kHz}$ 为信号采样频率的一半。

子带能量:信号的频谱被分为 4 个子带，$[0,(\omega_o/8)]$，$[(\omega_o/8),(\omega_o/4)]$，$[(\omega_o/4),(\omega_o/2)]$，$[(\omega_o/2),\omega_o]$在每一个子带中计算子带总能量的对数，

$$P_j = \log\left[\int_{L_j}^{H_j}|F(\omega)|^2\right]d\omega。$$

谱质心:

$$\omega_c = \frac{\int_0^{\omega_o}\omega|F(\omega)|^2\,d\omega}{\int_0^{\omega_o}|F(\omega)|^2\,d\omega}$$

带宽：信号频谱成分与谱质心之差以能量进行加权的均值,

$$B = \sqrt{\frac{\int_0^{\omega_o}(\omega-\omega_c)^2|F(\omega)|^2\,d\omega}{\int_0^{\omega_o}\omega|F(\omega)|^2\,d\omega}}$$

基音频率：基音频率是指发浊音时声带振动的频率，它已广泛地应用于语音识别和分类领域。笔者采用两种方法求单帧信号的基音频率并取它们的均值来代表该帧信号的基音频率特征。第一种方法[6]采用 Yule-Walker 法估计经低通滤波后信号的 AR 谱，然后在一定范围内(50~500 Hz) 进行峰值检测，当峰值超过一定门限时，记录峰值位置为基音频率；第二种方法[7]对经低通滤波后信号进行中心削波，然后计算削波后信号的归一化自相关函数，并对自相关函数在一定范围内(50~500 Hz)进行峰值检测，计算基音频率。

Mel 倒谱系数(Mel-Frequency Cepstral Coefficient,MFCC)：MFCC 考虑了人耳的听觉特性，将线性频率变换为基于 Mel 频率的非线性频谱(1 kHz 以下呈线形关系，而

在 1 kHz 以上时呈对数关系)，然后再将其转换到倒谱域上。其计算过程为：求出能量谱，并用 K 个 Mel 带通滤波器(类似于耳蜗作用的滤波器组) 进行滤波，将每个滤波器频带内的能量进行叠加，假设第 k 个滤波器输出功率谱为 x(k)，则相应 MFCC 系数为(L 为 MFCC 的维数)

$$c_n=\sqrt{\frac{2}{K}}\sum_{k=1}^{K}[\lg x(k)]\cos[n(k-0.5)\pi/K],$$

$$n=1,2,\cdots,L$$

图 6.4 Mel 倒谱系数提取流程

6.4 使用 LDA 对特征向量进行降维

在每一段音乐中提取出代表该音乐的特征向量之后，对于一般的音乐分类方法，它们直接使用这些向量对分类器进行训练，并把这种向量用于最终的音乐分类，即性能测试阶段中。但是此时得到的特征向量维数很高，若直接用于分类的话，不但会碰到计算复杂度太高或存储空间需求太大的问题，而且还可能会遇到维数冗余(excessive dimensionality)的困难。针对这些问题，一般有两种可行的解决方法：一是重新设计特征提取器，选择现有特征的一个子集用于音乐分类：二是通过某种方式(如线性组合)来组合现有特征，得到全新的低维特征集。

使用线性组合的方法对现有特征进行降维是计算机视觉中最常见的处理方法，这是因为它不但易于计算而且易于分析"实际上，线性方法将高维空间的矢量投影到较低维的空间中，也就是利用已有特征参数构造一个较低维数的特征空间，将原始特征中蕴含的有用信息映射到少数几个特征上，忽略多余的不相干信息"从数学意义上讲，就是对一个 n 维向量 $X=[x_1,x_2,\cdots,x_n]^T$ 了进行降维，变换为低维向量 $Y=[y_1,y_2,\cdots,y_m]^T$，$m<n$其中 Y 确实含有向量 x 的主要特性。当前已有两种经典方法用于寻找降维中有效的线性变换：

- 主成分分析(Principal Component Analysis or PCA)，寻找一个投影方向使投影后的数据能在最小均方意义上最好的代表原有的数据：
- 线性判别分析(Linear Discriminative Analysis or LDA)，寻找一个投影方向使投影后的数据在最小均方意义上得到最好的分离。

下面介绍这两种方法在线性降维处理中的具体实现。

a) 主成分分析（PCA）

PCA 方法也称为 Karhunen-Loeve 方法，它选择一个降低维数的线性投影，使得所有投影样本的离散度(scatter)最大。

假设在样本集中有 N 个样本向量行$\{\overrightarrow{x_1},\overrightarrow{x_2},\cdots,\overrightarrow{x_N}\}$它们的维数为 n，而每一个样本向量都属于 c 个类别行$\{\overrightarrow{X_1},\overrightarrow{X_2},\cdots,\overrightarrow{X_c}\}$中的一类。假设存在一线性变换将 n 维的原样本向量映射到 m 维的特征空间中，m<n。则新的样本向量$\{\overrightarrow{X_1},\overrightarrow{X_2},\cdots,\overrightarrow{X_c}\}$可以表示为：

参赛队号#2854

$$\overrightarrow{y_k} = W^T \overrightarrow{x_k} \qquad k = 1,2,\cdots,N$$

其中$W \in R^n$称为变换矩阵，W 的所有列之间都是正交的。总体离散度矩阵(total scatte matrix) S_T：定义为：

$$S_T = \sum_{k=1}^{N} (\vec{x}_k - \vec{\mu})(\vec{x}_k - \vec{\mu})^T$$

其中 N 是样本的数目，$\vec{\mu} \in R^n$是所有样本的均值。

对原样本向量实施线性变换$\vec{\mu} \in R^n$得到变换后向量伽，$\{\overrightarrow{y_1}, \overrightarrow{y_2}, \cdots, \overrightarrow{y_N}\}$的离散度矩阵为$W^T S_T W$. 在 PCA 中，投影方向$W_{opt}|$的选择是使得投影后向量总体离散度矩阵的行列式最大的矩阵 W，即：

$$W_{opt} = \arg\max_{W} |W^T S_T W|$$

$$= [\vec{w}_1 \quad \vec{w}_2 \quad \cdots \quad \vec{w}_m]$$

其中$\{\overrightarrow{w_t} | i = 1,2,\cdots,m$是对应于$S_T$的 m 个最大特征值的 n 维特征向量.

这种方法的缺点是总体离散度的最大化不仅要求类与类之间离散度的增大，而且要求类内离散度的增大，因此对于分类目的来说，这是一个障碍。由此可知，PCA 投影是由投影后特征恢复原特征的最优化解，但从分类的角度上说，PCA 并不是最佳的投影选择. 下面介绍本文使用的降维方法——线性判别分析.

b) 线性判别分析（LDA）

LoA 也称为FLo(FISherLinearoiscriminative)[8],它基于这样的一个假设:在理想情况下,类内的变化都存在于原特征空间的一个子空间中,因此各类是凸起并线性可分的,可以在使用线性投影的方法实现降维的同时保持类间的线性可分性。这就是人们喜欢在模式分类中使用线性方法进行降维的一个很重要的原因。LDA 的目标是找到一个投影方向,使得投影后的子空间中不同类间样本的分离度尽可能大一些,而类内样本尽量密集,即类内离散度越小越好。这样在经过LDA的处理后,将降维后的特征用于音乐分类,可得到更高的分类精确度。

定义类间离散度矩阵S_B为:

$$S_B = \sum_{i=1}^{c} N_i (\vec{\mu}_i - \vec{\mu})(\vec{\mu}_i - \vec{\mu})^T$$

定义类内离散度矩阵S_W为:

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (\vec{x}_k - \vec{\mu}_i)(\vec{x}_k - \vec{\mu}_i)^T$$

其中μ_i是类X_i中特征向量的均值，N_i是类X_i中的特征数。若S_W是非奇异的,则最佳投影W_{opt},的选择是使得投影后类间离散度矩阵的行列式与类内离散度矩阵的行列式比值最大的 W,即:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} = [\vec{w}_1 \quad \vec{w}_2 \quad \cdots \quad \vec{w}_m]$$

其中$\{W_t | i = 1,2,\cdots,m$的广义特征向量,它们分别对应于所有广义特征值中 m 个最大的广义特征值$\{\lambda_i | i = 1,2\cdots,m\}$,即:

参赛队号#2854

$$S_B\vec{w}_i = \lambda_i S_W\vec{w}_i \quad i = 1,2,\cdots,m$$

在上面这个式子中最多存在 c - 1 个非零的广义特征值,因此降维后特征向量的维数上限是 c - 1,c 是待分类的类别数。

图 3-12 是一个两维分类的例子,其对比了 PCA 和 LDA 在投影方向上的不同,图中 N = 20,n = 2,m = 1 由图中可见,PCA 和 LDA 都将特征点从两维投影到一维,具体比较两者的不同,发现 PCA 虽然使得点与点之间的距离变大,但实际上是把两类点混淆了在一起,导致不能用线性的方法将它们分开;而 LDAF(LD)不但保留了类别间的可分性,而且取得了更大的类间离散度,因此使得最终分类更加简单，而可以期望最终分类的精确率也会得到提高。

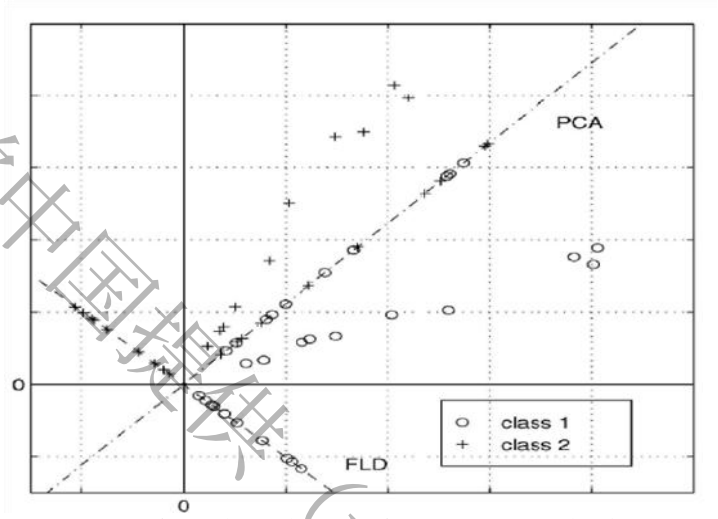

图 6.5 在两类分类问题中 PCA 及 LDA 的应用

在实际的实验计算过程中，可能会遇到类内离散矩阵S_W是奇异的情况，例如，若用于训练的样本数较少就可能导致这种情况。为了解决S_W奇异的这种情况，可以先将样本向量投影到一较低维空间（N - c 维），使得在该空间中类内离散矩阵S_W为非奇异的，这可以通过 PCA 来实现；然后再通过标准的 LDA 将所得特征向量降维至 c - 1 维。具体公式如下：

$$W_{opt}^T = W_{lda}^T W_{pca}^T$$

其中：

$$W_{pca} = \arg\max_W \left|W^T S_T W\right|$$

$$W_{lda} = \arg\max_W \frac{\left|W^T W_{pca}^T S_B W_{pca} W\right|}{\left|W^T W_{pca}^T S_W W_{pca} W\right|}$$

本论文采用 LDA 对从原始音乐提取的高维特诊向量进行降维，得到低维的输出变换向量，如下图：

参赛队号#2854

图 6.6 LDA 用于高维音乐特征

原特征向量为 n 维，变换后的向量为 c 维，n > c，变换后的 c 维向量将代替原 n 维向量进入分类器进行音乐分类。由于投影后的 c 维子空间中不同类间样本的分离度较大，而同类的样本尽量密集，因此能有效的提高最终的音乐分类精确率。

6.5 使用 SVM 对特征向量进行分类

a) 针对线性不可分问题的非线性映射问题

音乐分类问题可视为一多维空间中的线性不可分问题。为了解决线性不可分问题，V.Vapnik 引入了核空间理论:将低维输入空间中数据通过非线性函数映射到高维属性空间 H（也称为特征空间），将分类问题转化到属性空间进行。可以证明，如果选用适当的映射函数，输入空间线性不可分问题在属性空间将转化为线性可分问题。如下图 6.7 所示：

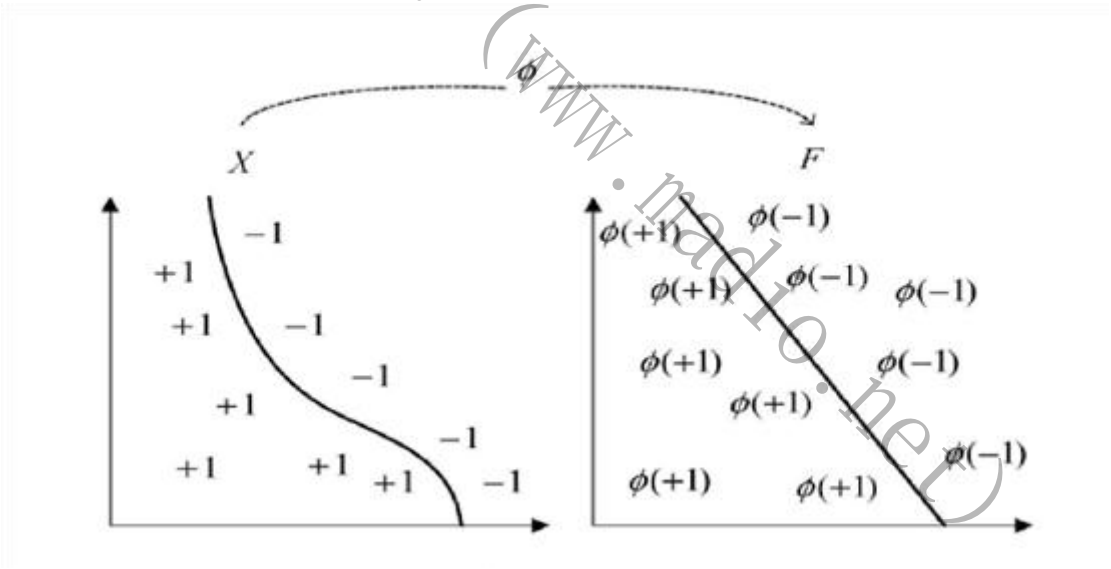

图 6.7 线性不可分问题的非线性映射问题

其中$\emptyset(1)$代表一非线性变换，它将输入空间 X 映射到高维的特征空间 F，由图可见，原来在 X 中线性不可分的特征点映射到 F 后变为线性可分。

不同的非线性变化$\emptyset(1)$对应着下式中不同的核函数$K(x,\overline{x})$，目前经常用于模式分类中的核函数有：

1.ERBF 函数：

$$K(x,\bar{x}) = \exp\left(\frac{-|x-\bar{x}|}{2\sigma^2}\right)$$

2. 高斯函数：

$$K(x,\bar{x}) = \exp\left(\frac{-|x-\bar{x}|^2}{2\sigma^2}\right)$$

3. 多项式函数：

$$K(x,\bar{x}) = \left(\langle x,\bar{x}\rangle + 1\right)^d$$

b) 由两类分类器组合成为多类分类器的二叉树结构

典型的 SVM 是两类分类器，能将测试样本分入正类（+1）或负类（-1）中。因此，为了实现多类分类，必须采用特定的策略来组合数个两类分类器以得到多类分类器。

本论文采用一种由低向上的二叉树分类策略，二叉树的每一个叶节点代表一个音乐类别，对于每一个非叶节点，在它的两个子节点之间使用 SVM 进行两类分类，胜利者被保留并进入上一层的分类，分类不断的向上进行直到抵达树根处并得到最终的分类结果，具体过程如下图 6.8 所示

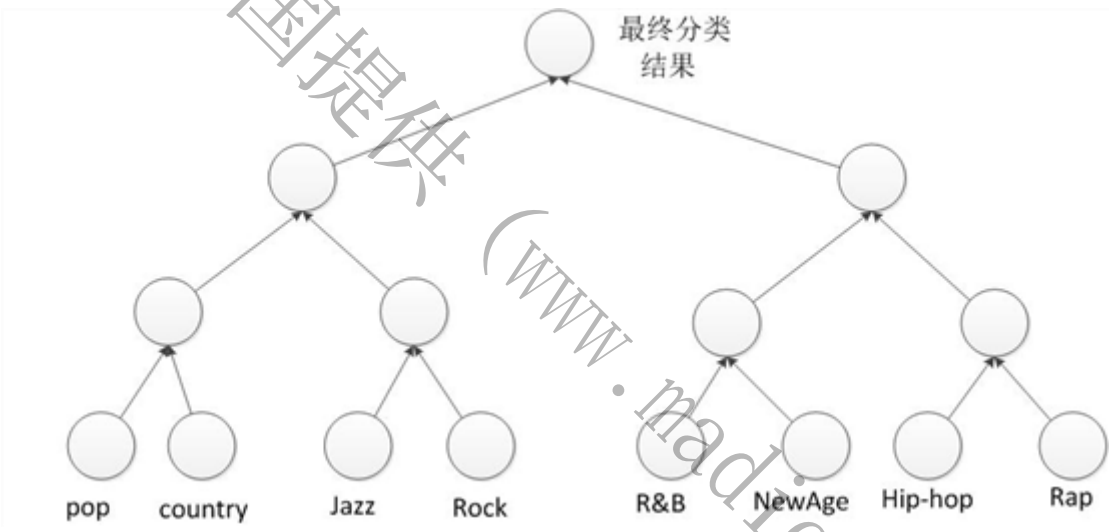

图 6.8 由低向上的二叉树分类策略

可以证明这种由底向上的二叉树分类方法不但能取得最佳的分类精确率，而且也能实现最好的时间复杂度。

6.6 数据分析和结论

a) 算法实现流程图

参赛队号#2854

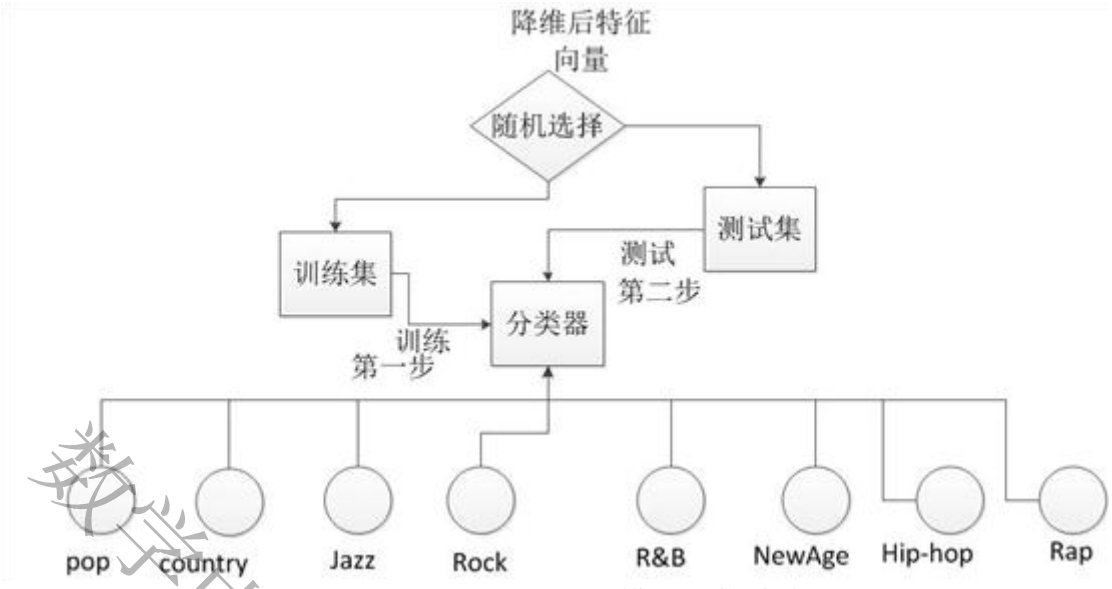

图 6.9 基于 LDA 和 SVMs 模型的算法流程图

b) 算法实验过程

1. 对每一首音乐进行分帧处理,帧长 256asmpel(s 约 23m)s,相邻帧之间有 128 点(50%)的重叠,这样就基本保证了语音信号在每一帧内保持短时平稳;
2. 为了信号的频谱变得平坦,便于下面进行频谱分析和声道参数的提取,在预处理中进行预加重处理,这通过一阶的预加重数字滤波器来实现(参数为 0.96):

$$s'_n = s_n - 0.96 \times s_{n-1} \qquad n = 1,\cdots,255$$

其中s_n为每帧信号中的第 n 个采样值,而$s'_0 = s_0$;

3. 为了克服对原语音信号加矩形窗所带来的频谱泄漏现象,对每一帧信号加汉明窗处理:

$$s_i^h = s_i' * h_i \qquad n = 1,\cdots,255$$

其中汉明窗 $h_i = 0.54 - 0.46 \times \cos(2\pi i/255)$。

4. 进行静音帧判别,并抛弃静音帧(即短时帧能量低于某一网值的音乐帧)。
5. 经过上述步骤的处理后,可从每一段音乐中提取音乐特征组成特征向量,并用于后继的分类器中进行音乐分类。具体提取的特征如表 6.1 所示:

表 6.1

特征		特征数
感知特征	帧能量	1
	子带能量	4
	谱质心	1
	带宽	1
	基音频率	1
Mel倒谱系数		L

由表 6.1 可知，由每一帧非静音信号可提取 8 个感知特征和 L 个 MFCC 系数（L 可根据实际最终分类精确类进行调整），假设在每一段音乐中（5 秒）有 N 帧非静音，则意味

着可以在该段音乐中提取出 N 个 8 维的感知特征向量和 N 个 L 维的 MFCC 特征向量，针对这两种不同特征的具体处理如下描述：

- 感知特征：计算每一段音乐 8 维感知特征向量的均值和标准差，并将它们级联到一个 8X2=16 维的特征向量，再加上基因比率特征（检测出有基音的帧数/总帧数）就得到了代表感知特征的 17 维特征向量。接着，在整个训练集中对每一维特征进行归一化：

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \qquad i = 0,1,\cdots,16$$

其中X_i是感知特征向量中的第 i 个元素，μ_i是训练集中所有感知特征向量的第 i 个元素的均值，σ_i是训练集中所有感知特征向量的第 i 个元素的标准差。这样就得到了最终的感知特征向量集（17 维）。

- MFCC 特征：同 1 中的步骤计算每一段音乐信号 L 维 MFCC 系数的均值和标准差，但是并不进行归一化处理，这样就得到了 2L 维的 MFCC 特征向量。

对感知特征进行归一化而 MFCC 特征向量并不进行归一化的原因是：按照一般经验，针对感知特征向量的归一化能提高最终的音乐分类精确率，而与此相对，MFCC 特征向量的归一化将导致精确率的下降。如今已分别得到了感知特征和 MFCC 特征向量，下面的问题就是如何将它们按照一定的权重级联成一个长的向量来代表它们所对应的那一段音乐。具体的级联过程如下：

由于存在 17 个感知特征及 2L 个 MFCC 特征，而感知特征已进行归一化处理，所以其标准差之和为 s1=17x1，而 MFCC 特征未经归一化，因此其标准差之和为$s_2 = \sum_{i=1}^{2L} \sigma_i$，其中$\sigma_i$为 MFCC 特征向量中第 i 个元素的标准差。考虑到这两个不同特征集的相对可靠性，最终联级的形式为（感知特征向量/s1）异或（MFCC 特征向量/s2），由此便得到代表原音乐的（17+2L）维的特征向量。

c） 数据分析

研究选用 4 种音乐类型:古典、爵士、流行和摇摆。实验数据库包含 400 首音乐录音,16 位 PCM 格式,采样速率 22.050KHz,每首截取 30 秒长度片断。400 首录音出自不同作品。每种音乐类型 100 首录音。数据库录音来自 CD、mp3、数据库和无线电广播。每类 50 个样本作为训练集, 50 个样本作为测试集。由于小波系数值较大,提取特征时需要作归一化处理,提高识别精度。训练测试 SVM,循环迭代参数 C 和 D,从实验中得到精度最高的 C 和 D。当存在多个相同的最高测试率时,取支持向量少的那组参数作为最优参数。窗口长度 4096 个采样点。对每种音乐类型提取特征向量,使用不同的子带和子段的数目进行测试。实验中使用 6 种组合形式用 SVM 分类。实验结果见表 1。整个实验在 VisualC++ 6. 0 平台下开发完成。

表 6.2

参赛队号#2854

特征方法	子段数	子带数	正确率%					
			古典/流行	古典/摇摆	古典/爵士	爵士/流行	爵士/摇摆	摇摆/流行
1	1	16	83.7	93.3	77.7	73.3	85.9	81
	1	32	83.1	93.3	80	73.7	88.1	85.9
	2	64	82.5	93.0	79.0	73.6	89.7	83.8
	4	64	80.4	92.8	80.4	73.4	90.6	82.8
2	2	64	75.6	93.7	78.2	83.0	86.5	82.8
	4	64	71	93.6	82.5	82.4	88.5	82.1

不同音乐类型间组合的正确率， SVM 分类音乐类型,分类正确率可达 86%,尤其对古典/摇摆这对类型分类表现最好。表明该方法合理、有效。

d) 结论

通过对实验数据的分析，证明了本论文所提出的音乐分类结构对最终音乐分类性能的提升作

用,并显示 LDA 降维处理和 SVMS 分类器在音乐模式分类方面的优良性能。具体来说,实验数据说明了:

1 对原音乐特征进行 LDA 降维处理有效的提高了最终的音乐分类正确率,这个结论对所有的分类器都成立。

2.提取的 MFCC 系数的维数对分类的结果有相当影响,可根据具体实验数据及结果选择。

3.音乐分类的正确率随待分类音乐的内在相似性改变,若待分类音乐之间本来就存在相似性,则音乐分类系统的分类正确率随之下降。

七、基于分形维数的流行音乐风格分类模型

7.1 音乐与 1/f 噪声

在音乐创作中，作曲者通过变奏、反复、发展等方法对主旋律进行多次变换，从而构造出一首歌曲的节奏、音程、力度及音调等音乐基本特征。而从数学的角度看，这些变换相当于随机自仿射变换，这样就构造出了某种程度的自相似性。Voss R.F.发现几乎所有的音乐都在模仿$1/f$噪声（其中 f 为噪声的频率）[9-10]。

对于不同风格和文化的流行音乐，目前还不清楚其低于 20Hz 区域$1/f$噪声的特点。我们将分析在七个不同风格的流行音乐（classic, hip-hop, newage, jazz, rock, rap and pop）中的$1/f$噪声。为了便于分析，我们用了 20 首不同风格的音乐，记录格式为 16KHz，16 位和单声道类型。我们通过分析低于 20Hz 区域的$1/f$噪声发现，在该区域的$1/f$噪声可能不会显示音乐的文化特点，但可以明显的区分音乐风格。这意味着我们可以通

参赛队号#2854

过分析低于 20Hz 区域$1/f$噪声的特点来区分音乐的风格。

a) 背景

自然有各种各样的声音。其中一个被称为噪音。它是已知的，根据其行为的频率[1]，噪声有三种不同的类型，如$1/f^0$，$1/f^1$和$1/f^2$。DNA 序列也显示$1/f^a$的结构特性。Voss 等人发表的论文，分析了 1 / f 在音乐和语音中的结构特性 。在 Voss 等人的论文发表后，已经有 一些关于$1/f$和音乐之间的关系研究的。事实上，$1/f$意味着，有一个长程相关性如 bigerelle 等。关于$1/f$的分析可以用于对音乐进行分类，在整个频率范围内进行的$1/f$分析的资料有很多，而且在低频率范围的$1/f$分析更加重要。此外，还不能确定低于 20 Hz 的$1/f$是否能区分不同文化和风格的音乐。

b) 噪声功率谱分类

- 白噪声：
  白噪声是其相关处处为 0。这意味着，在白噪声，其不包含任何信息。它在整个频率范围内具有相同的功率谱。因此，它表现为$1/f^0$。热噪声实际上是与已知的白噪声有非常相似的结构特性。
- 布朗噪声：
  布朗噪声由布朗运动产生，是小颗粒在液体中的运动。有时它被称为红色的噪音。布朗噪声显示光谱行为$1/f^2$，这意味着，这将非常强烈依赖与它的过去状态。
- 粉红噪声（$1/f$）
  粉红噪声即$1/f$噪声。并且功率谱介于白噪声和棕色的之间，这意味着它可能会显示出长程相关性的特点。

c) $1/f$噪声分析

为了便于分析，我们用了 20 首不同风格的音乐，记录格式为 16KHz，16 位和单声道类型。为了得到长期的$1/f$结构特性，分析了音乐的整个波形文件。得到每个波形文件的功率谱的傅立叶变换。

图 7.1 显示了 classic 的功率谱，它清楚地表明了在低频区的$1/f$结构特性。然而图 7.2 rap 的功率谱，并不能显示在低频区的$1/f$结构特性。这意味着，一些音乐显示了$1/f$低频区域的结构特性，一些没有。

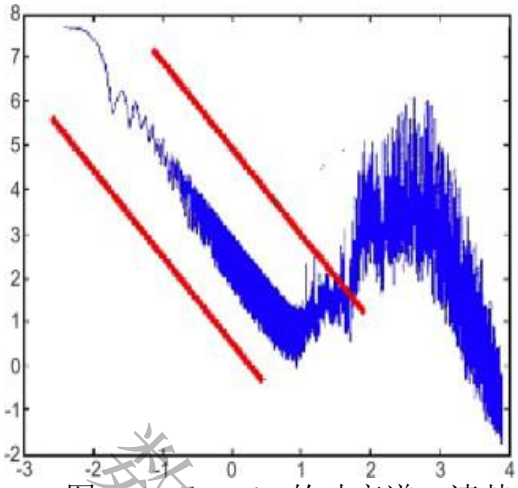

图 7.1 classic 的功率谱，清楚地表明了在低频区的$^1/_f$结构特性

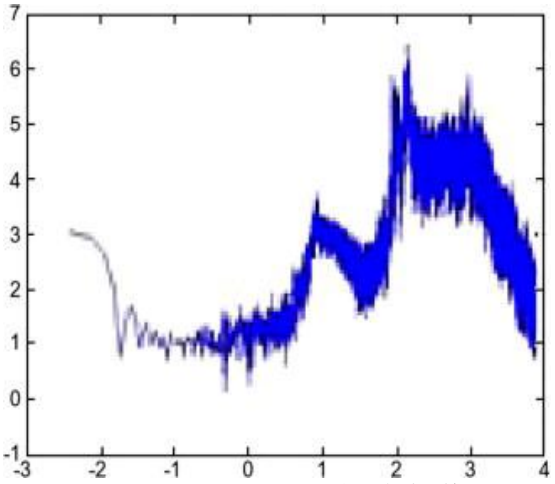

图 7.2 rap 的功率谱

图 7.3 和图 7.4 为 pop 和 rock 风格的音乐显示的功率谱。它清楚地表明，pop 显示的$^1/_f$结构特性在低频区相对较好，但 rock 风格的音乐没有明显的$^1/_f$结构特性。

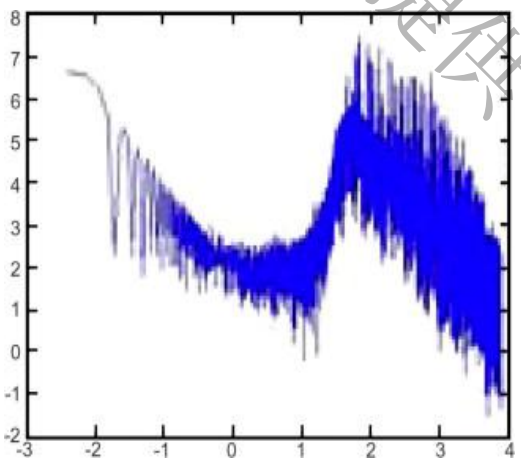

图 7.3 pop 的功率谱

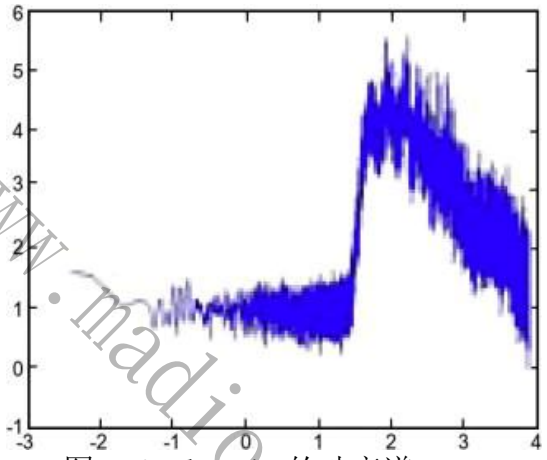

图 7.4 classic 的功率谱

因此，要想知道在不同风格的歌曲的$^1/_f$结构特性的倾向，我们共分析 140 首不同风格音乐，评估出$^1/_f$在每首音乐中低于 20 Hz 区域的结构特性。图 7.5 显示了七个不同风格的音乐低于 20 Hz 区域的$^1/_f$的结构特性（垂直轴表示的程度（%），横轴的歌曲数量）。正如预期的那样，classic 较好的显示了$^1/_f$结构特性。

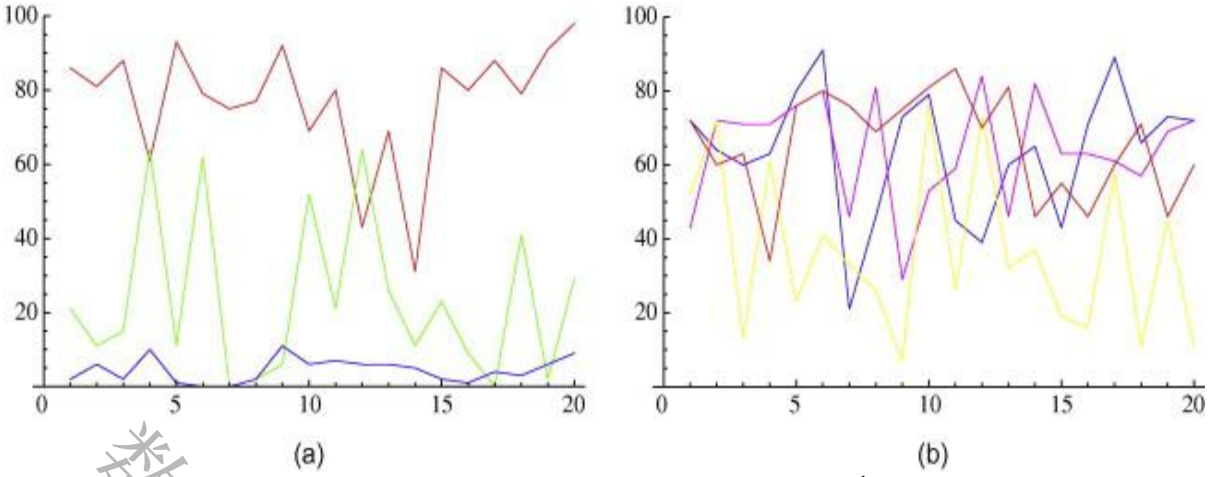

图 7.5 七个不同风格的音乐低于 20 Hz 区域的$1/f$的结构特性

由表 7.1 中所示的关于不同风格音乐在低于 20 Hz 区域$1/f$结构特性的平均值，可以看到，classic 表现了最好的$1/f$噪声分析，pop 表现第二，rap 表现最差。这意味着，在低于 20 Hz 区域的$1/f$噪声可能不会显示音乐的文化特点，但可以明显的区分音乐风格。

表 7.1

排名	类型	度的$1/f$行为在该地区低于20赫兹（%）
1	classic	77.3
2	pop	65.4
3	jazz	63.9
4	newage	63.6
5	rock	36.6
6	Hip-hop	23.5
7	rap	4.45

关于不同风格音乐在低于 20 Hz 区域$1/f$结构特性见附录

d) 结论

我们分析了在七个不同风格的流行音乐（classic, hip-hop, newage, jazz, rock, rap and pop)中的$1/f$噪声。通过分析低于 20Hz 区域的$1/f$噪声发现，在该区域的$1/f$噪声可能不会显示音乐的文化特点，但可以明显的区分音乐风格。这意味着我们可以通过分析低于 20Hz 区域$1/f$噪声的特点来区分音乐的风格。

7.2 关于 1/f 噪声的分形性质

由上文可知，我们可以通过分析低于 20Hz 区域$1/f$噪声的特点来区分音乐的风格，此方法的在低于 20Hz 的区域效果明显，然后不具有普遍性，在此，我们通过对$1/f$噪声的自身性质进行进一步研究。从数学角度看，由于音乐主旋律的多次变换相当于随机自仿射变换，这样就构造出了某种程度的自相似性。Voss R.F.发现几乎所有的

音乐都在模仿 $1/f$ 噪声，同样的，音乐也具有$1/f$噪声的随机性和可预见性，而$1/f$ 噪声是具有分形性质的，因此音乐也应该具有分形性质，这便为我们从分形的角度研究音乐的分类奠定了基础。下文利用分形维数对音乐力度变化的复杂程度进行分析，从而区分不同音乐风格的内在差异性。

7.3 音乐的分形维数计算

分形维数实质上是 Hausdorff 豪斯道夫维数，但由于 Hausdorff 豪斯道夫维数极难计算，所以一般采用近似计算方法，如盒维数、自相似维数、关联维数等。本文采用盒维数的计算方法，对音乐力度变化的复杂度进行计算，从而利用分形维数的大小来衡量音乐变化的复杂程度大小。盒维数的定义如下：

设 $F\subset Rn$ 为任意非空有界集。用 $N\delta(F)$ 表示直径最大为 δ，且覆盖 F 的集的最少个数，则 F 的上、下盒维数定义为

$$\overline{\mathrm{DIM}_B(F)}=\lim_{\delta\to 0}\sup\frac{\log N_\delta(F)}{-\log\delta}$$

和

$$\underline{\mathrm{DIM}_B(F)}=\lim_{\delta\to 0}\sup\frac{\log N_\delta(F)}{-\log\delta}$$

如果 F 的上、下盒维数相等，就把这相等的值称为 F 的盒维数。

$$\mathrm{DIM}_B(F)=\lim_{\delta\to 0}\frac{\log N_\delta(F)}{-\log\delta}$$

具体的盒维数计算方法又分多种，我们采用最常用的网格法作为音乐的盒维数计算方法。盒维数估计算法没有考虑落在每一个格子中点的个数，因此，分维数的估计值均小于理论值 ，但多尺度盒维数较单尺度盒维数的估计精度能提高一些，说明多尺度盒维数能更好地反映信号的复杂程度。所以我们用多尺度盒维数来估算。其具体方法如下：

- 选取合适的网格的最大边长一般为 2n（n 为正整数）。
- 对输入的数据进行重采样使采样过后的数据点数等于网格的最大边长加 1。
- 把网格大小 e 分别设为 2，4，8，…直到我们设的最大边长。
- 分别在每种边长 e 下，计算数据占有的网格数并把这些格子数相加得出总的格子数 N(e)。
- 对以上数据进行 n+1 次迭代。
- 设常系数 k=1，对 log(N(e)) 和 log(k/e) 进行最小二乘的一次曲线拟合，斜率就是盒维数 D。

$$D=\lim\frac{\log N(e)}{\log(k/e)}$$

对于自相似的数据集合，上式是常数，而对于实际数据集，在以 log(k/e) 为横坐标，以 log(N(e)) 为纵坐标的直角坐标系中描点，其曲线的近似斜率即为该数据集的分形维数。研究证明数据集的分形维数是数据集的固有维数的一个精确度量[11]。所以本文的音乐分类方法是对音乐特征的相似性进行分类，而音乐特征的相似不是指在某个时间段内有相同的音乐特征，而是总体感觉相似。因此，对一首音乐的整体特征维数的计算，更能反映人的感受，这也符合实际应用需求。基于以上分析，

构建音乐特征数据集如下：一首音乐为一个数据，提取每一首音乐的分形特征构成一条记录。一个数据记录是否具有分形特征尚未有理论上的鉴定，但是通过对具体数据记录的分形维数计算可以得到反演认定。

7.4 基于分形维数的音乐分类方法

为了检验提出的基于维数的音乐分类方法的有效性,选取 wav 格式的音乐 302 首，包括 classical 音乐，country 音乐，hip-hop 音乐风格的音乐各 100 首左右，数据采用单声道、采样率为 44.1kHz、存储精度为 8 位格式。所选音乐要求是各类音乐中比较有代表性的。

这 302 首音乐具体分为，96 首 classical 音乐，100 首 country 音乐，106 首 hip-hop 音乐。

a) 训练过程

以 classic 音乐为例，其训练过程如下：

- 先把 96 首 classic 音乐划分成两部分，一部分 40 首左右作为训练集合，其它的为另一部分。现只对第一部分 40 首左右的 classic 音乐进行如下流程处理：随机选取 4 首音乐作为第一小组。
- 对每一组音乐进行维数计算，得出最大值和最小值。如表 1 中的第二列，第三列。对上步中的最大值与最小值做差，得出差值。如表 1 中的第四列。
- 用下一组的最大值与最小值的差减去上一组的最大值与最小值的差，得出相对差值。

  注意 1：如果是第一组，则第一组的最大与最小值差值要是大于等于 0.1 则抛弃这组，重新随机选取 4 首音乐作为一组重复第 2、3、4 步 。如果这个差值小于 0.1 则第一组的相对差值默认为 0。

  注意 2：如果不是第一组，则若新算出来的相对差值超过上一组的一倍或者超过 0.1 则把新算出来的这组抛弃，进行下面的步骤，否则记录这组，如表 1 中的第 5 列。
- 再在去除上组后的第一部分中随机选取 4 首音乐加在第一组上形成第二组。重复上面 2 到 4 步。
- 对算出来的数据在直角坐标系中画出，直角坐标系中横坐标为训练音乐数量，纵坐标为相对差值（如图 1）。
- 直到有至少两个连续的 0 点出现（如图 7.6）才完成此类音乐的训练，得出此类音乐的维数范围。如（表 7.2 ）中最后一行第二列，第三列即为此类音乐的维数范围。

表 7.2

训练音乐的数量	最小值	最大值	最大值与最小值差值	相对差值
4	1.5945	1.6476	0.0571	0
8	1.5892	1.6516	0.0624	0.0053
12	1.5892	1.6599	0.0707	0.0083
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
40	1.5892	1.6709	0.0097	0

得出第一类 classic 音乐的维数范围是 1.5712 ～ 1.6706。第二类 country 音乐对

维数范围的确定同 classic 音乐方法相同。结果如图 7.7，维数范围是 1.6715 ～ 1.7296。第三类 Hip-hop 音乐，维数范围的确定与 classic 音乐方法相同。结果如图 7.8，维数范围是 1.7302 ～ 1.8233。

b) 测试过程

仍以 classic 音乐为例。其它两类的测试过程与 classic 音乐相同。把这 96 首音乐作为测试集进行测试，测试过程如下：用程序计算所有 96 首 classic 音乐的维数，用上面测试集算出的 classic 音乐维数范围来划定这 96 首 classic 音乐有多少首是正确划分，有多少首划分到错误的类型中去。其它的两类测试过程与此类相同。最终的测试结果如（表 7.3）。

表 7.3

分类结果	音频总数	classic	country hip-hop	Hip-hop	准确率
classic	96	81	13	2	84.375%
country	100	6	90	4	90.000%
Hip-hop	100	1	12	93	87.736%

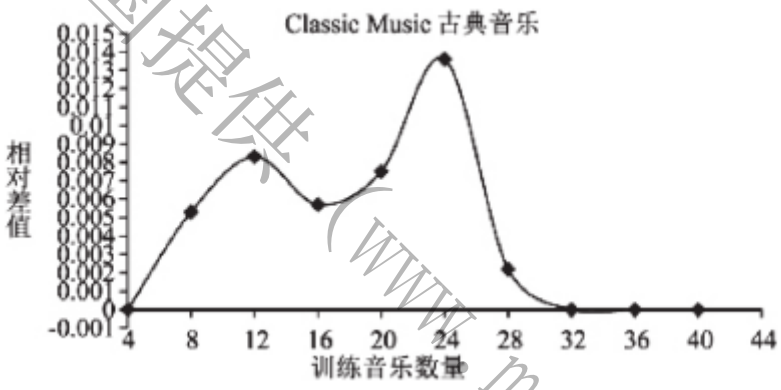

图 7.6

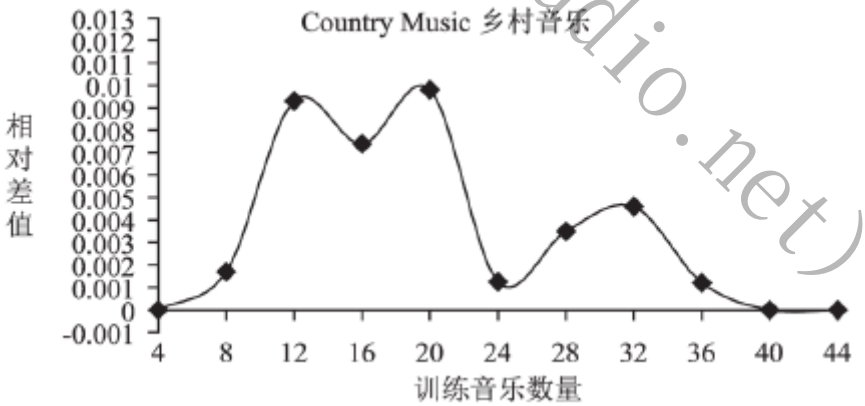

图 7.7

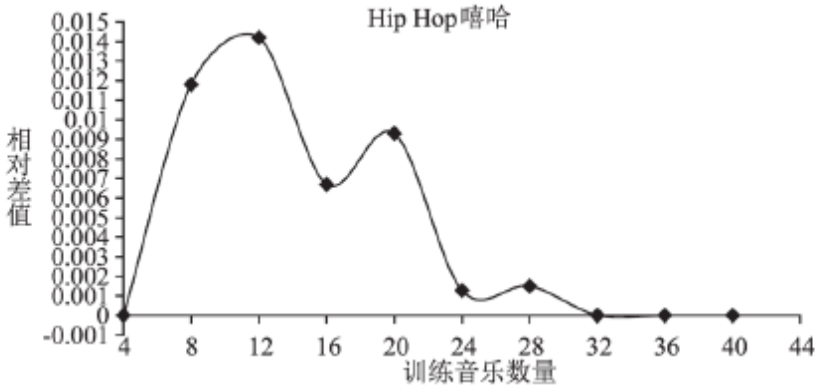

图 7.8

7.5 数据分析和结论

可以看出，三类音乐的分形维数范围是从小到大排列的，classic 音乐最低，Hip-hop 音乐最高。且本方法对于 country 音乐的区分度准确率最高，Hip-hop 音乐次之，classic 音乐最低。

在 classic 音乐的测试中，绝大部分分类错误的音乐都分类在了 country 音乐中，只有少部分音乐分在了 Hip-hop 音乐中。同样的在 Hip-hop 音乐的测试中绝大部分分类错误的音乐都分类在了 country 音乐中，只有极少部分音乐分在了 classic 音乐中。再综合三类音乐的分形维数递增，这就可以说明此三种音乐的分类想要跨越中间音乐类型不是很容易。

本方法对三种音乐类型的分类结果准确率较高，且本身方法简便，实现程序运算速度快，可以实现音乐的快速自动分类。

本文提出利用音乐的分形维数作为音乐的特征进行音乐分类的方法，并着重将整首音乐进行统一处理的研究思路，其优势在于只用一维特征就能区分音乐的不同类型，即分形刻画了音乐的内在特征——部分与整体的相似性。该方法具有应用简单、分类准确度高、速度快等优点。然而，该方法对音乐的分类还存在不足，如分类不够精细，部分音乐的维数很接近等。所以在以后的工作中要对除振幅外的其他音乐特征进行进一步研究，再适当结合其他分类方法的基础上，充分利用多维向量来进行音乐或音频的分类，以便提高音乐分类的精细程度，达到更好的分类效果。

八、模型评价及优化

- 基于标签的流行音乐风格分类模型
  基于标签的方法对流行音乐风格进行分类的模型是一种比较好的分类方法，仅仅使用底层声学特征而进行的风格分类往往很难得到好的结果，由此对于网络资源在音乐风格分类的贡献不可忽视。但是，基于标签的方法对流行音乐风格进行分类也有它的局限性，例如对于没有标签的原创歌曲，只能用传统的底层声学特征来进行风格分类。

- 基于互联网语义的流行音乐风格分类模型
  基于互联网语义关系来对流行音乐风格进行分类是非常好的一个研究方向，由于时间有限，本论文此次只是提出这个模型，实验结果虽然较好，但是不能以

一概全，需要用大量数据进行佐证。本模型可研究的潜力非常之大，它不仅与语义网络紧密相连，同时又和数据挖掘技术密切相关，这无疑为我们在以后的研究中提供了一个新的方向。

- 基于 LDA 和多类 SVM 的流行音乐风格分类模型
  目前绝大多数音频分类算法集中在两方面——音频的特征提取以及根据音频特征进行分类。现有的音频特征算法有：短时过零率、时域的短时能量、谱质心分析、频域带宽等，还有基于听觉感受的 MFCC(Mel-frequency cepstral coefficients)梅尔倒频谱系数等。另一方面，分类算法可利用模式识别和模式分类中已知算法，如 CMM(Gaussian mixture model) 高斯混合模型、NN(Neural Network) 神经网络、HMM(Hidden Markov Model) 隐马尔可夫模型等。本文通过对主流的特征提取及分类的算法进行改进，使用 LDA 方法对特征进行降维，使其更有利于分类，再采用由底向上的二叉树分类策略，不但能取得最佳的分类精确率，而且也能实现最好的时间复杂度。但此模型算法复杂，数据庞大，开销高,不具有深度推广的潜质,如何在准确率和开销之间做出权衡十分必要。

- 基于分形维数的流行音乐风格分类模型
  利用音乐的分形维数作为音乐的特征进行音乐分类的方法，着重将整首音乐进行统一处理的研究思路，其优势在于只用一维特征就能区分音乐的不同类型，即分形刻画了音乐的内在特征——部分与整体的相似性。该方法具有应用简单、分类准确度高、速度快等优点。然而，该方法对音乐的分类还存在不足，如分类不够精细，部分音乐的维数很接近等。所以在以后的工作中要对除振幅外的其他音乐特征进行进一步研究，再适当结合其他分类方法的基础上，充分利用多维向量来进行音乐或音频的分类，以便提高音乐分类的精细程度，达到更好的分类效果。

流行音乐的风格的自动分类是一项非常具有挑战性的工作，虽然之前的研究使得风格分类的准确率有了一定程度的提高，但仅仅使用底层声学特征对音乐进行风格分类往往不能得到令人满意的结果，而通过标签和语义的分类方法需要海量数据的支持才能提高准确率，不能满足实际音乐信息检索系统的需要。由此可以通过多模态音乐风格分类的方法，即使用多种分类模型得出多种结果，最后将结果进行加权结合以提高风格分类的准确率，多模态音乐风格分类的框架如图 8.1 所示。由于基于分形维数的模型没有传统的声学特征提取方法成熟且完善，在此可对参与结果综合的模型进行选择，适当增加或减少。

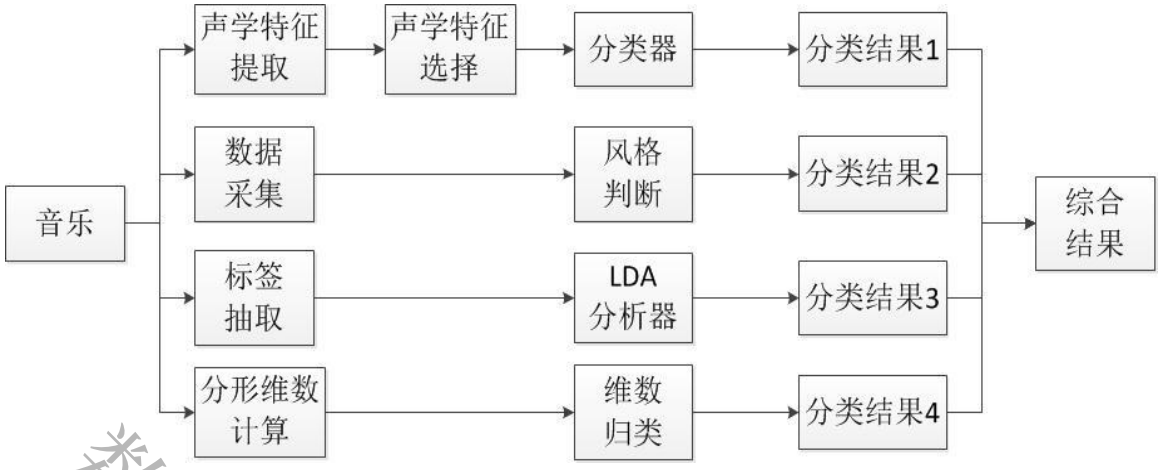

图 8.1 多模态音乐风格分类框架

九、模型的推广

本次流行音乐分类的问题，我们运用了四种模型进行解答，其中的每一个模型都可以推广到现实生活中去，这就很好的体现了数学建模的意义所在：我们可以通过对一个问题的解答，而将其运用到更多的现实事件中。基于标签的流行音乐风格分类模型和基于互联网语义的流行音乐风格分类模型由于其本身依赖于互联网，所以把它们移植到网络电台的推荐功能也是合理且可行的。而基于 LDA 和多类 SVM 的流行音乐风格分类模型和基于分形维数的流行音乐风格分类模型则能在一定程度上弥补前两种模型的局限性，能够很好的利用流行音乐底层声学特征对音乐进行合理自然的分类，如果能把这四种模型有机地结合起来，也就是我们前文提出的多模态流行音乐分类方法，则可以以最大准确性给用户推荐其所偏好的音乐。多模态流行音乐风格分类方法通过对数据科学的处理，不仅适用于网络电台的推荐功能，而且在对流行音乐市场的分析、基于流行音乐的大众审美研究等提供事实依据。

十、参考文献

[1]. Fine S, Navratil J, Gopinath R A. A hybrid GMM /SVM approach to speaker identification[C] //Proceedings of 2001 IEEE International Conference on Acoustics, Speech and SignalProcessing, 2001.
[2]. 鲍玉斌，王琢，孙怀良等 . 一种基于分形维的快速属性选择算法东北大学学报，2003，24（6）：527-530.
[3]. Burred J J ; Lerch A A hierarchical approach to automatic musical genre classification 2003
[4]. Li Xin ; Guo Lei; Zhao Yihong Tag-based social interest discovery 2008
[5]. Sordo M; Celma O; Blech M The quest for musical genres: Do the experts and the wisdom of crowds agree 2008
[6]. McKay，C.，and I Fujinaga.2007.”jWebMiner：A web-based feature extractor. ”Accepted for publication at the 2007 International Conference on Music Information Retrieval
[7]. 甄超，郑涛，许洁萍，《基于音乐语义信息的音乐风格分类研究》,第五届全国信

息检索学术会议 CCIR2009，2009
[8]. Tzanetakis，G and Cook,P.，“Musical Genre Classification of Audio Signals”. IEEE Transactions on Speech and Audio Processing,vol.10 no.5，PP.293-302，2002
[9].Voss R.F,Clarke J.1/f noise in music and speech[J].Nature,1975,258:317-318.
[10].Voss R.F,Clarke J.“1/f noise” in music:Music from 1/f Noise[J].J.Acoust.Soc.Am.1987.63(1):258-263 .
[11].郭平，陈其鑫，王燕霞。基于分形维数的属性约减[J].计算机科学，2007,34(9)：189-190。