单位代码：10293　密　级：公开

南京郵電大學

硕士学位论文

论文题目：基于 SVM 主动学习的音乐分类

学　　号　1012010519

姓　　名　姚磊

导　　师　邵曦

学 科 专 业　信号与信息处理

研 究 方 向　音乐信息检索

申请学位类别　工学硕士

论文提交日期　2015.03

Music Classification Based on SVM Active Learning.

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Engineering

By

Yao Lei

Supervisor: Prof. Shao Xi

March 2015

南京邮电大学学位论文原创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京邮电大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

本人学位论文及涉及相关资料若有不实，愿意承担一切相关的法律责任。

研究生签名：_____ 日期：_____

南京邮电大学学位论文使用授权声明

本人授权南京邮电大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档；允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索；可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。本文电子文档的内容和纸质论文的内容相一致。论文的公布（包括刊登）授权南京邮电大学研究生院办理。

涉密学位论文在解密后适用本授权书。

研究生签名：_____ 导师签名：_____ 日期：_____

摘要

近年来互联网技术与多媒体信息技术的快速发展，宣告了二十一世纪正式开始步入大数据时代，如何从海量的数据信息中检索出有用的信息将非常具有研究意义。互联网的多媒体信息中就包括数量增长迅速的数字音乐，大量歌手涌现，海量专辑和网络歌曲纷纷面世，另外受世界文化发展多元化的影响，各式各样的音乐风格也随之产生，为满足人们根据自己不同的喜好来准确而又快速的找到自己想要查询的歌曲，这就要求音乐检索系统更加高效和快速。然而传统的音乐分类都是先训练音乐样本得到分类模型，然后对未知的音乐样本进行预测，这种传统的分类方法所得到的分类器效果往往取决于训练样本的数量。对海量的训练样本全部进行人工标注显然是不现实的，主动学习方法可以很好的解决这个问题。

支持向量机（Support Vector Machine）是一种性能十分优良的机器学习方法，本文将主动学习方法与 SVM 相结合，并将其应用于音乐流派分类当中。传统的基于 SVM 的主动学习方法的样本选择策略往往只局限于样本的不确定性，即认为距离分类超平面最近的样本点所包含的价值也就最大。基于这种算法本文做出了如下改进：

（1）在选取最有价值样本时，考虑选取那些距离分类超平面较近的样本的同时也考虑保证样本的多样性。由于音乐样本的特征维度较高，本文选取样本之间的角度来作为样本多样性的衡量标准，并由此制定了最终的样本价值评判标准 *score*；

（2）“一对其余”方法是 SVM 应用于多分类时的常用方法，然而这个方法人为的造成了数据集的偏斜，这会对最后的分类效果产生一定的影响，所以本文在进行价值样本选取时，制定了样本平衡性判断标准参数 b，当主动学习方法选取的价值样本数量不满足平衡性条件时就对其进行平衡性调整。

关键词： 支持向量机，主动学习，样本多样性，音乐分类

Abstract

In recent years, the rapid development of the internet technology and multimedia information technology officially declared that the age of big data has been coming in twenty-first Century , how to retrieve the useful information from the information of the mass data will have great research significance.Multimedia information on the Internet certainly includes the digital music with the rapid growth of the number, a lot of singers is emerging, the massive album and network songs have become available, also influenced by the development of world cultural diversity, many kinds of music style have emerged too, in order to satisfy the people according to their different preferences to accurately and quickly find which song they want to query, this requires a music retrieval system witch is more efficient and fast. However, the traditional music classification system will train music samples firstly to obtained the classification model, and then predict the class of unknown music samples,the effect of this classifier with this kind of traditional classification method often depends on the number of training samples. To label the mass of training samples manually is clearly not realistic, active learning can be a very good solution for this problem.

SVM(Support Vector Machine) is a kind of excellent machine learning methods, active learning method is combined with SVM in this paper, and this method is applied to the music genre classification. The traditional sample selection strategy of active learning method based on SVM is only confined to the uncertainty of sample , that if a sample is nearest to the classification hyper plane, means the value of this sample is maximum.The improved method proposed in this paper based this algorithm is described as follows:

(1)When we select the most valuable samples, we will consider to choose the samples those distance to classification hyper plane is near, and at the same time ensure the diversity of them. Because the dimension of music samples characteristics is relatively high, the angle between samples is selected as the sample diversity measure in this paper, and thus established the ultimate criterion of sample value: score ;

(2)"1-v-r one versus the rest" is a common method of applying SVM to multi-classification , but this method will cause the deflection of data sets artificially, it will have a certain bad influence to the final classification effect, so the sample balance judgment standard parameter:b is established in this paper,when the number of valuable samples selected by

the active learning method do not satisfy the equilibrium conditions, the balance adjustment for them will be done.

Key words: support vector machine, active learning, sample diversity, music classification

目录

专用术语注释表

符号说明：

MFCC	Mel-Frequency Cepstrum Coefficients	梅尔倒谱系数
PLP	Perceptual Linear Prediction	感知线性预测
FFT	Fast Fourier Transform	快速傅里叶变换
DCT	Discrete Cosine Transform	离散余弦变换
SVM	Support Vector Machine	支持向量机
RBF	Radial Basis Function	径向基函数
DAG	Directed Acyclic Graph	有向无环图
RASTA	Relative Spectral	相对频谱
MIR	Music Information Retrieval	音乐信息索引
1-v-1	One Versus One	一对一
1-v-r	One Verse the Rest	一对其余

第一章 绪论

1.1 研究背景及意义

近年来互联网行业获得了飞速发展，网络所负载的信息量也正以指数级的规模增长，人们可以做到足不出户的在因特网上获得大量信息，但同时也面临着一个严峻的问题，那就是对于用户而言从互联网上获得的信息并非都是有用的，很多有意义的信息往往会被大量的无用信息所淹没，这个时候就需要建立一种高效新颖的信息分类检索系统，它可以更加高效的管理冗杂的网络信息，该方向自然而然也就成了当今研究的热点之一。机器学习正是通过利用计算机技术对收集到的信息进行智能化处理，使计算机能够模拟甚至实现人类的学习行为从而获取各种新的知识或技能，保存并应用到各种场景中去的一种技术。

音乐作为网络多媒体信息载体之一，更是世界文化的一种存在形式，其在人们的日常生活中已经成了一种必不可少的调味品。互联网时代的到来，信息的共享，人们可以很方便的获得越来越多的音乐资源，也推动了音乐研究的迅速发展。音乐研究是一种涉及了音乐理论、信息处理、人工智能、机器学习等多领域的较新学科。音乐分类作为音乐研究领域的一个重要分支，因其广泛的应用前景及重要的学术价值，已经成为了当今研究热点之一。

传统的音乐信息检索通常是简单基于文本的，也就是用户通过音乐曲目的名称、歌手、歌词或者词曲作者等标注的文本信息进行检索。这种检索方式虽然有着操作简单、检索速度快、实现简单等诸多优点，但同时也有着明显的缺陷。首先，这种基于文本的音乐信息检索是需要人工标注的，尽管现在绝大多数音乐网站对于音乐流派、情感等热门类别的人工标注已经取得了一定的成果，但是这同时也是以牺牲大量的时间、人力和资源为代价的。另外，随着现在科技的发展，人们也越来越追求自身的精神享受，不再仅仅满足于传统的根据已知信息来检索固定曲目，而是希望计算机可以根据自身的喜好来自动的推荐一些未知歌曲，或者说希望当自己对一首曲目除了旋律之外一无所知时计算机可以自动的检索出自己想找的这首歌曲。这就要求音乐分类需要从原来基于文本的人工化分类向自动化分类过渡转变。为此，对于音乐自动分类的研究也就越来越多。基于内容的音乐信息检索是目前音乐自动分类的主流[1][2][3][4]，其原理是从符号化的音乐文件里提取出类似于音色、音高和节奏等自身的声学特征，然后根据这些特征进行分类。这种技术不需要人工标注，可以根据音乐自身的声学特征进行自动标注分类，大大节省了人力物力，这对于步入大数据时代的今天意义重大。而且这种技术不需要听歌人提供歌曲名、歌手等信息，而只需要知道某段旋律就可以进行检索，并

由此衍生出了各种音乐推荐系统。对用户进行音乐喜好的量身定制，大大丰富了用户的体验。

传统的机器学习方法需要专家对原始样本集进行组织整理以形成训练集，最终分类模型的好坏往往也取决于已标注的样本集规模的大小。但是不管在哪个领域，都需要专家具备相当强的专业背景，同时标注样本类别也会消耗大量的时间、精力和资金，这种机器学习方法所消耗的成本自然是巨大的。因此，人们开始关注那些大量的未标注样本，这些样本往往也蕴含了大量的信息。于是，主动学习方法应运而生。

主动学习的相关概念由 Simon 于 1974 年首次提出[5]，是机器学习十分重要的研究范围之一。其与传统的被动学习最大的不同在于被动学习是从样本中随机的选取训练样本，被动的对样本的类别信息进行学习，而主动学习的核心是根据学习过程，主动的去选取那些对于当前分类器影响最佳的样本进行学习，所谓最佳样本也就是最有价值样本，它可以最大化的修正分类器。主动学习的具体实现步骤如下：首先从全部样本中通过聚类算法或者随机的选取较少的样本并人工标注它们所属的类别，然后进行训练，得到一个初始分类器；然后，制定某种选择策略从剩余的未标注样本中选取一部分对当前分类器最有价值的样本加入到原来的训练集，重新训练得到一个更精准的分类器；如此循环往复，直至达到设定好的迭代停止条件。

主动学习方法与传统的学习方法相比，有着极为显著的优势：

（1）主动学习方法通过制定合适的样本选择策略从未标注的样本池里有目的的选取对当前分类器最有价值的样本，这种让分类器根据自身进行选择的方式有效避免了标注者的主观倾向或个人偏好。另外，这种有方向性地选取信息量最大的样本可以使分类器的性能的收敛速度远远快于传统的随机采样系统；

（2）主动学习是一个不断反馈的过程。每次迭代都是一次反馈，主动学习能够根据这一反馈来判断标注者的个人喜好，这种特性使得主动学习可以在个人定制化的系统里得到很好的应用，比如根据就本文实验中针对的音乐分类系统，可以衍生出用户的个人音乐推荐系统，这种人机交互的感受可以大大提升用户的满意度。

（3）主动学习的“主动性”大大提升了学习系统对外界环境的“感知力”。这种特性可以使得学习系统根据不同的用户，不同的应用场景来建立相应的训练集。当对象类别和测试分类随着时间、地点有了改变，主动学习可以很好的进行调整，减少环境变化带来的不必要标注代价。

总之，主动学习可以大大减少人工标注样本所耗费的人力物力，同时也大大缩减了训练样本的数量，而且其最终获得的分类器性能与对全部样本进行标注训练得到的分类器性能相差无几。因此，近年来越来越多的学者对于未标注样本在机器学校过程中的价值越来越重视，

同时也成为了一个新的研究热点。

1.2 国内外研究现状

音乐分类属于多媒体信息分类的相关范畴，对音乐进行初步的分类是为了方便接下来的音乐信息检索。音乐是音频的一种，音乐分类的发展历程也就可以从音频信息的相关检索技术讲起。音频信息检索技术的研究最早起始于上世纪 90 年代，主要研究方向是利用音频中的时域或频域等相关物理特性对音频信息进行基于内容的检索。音频信息检索最初目的是为了检索出音频中的语音信息，将语音信号转换成文本的表达形式，最后通过检索文本中的相关关键字来间接的对语音信息进行检索，然而这种方法并没有实际的应用到音频分类技术当中。现如今音乐检索已经成为了音频信息检索领域当中研究十分活跃、也非常富有成果的领域之一，其检索方式有根据哼唱进行的检索和根据节拍进行的检索等，音乐库和音乐系统的构建也都用到了音乐分类技术，这使得其检索空间相应减少，检索速度也大大加快，明显提升了效率。

目前关于音乐自动分类方法的研究，主要集中在以下三个方面：

一是在音乐特征提取的选择上以及特征向量的维度上进行相应的改进[6,...,21]。关于音乐样本的主要特征包括过零点率、短时能量以及基音频率等，这些音乐特征可以在某种程度上反映响度、音调、节奏等音乐感知特性，然后计算各帧之间的均值、方差、自相关系数等统计特性，形成最后的多维特征向量用于接下来分类器需要的训练集。美国加利福尼亚研究的 Muscle Fish[22]系统正是基于这种方法的音乐检索系统。Li[23]提取了 MFCC 系数之后再提取基音频率、谱质心、子带能量等感知特性，最后级联形成高维特征向量。Lin 则在 Li 的基础上进行了改进，他使用了小波变换的方法来提取音乐相应的感知特性，实验结果表明通过小波变换提取的音乐特征明显比其他方法准确。Tzanetakis[24]则是提取了音乐样本中的节奏、音色和音调等特征，构成的特征向量也达到了几十维。

二是在分类算法的选择上进行改进[25][26][27][28]。目前应用与音乐分类领域的分类算法主要有基于高斯模型（GM）的分类算法、基于决策树（DT）的分类算法、基于隐马尔可夫模型（HMM）的分类算法、基于神经网络（NN）的分类算法和基于支持向量机（SVM）的分类算法等，实际音乐分类时选用哪种分类算法很大程度上影响着最终系统的性能，而且这些分类算法在本身基本原理的基础上都可以进行相应的改进来进一步的优化分类系统；

三是在多分类问题的解决方法上进行改进[29]。比如在支持向量机算法中，处理多分类问题时，有“一对一法”、“一对其余法”、“有向无环树法”、“分层法”、“树形结构法”

等方法，几种方法各有利弊，可根据实际情况和需要来选用。另外，在这些基本方法的基础上依然可以作出进一步的改进。

而在目前已有的分类算法中，很大一部分采用的是被动监督学习方法，所谓被动监督学习也就是人为地对需要训练的样本一一进行标注，然后训练学习得到学习器，最后对未知的样本数据集进行预测分类，这种被动学习方法的分类效果只有在标注好的训练样本数量足够多时才能使分类器的效果有所保证，所以该方法所具备的最大问题就是标注这些训练样本需要耗费的人力和时间是相当巨大的。因此，如何将那些容易获得的未标注样本充分的利用起来以得到训练分类器具有十分重要的研究意义和应用价值。目前在利用未标注样本进行训练学习的众多方法中，主动学习是最行之有效的方法之一。其原理是主动学习制定了某种样本选择策略，要求分类器可以自动地从给定的未标注的样本中选择出对当前分类器最有价值的样本，这部分样本可以最大程度的影响分类器，将这部分样本提供给用户进行人工标注，然后将新标注的这部分样本加入到原本的训练集再次训练得到学习器，如此循环往复，从而在多次迭代中不断地改善学习器性能。一般只要采用合理的选择策略，那么就能在达到同样甚至更好的分类效果的前提下，使训练集的规模更小，从而有效地减少人工标注样本所耗费的代价。

目前无论国内还是国外，关于在音乐分类中应用主动学习的研究相对较少，但是主动学习方法本身已被广泛应用于模式识别和人工智能等领域[30,...,34]，理论上它能够适用于各类学习器，比如支持向量机[35]，隐马尔可夫模型[36]和逻辑回归[37]等等。关于主动学习自身的研究，Seung、Opper 和 Sompolinsky 提出了基于委员会投票的主动学习算法[38]，这种算法有两大模块，即训练模块和询问模块。训练模块负责训练已标注好的样本集来形成分类器，询问模块则从给定的未标注样本中选择最优价值样本标注后作为下一个输入。顾名思义，这个算法需要一个委员会来对样本类别进行投票，委员会由 2k 个委员组成，每个分类器都是一个委员，询问模块选择的是委员会最具争议的样本。与此类似的采用委员会投票的方法还很多[39]。Nguyen 和 Smeulders 提出了一种主动学习与预聚类相结合的算法[40]，主要针对的是二分类问题。他们发现与传统算法相比，如果能将未标注样本的先验分布知识加入到训练过程，将有更好的学习效果。中科院的孙功星、戴贵亮[41]等人将主动学习与神经网络算法相结合，使得在样本规模较大、信息冗余严重时仍能保持良好的学习效果。另外关于支持向量机和主动学习方法结合的相关研究也比较多。Tong 较早就提出了 SVM 主动学习算法[42]，它选择每次迭代之后距离分类超平面最近的一个样本进行标注并认为这个样本的类别最不确定，信息量最大。这种一次只标注一个样本的策略在进行文本分类时可以获得比较不错的效果。后来 Tong 又提出了一种基于版本空间缩减的 SVM 主动学习策略[43]，同样在文本分类问题上取得了不

错的效果。其它的还有 Mukerjee[44]、Schohn[45]、Vlachos[46]、周艳丽[47]、解洪胜[48]、Wu 与 Huang[49]、张健沛[50]等人也都在这个领域做了深入而广泛的研究。总之，SVMactive 由于其自身的优良性能，已经被广泛应用于障碍物检测[51]、图像检索[52]、规则训练[53]等各个领域。

1.3 本文研究内容与论文结构

本文对音乐分类中所涉及的相关理论知识和技术进行了总结探讨，并将基于 SVM 的主动学习方法应用于音乐分类当中，同时对传统的基于 SVM 的主动学习方法在样本选择策略上进行了相应改进，即传统的基于 SVM 的主动学习在分类器迭代时，只选取那些距离分类超平面最近的那部分样本点，认为这部分样本点具备最大不确定性，它们加入到训练集可以最大化修正分类超平面。根据这个原理，本文提出了一种改进的基于 SVM 主动学习方法，认为在选取价值样本时除了考虑样本的不确定性，应该同时保证样本的多样性以避免过学习现象。另外，本文针对音乐的分类属多分类，使用 SVM 分类器对其进行分类时采用了常用的“一对其余”的多分类方法。但这种方法人为的造成了数据集的偏斜问题，因此本文在对样本进行选择之后又进行了一次平衡性调整，进一步改善了分类器的性能。综上，本文的论文结构安排如下：

第一章 绪论。简单介绍音乐分类及主动学习的研究意义和背景，同时简述了两者的国内外研究现状。

第二章 音乐分类与主动学习相关介绍。首先介绍音乐分类的基本原理，针对音乐分类的特征提取部分进行详细介绍。然后介绍主动学习的基本概念与框架，简述主动学习的过程和步骤。针对主动学习中的两个核心模块：学习引擎和选择引擎，分别进行较为详细的介绍。学习引擎主要有 K-近邻学习方法、神经网络学习方法和支持向量机学习方法等几个较为常见的方法。选择引擎作为主动学习的核心部分，文章详细介绍了基于不确定度缩减、基于版本空间缩减和基于误差缩减的三种不同的选择策略，并对各自的优缺点进行了讨论。

第三章 基于支持向量机的主动学习方法介绍。概述统计学习的相关理论和基础知识，详细介绍支持向量机方法的理论基础，并介绍支持向量机方法法在进行多分类时常用的几种方法。详细介绍支持向量机与主动学习相结合的几种常用的典型方法。针对传统的 SVM 主动学习提出两种改进策略，即兼顾样本不确定性和多样性，同时对选择的价值样本进行平衡性调整来进一步保证分类器性能。

第四章 系统设计和实验仿真。基于本文提出的主动学习的两个改进策略，进行两组实验进行对比，结果证明，本文提出的改进的主动学习方法是有效的。

第五章 总结与展望。对本文进行的工作研究进行总结，提出存在的不足以及今后努力的方向。

第二章 音乐分类与主动学习

2.1 音乐分类概述

所谓音乐分类，顾名思义就是按照某个标准将不同的音乐进行归类，目前国际上比较流行的分类标准是基于流派和情感的，而由于每个人对音乐的主观感知不同，后者的研究还处于尚不完善的阶段。不管是音乐分类还是其他分类，其本质上还是属于模式识别的范畴，一般包括特征提取和分类识别两个主要部分。所以，音乐分类的整体框架结构如下所示：

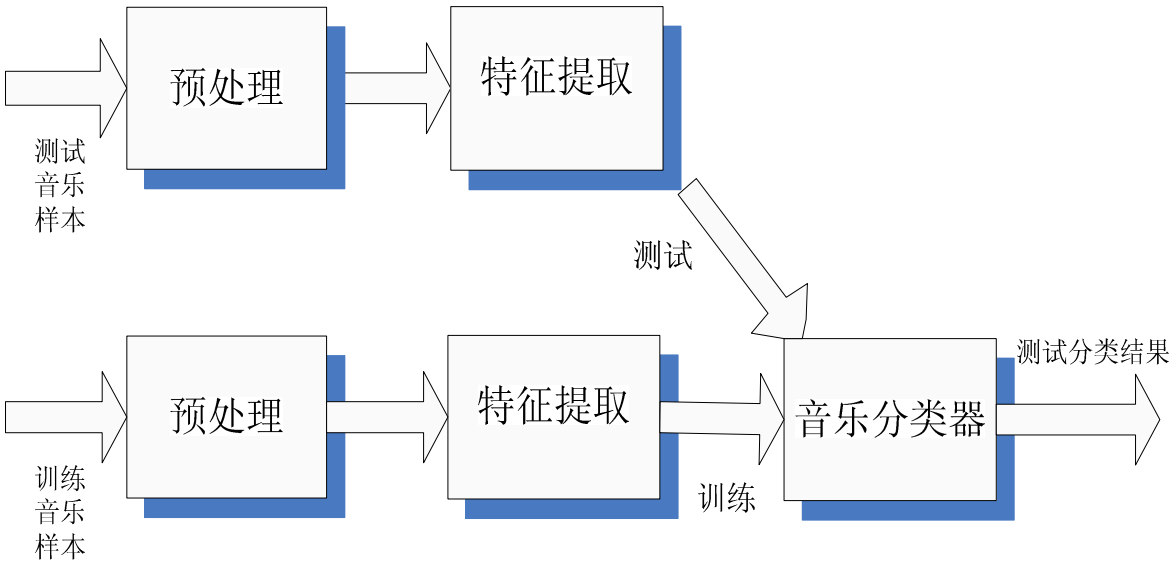

图 2.1 音乐分类系统框架

由上图不难看出，音乐分类的主要步骤为：先对音乐信号进行预处理，然后对音乐样本特征提取，经过训练后得到学习器再进行样本测试，系统各模块功能如下：

（1）信号预处理：主要对音乐信号先进行分帧，然后预加重处理，最后进行加窗，旨在滤除原始信号中无用的信息和干扰噪声。

（2）特征提取：此模块是整个分类系统的关键模块，因为特征的选择和提取很大程度决定了分类系统的效果好坏，其旨在提取出能够准确反映信号特征的关键特征参数。音乐分类领域常用的特征参数有能量、过零点率、短时频谱、线性预测系数（LPC），以及梅尔倒谱系数（MFCC）等。

（3）训练：同样是能够决定分类器最终性能的关键模块，对构建的训练样本集进行特征提取然后训练得到学习器，并确定学习器中需要的参数，进而获得最终的学习器。

（4）测试：用于测试分类器效果。新的待分类音乐样本进入分类系统，使用训练样本集

训练后得到的音乐分类器对未知的测试音乐进行分类并预测，判断其是否被分类到它本应该属于的类别，最后统计系统的总分类准确率。

2.2 音乐特征提取

无论选用哪种分类器，在对每个测试样本进行分类之前，都需要先将每个样本转换成多维的特征向量，而如何选用以及选择哪些样本数据中的特征将会对最终的分类效果帮助很大。通常情况下，音乐特征提取之后具备的特征维度越高，分类器越容易辨别，类与类也就越容易分开。下面重点详细介绍信号特征提取前的预处理过程以及本文选取的音乐特征：梅尔频率倒谱系数（MFCC）和相对谱-感知线性预测（RASTA-PLP）。

2.1.1 信号预处理

为了便于分析，在提取音乐信号特征之前需要对音乐信号进行预处理，整个过程可归纳为如图 2.2 所示：

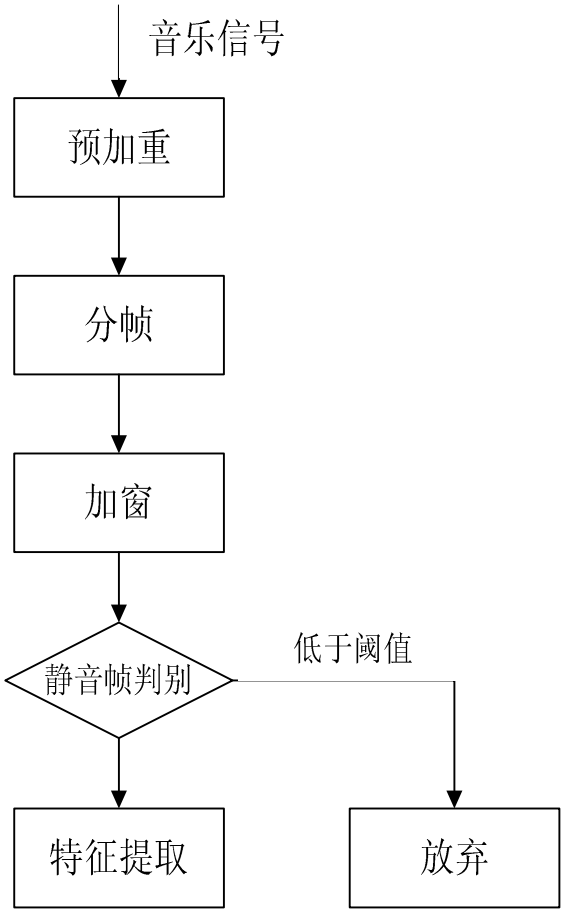

图 2.2 音乐信号预处理过程

（1）预加重

音乐信号不同于一般的语音信号，它不仅包含了乐器演奏的乐声信号还包含了人声演唱的部分，而人声演唱的部分则可以认为是普通的语音信号，另外音乐信号里高频部分比例远大于低频部分，频谱不平坦，那么为了对音乐信号进行解析，就得事先进行预加重去噪，提升高频信号部分。

预加重通过一阶高通数字滤波器来实现。其表达式为：

$$H(\mathrm{z}) = 1 - \mu\, \mathrm{z}^{-1} \tag{2.1}$$

式中预加重系数 μ 取值为[0.9，1.0]。

信号方程为：

$$y(n) = x(n) - \mu x(n-1) \tag{2.2}$$

一阶高通数字滤波器如图 2.3 所示，其中 x(n) 是数字化后的音乐信号，$y(n)$ 是处理后输出的信号：

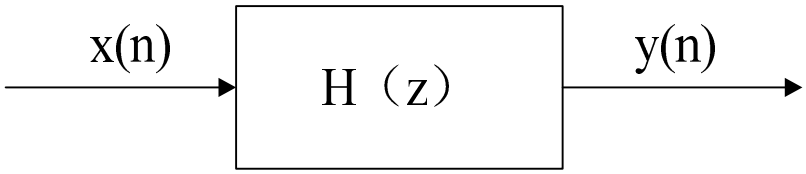

图 2.3 预加重滤波器示意图

经过滤波器预加重输出的信号需要进行归一化处理。

（2）分帧与加窗

由于音乐信号自身也有着长时非平稳特性，但是在 10ms-30ms 内可近似认为其特性平稳，是具备短时平稳特性的准稳态过程。因此处理音乐信号时要截取几个片段，每个片段作为一帧，为了能够平滑过渡，帧与帧之间会有重合，这部分就是帧移。在后面实验中，音乐样本采样率设置为 16Khz，加 32ms 汉明窗，帧移 16ms。

后面将介绍如何提取 MFCC 和 RASTA-PLP 特征，提取这两个特征都要进行傅里叶变换，此过程中会出现吉布斯效应，为了避免这种情况音乐信号在进行分帧后需要通过加窗处理来实现平滑滤波。一般常用的窗函数有：矩形窗、三角窗、汉宁窗、汉明窗、布莱克曼窗等。

矩形窗：

$$w(n) = \begin{cases} 1\,, & 0 \le n \le N-1 \\ 0\,, & \text{其他} \end{cases} \tag{2.3}$$

汉宁窗：

$$w(n)=\begin{cases}0.5-0.5\cos(\frac{2\pi n}{N-1}),\ 0\le n\le N-1\\ 0,\ 其他\end{cases} \tag{2.4}$$

汉明窗：

$$w(n)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{N-1}),\ 0\le n\le N-1\\ 0,\ 其他\end{cases} \tag{2.5}$$

布莱克曼窗：

$$w(n)=\begin{cases}0.42-0.5\cos(\frac{2\pi n}{N-1})+0.08\cos(\frac{4\pi n}{N-1}),\ 0\le n\le N-1\\ 0,\ 其他\end{cases} \tag{2.6}$$

2.2.2 MFCC 特征提取

MFCC 是一种基于人听觉特性的音频特征参数，是在自动识别和分类系统中应用最为广泛的特征参数之一[54][55]。某种程度上，MFCC 更能近似代表人的听觉特性，它能够对人的听觉模型逼近模拟，即使是在音乐特征方面，MFCC 也要比其他的短时特征更能准确的代表信号特性[56][57]。关于 MFCC 的计算流程如图 2.4 所示：

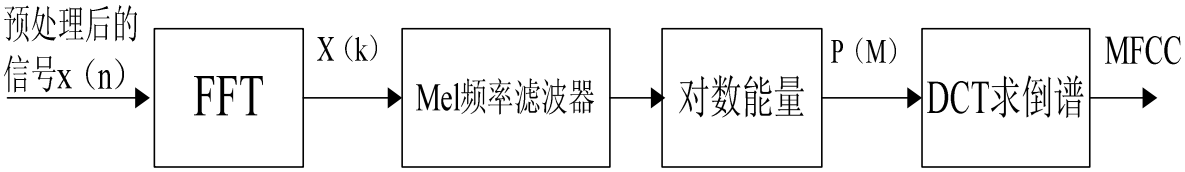

图 2.4 MFCC 计算流程

1）预处理。也就是对输入的原始音乐信号分帧加窗，得到一帧音乐信号；

2）快速傅立叶变换。本文采样率设置为 16KHz，加 32ms 汉明窗，帧移 16ms。对一帧音频信号的 N 个点进行快速傅立叶变换，得到一帧音乐信号的频谱表达：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N},(0\le n,k\le N-1) \tag{2.7}$$

得到信号频谱值，信号频谱的平方值就是能量谱。

3）设计 Mel 滤波器组。由在 Mel 频率尺上定义的 m 个三角形滤波器组成，实验中设置 m=19，各个三角形滤波器的中间频率在 Mel 频率轴上是相等距离间隔分布，在频率轴上是随着 m 的增大而增宽的，其关系如图所示。三角滤波器的频率响应如下：

$$H_m(k)=\begin{cases}\dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, (f(m-1)\le k\le f(m))\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m+1)-f(m))}, (f(m)\le k\le f(m+1))\\ 0, k<f(m-1)\cup k>f(m+1)\end{cases} \tag{2.8}$$

其中 $f(m)$ 为第个三角滤波器的中心频率：

$$f(m)=(\frac{N}{F_S})B^{-1}(B(f_l+m\frac{B(f_n)-B(f_l)}{M+1})) \tag{2.9}$$

其中 f_l 和 f_h 分别是三角滤波器组的最低频和最高频，N 为进行 FFT 时的点数，M 为滤波器组中滤波器的个数，F_s 为音乐信号的采样频率，B^{-1} 为 Mel 频率到时域频率的转换公式：

$$B^{-1}(b)=700\ (e^{b/1127}-1) \tag{2.10}$$

4）对输出信号取对数。这是为了得到鲁棒性比较好的谱估计误差，短时能量谱通过 Mel 滤波器组后，对输出的信号进行取对数：

$$S(m)=\ln(\sum_{k=0}^{N-1}|X(k)|^2 H_m(k)),\ (0\le m\le M) \tag{2.11}$$

5）对由 4）求得的对数能量进行离散余弦变换。变换为倒谱：

$$c(i)=\sum_{m=1}^{M-1}S(m)\cos(\frac{n\pi(m+0.5)}{M}),\ 0\le m\le M \tag{2.12}$$

然后对每一帧取其前 20 个系数，这就是提取的音乐信号的 MFCC 特征参数。

2.2.3 RASTA-PLP 特征提取

传统的感知线性预测 PLP 的特征提取是基于短时频谱上的，帧移频谱的变化会在进行短时分时引入到特征参数中。而 RASTA（RelAtive SpecTrA）技术其实是在传统的 PLP 特征提取过程中的临界频谱估计时对每一个频带部分加入了一个带通滤波器，这样可以有效抑制 PLP 特征提取过程帧与帧之间频谱的快速变化。常用到的 RASTA 滤波器主要有 Log-RASTA

和 J-RASTA 两类，其差别在于对每一对临界频谱添加带通滤波器前后的处理不同。本文使用的是 Log-RASTA 滤波器，RASTA-PLP 特征提取流程如图 2.5 所示。

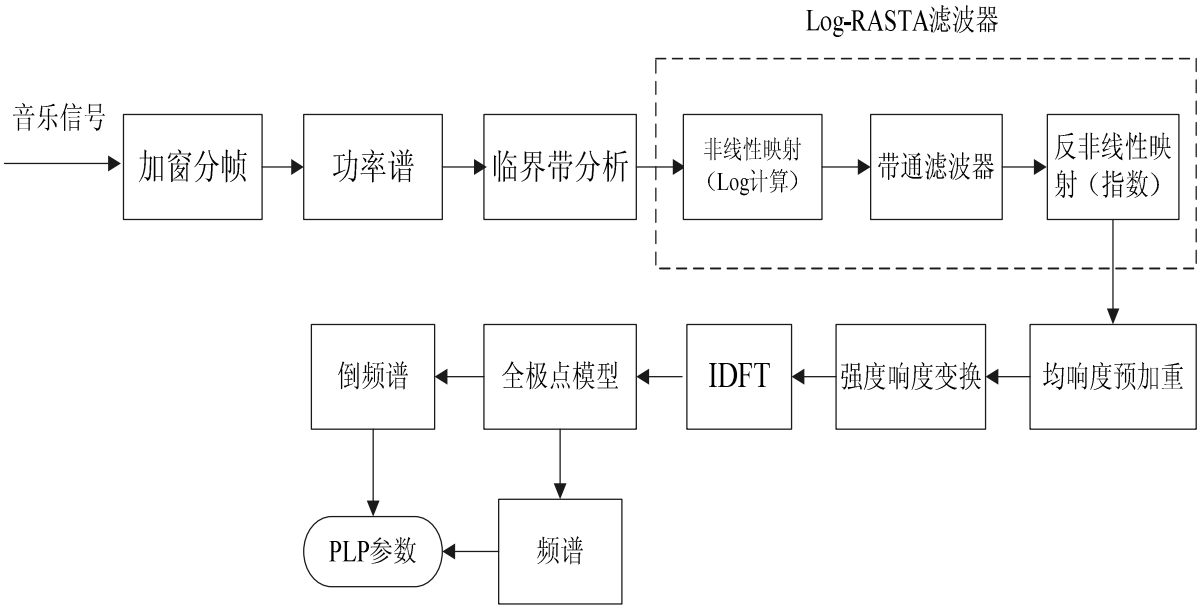

图 2.5 RASTA-PLP 提取过程

1）对音乐信号预处理，求功率谱并进行临界频带分析。人类听觉频谱特性是在 800-1000HZ 之间大致成线性增长，高出此范围会增长会减慢。PLP 通过将频率轴映射到 Bark 域并通过整合临界带能量来产生临界带频谱的近似。

首先由公式（2.10）进行频率的转换。

$$B = 6\lg\left(\frac{f}{600}+\sqrt{1+\left(\frac{f}{600}\right)^2}\right) \tag{2.13}$$

在本文中音乐信号的抽样率是 16KHZ,转换到 Bark 轴得到了 20 个临界频带，进一步得到中心频率 B0(k)=0.98k Bark（k=1,2,…,20），临界频带的加权系数也可也求出：

$$\psi(B-B_0(k))=\begin{cases}0, & B-B_0(k)<-1.3\\ 10^{(B-B_0(k)+0.5)}, & -1.3<B-B_0(k)\le -0.5\\ 1, & -0.5<B-B_0(k)<0.5\\ 10^{-2.5(B-B_0(k)-0.5)}, & 0.5<B-B_0(k)\le 0.5\\ 0, & B-B_0(k)>2.5\end{cases} \tag{2.14}$$

同时由（2.14）的反变换可以求出每个 Bark 刻度带到频率轴上的低频 $f_l(k)$ 和高频 $f_h(k)$：

$$f_i(k)=\frac{1}{2}\times 600\times\left(10^{\frac{B_0(k)-2.5}{6}}-10^{-\frac{B_0(k)-2.5}{6}}\right) \tag{2.15}$$

$$f_h(k)=\frac{1}{2}\times 600\times\left(10^{\frac{B_0(k)+1.3}{6}}-10^{-\frac{B_0(k)+1.3}{6}}\right) \tag{2.16}$$

得出了每一 Bark 带的最低和最高频也就可得出每个点的加权系数值，将加权系数与前面得到的短时功率谱相乘即可得出临界带频谱 $\theta(k)$ 。

$$\theta(k)=\sum_{N=n_l(k)}^{n_h(k)} p(f(N)\psi(B-B_0(k))) \qquad (k=1,2,...,20) \tag{2.17}$$

2）对得到的临界带功率谱幅值进行非线性映射，即取对数计算，将对临界带频谱的处理变换到对数域的处理，这样能够将频域的一些乘性失真变换为加性失真从而被过滤掉。

3）通过 RASTA 带通滤波器。此处的带通滤波器，功能等效于一个 IIR 滤波器，其函数表达式表示如下：

$$H(z)=z^4\frac{0.2+0.1z^{-1}-0.1z^{-3}-0.2z^{-4}}{1-\rho z^{-1}} \tag{2.18}$$

这里取 $\rho=0.94$ 。

4）对得到的数据进行反非线性映射，即进行指数扩展变换。

5）等响度曲线预加重。通过等响度曲线来抑制低频和高频部分，其目的是将音乐信号控制在 400-1200HZ 频段，这是因为人耳对这部分相对敏感。

$$\Gamma(k)=E[f_0(k)]\cdot\theta(k),(k=1,2,...,20) \tag{2.19}$$

其中 $f_0(k)$ 是第 k 个临界带频谱对应的中心频率，$E[f_0(k)]$ 是权值系数，体现了人耳感知频率的敏感度。

$$E[f_0(k)]=\frac{(f_0^2(k)+1.44\times10^6)f_0^4(k)}{(f_0^2(k)+1.6\times10^5)^2\times(f_0^2(k)+9.61\times10^9)} \tag{2.20}$$

6）幅值立方根压缩—强度响度变换。由于响度的强度和感知响度的强度是非线性关系，需要乘以 0.33 次方来模拟听觉的幂法则，数学表达式如下：

$$\Phi(k)=\Gamma(k)^{0.33} \tag{2.21}$$

7）由全极点模型求解线性预测系数。将线性预测系数应用于音乐信号的基本原理是：采样的音乐信号之间有一定相关性，一个采样信号可以利用多个音乐采样信号，通过线性组合的方式来模拟。

8）计算倒谱。由 LPC 之后得到的线性预测系数可以计算音乐信号包含的倒谱特征。

2.3 主动学习概述

前文已经介绍，音乐分类属于机器学习的一个领域。机器学习作为人工智能中最为核心的研究方向之一，一直备受关注。然而传统机器学习研究发展至今，有一个不得不面对的问题：大量的未标注样本极易获取，而稀有的已标注样本很难获得。传统机器学习属于被动的监督式学习，其训练得到的分类器性能完全依赖于已标注样本，而且往往很大程度上依赖的是已标注样本的数量。相关研究已经表明，对未知样本进行类别标注不仅需要标注者有着相应领域极为专业的知识背景，更要花费大量而宝贵的时间。互联网的时代到来，大数据、云计算等研究领域所面临的该问题尤为明显。于是现实情况与此形成了不可调和的矛盾。这种情况下，半监督式学习和主动学习方法应运而生，这些年更是发展迅猛成为了解决上面的问题的重要技术。

半监督式学习和主动学习的目的是一样的，都是希望不仅能够利用已标注样本来提升学习器性能，也可以充分利用未标注样本中所包含的样本信息以进一步改善学习器。区别在于，主动学习的过程更类似于人类的学习过程，通过不停迭代的方式来提高学习器的泛化能力。

总之，主动学习最终目的就是最大化的利用未标注样本，希望从未标注样本中可以有目的性的选取最有价值样本，然后由专家进行标注后加入训练集训练后得到分类器，期望通过多次迭代过程来不断完善分类器的性能。

2.3.1 基本步骤

主动学习的基本步骤可概括如下：候选样本中所有样本都没有被用户标注类别，通过聚类算法或者随机从候选样本集中选取极少量的样本作为初始训练样本集，同时人工标注它们的类别，确保初始样本集中至少包含一个正类样本和一个负类样本。根据这些已标注的初始训练样本集经过训练后得到一个初始分类器，根据当前分类器，利用某种先前制定好的选择策略从候选样本集中选取一部分对当前分类器最有价值的样本，人工标注它们的类别并添加到原本的训练样本集当中，重新训练学习器，利用新的分类器再次从剩余样本中筛选出对当前分类器最有价值的样本。重复以上步骤，直到原本的候选样本集变为空或者达到事先设定的某个指标。如图 2.6 所示

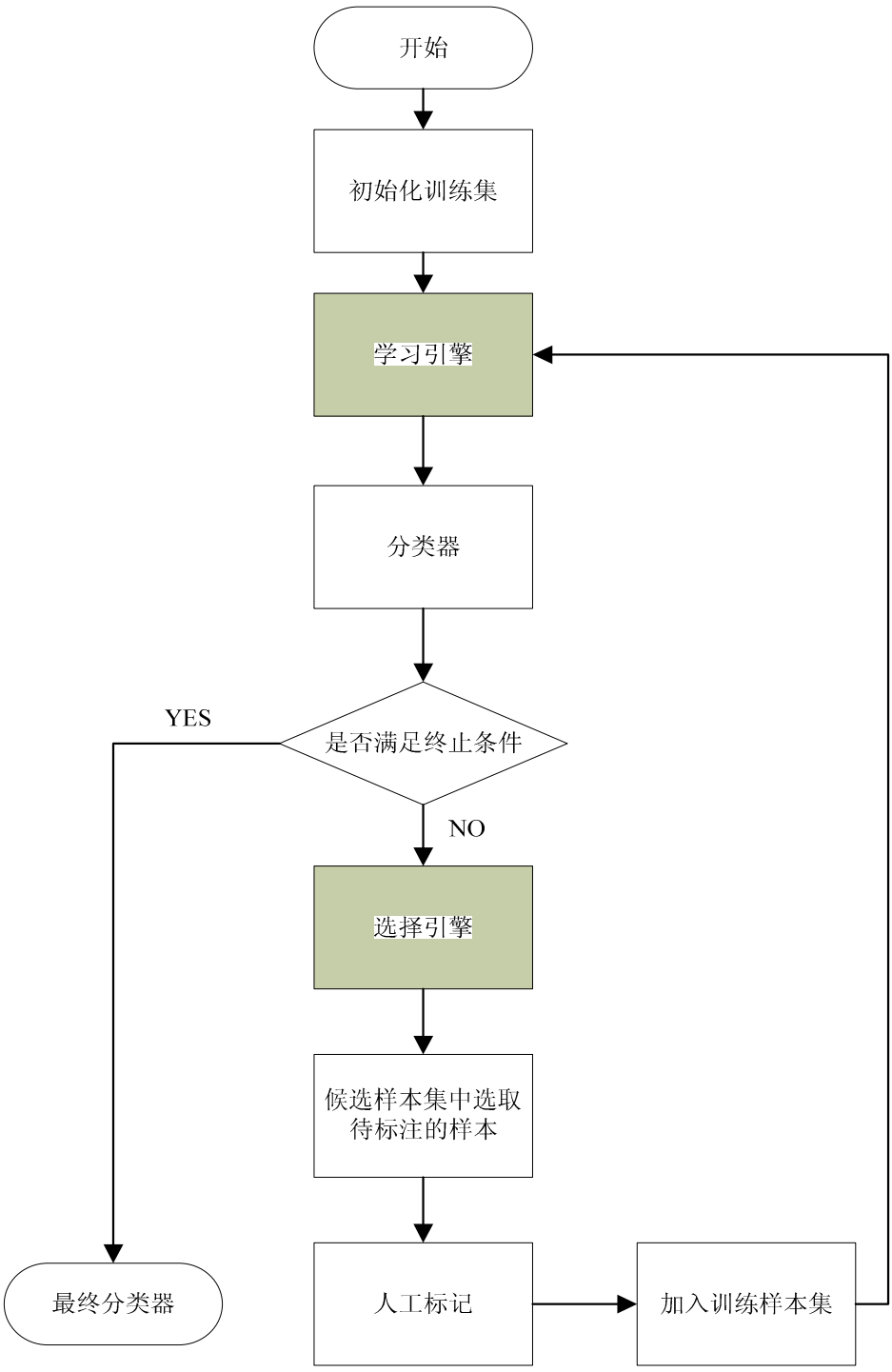

图 2.6 主动学习示意图

为了更加简明直观的说明上述过程，可以作出如下假设，集合U为尚未标注的候选样本集；集合S中的元素为每次迭代批量选取的需要标注的最有价值样本，其初始值为ϕ，且在每次迭代前都要清空；L中的样本为人工标记过的所有样本，用作训练集，其初始值同样是ϕ。

那么主动学习的基本步骤可以表示如下：

1）初始状态时，从样本池U 中选取一小部分样本作为初始样本经过标记得到L和新的U（$L << U$），保证L中每个类别至少有一个样本存在；

2）根据训练集L训练得到分类器；

3）根据当前分类器，采用某种样本选择策略对U 中的每个样本计算其价值，选择价值最大的样本集S；

4）对S 中的样本进行标注，然后加入训练集L，同时执行$U = U - S$；

5）当满足某个迭代停止次数或条件时，停止学习，得到最终分类器，否则跳转至（2）步，继续迭代。

对于主动学习效果的好坏，一般有两种方式进行评价，一种是达到同样分类准确率时所减少的需要标注的样本数比较，一种是同样训练样本数量下的分类准确率比较。

从上面给出的主动学习基本框架图中可以看到，主动学习两个最为重要的模块：学习引擎（Learning Engine，LE）和选择引擎（Sampling Engine，SE）。针对这两个主要部分，下面进行详细介绍。

2.3.2 学习引擎

学习引擎顾名思义，这部分用于构建分类器，这部分依然是依赖于已标注样本，只不过引入主动学习的概念之后，这里变成了一个迭代循环的过程，当分类器的精度到实验设置的要求就停止迭代输出结果。其本身还是一个传统的监督式学习过程，简称分类器。

常用的分类器有如下几种：

（1）K-近邻法（KNN，K-Nearest Neighbor）

K-近邻法是模式识别中应用最广泛的方法之一[58]，利用的是待分类样本与训练样本的距离远近来预测其类别，十分直观，其基本原理也非常简单：即当经过特征提取的待分类样本进入分类系统，系统将计算该样本与所有样本之间的距离，选择距离该样本最近的 K 个样本，然后判断这 K 个样本的类别，占多数的类别将被判定为待分类样本的类别，如图 2.7 所示。KNN 算法虽然理论简单，实现方便，但也有着其固有的缺点，即由于要存储所有训练样本，同时还要计算待分类样本到所有样本点的距离，将消耗相当大的存储量和计算量，强力搜索算法的复杂度达到了$O(k^2N^2)$，当处理的样本又是音乐这种高维的特征向量时，这个问题也将尤其的严重。

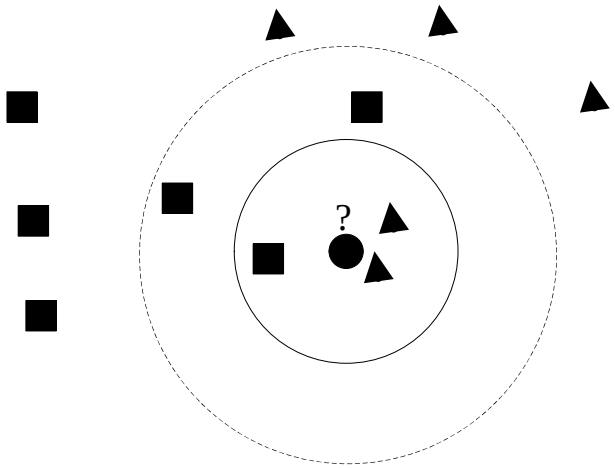

图 2.7 K-近邻法

（2）神经网络模型（NN，Neural network）

神经网络是人工智能领域的一门新兴学科，它通过模拟人类神经细胞能够存储和分布消息的特征，具备着自学习性强、容错性强、鲁棒性强等诸多优点。基于神经网络的分类模型[59][60][61]中，分类所需要的阴式信息存储在神经元连接的权值上，权值向量由使用的迭代算法确定，网络输出无误的时候权值不变，否则增加或降低。

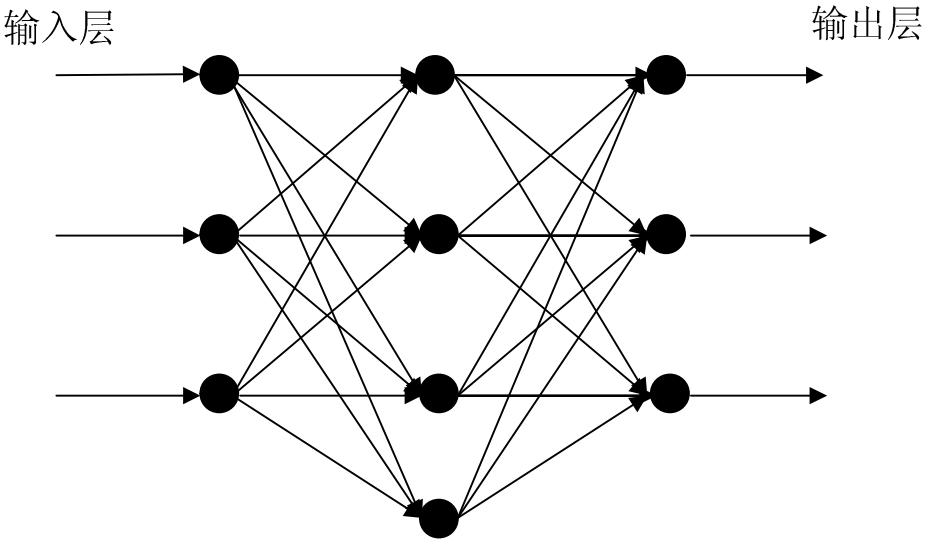

图 2.8 BP 神经网络结构

误差反向（BP）神经网络是神经网络中最为常用的一种。采用 BP 网络算法的分类器一般都具备三层结构，即输入层、隐含层和输出层，如图 2.8 所示。BP 神经网络分类器的基本原理步骤如下：

首先，根据问题规模大小，确定输入输出神经元数目，神经元数目对应样本特征向量维度。

其次，确定网络参数，通过训练确定最优网络参数。

尽管神经网络方法有着诸多优点，但是其也有着明显的不足，如不同结构的网络特征差异

很大，容易陷入局部最优解等问题。

（3）高斯混合模型（GMM， Gaussian Mixture Model）

在 GMM 模型中，需要假定每个类别的分布均服从正态分布，如图 2.9 所示。对 K 个高斯概率密度分布函的加权求和就是 K 阶的高斯混合模型。

$$P(z_n \mid K=k)=\frac{1}{\sqrt{(2\pi)^d \mid \sum k}}\exp(-\frac{1}{2}(z_n-\mu_k)^T(z_n-\mu_k)) \tag{2.21}$$

式中 z_n 是信号的特征向量，K 是 GMM 的阶数并有：

$$P(z_n \mid C=c)=\sum_{k=1}^{K} p(z_n \mid K=k)P(K=k \mid C=c) \tag{2.22}$$

式中的均值 μ_k，协方差矩阵 $\sum_k$ 等参数的求取一般用最大似然估计（ML Maximum Likelihood）方法。

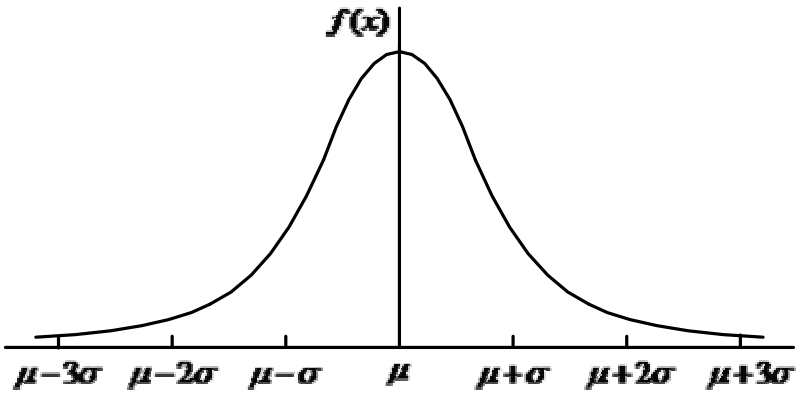

图 2.9 正态分布图

GMM 用于分类时，首先要对利用典型的样本训练出 GMM 分类模型，根据得到了训练模型对某一样本进行测试时，输出的是测试样本归属为某一类的概率大小。GMM 在音乐信息提取的过程中应用的比较广泛。Pampalk 等人提出了用决策树理论，GMM 树状结构进行音乐风格的分类。West 和 Cox 使用的决策分类是二叉树，所有的类别都包含在二叉树的根节点，根节点下的两个子树则分别代表不同的子类别，每个子类别又有两个子类别，直到完成完全分类。

（4）隐马尔科夫模型（HMM ，Hidden Markov Model）

HMM 是建立在一阶 Markov 链的基础上的，它们的概率特性基本相同，不一样的是一阶 Markov 没有双内嵌式随机过程，而 HMM 则有。一个随机过程描述状态和观察值之间的统计关系，解决了用短时模型来描述平稳段信号的问题；另一个描述的是状态的转移，来说明相邻短时平稳段之间是如何转变的，如图 2.10 所示。可以定义一个 HMM 模型 $\lambda=(A,B,\pi)$，其

中 A 是状态转移概率分布，B 是观察值的概率分布，π 是初始状态分布。所以 HMM 模型可以对时间序列建立模型。HMM 中经常用到的算法有前向-后向算法、维特比算法和 Baum-Welch 算法。

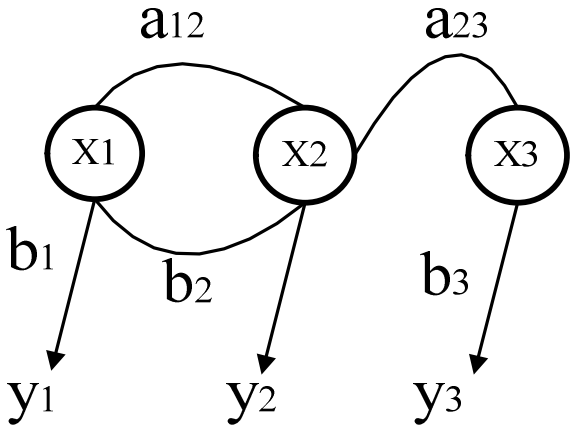

图 2.10 隐马尔科夫模型状态变迁图

（5）支持向量机（SVM，Support Vector Machine）

支持向量机方法是一种建立在 VC 维理论和结构风险最小化基础上的机器学习方法，它试图根据仅有的样本数据在模型的复杂度（指对特定训练数据的学习精度）和学习能力（指准确地识别任意样本类别的能力）之间寻求某个折衷点，以期望能够得到最好的推广能力。关于 SVM 的基本理论下一章本文会进行较为详细的介绍，这里简单讲其最大的特点就是当样本在低维空间线性不可分时，通过松弛变量以及核函数等技术，将其映射到相应的高维空间便变成线性可分的了。

关于支持向量机的入门，有一个经典的例子，如图 2.11（a）所示：横轴上面的两个端点，a 与 b 之间红色部分所有点认为是正样本，而两边黑色部分的点则认为是负样本。明显无法找到一个线性函数将这两类正确分开，但是再看图 2.11（b），可以找到一条曲线，这条曲线可以把这两类完全划分开，曲线上方为正类，下方为负类。这个曲线的表达式通常写成如下形式：

$$g(x) = c_0 + c_1 x + c_2 x^2 \tag{2.23}$$

虽然它并非一个线性函数，但可以将其写成另外的一种形式：

$$g(x) =< a, y >= ay \tag{2.24}$$

其中 $\vec{a} = (c_0, c_1, c_2)$，$\vec{y} = (1, x, x^2)$，不难看出，在任意维的空间当中，这种形式表达的函数都是线性的，只是其中的 a 和 y 都变成了多维向量罢了。

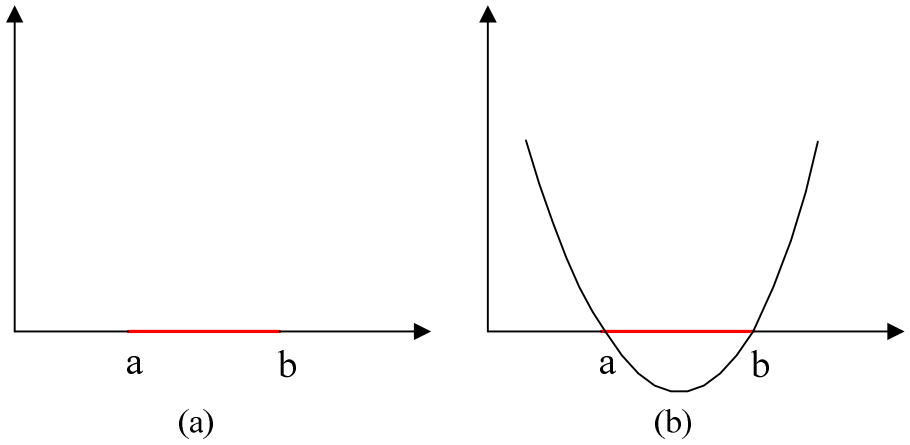

图 2.11 SVM 原理经典举例

虽然 SVM 最初只是用来针对解决二分类问题的，但是由于其优良的性能，也同样被用于解决各个领域的多分类问题，最基本的方法便是组合多个二分类器来实现，常用的有“一对一”、“一对多”和有向无环图等方法，下一章将详细介绍。

2.3.3 选择引擎

选择引擎是主动学习框架当中最为核心的模块。换句话说，主动学习方法不同于其它算法的地方就在于选择引擎部分。主动学习算法在执行每次迭代时，选择引擎负责从未标注的样本池中使用不同的选择策略选取对当前分类器最有价值的样本，人工标注后加入到原本训练集重新训练得到分类器，如此循环往复直到达到迭代停止条件。其目的是在牺牲最少代价的前提下，最大化提升学习器性能。

选择策略是选择引擎的核心，同时也直接影响着主动学习方法应用到分类当中效果的好坏。所以本文的研究重点也是针对传统的主动学习选择策略进行相关的探究，进而在其基础之上做出相应的改进，并将新的选择策略应用于音乐流派分类当中。接下来首先就传统主动学习选择策略进行详细介绍。

2.4 主动学习选择策略

按照未标注样本的获取方式不同，通常将主动学习选择策略大致分为基于成员查询的选择策略，基于池（Pool-based）的选择策略和基于流（Stream-based）的选择策略。其中基于成员查询的选择策略是最早被提出来的，它的基本思路是设想学习器在某种程度上可以控制周围环境，能够向标注的人进行提问，通过这种方式来确定标记哪些样本，以期望可以获得

尚不可知的概念。但是基于成员查询的选择策略最大的缺点就是缺乏目的性，不够“主动”，它还是需要将所有未标注的样本都反馈给标注者，不能结合样本的实际分布进行选择性的提供。针对这种不足，后面两种选择策略衍生出来。二者最大的不同之处在于，基于流的选择策略要处理的未标注样本是逐次逐个提供的，也就是说在该策略中，无法对未标注样本的价值进行整体的衡量和比较，只能通过在算法中设定一个阈值，在对样本进行选择时，先通过某个标准估计出这个样本的价值，然后与阈值相比较，达到条件则将其交由专家标注后加入训练集，否则丢弃。这种策略对阈值的设置要求很严格，因为阈值直接影响着学习器性能，而且不同数据不同问题,阈值的选取也不同，由于种种类似于此的局限性，基于流的策略应用并不广泛，目前大多是基于池的选择策略。在基于池的选择策略的主动学习方法中，目前应用最广泛的有基于不确定度缩减，基于版本空间缩减和基于期望误差缩减等方法。基于不确定度缩减的主动学习方法认为未标注样本中类别最不容易被确定的样本是最有价值的，基于版本空间缩减的主动学习方法则认为能够在每次迭代时最大幅度减少当前版本空间规模的样本为最有价值样本，基于期望误差缩减方法的核心思想类似于贪婪算法。针对这三种选择策略下面进行详细介绍。为了便于描述，这里使用 (x, y) 来表示已标注的样本及其对应的类别标签，$\bar{x}$ 则表示未标注的样本，$\bar{y}$ 是对于 $\bar{x}$ 的预测类别。

2.4.1 基于不确定度缩减的选择策略

基于不确定缩减的选择策略是目前应用最为广泛，原理也最为简单明了的一种主动学习选择策略。这里用 $p(\bar{y}\,|\,\bar{x})$ 来表示一个未标注的样本 $\bar{x}$ 属于类别 $\bar{y}$ 的概率。那么，基于不确定缩减的选择策略就是希望选择出那些能够使得 $p(\bar{y}\,|\,\bar{x})$ 最接近 0.5 的未标注样本，这和一开始学习信息论基础时的概念是一样的，两类情况时一个样本属于某一类的概率接近 0.5 时，其具备最不确定性，信息熵最大，包含的信息量也最多。研究表明，这种选择策略可以适用于大部分的学习器，在大幅减少人工标注耗费的人力和时间代价的同时还能提升学习器的分类准确率和推广能力，也是目前研究最为成熟的一种主动学习选择策略。

设 $f(\bar{x}, \bar{y})$ 是用于衡量未标注样本不确定性的函数，根据上面说的基本思路，自然而然的是希望学习器根据 $f(\bar{x}, \bar{y})$ 来直接估算出 $p(\bar{y}\,|\,\bar{x})$ 的值，然后判断哪些结果最接近 0.5 的

未标注样本为价值样本。Lewis 和 Catlett[62]最早将这种方法应用到了决策树的模型当中，他们定义 $f(\bar{x},\bar{y})$ 形式如下：

$$p(c\mid w)=\frac{\exp\left(a+b\sum_{i=1}^{d}\log\frac{p(w_i\mid c)}{p(w_i\mid \bar{c})}\right)}{1+\exp\left(a+b\sum_{i=1}^{d}\log\frac{p(w_i\mid c)}{p(w_i\mid \bar{c})}\right)} \tag{2.25}$$

其中 c 表示样本的正类别，$\bar{c}$ 表示样本的负类别，w_i 表示样本的特征。目前该方法已经被应用到了隐马尔科夫模型和结构化分类模型当中，并取得了不错的效果。

现实情况中，面临的问题往往复杂的多，很难找到一个合适的函数 $f(\bar{x},\bar{y})$ 可以直接估计出 $p(\bar{y}\mid\bar{x})$ 的值，因此，用 $f(\bar{x},\bar{y})$ 来表示信息熵的方法得以引入。信息熵的概念通俗易懂，且容易计算，可以轻松地应用到各种复杂的现实场景中去。但是关于针对各自场景，如何定义信息熵才能更好的提升主动学习方法的性能目前尚未有统一的标准，缺少指导性原则。

除了上面描述的两种情况之外，还有一种是利用分类器的几何特性，比如针对支持向量机方法提出的主动学习。前面已经简单介绍了支持向量机方法，因为其自身在小样本分类中良好的表现，将 SVM 和主动学习结合的算法也日益增多。由于支持向量机方法的目的是寻找一个最优分类超平面并以最大的几何间隔将两类样本分开，根据这种几何特性，不难想到那些距离分类超平面最近的样本点就应该是对于当前分类器最难判定的点，具备最不确定性，把这部分样本点筛选出来用于人工标注的选择策略也叫最近边界策略（Closed-to-Boundary）。该方法已经被广泛应用于各类多媒体信息检索当中，并效果显著。但这种方法不可避免的会选到离群点，也叫奇异点，这种样本点虽然距离超平面真的很近，但是由于自身特殊性无法代表一类样本的特性，把这样的样本点加入到训练集对学习器而言其实是一种噪声，某种程度上会影响分类器性能。

2.4.2 基于版本空间缩减的选择策略

基于版本空间缩减的选择策略是在主动学习迭代过程中每次选出的价值样本可以最大程度的缩减当前的版本空间。基于委员会投票（Query-By-Committee，QBC）的选择方法就是基于这种选择策略应用最为广泛的一种主动学习算法。它的基本思想是这样的：首先，从当前的版本空间中选择若干个假设来构建一个“委员会”，然后，使该委员会的各个成员假设

对未标注样本的类别进行预测，最后统计每个未标注样本对应的预测结果，各个假设最不一致的样本被认为是可以最大程度缩减当前版本空间的样本。简单讲，就是根据构造的多个分类器，选择各个分类器都不确定的样本进行标注，本质和基于不确定缩减的选择策略相似。这种样本选择策略的核心是如何构建一个具备较好的推广能力且高效率的委员会。

关于委员会的构建方法，可以看做是多个分类器的集成问题。目前比较经典的有 Boosting-QBC 和 Bagging-QBC 两种建立委员会的策略。这两种方法都是先利用重采样技术得到了若干待选的假设，然后从未标注的样本中随机地选择一部分样本对当前待选的假设进行区分，针对这些假设，选取那些不确定度最大的样本从而达到缩减版本空间的目的。区别在于 Bagging-QBC 是为了减少假设偏置，实现简单，而 Boosting-QBC 则是为了增强弱分类器假设的性能，实现困难。委员后构建完成之后面临的问题就是以哪种指标来判断成员对未标注样本的分歧程度，选举熵（VE）是一种应用相对广泛的指标。公式如下：

$$VE(\bar{x}) = -\sum_{i\in|M|}\frac{V(\bar{y},\bar{x})}{|M|}\log\frac{V(\bar{y},\bar{x})}{|M|} \tag{2.26}$$

上式的$V(\bar{y},\bar{x})$是指委员会中的成员假设对未标注样本的投票结果，$|M|$则表示委员会成员个数。其实选举熵的理论概念归根结底还是在基于不确定度缩减的选择策略中定义熵的范围内，只不过在原来的基础上有了优化。还有一种方法并不是针对每个样本来计算出熵值，而是选择了得到一个区间，将落在这个区间内的样本选择出来由人工标注。这种方法的实现公式可表达如下：

$$KL(p_1(x), p_2(x)) = -\sum_{i\in|L|} p_1(x_i)\log\left(\frac{p_1(x_i)}{p_2(x_i)}\right) \tag{2.27}$$

综上所述，这种以缩减版本空间为目的的基于委员会投票的选择策略在实际应用中也得到了广泛应用，尤其在自然语言处理中效果显著。

2.4.3 基于误差缩减的选择策略

基于误差缩减的选择策略是希望选择出这样一批未标注样本，它们可以最大程度缩减分类器误差并提升学习器泛化能力。这种选择策略由于自身强大的统计理论基础，可以根据样本的实际分布来找寻可以使学习器未来泛化误差大幅缩减的未标注样本。它有效的避免了不确定缩减方法中容易遭遇的离群点的干扰，但是由于模型参数不断变化使得每次筛选时都要重新训练一次，使得自身的计算量十分庞大。算法基本步骤可概括如下：将候选样本集中的

每一个未标注样本进行人工标注后加入到已标注训练集得到分类器，计算分类器在进行训练后误差的变化，最后根据误差值选择出能最大程度缩减分类器误差的样本进行人工标注。

基于误差缩减的选择策略虽然不如前面两种方法应用广泛，但对其的相关研究却起始很早。显然这种方法的核心部分是找寻出能够衡量模型误差的函数。费舍尔信息函数就是用来判别分类模型里样本的标记对误差的影响程度的一种函数，它可以有效地判断当前分类器对未标注样本的不确定程度。计算公式表示如下：

$$I(\lambda) = -\iint p(\bar{y} \mid \bar{x}, \lambda) \frac{\partial^2}{\partial \lambda^2} \log p(\bar{y} \mid \bar{x}, \lambda) d\bar{x} d\bar{y} \tag{2.28}$$

还有另外一种应用较多的误差缩减策略，是用未标注样本的未来期望误差来替代。首先将每个未标注样本进行人工标注并加入到已标注训练集，算法会估计出当前的概率分布 $\overline{p}(x, y)$，然后计算 $\overline{p}(x, y)$ 与实际概率分布 $p(x, y)$ 的期望误差，最后能使该误差最小的未标注样本将被筛选出来作为价值样本。

上述三种常用的选择策略，各有各自的优缺点，可根据应用场景的不同进行选择。

2.5 本章总结

本章节首先简单介绍音乐分类的基本原理，并详细说明了 MFCC 和 RASTA-PLP 音乐特征的提取过程，接着介绍主动学习方法的基本概念与框架，简述了主动学习的基本步骤。针对主动学习中的两个核心模块：学习引擎和选择引擎，分别进行了较为详细的介绍。学习引擎主要有 K-近邻学习方法、神经网络学习方法和支持向量机学习方法等几个较为常见的方法。选择引擎作为主动学习的核心部分，文章详细介绍了基于不确定度缩减、基于版本空间缩减和基于误差缩减的三种不同的选择策略，并对各自的优缺点进行了讨论，下面一章将对主动学习在支持向量机上的应用进行详细展开。

第三章 基于 SVM 的主动学习

3.1 支持向量机

支持向量机（SVM）是 Cortes 和 Vapik[63]在 1995 年提出的一种建立在 VC 维理论和结构风险最小化基础上的机器学习方法，并且性能十分优良，它在解决小样本、非线性以及高维模式识别等问题当中能够表现出它自身特有的一些优势[64]。前面已经介绍，支持向量机方法能够根据仅有的样本数据在模型的复杂度（指对特定训练数据的学习精度）和学习能力（指准确地识别任意样本类别的能力）之间寻求某个折衷点，以期望能够获得最好的推广泛化能力。简单讲，支持向量机方法的目的是找到一个最优分类超平面，它能够以最大间隔将两类数据完全分开。不管在二分类还是多分类问题上 SVM 都能有良好的学习效果。

3.1.1 统计学习理论

统计机器学习方法和传统机器学习方法的不同，本质上在于统计机器学习方法能够比较准确的给出学习的效果，同时又能解答实际需要的样本数等问题，相对于这种缜密的思维，传统机器学习方法目前尚还缺乏相应的指导和原则。所谓 VC 维[65]可以认为是对函数类的一种度量，通常可以简单的理解为一个问题的复杂程度，即 VC 维度越高，一个问题相应的也就越复杂。同时正是由于 SVM 关注的 VC 维，其在解决问题时和样本的维度是无关的，这个特性也使得 SVM 在处理很多类似于音乐或者文本分类这种高维分类问题时能有非常不错的效果。

本质上来讲，机器学习就是一种对某个问题真实模型的一种逼近，而真实模型显然是无论如何都没办法确切知道的，那么真实模型和假设的近似模型之间的自然会存在着一个误差，这个误差叫做风险。在获得一个近似模型之后（假设就是本文中的音乐分类器），真实误差显然无法知道，但可以选择能够掌握的量来尽可能的逼近它，比如使用样本数据通过分类器得到的分类结果与其对应的真实结果之间的误差值来表示。而准备的测试样本本来是已经标注过的，那么真实结果也就是知道的。这个差值称为经验风险 $R_{emp}(w)$。传统机器学习方法通常把经验最小化作为想要实现的目标，但由于其自身的泛化能力太差，因此而产生的结果往往是其在测试样本集上可以达到很好的分类效果而一旦进行真实分类效果就大打折扣。这是由于经验最小化原则能够适用的前提是计算得到的经验风险要真的可以逼近真实风险才行，而实际情况是获得的实验样本集相对于现实中的数据集，占的比例实在是太小了，根本无法

做到很好的逼近。

为了解决上面的问题，泛化误差界的概念被引入到机器学习当中，它指出真实风险应该由经验风险和置信风险两部分内容衡量，经验风险代表的是分类器在给定测试样本上的误差；置信风险则代表的是在多大程度上能够相信学习器对未知样本进行分类后得到的结果。然而由于后者根本没有办法十分精确的计算，所以只能给出一个估计的大概区间值，整个误差也就只能计算它的上界。置信风险主要和样本集规模和分类函数的 VC 维两个因素有关，显然给定的样本集规模越大，分类器的学习效果可能越好，相应的置信风险也就跟着变小；VC 维越高，泛化能力也就越差，置信风险会跟着变大。泛化误差界公式表示如下：

$$R(w) \le R_{emp}(w) + \phi(n/h) \tag{3.1}$$

其中 $R(w)$ 为真实风险，$R_{emp}(w)$ 为经验风险，$\phi(n/h)$ 为置信风险。综上，统计学习的目的便从原来的最小化经验风险变成了现在的期望经验风险和置信风险之和最小，也就是结构风险最小。支持向量机方法正是这样一种努力使结构风险最小化的算法，如图。

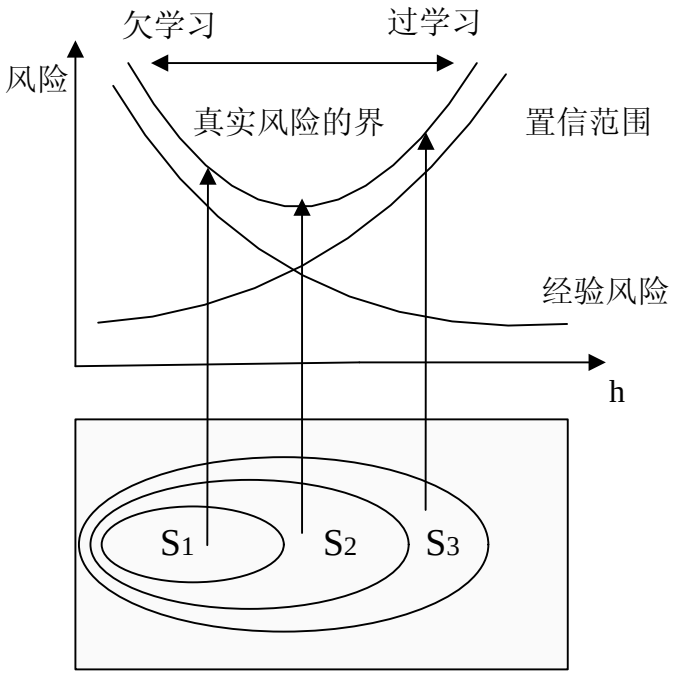

图 3.1 最小化结构风险原理图

3.1.2 SVM 基本原理

SVM 方法最初是用来解决二分类问题的，关于其在多分类中的应用将在下一节详细介绍。下面对其在二分类中的基本原理进行详解。

给定训练样本集 $X = \{x_1 \cdots x_n\}, X \in R^d$ 。对应的类别标注为 $\{y_1 \cdots y_n\}, y_i \in \{1,-1\}$ 。设

训练样本特征向量的维度为d，样本数量为n。

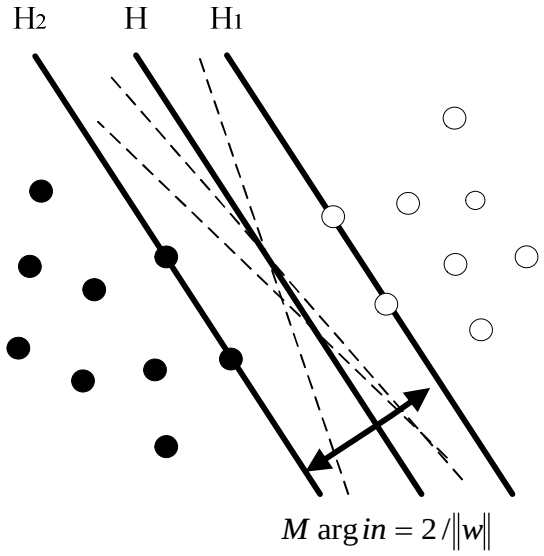

图 3.2 SVM 原理图

（1）线性支持向量机

对于线性可分的问题，二分类问题可以构造一个分类超平面使的正负样本能被完全分开。如图 3.2 所示。左边的实心样本点代表着正样本，右边的空心样本点代表了负样本。H_1和H_2之间存在若干分类面，它们都能够将正样本和负样本完全分开。如果其中一个分类面不仅可以将正负两类样本完全分开而且还能最大化几何间隔，那么这个分类线被称为最优分类超平面。所谓几何间隔就是H_1和H_2之间的距离。H是分类面，H_1和H_2是与H平行，并且同时经过距离H最近的两类样本的直线。而恰好落在H_1和H_2上的样本点也就是我们说的支持向量。正是这些支持向量共同构建起了得到的最优分类超平面。假设线性判别函数为$g(x)=wx+b$。通过归一化使$\{x_1 \cdots x_n\}$满足$g(x)\geq 1$，此时，分类间隔为$2/\|w\|$。所以，当

$$y_i[wx_i+b]-1\geq 0, i=1,\cdots,n \tag{3.2}$$

成立时，此分类器可以正确标注所有样本。显然，最大化分类间隔其实就是使$\|w\|$最小化。所以，最优分类超平面应该同时满足式（3.3）和使$\|w\|$最小。而支持向量机就是令式（3.3）中等式成立的样本。综上所述，最优分类超平面的求解问题便等价于求以下的约束优化问题：

$$\begin{aligned} &\min \quad \|w\|^2 / 2 \\ &s.t. \quad y_i[wx_i+b]-1\geq 0, i=1,\cdots,n \end{aligned} \tag{3.3}$$

这样就将 SVM 的求解最后转化为求解二次规划问题，因此从理论上讲 SVM 的解就是全

局唯一的最优解。

首先，构造拉格朗日函数：

$$M \arg in = 2/\|w\|$$

$$L(w,a,b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i y_i (x_i \cdot w + b) + \sum_{i=1}^{n} a_i,\ a_i \geq 0, i = 1,2,...,n \tag{3.4}$$

式中 a_i 为拉格朗日因子，然后分别对上式中的 w 和 b 求偏微分并令它们等于 0，得到 $w = \sum_i a_i y_i x_i$ 和 $\sum_i a_i y_i = 0$，把原来的最优化问题转化为对偶问题：

$$\begin{aligned} &\max\ W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j a_i a_j (x_i \cdot x_j) \\ &s.t.\quad \sum_{i=1}^{n} y_i a_i = 0 \quad a_i \geq 0,\ i = 1,2,3,...,n \end{aligned} \tag{3.5}$$

求解上式可以得到每个样本对应的 a_i 值，得到的解就是优化问题的最优解。只有不为 0 的 a_i 对应的样本才为支持向量，通常只有很小一部分的样本的 a_i 不为 0。最后的分类函数判别式如下：

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b\right] \tag{3.6}$$

$$b = \frac{1}{2}\left[\sum_{i=1}^{n} a_i y_i x_i \cdot x_r + \sum_{i=1}^{n} a_i y_i x_i \cdot x_s\right] \tag{3.7}$$

上式计算的 b 为偏斜量，式中的 a_i^* 不为 0 时，x_r 和 x_s 代表的是两类样本中任意一对支持向量。

现实情况中，往往由于受噪声影响，而使得分类样本不能被线性分开，从而也就无法得到一个无误的分类超平面。这里的噪声可以认为图 3.3 中最右边的黑色点，明显它实际上是负类的一个样本，这奇异的一个样本使得原本线性可分的问题变的线性不可分了，通常把这类问题叫做“近似线性可分”。对于这类问题，我们通常的处理方法是认为那个样本点本来就是用户标注样本时不小心标注错的，是干扰，是噪声，应该忽略掉它的存在。但是它的存在又确实造成了问题的不可解，于是针对这种情况，我们处理的方法是允许一小部分样本点到分类超平面的距离不必满足原来的要求，也就是说原来我们要求所有样本点到分类超平面的间隔都应该起码大于 1，现在加入容错性，允许给 1 这个硬性的阈值加上一个松弛变量，即允许某些样本点落到几何间隔内，表达式变为如下形式：

$$y_i[wx_i+b]\geq 1-\xi_i$$
$$\xi_i\geq 0, i=1,\cdots,n \tag{3.8}$$

松弛变量是非负的，就是说最终得到的结果是样本的间隔允许比 1 小，当计算出这部分样本点的间隔小于 1 时，意味着分类器放弃了对这些奇异点的精确分类，虽然这本身会给分类器带来些许的损失，但是同时也使得分类超平面不必再受这少数样本点的影响而向这些样本点移动，可以得到更大的几何间隔。所以这两者之间需要多加权衡。

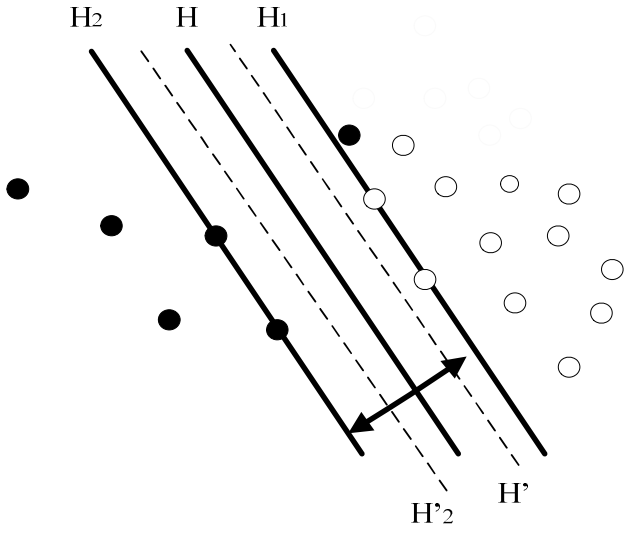

图 3.3 奇异点

由上文已经知道$\|w\|^2$是目标函数，期望它的值越小越好，所以损失应该是一个能使$\|w\|^2$变大的量。通常有两种方法来衡量损失，第一种是二阶软间隔分类器：

$$\sum_{i=1}^{n}\xi_i^2 \tag{3.9}$$

另一种是一阶软间隔分类器：

$$\sum_{i=1}^{n}\xi_i \tag{3.10}$$

把损失加入到目标函数需要一个惩罚因子，于是原来的求优问题可以写成如下形式：

$$\begin{aligned} &\min \quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad y_i\left[wx_i+b\right]\geq 1-\xi_i,\ \xi_i\geq 0 \quad i=1,2,3,...,n \end{aligned} \tag{3.11}$$

（2）非线性支持向量机

上面介绍了支持向量机在解决线性可分问题以及“近似线性可分问题”的基本原理。但

是在现实世界里，问题要复杂的多。即很多时候，在原本的低维样本空间，样本是极度不可分的，无论怎么样来寻求分类超平面，总有很多的奇异点不满足要求，这个时候就需要将在低维空间里线性不可分的样本数据向高维空间映射，虽然映射之后也还不是完全线性可分的，但起码是“近似线性可分”的，然后借助松弛变量来处理那少部分奇异点就可以达到非常不错的效果。将样本从低维空间向高维空间的映射需要借助核函数来实现，令核函数为：

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{3.12}$$

核函数本身必须满足 Mercer 条件，它的基本作用就是输入两个低维空间里的向量，进而能够计算出经过某个变换后的高维空间的向量内积值。因此原问题可以转换成如下形式：

$$\begin{aligned} &\max \ W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j a_i a_j K(x_i \cdot x_j) \\ &s.t. \quad \sum_{i=1}^{n} y_i a_i = 0 \quad 0 \le a_i \le C, \ i = 1,2,3,..., n \end{aligned} \tag{3.13}$$

判别函数变为：

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} a_i^* y_i K(x_i \cdot x) + b \right] \tag{3.14}$$

$$b = \frac{1}{2}\left[\sum_{i=1}^{n} a_i y_i K(x_i \cdot x_r) + \sum_{i=1}^{n} a_i y_i K(x_i \cdot x_s) \right] \tag{3.15}$$

3.1.3 核函数介绍

核函数使得支持向量机在处理非线性可分问题时能有良好的表现。不同核函数构造出来的非线性分类器也是不同的，在处理实际问题时，对于核函数的选择目前尚缺少指导原则，更多的需要通过实验验证之后选取最佳核函数。目前常用的核函数有以下几种：

（1）线性核函数：

$$K(x, x_i) = (x_i \cdot x) \tag{3.16}$$

（2）多项式核函数：

$$K(x, x_i) = [p(x_i \cdot x) + s]^q \tag{3.17}$$

（3）*Sig* mod 核函数：

$$K(x, x_i) = \tanh(\mu(x_i \cdot x) + c) \tag{3.18}$$

（4）径向基（RBF）核函数：

$$K(x, x_i) = \exp(-\gamma |x - x_i|^2) \tag{3.19}$$

以上几种核函数应用最为广泛的是径向基核函数，它既有较广的收敛域，又适用于低维、高维、小样本、大样本等各种情况。本文在进行音乐分类时也是选取了表现最好的径向基核函数，γ 取值为 8。

3.1.4 多分类介绍

SVM 算法最初只是用来针对解决二分类问题的，然而在包括音乐分类在内的实际应用中，分类问题几乎全部是多分类问题。为了解决这个矛盾，目前主要有两种策略：一种是确定多个分类面，对每个分类面的求解过程整合为比较大型的最优化求取问题，对这个最优化过程求解来实现多类分类。然而在进行优化求解时会比第一种方法处理的变量要多的多，尤其是当分类类别比较多时，计算量会大到无法实用的地步。于是人们更多的采用是第二种策略，其基本思想是把多分类问题通过某种方式分解为多个二元分类器，通过组合这些二元分类器来实现多分类，有如下几种常用方法：

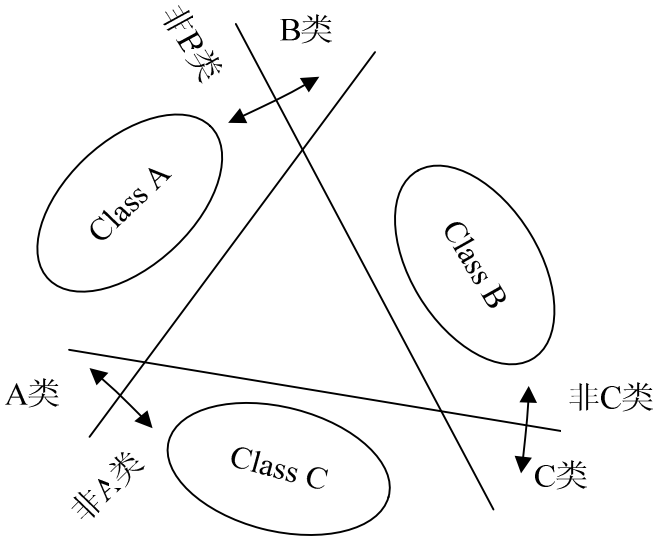

图 3.4“一对其余”方法

（1）一对其余分类策略（1-v-r one versus the rest）

一对其余的分类策略是最早被提出来解决 SVM 多分类问题的一种思想十分简单的方法，基本思想就是第一次将第一个类别的样本定为正样本，其余所有样本都被认为是负样本，这样便得到了一个二元分类器，它可以判断样本是不是第一类的；第二次将第二个类别定为正样本，其余所有样本定为负样本，同样得到一个二元分类器，可以判断该样本是不是第二类的，依次类推。不难得出，如果有 n 个类别，最终便可以得到 n 个分类器，如图 3.4 所示。这样当需要对某个样本分类时，只需要用这 n 个分类器依次判断其是不是属于自己那一类就

可以了。这种分类策略的优点是针对每个优化问题的规模都很小，而且由于调用的分类器数目较少，因此分类时候的速度也很快。但其也有着明显的缺点，即可能导致分类重叠或分类不可分问题。分类重叠问题就是调用完所有的 n 个分类器，有两个或多个分类器都判定其属于自己那一类，这种问题相对容易解决，一般将类别判别给距离各个超平面的距离最远的那个分类器。分类不可分问题相对就比较麻烦，因为所有的分类器都判定待分类样本不是自己那一类。另外，这种分类策略人为造成了数据集的偏斜问题，样本类别越多越严重。

（2）一对一分类策略（1-v-1 one versus one）

一对一分类策略同样是将多分类问题拆分成多个二分类问题，只不过其每次选一类样本作为正样本，同样只选另一类作为负样本，训练得到一个分类器，这样对于包含 n 个类别的样本，最终可以得到 $n(n-1)/2$ 个二元分类器，如图 3.5 所示。虽然分类器数目相对于一对其余策略多了不少，但是在训练时候所用的总时间相对于一对其余策略少得多。当待分类样本进入分类系统，第一个分类器会投票判断其属于哪一类，让每个分类器都进行投票，最终统计每个样本的得票数，对应每个样本的得票如果哪个类别的投票最多，就判断这个样本属于那一类。虽然这种分类策略还是可能会出现分类重叠问题，但是避免了不可分问题。然而当类别数目过多时，这种分类策略在进行分类时会变得很慢，因为需要调用的分类器数目是平方级增长的。

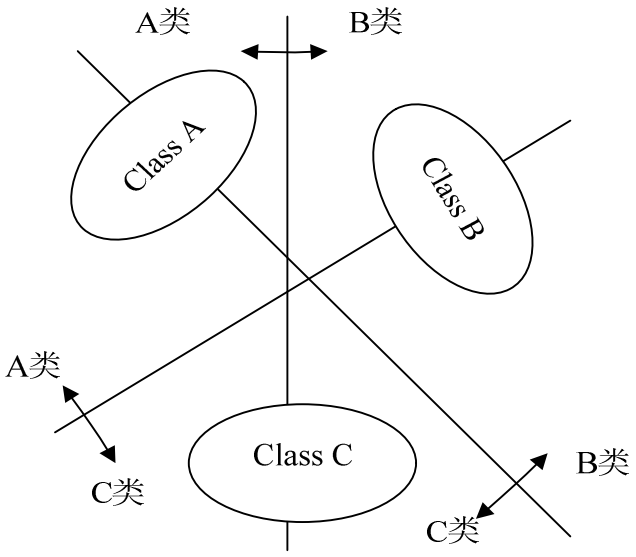

图 3.5 “一对一”方法

（3）有向无环图策略（DAG Directed Acyclic Graph）

有向无环图策略和一对一策略一样，也需要构建 $n(n-1)/2$ 个分类器，但是在测试分类时它采用的是有 $n(n-1)/2$ 个节点和 n 个分支的有向无环图方法，如图 3.6 所示，每次样本点测

试都是从根节点开始，下面的每个节点都是一个二元分类器，根据输出值决定是向左还是向右移动到下一个节点，直到最后得到输出结果。这个过程其实仅仅调用了 $(n-1)$ 个分类器，分类速度很快，而且既没有分类重叠问题，也没有不可分问题。其局限在于对根节点分类器的分类准确率要求非常高，因为一旦根节点分类器判断错误，后面无论如何也无法纠正这个错误了，即错位向下累积现象。因此，在对根节点分类器的选取时，往往会选择差别特别特别大的两个类别。大到分类器把样本分错的概率很低。

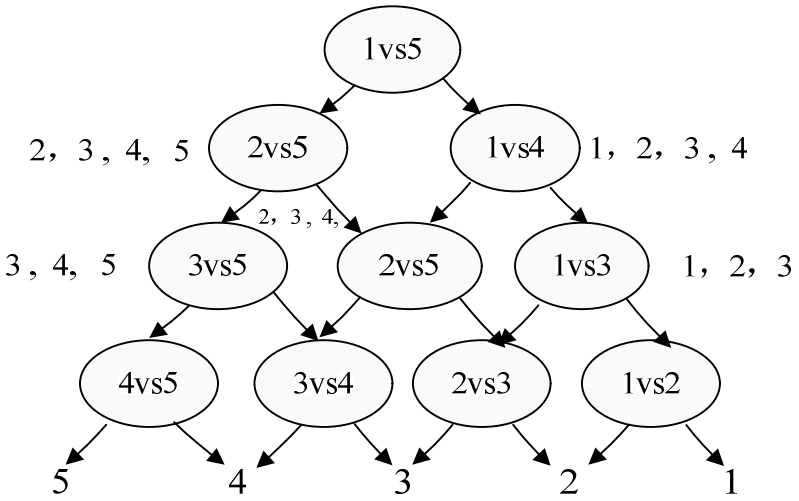

图 3.6 有向无环图方法

经过上面简单的介绍，表 3.1 对这个三种方法在分类速度、构建分类器数目等性能上进行一下比较。

表 3.1 多类 SVM 方法比较

	1-v-1	1-v-r	DAG
分解方法	一对一	一对其余	一对一
输出判决策略	投票法	最大输出法	有向无环图
二分类器数量	N(N-1)/2	N	N(N-1)/2
训练快慢	慢	慢	慢
分类快慢	慢	慢	快
分类精度	高	高	高
泛化误差	无界	无界	有界

本文在后续的实验仿真过程中选用了原理相对简单的“一对其余”多分类方法，因为本文的实验目的并不是减少运算时间，而是减少人工标注的工作量，增加的些许机器运算时间与大量的人工标注时间相比是可以忽略的。

3.2 常用的 SVM 主动学习方法

在上一章已经详细介绍了主动学习方法核心模块：选择引擎中常用的几种样本选择策略，这一章的前半部分也详细介绍了目前最流行的学习器之一：支持向量机的相关理论知识基础。下面本文将接着把支持向量机这一高效的学习器与主动学习算法相结合，先介绍几种常用的 SVM 主动学习方法，然后基于传统的 SVM 主动学习方法做出相应的改进。

3.2.1 基于不确定度缩减的 SVM 主动学习

前面已经说过，通俗讲基于不确定度缩减的主动学习方法认为类别最不确定的样本本身具备的价值也最大。而样本的不确定性通常有两种方法来衡量：第一种，就是利用信息学中的信息熵来判断样本的价值，因为样本不确定性与信息熵是成正比的，信息熵越大样本的类别越不确定，其所包含的信息量也就越大，也就越有价值，反之亦然。另外一种是从几何学角度来看。这种方法认为距离分类超平面最近的样本点的类别最模糊不清，当分类超平面发生移动时，这些样本最容易被影响。基于这种理论，将这种选择策略应用到 SVM 上面就是说样本的不确定性与样本到分类超平面的距离成反比。因为 SVM 训练得到的分类器，其两端的间隔线由少数的支持向量决定了，间隔外的样本可以被正确分类，这些样本即使加入到原来的训练集也不能替代原来的支持向量，分类超平面也不会被改变，而处在分类间隔内的样本点则有可能被分错，把这部分样本点加入到原来的训练集后支持向量就会发生变化，从而分类超平面也会跟着变化。便有了如下形式：

$$dis = \left| w \cdot x + b \right| = \left| f(x) \right| \tag{3.20}$$

这种方法是原理最简单的主动学习方法，充分利用样本点包含的隐藏信息，每次主动选取那些距离分类超平面最近的样本点作为新的样本点进行训练，这些样本点对于当前分类器来说都是最不确定的样本点，它们的加入可以较大幅度的修正分类超平面。随着迭代次数的增加，剩余的样本信息量也就越来越少，对分类超平面的修正幅度也会逐步减小。这种策略所基于的核心思想正是 SVM 的分类超平面仅仅与支持向量有关而与其他向量无关。

上述方法思想浅显易懂，但是也有着自身的局限性。关于未标注样本信息熵的定义还不是很完善，仅仅依靠样本到分类超平面的距离来选取价值样本又容易选到奇异点。

3.2.2 基于版本空间缩减的 SVM 主动学习

Simon Tong 于 1998 年提出了基于版本空间的主动学习的相关理论。给定一个已标注的

训练集 L 和一个 Mercer 核函数 K，在特征空间 F 可以得到一系列分割超平面，称这一系列的超平面为特征空间。换句话说，版本空间中的任意一个分类超平面 f，对于每一个训练样本 x_i，其标签为 y_i，都有 $y_i=1$ 时 $f(x_i)>0$，$y_i=-1$ 时 $f(x_i)<0$。这一系列可能的超平面可以写成如下形式：

$$\mathrm{H}=\{f \mid f(x)=\frac{w\cdot\Phi(x)}{\|w\|} \quad w\in W\} \tag{3.21}$$

简单讲这里的参数空间 w 也就是 F，特征空间 V 可以定义如下：

$$V=\{f\in H \mid \forall i\in\{1...n\} \quad y_i f(x_i)>0\} \tag{3.22}$$

注意到 H 其实是一系列的超平面，在向量 w 和超平面 f 之间有着一一对应的关系，因此对 V 重新定义：

$$V=\{w\in W \mid \|w\|=1,\ y_i\cdot(w\cdot\Phi(x_i))>0,\ i=1,2,...,n\} \tag{3.23}$$

W 为参数空间，定义 $Area(v)$ 为版本空间大小，代表 $\|w\|=1$ 时得到的表面积，不难看出版本空间只有在训练数据集在特征空间是线性可分时才存在。也就是说特征空间 F 和参数空间 W 之间存在对偶关系：在特征空间 F 的点映射到参数空间 W 中其实是一个超平面，反之亦然。

假设给定一个新的训练样本 x_i，以及对应的标签 y_i，那么对于任何分类超平面都要满足 $y_i(w\cdot\Phi(x_i))>0$。现在不再认为 w 是特征空间 F 中的一个普通向量，而是认为 $y_i\Phi(x_i)$ 是参数空间 W 中的一个普通向量，因此 $y_i(w\cdot\Phi(x_i))=w\cdot y_i\Phi(x_i)>0$ 定义了 W 中的一半空间，$w\cdot y_i\Phi(x_i))=0$ 则定义了 W 中的一个超平面，它代表了版本空间 V 中的边界之一。

根据上节对支持向量机的介绍，对应版本空间可以写成如下形式：

$$\begin{aligned} \max imize_{\ w\in F} \quad & \min_i\{y_i\cdot(w\cdot\Phi(x_i))\} \\ s.t. \quad & \|w\| \ =1 \\ & y_i\cdot(w\cdot\Phi(x_i))>0,\ i=1,2,...,n \end{aligned} \tag{3.24}$$

希望可以在版本空间找到一个 w^* 点，这个点可以使到任何一个超平面的最小值最大化，支持向量机方法需要找到一个超球，其中间恰好在版本空间上且其表面没有被已标注样本对应的超平面分割，超球半径为球心到任意分割面的距离。

版本空间越大，相应的假设超平面也就越多，分类器效果也就越差，所以希望选取的样

本可以尽可能快的收缩版本空间的大小。Simon Tong 的选择策略是每次都去选取那些可以平分版本空间的样本，也就是说每加入一个新样本重新训练后得到的新的版本空间大概是原来版本空间的一半，即：

$$Area(v_{i+1}^{*}) = \frac{1}{2} Area(v_i^{*}) \tag{3.25}$$

然而现实情况是每次计算版本空间大小并不实际，对此 Tong 提出了三种解决方法：

（1）简单间隔：

给定一些样本数据$\{x_1, x_2, x_3, ..., x_i\}$以及对应标签$\{y_1, y_2, y_3, ..., y_i\}$，通过 SVM 方法训练得到的$w_i$差不多可以认为是当前版本空间$V$中得到的最大超球的球心。如图所示。然后开始对尚未标注的样本池中的每一个样本进行测试，不难发现参数空间中的超平面距离 w_i 越近，那么这个超平面也就距离版本空间的中心越近，也就越有可能最大化的平分版本空间。于是，此时选择策略也就是选取参数空间W中那些距离 w_i 最近的超平面所对应的样本点。参数空间W中的超平面到 w_i 的距离可以近似认为是特征空间中特征向量 $\Phi(x)$ 到超平面 w_i 的距离，这个距离可以由$|w_i \cdot \Phi(x)|$简单计算得到。根据这种选择策略，图 3.7 中，选取的价值样本点将会是样本 a。

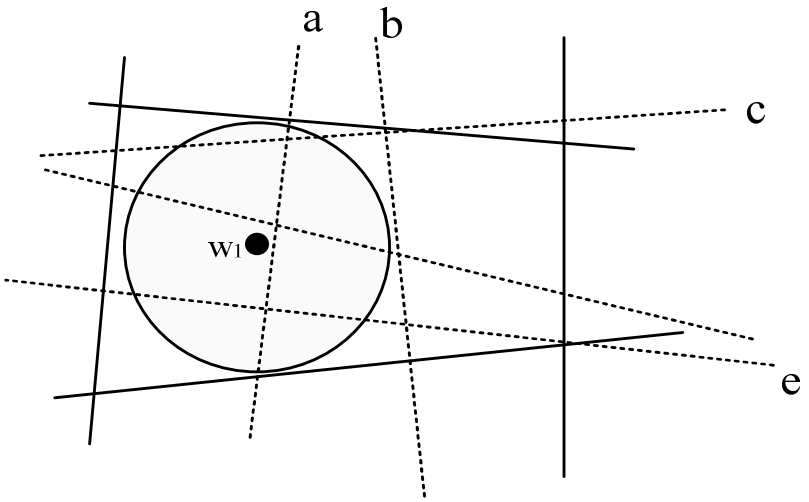

图 3.7 基于版本空间缩减的 SVM 主动学习（简单间隔）

（2）最大最小间隔：

简单间隔选择策略的效果好坏很大程度上需要依赖于得到的 w_i 有多大程度的靠近超球的中心。如果这个条件不能严格成立，在选取价值样本时，甚至可能会出现选取的样本和超球都不相交的情况，更别说平分版本空间了。针对这个问题，Tong 提出了最大最小间隔。假

设 m_i 是当前版本空间超球的半径，此时从未标注样本池中选取样本，首先将一个未标注的样本 x_i 标注为 1，形成的新的版本空间的半径为 m^+，然后将其标注为-1，形成的新的版本空间超球的半径变成了 m^-。此时希望新加入的这个样本可以最大化的将原来的版本空间平分，即希望 $Area(V^+)$ 和 $Area(V^-)$ 尽可能的相同。因此，选取的样本需要满足 $\min(m^-, m^+)$ 条件。按照这种选择策略，图 3.8 中，选取的价值样本将会是样本 b。

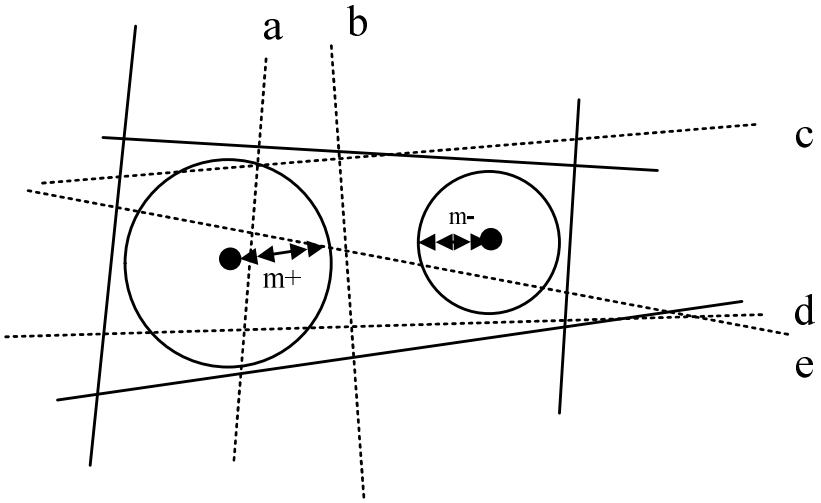

图 3.8 基于版本空间缩减的 SVM 主动学习（最大最小间隔）

（3）最大比例间隔：

最大化比例间隔方法与最大最小间隔方法类似。同样是对新加入的样本分别标注 1 和-1，得到相应的 m^+ 和 m^-。考虑到初始版本空间可能过于细长，对应形成的 m^+ 和 m^- 都太小，于是，将最大最小间隔方法的选择策略变为 $\min(\frac{m^-}{m^+}, \frac{m^+}{m^-})$。按照这种选择策略，图 3.9 中，选取的价值样本将会是 e。

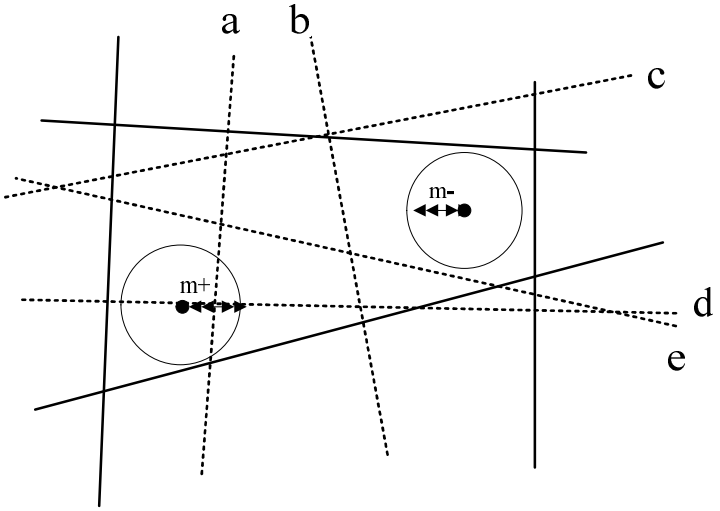

图 3.9 基于版本空间缩减的 SVM 主动学习（最大比例间隔）

总的来说，基于版本空间缩减的这种方法以样本在每次迭代中能后减少版本空间规模的大小幅度为其价值的衡量标准。该方法在处理二分类问题时可以取得相当不错的效果，但是在多分类问题中，目前的研究相对较少。

3.3 本文的 SVM 主动学习方法

3.3.1 样本多样性

上文介绍到的基于不确定度缩减的主动学习中所谓“最有价值”的样本实际上就是分类器最不确定的样本。而 SVM 分类器的目的是找到一个最优分类超平面以最大间隔将两类数据分开。而每次迭代时，未标注样本中选取几何间隔内的样本加入训练后，所得到的新的分类器的分类面位置最有可能变化，而几何间隔外的样本对新的分类器的分类面位置影响并不大。所以，在传统的 SVM 主动学习方法中认为对于 SVM 分类器而言，那些距离最优分类超平面最近的样本点就是最有价值的样本点，但是如果仅仅以此作为判断样本是否是最有价值样本的标准可能会面临重复学习问题，因为实际情况中样本集的数量往往是相当庞大的，所以每次迭代选取价值样本时需要批量选取，而每次批量选取的价值样本很可能因为相互之间的相关度太大而出现信息冗余，这将导致重复学习。换句话说，希望分类器每次迭代选取的样本集不仅仅具备最不确定性，同时还希望这个样本集保持多样性，从而最大化修正分类超平面。

也就是说，样本多样性的保持，其实就是希望样本相互之间的相关度越低越好。关于相关度的衡量方法，比较简单的就是使用欧氏距离来衡量，欧氏距离计算方式如下：

$$d(x_i, x_j) = sqrt(\sum (x_{il} - x_{jl})^2), l = 1,2,..., n \qquad (3.26)$$

其中 x_i 和 x_j 代表任意两个样本的特征向量，x_{il} 和 x_{jl} 分别表示特征向量 x_i 和 x_j 的第 l 维特征参数。

两个样本的欧氏距离越大说明两个样本之间的相关度越低，样本的多样性也就越好。然而音乐分类中的音乐样本在进行特征提取之后，样本的特征向量维度已经达到了 100 维以上，欧氏距离已经不再能够准确衡量高维向量之间相关度了。针对高维特征向量，样本的多样性可以通过样本间的角度来衡量，将特征空间的样本点映射到版本空间其实是一个超平面。那么两个样本间的角度可以用其对应的两个超平面 h_i 和 h_j 间的角度表示，根据核函数 K 可以写成如下形式：

$$\left|\cos(\angle(h_i,h_j))\right| = \frac{\left|\Phi(x_i)\cdot\Phi(x_j)\right|}{\left\|\Phi(x_i)\right\|\cdot\left\|\Phi(x_j)\right\|} = \frac{\left|K(x_i,x_j)\right|}{\sqrt{K(x_i,x_i)K(x_j,x_j)}} \quad (3.27)$$

式中 $\Phi(x_i)$，$\Phi(x_j)$ 表示样本对应超平面的法向量。公式（3.27）求得的是两个样本夹角的余弦值的绝对值，其值越小意味着两个样本间的角度越大，也就是相关度越小。

为保证每次批量选取样本集的多样性，首先根据已标注的初始样本集训练得到一个分类超平面 h，可以计算每个未标注样本 x_i 到这个超平面的距离。同时定义 S 为每次批量选取的样本集，其初始为 ϕ。 未标注样本 x_i 和当前选取样本集 S 的角度定义为 x_i 和 S 中全部样本角度中的最大角度，这个角度衡量了最终选取的样本集的多样性，并以此作为样本是否被选为价值样本的判断因素。那么样本 x_i 多样性判断标准可写成如下形式：

$$div(x_i) = \min_{x_j \in S} \frac{\left|K(x_i,x_j)\right|}{\sqrt{K(x_i,x_i)K(x_j,x_j)}} \quad (3.28)$$

这里需要注意的是，主动学习本身已经是个迭代过程，即每一次训练就代表了一次迭代，而每次迭代前在对未标注样本进行批量选取时 S 中的样本是逐一递增的，直到达到设定的批量值，也就是说样本集 S 的形成本身也算是一个内部迭代过程。因为 S 的初始为空，其加入的第一个样本就认为是距离分类超平面最近的样本，后续样本的添加则按照公式（3.28）计算得到最小值的样本进行添加。详尽的批量选取步骤将在后面的主动学习算法中给出。

既然选出的样本点需要兼顾到样本到超平面的距离（即自身的不确定性）和样本本身的多样性两点，那么这里引入参数 α 来对以上两个选取条件进行权衡，并制定了样本 x_i 价值的评价标准如下所示：

$$score = \alpha\cdot\left|f(x_i)\right| + (1-\alpha)\cdot\left(\min_{x_j \in S} \frac{\left|K(x_i,x_j)\right|}{\sqrt{K(x_i,x_i)K(x_j,x_j)}}\right) \quad (3.29)$$

上式前半部分函数 $f(x_i)$ 计算了未标注样本 x_i 到当前分类超平面的距离，这是对样本不确定性的判断因素，希望它越小越好。上式后半部分是按照上文提出的 S 形成策略计算了未标注样本和 S 的角度的余弦值，希望它们之间的角度越大越好，那么相应的自然是式子本身求得的余弦值越小越好。α 在其中则起到了权衡的作用，其取值范围为 0 到 1，如果取值大于 0.5 意味着更看重样本的不确定性，反之意味着更看重样本的多样性。综上，$score$ 值越小意味着样本的价值越大。

3.3.2 样本平衡性

由于本文在进行音乐分类时采取了“一对其余”的多分类方法，也就不可避免的人为造成了数据集的偏斜，即某一类样本的数量要远多于另外一类样本，如图 3.10 所示。如果不对样本集进行处理，将会使算法的学习能力下降从而影响最终的分类准确率。

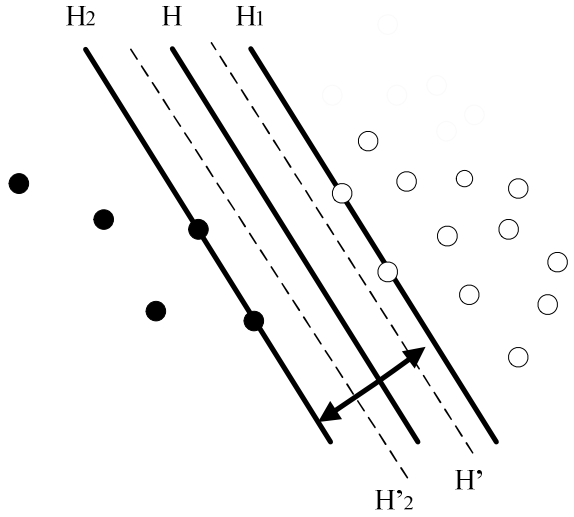

图 3.10 数据集偏斜

为了避免出现数据集不平衡的问题，在每次迭代使用本文提出的样本选择策略选出价值样本集后都对样本集的平衡度 b 进行检测，定义如下：

$$b=\begin{cases} pos/neg & pos\le neg \\ neg/pos & else \end{cases} \tag{3.30}$$

其中， pos 为正样本数目， neg 为负样本数目。同时设定一个阈值 ε ，当 b 不大于 ε 就认为数据集是偏斜的，此时就将多数的那类数据进行聚类，聚类个数与少数样本的数目相同，然后将与聚类中心最接近的多数类样本与少数类样本加入训练集，将多数类的其他样本除去。

综合考虑上面两处改进，做出如下假设：

集合 U 为尚未标注的候选样本集；集合 S 中的元素为每次迭代批量选取的需要标注的最有价值样本，其初始值为 ϕ ，且在每次迭代前都要清空；批量选取值设为 m ； $Train$ 中的样本为人工标记过的所有样本，用作 SVM 训练集，其初始值同样是 ϕ 。

基于以上假设，本文提出的改进的 SVM 主动学习的具体步骤如下：

1） 从候选样本集 U 中通过聚类算法选取 n 个样本并标注其类别，将其作为初始训练样本集 $Train$ ，并保证 $Train$ 中至少包含一个正类样本和一个负类样本，令 $U=U-Train$ ；

2） 用 SVM 算法训练 $Train$ 得到最优分类超平面；

3） 构建批量选取的样本集 S；

3.1）判断样本集 S 中的样本个数是否小于 m，如果小于 m，跳至 3.2），否则跳至 4）步；

3.2）按照公式 （3.29）对 U 中的所有样本计算它的 $score$ 值，选取 $score$ 值最小的样本 x_s，并执行 $S = S \cup \{x_s\}$，跳至 3.1）步；

4）人工标注中样本点的 S 类别，按照公式计算当前 S 集的 b，$b < \varepsilon$，执行上文提出的平衡性调整方法；

5）执行 $Train = Train \cup S$，$U = U - S$，当达到设定的迭代次数或条件跳至第 6）步，否则返回到第 2）步。

6）算法结束。

在后面的实验中会发现，每次迭代都要重新训练分类器，这样虽然减少了人工标注，但是增加了训练的次数也就是机器的运算时间，不过在实际应用中，因为每次选取最有价值的样本只需要计算所有样本到分类超平面的距离和相互之间的角度，并没有增加太多计算量，而且与人工标注全部样本所耗费时间和人力相比，增加的少许机器运算时间是完全可以接受的。

3.4 本章小结

本章首先详细介绍了 SVM 的相关理论基础，接着介绍了 SVM 在解决多分类问题时常使用的“一对其余”、“一对一”、“DAG”等方法的基本原理以及各自的优缺点，然后介绍了几种常用的传统主动学习方法并指出各自的优点与不足，并同时提出了本文在传统方法的基础上改进过的基于 SVM 主动学习方法，即在考虑样本不确定性的同时也兼顾样本的多样性，避免选取到奇异点，同时针对多分类中“一对其余”方法人为造成的数据集偏斜问题进行了进一步的平衡性调整处理。

第四章 实验结果与仿真分析

前面几章主要介绍音乐分类和基于 SVM 主动学习理论的基本原理，并提出了一种改进的 SVM 主动学习方法。本章将通过实验仿真，将上文提出的 SVM 主动学习方法应用到音乐流派的分类当中，根据分类效果进一步验证上一章提出的理论，并与传统的理论方法进行比较。

4.1 实验数据构造与系统框架

4.1.1 实验数据

本实验建立的音乐数据库包含了 5 种音乐风格类别：舞曲、抒情、爵士、民乐和摇滚。数据库中的全部歌曲都下载于互联网的大型音乐网站，下载格式为 mp3。用于训练和测试的音乐数据集都是人为主观标注的。其中每种风格歌曲有六百多首共三千多首歌曲，邀请了 50 名学生分为 10 组，每组有 5 人，让这 5 人对该组约三百多首歌曲进行分类标注。在分类标注不确定的时候允许标注者反复听，直到正确给出标注。在标注结束后，每首歌曲有 5 个标注标签，只有一首歌曲中的 5 个分类标签中存在多于或等于 3 个相同标签时才标注为此类，才允许将该歌曲放入音乐数据库中。最后得到的音乐训练库与测试库的构成如下表所示：

表 4.1 音乐训练和测试数据库

	舞曲	抒情	爵士	中华民乐	摇滚
训练库歌曲数	500	500	500	500	500
测试库歌曲数	182	182	178	124	135

上文中已经提到，在后面的实验仿真过程中，决定选用 wav 这可以记录音乐文件原始波形的音乐格式。所以在对音乐文件进行分类之前要对音乐样本库中的音乐进行格式转换，这里选用了 Format Factory 格式转换工具把数据库中的所有 mp3 音乐文件转换成了 wav 格式。另外，由于完整的一首歌的时间比较长，对其提取特征向量的数据会比较庞大，而且往往一首歌的一部分已经足够反映一首歌的音乐特征，所以截取每首歌曲的中间部分 30 秒作为音乐样本，并取单声道，抽样率设置为 16KHZ。

4.1.2 系统框架

本文的仿真环境是基于 MATLAB 环境下的，首先对训练音乐样本数据库提取特征并进行训练得到 SVM 分类器模型，然后对测试音乐库进行特征提取并利用已经得到的 SVM 训练模型对测试音乐进行分类，其对测试集音乐样本分类的准确率将作为评估 SVM 主动学习应用于音乐风格分类系统有效性的指标之一。其中音乐风格分类准确率的定义为：

$$C_{correct} = \frac{N_{correct}}{N_{total}} \times 100\ \% \tag{4.1}$$

其中 $N_{correct}$ 是被正确分类的测试集音乐样本数，N_{total} 是测试集音乐样本数。

除了准确率这个指标之外，为了验证 SVM 主动学习的有效性，还需要另外一个指标，即训练样本的标注数 L_{train} ，本指标需要在固定某个准确率的情况下进行验证。

根据前面介绍，本文的 SVM 主动学习音乐分类系统的结构设计如下：

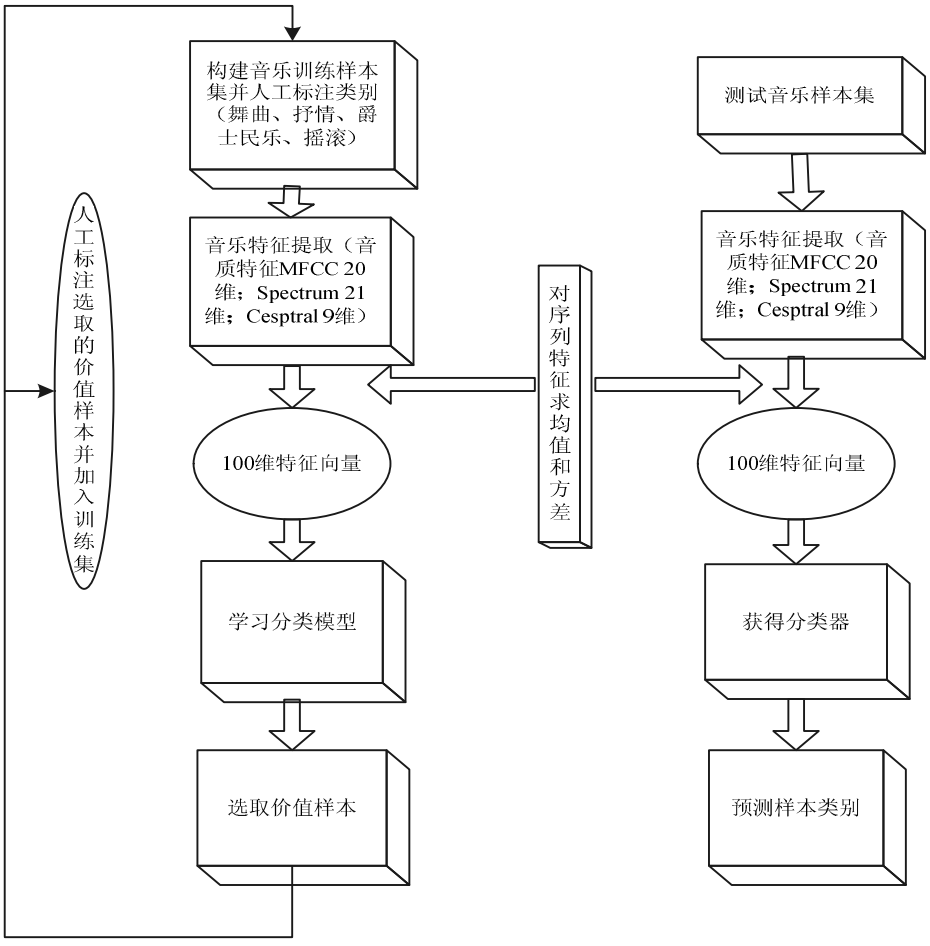

图 4.1 基于 SVM 主动学习的音乐分类系统

将 SVM 主动学习方法应用于音乐流派分类的详细步骤描述如下：

首先将音乐库中事先下载下来的 MP3 格式的音乐文件转换成 wav 格式，然后截取中间部分的 30 秒音乐片段，采样率设置为 16KHZ，声道设置为单声道，基于 SVM 主动学习的音乐风格分类系统可以简单概括为如下：

（1） 对处理后的音乐片段进行预处理：对 30 秒的音乐片段添加 32ms 的汉明窗，其采样率设置为 16KHZ，所以一帧有 512 个采样点，帧移为 16ms，即 256 个采样点，这样一首歌曲得到 1875 帧。

（2） 提取音乐样本特征向量：对经过预处理的每一首歌曲提取其前 20 维 MFCC；取 RASTA-PLP 倒谱 9 维，RASTA-PLP 频谱 21 维，然后对得到的 MFCC 和 RASTA-PLP 特征向量分别求其均值和方差，这样每首音乐片段样本由 100 维的特征向量表示。

（3） 音乐分类模型的形成：从已经特征提取完的全部训练样本集中选取极少部分的音乐样本标注后加入训练集作为初始训练样本集，并通过 SVM 算法训练形成初始分类模型。根据本文提出的 SVM 主动学习方法选出剩余未标注样本中的价值样本重新加入训练集形成新的分类模型，如此迭代往复直到达到停止指标；

本文中音乐分类系统中的 SVM 器选用了 SVM-light 分类器，关于该分类器下一节将详细介绍。针对本文的音乐多分类问题，选用了 SVM 处理多分类问题时的“一对其余方法”，每个音乐样本保证只标注一次，已经标注过的样本不再重复标注，即不加入标注次数指标。

4.2 SVM-light 分类器

本文所使用的 SVM 分类程序 SVM-light 是由 Thorsten Joachims 基于 C 语言实现的，有着如下特点：使用了快速优化算法；可以解决分类、回归以及排序问题；可以高效的处理成千上万的支持向量；可以处理几十万的训练样本；使用了稀疏向量等。这是一个操作非常简单，并且使用起来十分快速有效的 SVM 软件包。SVM-light 包含两大功能模块：学习模块（svm_learn.exe）和分类模块（svm_classify.exe）。同时这个软件包提供了各种主流计算机语言的实现接口。本文选用了基于 matlab 实现的接口版本。在 matlab 中主要调用的两个函数如下：

SVM 训练的格式为：

svm_learn(MySVMOptions,filename,modelname);

输入参数：

MySVMOptions：训练参数选项；

filename：保存有训练特征及其标注的文件名；

输出参数：

modelname：该类的模型参数保存在 Modelname 文件中。

SVM 分类的格式为：

svm_classify(MySVMOptions,filename,modelname,predfilename);

输入参数：

MySVMOptions：训练参数选项；

filename：保存有预测样本特征及其标注的文件名；

modelname：该类的模型参数保存在 Modelname 文件中。

输出参数：

prefilename：保存预测样本的类别信息。

4.3 实验仿真与分析

为了验证本文提出的 SVM 主动学习方法的有效性，做了两组实验。第一组实验，与简单以样本到分类超平面的距离为判断标准的传统 SVM 主动学习方法和 SVM 随机采样方法在分类准确率的收敛速度上进行了对比；第二组实验，固定想要达到的准确率，对需要标注的样本数目进行了对比。

（1）第一组实验 分类准确率收敛速度

为了更好的对三种学习方法（实验中把本文提出的兼顾多样性主动学习算法又拆分成平衡调整前和调整后两种情况）的效果进行对比，从 2500 个训练样本中随机选取 100 个样本作为初始训练样本进行标注，每次试验三种算法的初始样本相同，实验设置了不同的批量选取值 m，并统计每次迭代的分类准确率，同时把平衡性调整参数 b 的阈值 ε 设置为 0.5，对前 20 次实验记录的准确率做了平均，图 4.2 直观的将对比结果进行了展示：

仔细观察图 4.2 发现，本文提出的 SVM 主动学习方法在整个迭代过程中，无论是分类准确率还是收敛速度都要优于其他两种方法，对选取的样本进行平衡调整后的准确率也要比不进行调整的准确率高。另外观察前面几次迭代，这几种学习方法的差距并不大，传统的 SVM 主动学习方法甚至不如 SVM 随机采样方法，猜测可能是由于初始样本过少，导致其与全部样本的特性偏差较大，随机采样算法更有机会选到那些相互之间相关度小的样本，从而验证了之前所说的并不是距离超平面最近的样本点就一定是信息量最大的样本点，需要考虑样本信息冗余的情况；后面随着迭代次数的增加，分类超平面一次次的修正，分类器越来越能代表

全部样本的特性，两种主动学习分类器分类准确率的收敛速度便开始明显快于随机采样 SVM 分类器，而本文提出的改进 SVM 主动学习方法的效果最优。

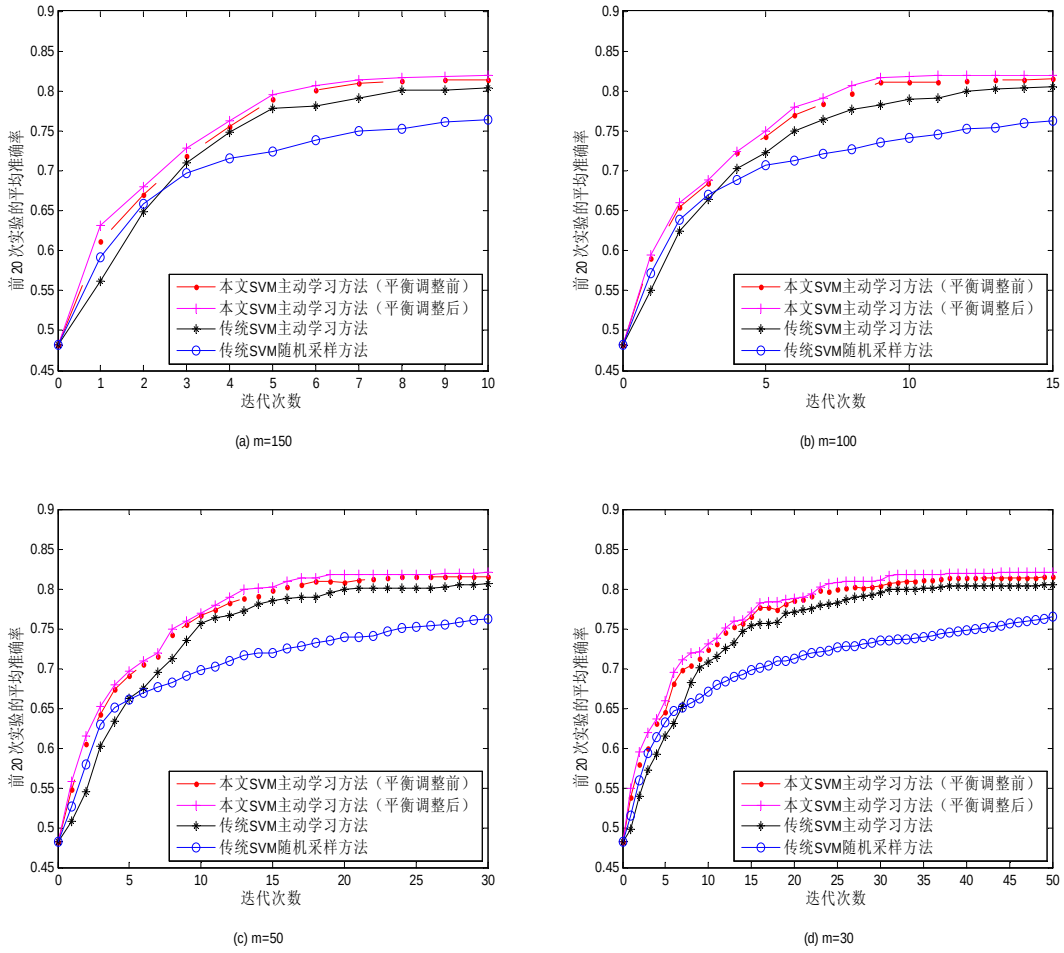

图 4.2 几种方法分类准确率收敛速度对比

（2）第二组实验 标注样本数

为了验证想达到同等准确率 SVM 主动学习可以大大减少需要标注的样本数，首先标注了全部样本并将其加入训练集训练，得到了分类准确率如表 4.2 所示。

表 4.2 全部样本参与训练结果

类别	错误样本数	准确率
舞曲	21/182	88.462%
抒情	52/182	71.429%
爵士	35/178	80.337%
民乐	34/124	72.581%
摇滚	8/135	94.074%
总准确率	150/801	81.273%

接下来为了说明本文提出方法的有效性，将迭代停止条件设置为分类准确率达到81%时迭代停止，同时统计对比两种主动学习方法已经标注的样本数（同样取20次实验的平均），每个样本只标注一次，即各个分类器多次迭代中如果选择了同一个样本则不计入新标注的样本数，最终统计结果如表4.3所示。

表4.3 迭代停止条件设为81%时

m设置	迭代停止时已迭代次数（改进/传统）	已标注样本数（改进/传统）
m=30	25/59	1099/1809
m=50	18/35	1137/1852
m=100	9/19	1232/1901
m=150	6/13	1255/1923

上表可以看出本文提出的SVM主动学习方法只标注了大概全部样本的一半就达到了和标注全部样本参与训练相同的准确率。 而传统的SVM主动学习方法标注的样本数虽然相对于全部样本数量有所减少但效果并不明显。另外因为在实际应用中，做到标注全部样本参与训练几乎是不可能的事情，所以退一步讲，如果牺牲些许的准确率是允许的话那么主动学习方法在减少标注样本数目方面的表现就会变得更为突出。比如在实验一中就已经可以看出，除了随机采样方法的准确率随着迭代次数增加一直趋于平缓的上升趋势，其他两种主动学习方法的准确率很快就收敛到了一个定值，也就是后面的迭代对分类准确率的提升已经非常有限。所以不妨牺牲一点准确率，将迭代停止条件设为准确率达到80%时。重新对比结果如表4.4所示。

表4.4 迭代停止条件设为80%时

m设置	迭代停止时已迭代次数（改进/传统）	已标注样本数（改进/传统）
m=30	22/31	761/1145
m=50	14/20	798/1177
m=100	8/12	823/1203
m=150	5/8	899/1242

实验结果表明在允许牺牲少许准确率的情况下，两种SVM主动学习方法都可以大大减少参与训练需要标注的样本数，本文提出的方法效果更优。同时观察到m值设置的越小，最终需要标注的样本数越少，这是因为迭代次数增加使得分类器有更多的机会去选择那些对分类器最有用的样本点，而m值设置越大需要的迭代次数自然也越少，所以可以根据实际情况

在两者之间进行权衡。

通过上述两个实验的对比，不难得出结论，在进行训练的时候，有目的性地选取价值样本进行标注迭代训练，可以使分类准确率迅速收敛，并且要达到同样的准确率需要标注的样本数目也少得多。

4.4 本章小结

本章首先介绍了实验使用的音乐数据库的建立以及基于 SVM 主动学习的音乐分类系统的基本结构。紧接着介绍了一种被广泛使用的 SVM 分类工具 SVM-light，并作为实验构造的工具。最后针对本文基于传统 SVM 主动学习方法提出了两点改进措施进行了两组实验进行算法有效性的验证。在主动学习中应用样本多样性和不确定性共同考虑的选择策略，同时人为对样本进行平衡性调整，无论在最终的分类准确率以及样本的标注数目上都有了较为明显的提升，基本达到了理论分析时所预测的效果。

第五章 总结与展望

全文总结

伴随着近年来互联网技术与多媒体信息技术的快速发展，网络负载的信息量也越来越庞大，音乐就是网络多媒体信息中最常见的一种。快速而准确的检索到自己喜欢的曲目也成为了人们生活中的一项基本需求，音乐自动分类技术的研究是音乐检索领域十分重要的研究方向，具有非常重要的实用意义。然而传统的音乐分类中，需要大量的人力和时间对训练样本进行标注，这样不可避免的会耗费大量的时间和人力。为了进一步减少人工标注的成本。提高分类效果和训练速度，主动学习应运而生。

SVM 是一种具备强大分类能力和优良泛化性能的分类器，因此被广泛应用于各个领域，然而如上面所示，SVM 算法优良的效果同样是需要依赖庞大规模的样本集，同样面临标注样本时人力耗费的巨大代价。那么自然而然的，将 SVM 这种优良的分类器和主动学习思想进行结合理论上可以取得非常好的效果。本文对基于 SVM 的主动学习进行了深入的研究并在实验中将其应用于音乐分类当中，并取得了一定的成果，总结如下：

（1）详细介绍了音乐分类与主动学习的相关理论和基本原理，针对主动学习中核心的两个部分：学习引擎和选择引擎，进行了详细的介绍，尤其详细介绍了选择引擎中常用的几种样本选择策略：基于不确定缩减的选择策略、基于版本空间缩减的选择策略和基于误差缩减的选择策略。

（2）分别详细介绍了 SVM 和主动学习的基本原理，并对现有的 SVM 主动学习策略做了简单的分析总结。针对其中一种基于不确定度缩减的主动学习方法本文提出了一种改进的方法。传统的方法认为距离 SVM 分类超平面最近的样本点具备最不确定性，包含的信息量最大，属于最有价值样本。本文提出的改进的 SVM 主动学习算法，把不确定性和多样性都作为样本是否是价值样本的判断标准，因为单参考不确定性有可能选到孤立点。同时针对“一对其余”方法中人为造成的数据集偏斜问题进行了平衡性调整。

（3）MATLAB 环境下通过实验仿真验证了本文提出的方法在分类准确率的收敛速度方面以及达到同等准确率需要标注的样本数目方面的优势。这充分证实了 SVM 主动学习方法确实可以大大减少人工标注样本所耗费的代价，在音乐分类中具有重要的意义。

未来展望

本文仅仅针对传统 SVM 主动学习的其中一种方法进行了初步的探索与研究，提出了进一步的改进，虽然在准确率和标记样本数等方面的验证都取得了一定的成果。但也有相应的局限性，还有很多方面值得深入研究。

（1）主动学习在二分类问题中的应用价值要远强于多分类，尤其体现在标记样本数的数量缩减上面。拿本文选用的“一对其余”方法来说，针对五种音乐流派进行了分类，那么对应了五个分类超平面，因为本文的选择策略是基于各个超平面的，所以在进行样本选择时针对各个分类超平面进行样本选择是独立进行的，主动学习迭代之前各自的未标注样本池是一样的，但是随着迭代进行，各自的样本池已经发生了变化，超平面之间，迭代之间会产生重复的样本，在实验中重复样本是没有做标注数的。但是显然随着迭代的进行需要标注的样本数还是会不断的增多，基本不会出现不满足条件的样本。是否可以制定一种信息熵，选出的样本并不是针对各个分类面独立的，每个分类面都对应着同一批样本，这个可以作为以后主动学习研究的一个可能的方向。

（2）一般主动学习算法都是人工设置参数来终止算法，是否可以改善算法的停止准则，使其可以自行在达到或近似达到最优分类精度时停止迭代。

（3）未标注样本的潜在价值仍然是未来工作的研究重点。

本文将一种改进的 SVM 主动学习方法应用于音乐分类系统中，把一种具备成熟理论的算法与具备广阔应用前景的音乐分类进行结合，有着十分重要的实用意义。

参考文献

[1] Logan B.Content-based playlist generation: Exploratory experiment. in:Proceedings of 3rd International Conference on Music Information Retrieval.Paris,France.2002

[2] Pauws S and Pats B.Eggen.Realization and user evaluation of an automatic playlist generator. Journal of New Music Research,June2003 32(2):179-92

[3] Chai W and Vercoe.Using user models in music information retrieval system. In:Proceedings of the 1st International Conference on Music Information retrieval.Plymouth,Massachusetts.USA.2000

[4] Logan B. Music recommendation from song sets. In:Proceedings of 5th International Conference on Music Information Retrieval.Spana.2004: 425-428.

[5] Simon H A, Lea G.Problem solving and rule education: a unified view knowledge and organization[J]. Knowledge and Cognition, 1974,15(2):63-73.

[6] Da-chuanWei. An improved feature extraction algorithm of humming music[C]// IEEE Transportation on 2011 International Conference on Mechanical, and Electrical Engineering (TMEE), 2011:2500-2503.

[7] Foucard R, Essid S, Richard G, Lagrange M. Exploring new features for music classification[C]// IEEE Transportation on 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS),2013:1-4.

[8] Liu H, Motoda H, Yu L. Feature selection with selective sampling.In:Sammut C,Hoffmann A G, Eds. Proceedings of the 19th International Conference on Machine Learning. San Francisco:Morgan Kaufmann,2002:395- 402.

[9] Das S. Filters, wrappers and a boosting-based hybrid for feature selection In: Brodley C E, Danyluk A P, EDs. Proceedings of the 18th International Conference on Machine Learning. San Francisco:Morgan Kaufmann,2001:74-81.

[10] Shan M K, Kuo F Fand Chen M F. Music Style Mining and Classification by Melody, in Proc. of IEEE ICMEO2, Lausanne, Switzerland, 2002.

[11] Chai W and Vercoe B. Folk Music Classification Using Hidden Markov Models, In Proc. oflC-AlOl, 2001.

[12] Dannenberg R B, Thom B and Waston D. A Machine Learning Approach to Musical Style Recognition, In Proc. of ICMC97, 1997.

[13] Soltau H, Schultz T, Westphal M and Waibel A. Recognition of Music Types, In Proc. of IEEE ICASSPYX, 1998:1137-1140.

[14] Pye D. Content-Based Methods for the management of Digital Music, In Proc. of IEEE ICASSPOO, , 2000:2437-2440.

[15] Cover T M. The best two independent measurement are not the two best.IEEE Transactions on System, Man and Cybernetics,1974,4(1):116-117

[16] Landley P. Selection of relevant features in machine learning. Proceedings of the AAAI Fall Symposium on Relevance. Menlo Park, CA.: AAAI Press, 1994: 140-144

[17] Jain A K, Zongker D. Feature selection: evaluation, application, and small sample performence . IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997,19(2):153-158.

[18] Xing E, Jordan M, Karp R. Feature selection for high-dimensional genomic microarry data. In: Brodley C E, Danyluk A P, Eds. Proceedings of the 18th International Conference on Machine Learning. San Francisco : Morgan Kaufmann, 2001: 601-608.

[19] Huang Yuan, Shian-Shyong Tseng, Wu Gangshan, etal. A Two Phase Feature Selection Method Using Both Filter and Wrapper. Proc of the IEEE International Conference on System, Man and .Cybernetics, 2:132-136

[20] Hall M A.Correlation-based feature selection for discrete and numeric class machine learning. In: Langley P, Eds. Proceedings of the 17th International Conference on Machine Learning. San Francisco:Morgan

Kaufmann,2000:359- 366.
[21] John G, Kohavi R, Pfeger K. Irrelevant features and the subset selection problem. In: Cohen W W, Hirsh H, Eds. The 11th International Conference on Machine Learning. San Francisc:Morgan Kaufmann,1994:121-129.
[22] Wold E, Blum T, Keislar D, and Wheaton J. Content-based classification, search and retrieval of audio, IEEE Mutimedia Mag,1996,3(3):27-36.
[23] Li S. Content-based classification and retrieval of audio using the nearest feature line method, IEEE Transportation Speech Audio Processing, 2000, 8:619-625
[24] Tzanetakis G and Cook P. Musical genre classification of audio signals, IEEE Transportation Speech Audio Process, 2002, 10(4):293-302.
[25] Bhat A S, Amith V S, Prasad N S, Mohan D M. An Efficient Classification Algorithm for Music Mood Detection in Western and Hindi Music Using Audio Feature Extraction[C]// IEEE 2014 Fifth International Conference on Signal and Image Processing (ICSIP),2014:359-364.
[26] Shen J, Shepherd J, Cui B, etal. A novel framework for efficient automated singer identification in large music databases[J]. ACM Transactions on Information Systems (TOIS), 2009, 27(3): 18.
[27] Little D, Pardo B. Learning Musical Instruments from Mixtures of Audio with Weak Labels[C]//ISMIR. 2008, 8: 127-132.
[28] Panagakis Y, Kotropoulos C, Arce G R. Music Genre Classification Using Locality Preserving Non-Negative Tensor Factorization and Sparse Representations[C]//ISMIR. 2009: 249-254.
[29] Gonzalez-Abril L, Angulo C, Velasco F, Ortega J A. A Note on the Bias in SVMs for Multiclassification[J]. IEEE Transactions on Neural Networks,2008,19(4):723-725.
[30] Zhu X, Lafferty J, Ghahramani Z. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions[C]. In: Proc of ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data. Menlo Park, CA:AAAI Press,2003:58-65.
[31] Donmez P, Carbonell J G. Proactive learning: cost-sensitive active learning with multiple imperfect or oracles[C]. In: Proc of ACM CIKM 2008, California, 2008:629-638.
[32] Sugiyamy M. Active learning in approximate linear regression based on conditional expectation of generalization error[J]. Journal of Machine Learning Research,2006,7:141-166.
[33] Balasubramanian K, Donmez P, Lebanon G. Unsupervised supervised learning II: training margin based classifiers without labels [J]. Journal of Machine Learning Research,2011,12:3119-3145.
[34] Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification[J] . Journal of Chemical Information and Computer Sciences,2004,44:1936-1941.
[35] 白龙飞，王文剑，郭虎生．一种新的支持向量机主动学习策略[J]. 南京大学学报（自然科学版）,2012,48(2):182-189.
[36] Scheffer T, Wrobel S. Active Learning of partially hidden markov models[C]. In:Proc of the ECML/PKDD, 2001, Berlin:Springer. 2001:1-15.
[37] David D, Lewis Willam, Gale A. A sequential algorithm for training text classifiers(Uncertainty Sampling)[C]. Proceeding of Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retreval, Springer-Verlag, London, 1994:3-12.
[38] Seung H S, Opper M, Sompolinsky H. Query by committee[C]. Proceedings of the 15th Annual ACM Workshop on Computational Learning Theory,California, 1992:287-294.
[39] Dagan I, Engelson S. Committee-based sampling for training probabilistic classifiers[C]. Proceeding of the 12th Int. Conf.on Machine Learning, 1995:150-157.
[40] Nguyen H T. Active learning using pre-clustering[C]. The 21th Int.Conf.on Machine Learning, Banff, Alberta, Canada, 2004:04-08.
[41] 孙功星，戴贵亮．神经网络主动学习的进化算法[J]. 计算机科学, 2002, 29(10): 61-63.
[42] Tong S. Active Learning: theory and application[R]. Stanford University, 2001: 1-168.

[43] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2: 45-66.
[44] Mukerjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time series using a support vector machine[C]. Principle J, Giles I, Morgan N. Processings of the 1997 IEEE Workshop on Neural Networks for Signal Processing, [S. I.]:IEEE Press, 1997:1125-1132.
[45] Schohn G, Gohn D. Less is more: active learning with support vector machines[C]. Proceeding of the 17th Int.Con. On Machine Learning, San Francisco:Morgan Kaufmann, 2000:45-66.
[46] Vlachos A. Active learning with support vector machines[D]. Master Science, School of Informatics, University of Edinburgh. 2004.
[47] 周艳丽. 基于主动学习 SVM 的智能车辆障碍物检测[D].南京理工大学, 2008, 6.
[48] 解洪胜, 张虹. 基于支持向量机的图像检索主动学习方法[J]. 山东师范大学学报, 2007, 22(4): 46-48.
[49] Wu H Y, Huang C H. Multi-path QOS routing in TDMA/CDMA ad hoc wireless networks[J]. Lecture Notes in Computer Science, 2004, 3252:609-616.
[50] 张健沛, 徐华. 支持向量机(SVM)主动学习方法研究与应用[J]. 计算机应用, 2004, 24(1): 1-3.
[51] 韩光, 赵春霞, 胡雪蕾. 一种新的 SVM 主动学习算法及其在障碍物检测中的应用[J]. 计算机研究与发 展, 2009, 46(11):15-20.
[52] 张玉芳, 陈卓, 熊忠阳, 刘君, 王银辉. 一种基于 SVM 和主动学习的图像检索方法[J]. 计算机工程与应用, 2010, 24:193-196.
[53] Ranganathan K, Lamnichi A., Foster I. Improving data availability through dynamic model-driven replication in large peer-to-peer communities[C]. CCGrid, 2002, 5:376-381.
[54]甄斌，吴玺宏，刘志敏，迟惠生.语音识别和说话人识别中各倒谱分量的相对重要性[J].北京大学学报（自然科学版）.2001,3,371-378.
[55] Rabiner L R and Juang B H, “Fundamentals of Speech Recognition”, Prentice-Hall Press, 1993.
[56] Bhat A S, Amith V S, Prasad N S, Mohan D M. An Efficient Classification Algorithm for Music Mood Detection in Western and Hindi Music Using Audio Feature Extraction[C]// IEEE 2014 Fifth International Conference on Signal and Image Processing (ICSIP),2014, 359-364.
[57] Gonzalez-Abril L, Angulo C, Velasco F, Ortega J A. A Note on the Bias in SVMs for Multiclassification[J]. IEEE Transactions on Neural Networks,2008,19(4),723-725.
[58] 乔玉龙, 潘正祥, 孙圣和.一种改进的 K-近邻分类算法.电子学报，2005,33(6):1146-1149.
[59] Scaringella N, Zoia G. On the modeling of time information for automatic genre recognition systems in audio signals[C]//Proceedings of the ISMIR 2005 6th International Conference on Music Information Retrieval, 12-15 September 2005. IEEE, 2005 (LTS-CONF-2005-057).
[60] Turnbull D, Elkan C. Fast recognition of musical genres using RBF networks[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(4): 580-584.
[61] Hamel P, Wood S, Eck D. Automatic Identification of Instrument Classes in Polyphonic and Poly-Instrument Audio[C]//ISMIR. 2009: 399-404.
[62] Lewis D, Catlett J. Heterogeneous uncertainty sampling of supervised learning[C] In:Proc of ICML 1994. San Francisco, CA:Morgan Kaufmann,1994:148-156.
[63] Cortes C, Vapnik V. Support vector networks Machine Learning,1995,20:273-295.
[64] 邓乃扬,田英杰. 数据挖掘中的新方法-支持向量机[M].北京：科学出版社，2004.
[65] Nvapnik V. Estimation of dependence based on empirical data.Springer-Verlag,Berlin,1982.

附录 1 攻读硕士学位期间撰写的论文

（1）邵曦、姚磊，《基于 SVM 主动学习的音乐分类》，计算机工程与应用，2014.12；

致谢

行文至此，才猛然醒悟到自己的研究生生涯即将结束，两年半时间，说长不长，说短不短，虽有些许遗憾，但也收获满满。两年半间，我收获的不仅仅是知识，更有学习的能力还有一群值得结识一生的好友，这一切都值得我用一生来回忆和珍惜。在这感伤的毕业季，我要衷心的感谢这几年在各方面都帮助过我的老师和朋友们，此生结识你们，万分荣幸！

首先，我最值得骄傲的是，读研期间我能跟随我十分敬仰和爱戴的邵曦教授。他不仅是我学业上的导师，也是我生活上，思想上的导师。他教会我的决不仅仅是知识，还有很多需要领悟一生的道理。邵老师的研究方向是多媒体信息检索，有着自己别具一格的思想见解和驾轻就熟的实践能力。在我写这篇毕业论文的时候给予了我很多受用的意见和帮助，邵老师是一个十分仔细认真的人，对于我论文中的许多小错误都要求我严谨的改过，秉承着自己在学术上一直以来的严谨性，在传授我专业知识的同时，也教会我做人的原则。生活中，邵老师更是像一个年长些的老朋友，从未给我们以距离感，天冷了问我们多加衣，找工作为我们多指导。千言万语汇成一句感谢！

其次我还要感谢实验室的同门：郁青玲、汪慧敏、陶凯云……，在我遇到困难时，你们毫不犹豫的向我伸出了援助之手，感谢你们！我会将你们永远铭记在心！

再次还要感谢我的父母和两个姐姐，他们虽然无法在学术上给我提供帮助，但是他们一直是我最坚实的后盾，让我可以放下一切一心一意的写论文，谢谢你们一直以来在我背后毫无条件的默默的支持我！

最后的最后，更要感谢专家导师们，感谢您们抽出宝贵时间审阅我的论文并提出无比珍贵的意见和建议，感谢您们！