## 目标

《基于多模态融合的音乐情感分类方法研究》

关键词：音乐情感分类；多模态；图学习；Hough森林；隐式空间

《神经网络在音乐分类中的应用研究》

关键词：音乐分类；神经网络；特征提取；隐含马尔科夫模型

## 收获

《基于多模态融合的音乐情感分类方法研究》

利用多模态来分析音乐的情感就如同利用人的多重感官来同时分析音乐，比单模态会拥有更高的准确率。

本文主要研究综合利用歌词和音频模态将会有效提高音乐情感分类的准确率，通过分析歌词与音乐模态与人类感知间的关系，挖掘两个模态间的内在关联性并进行相互补充与提高分类的准确率。

在音乐情感分类的领域，一些低阶特征包括MFCC、OSC以及一些反映音乐特性的特征包括音乐韵律、音乐节奏等对音乐情感分类有着较好的效果。

现根据russell提出的VA情感空间将音乐情感分为四个类别：happy，angry，sad和relaxed。其中Valence表示情感的极性（正面或者是负面情绪），Arousal表示情感的强度。



音乐情感分类主要部分：

1. 为了提取更具情感表现力的文本特征，我们提取具有情感区分度的词语集合对歌词进行过滤，同时由于歌词文本的稀疏性，我们是用来英文语言库对如上词汇表进行扩充以保留分布系数的情感词，此外还提出了基于半监督的RAE魔性的句子级别歌词特征提取方法，这种特征能够较好的反应出句子中每个词语所蕴含的情感信息。
2. 我们提取出了基于多模态投票的Hough森林情感分类方法对时间对齐的音频与歌词特征在Hough投票空间中进行融合。该方法通过将不同模态音乐数据的情感类别属性在时间上的相关性作为对应情感参数的投票值预测音乐的所属的情感类别，提高了音乐情感分类的准确率。
3. 由于音乐数据的两个模态特征在结构上的差异，使得研究者无法充分的探究其特征间的关联性。因此为了消除这种异构性，我们提出了针对音乐情感分类的扩展的多模态LPP算法为这两个模态的特征学习共同的隐藏空间，该隐式空间不仅能够使得数据保持其在原始空间中所具备分类特性，同时拉近了相同情感类别下的不同模态特征间的距离，使得我们可以自然地使用一个模态中蕴含的相关情感信息补充其他模态的情感信息，1为充分挖掘模态间在结构上的相关性创造了条件。利用上述隐藏空间，我们提出了基于多模态k近邻与图学习的情感分类方法，利用跨模态特征间的相似度传播情感类别标签。

多模态在各领域的发展

Rasiwasia提出了基于相关性与基于语义的方法学习文本与图像模态的共同表示，并利用这种表示提高了跨模态检索任务(利用文本搜索图像或者利用图像得到与其相关的文本)的准确率。该文章假设异构的特征空间中具有某种相关性，并使用CCA (Canonical correlation analysis)学习一个共同的子空间使得两个模态的相关性达到最大。此外，该文章还学习文本与图像数据和固定个数的语义概念间的映射，并利用这种映射将跨模态数据投影到共同的子空间中，从而消除跨模态数据间的语义鸿沟。Lm除音频与视频数据的异构性，解决了音频与视频数据相互检索的问题。他提出多媒体文档表示法将音频与视频特征量化为音频与视觉单词，而后使用潜在概率语义模型((PLSA)为两个模态学习共同表示，并利用基于相似度的传播模型修正不同模态特征间的相似性以提高跨模态检索的准确率。Ngiam[38]使用深度学习模型RBM(Restricted bo ltzmann Machines)与DBN(Deep Belief Network)为音频与视频模态学习共同的特征表示，并使用该融合后的表示进行语音识别任务，并在多个数据集上取德了不错的效果。Sharma为了学习在分类任务中更具区分度多种模态的共同表示，提出一种具有普适性的有监督框架。该框架通过将传统无监督的降维方法如PCA、LDA等进行扩展，将不同模态的数据投影到一个同构的子空间中，并拉近了相同类别数据间的距离。

多模态音乐特征表示

与传统的文本相比，歌词文本的词语分布更加稀疏，歌词句子通常较短，并且具有一定的重复性。这种歌词的高重复性以及词语的稀疏性导致了可用的情感语义信息更加稀少也因此增加了特征提取的难度。此外，由于音频特征与文本特征在结构上的异构性，使得两种特征所表达的情感之间存在难以逾越的鸿沟，这使得挖掘两种表示在情感表达上的相关性进行多模态分类成为了一个严峻的问题。

为了从音乐数据中挖掘出更加丰富的情感信息，我们提出了使用句子级别的特征表示方法以刻画音乐的情感。我们假设规范化歌词文件LRC中标识的起始时间戳为每句歌词对应的音频信息的起始位置，并一下一句歌词的起始时间戳为该句歌词的终止位置，从而保证歌词与音频信息的对齐。根据以上的假设我们将每首歌曲表示为一组保存了句子级别时间信息的特征向量集。

歌词文本特征表示

我们首先将简要介绍一种文件级别的歌词特征提取方法词袋模型((BOW)作为基准方法，并使用TF-IDF方法对词袋中的每个词进行重要程度的量化。而后，本节将重点介绍我们提出的基于句子级别的歌词特征提取方法。首先我们将歌词根据歌词文件的时间戳分成多个句子，并基于半监督的RAE[45]模型来学习句子级别的歌词特征，在得到句子级别特征表示

的同时保留了句子所包含的词语中所具备的情感信息。此外，为了获取更具分类能力的句子级别特征，提高特征在分类模型中的区分能力。我们提出了一种歌词文本预处理方法，该方法使用LM D模型提取更具情感类别区分度的词汇表，过滤仅含有少量情感语义信息的单词。进一步地，在过滤方法的基础上由于歌词的稀疏性问题，我们使用了基于大型语料库Wordnet的同义词扩展方法以保留含有较强情感语义信息但是过于稀疏的词语。

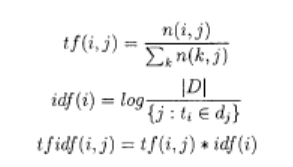
基于TF-IDF的词袋模型特征

TF-IDF是一种用于评估字词对于文档重要程度的统计方法。在一份给定

的文件中，使用TF (Term Frequency)描述单词在文档中的出现次数，使用IDF

(Inverse Document Frequency)用以描述单词出现的普遍性，该值与单词出现的

文档数成反比。该模型可用如下公式描述:



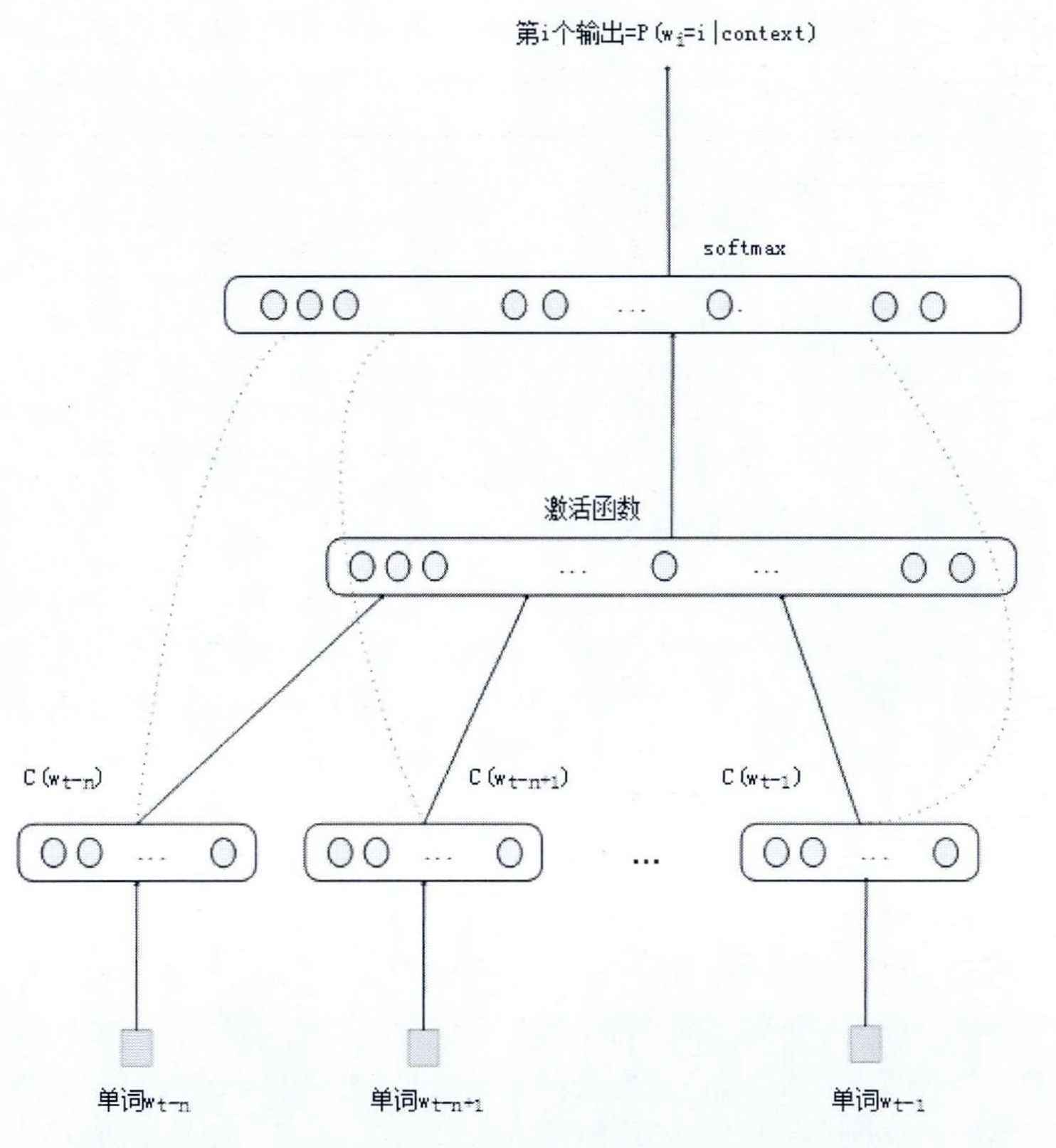
其中式1中n(i, j)表示单词wi在文档 dj中出现的次数，而其分母表示该文档的总词数，式2中|D|表示所有文档的个数，而分母则表示包含词语wi的文档个数。将式1与式3相乘我们得到最终单词的tfidf值。可以看到，该模型评判单词重要性的原理在于一个单词在文档中出现的次数越多，并且在其他文档中出现较少，则认定带单词具有较好的区分度并给予更高的分值。但这种基于词袋模型的特征表示法只关注歌词统计信息，致力于获取文档的局部内容而忽略了文档的结构化信息以及在在更细粒度上词语间的关联性，此无法更加准确地描述歌词的语义信息。例如’I love you very much’和’i hate you very much’两个句子，如果采用基于词袋模型的方法，由于两句话只存在一个词不相同因此在模型中所表达的含义是十分相近的，而事实上这两个句子表达了两种完全不同情感。同时由于歌词中单词分布的稀疏性，每个文档往往只拥有所有词汇表中的部分单词，而使用词袋模型将得到一个相当高维的特征，这使得该特征中大部分的值将被置为0，这将大大降低分类的效率以及有效性。

单词的分布式表示

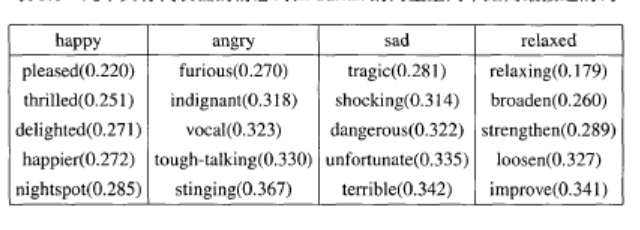
在传统的自然语言处理方法中，语言模型是为了计算文档中每个句子所可能出现的概率而创建的。这里我们假定S表示一个有意义的句子，而w表示句子中的每个词语，m表示词典的大小，因此每个句子可以由一串词语序列w1, w2, w3,…wt:来表示，因此每个句子出现的概率:P(S)=P(w1,w2, w3, ...wn)=P(w1)\*P(w2|w1)\*P(w3l w2, w1) ...P(wt|w1,w2,…wt-1),其中P(wi)表示第Z个单词出现的概率我们需要计算第i个词相对于前i一1个词的条件概率，复杂度将达到mi，为了简化计算的复杂度，n-gram模型假定每个词出现的概率只和它之前的n个词有关，与之类似，本文中介绍的基于神经网络的语言模型也是利用这种n-gram模型进行构建。

假定每个单词具有d维的特征表示，初始的词向量矩阵C为m\*d维矩阵，

每一行代表一个词向量。如图3.4所示，这是一个由三层神经网络构筑的语言模型，其中输入层由n\*d个节点组成，这些节点由wt-n,wt-n=1…wt-1这n个单词的d维向量表示拼接而成，中间层相当于普通神经网络的引才能够曾，而输出层则由m个节点组成，其中第i个输出表示第i个词出现的概率，这个神经网络的训练目标是得到一个较好的模型使得全局的P（wt|wt-n…wt-1）最优化。而使得该模型最优化的矩阵c将作为最终的词向量表示。基于上述思想，我们可以得到一个更加接近人类理解和感知方式的单词分布式表示模型，由于在本文中我们的歌词数据集数据相对较少，无法满足神经网络的学习要求，因此我们使用了Turian用一个半监督的神经网络模型训练了海量数据数周而得到了一个50维的词向量数据集，根据该数据及我们将离散的歌词数据转化为连续的向量表示作为我们后续特征提取模型的输入



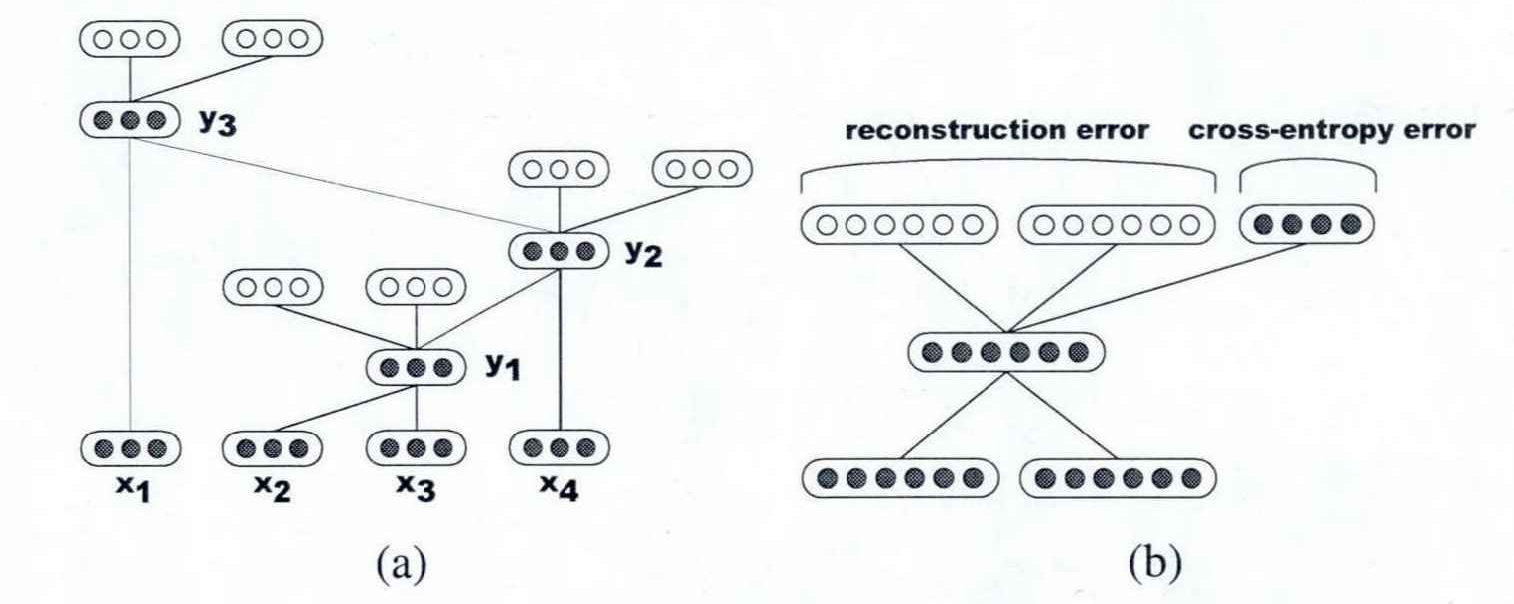
下表是部分单词在该向量空间的几个最近邻单词及其距离。



半监督的递归式Autoencoder

Autoencoder是一个非监督的多层深度学习算法，该算法的目的在于使得学习过后的输出层结果与输入层尽可能的相似，而其训练过后的隐藏层表示则可以用作输入层的简化表示。基于这种思想，每个句子S都可以表示为一棵递归树的形式。每一层递归都是一个三层结构的Autoencoder，将两个输入向量转化为输出向量，并得到一个能反应输入层信息的隐藏层表示，并递归地将这些隐藏层表示同初始化的输入一起进行学习，得到最终唯一的向量来表示S.这种RAE学习模型的优点在于，不仅能得到整个句子唯一的固定维度的特征表示，同时树中的每个节点都可以视作一段由多个词语组成的短语的向量表示，从而保证了在递归过程中不会丢失句子原本的语义。

一个简单的RAE模型如图所示：



图中S = (x1, x2, x3, x4)是一个由四个单词组成的简单句子，首先以x2, x3作为输入得到隐藏层的表示y1，继而用y1作为新的输入与x4一起得到y2，最终将y2,x1作为输入得到学习结果y3。通过这种方式，我们只需要合理地构建树形结构，就可以将一个含有多个词向量的句子不断地递归学习得到最终唯一的更具表现力的句子级别特征表示，

具体来说，我们假定每层Autoencoder的输入为v1, v2，输出为v1’，v2’，则每层Autoencode的重构误差可以表示为输入向量与输出向量的欧几里得距离。

对于每个句子来说，存在着多种树形递归结构，我们的目标是找到一个结构使得整个树上的重构误差最小，我们定义集合T（s）表示一个句子所有可能的结构，集合NL表示树中所有非叶节点，这个目标可以形式化的表示为：RAE（s）=

在RAE的基础上，为了使得最终学到的特征向量对于特定的情感更具区分度和表现力，算法在树中每个非叶节点上添加了一个简单的softmax层用以调整和预测结果在每个类上的分布

音频特征表示

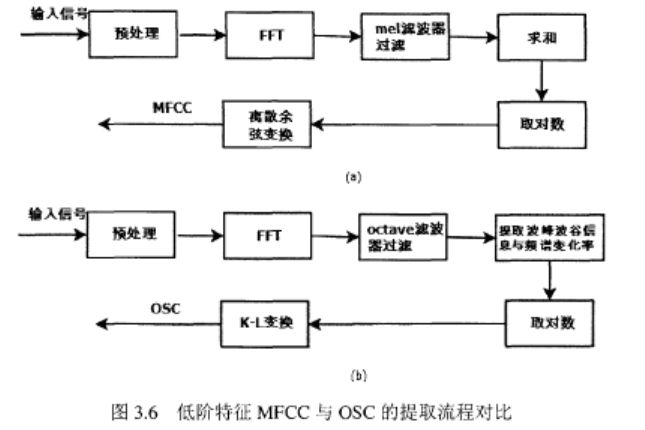
三种在音乐检索中十分热门的特征：1）MFCC 2）OSC 3）Chromagram feature

OSC：和MFCC相似，也是一种我与频谱分析得到的短时音色特征。OSC使用八分级制用来计算倒谱系数，这是由于歌曲中的声线通常与乐器的声线吻合，而他们的音调和谐波结构通常落在八度范围内。因此与MFCC相比，对于纯音乐来说，基于八度音阶的倒谱系数提取更具有表现力。OSC和MFCC的提取流程大致相同。不同点在于以下几个方面：

1. 使用的滤波器不同。OSC使用了octave范围的滤波器进行滤波，而MFCC则使用mel范围的滤波器，这种8分级制的频谱范围更加适合音乐的处理过程，并将在一定程度上反应音乐的结构化信息。
2. 对于每个子频带的幅度能量谱，OSC提取了波峰(peak)与波谷(valley)信息以及他们的变化率作为每个滤波器的最终输出，而不像MFCC采用求和的方式，从而提供了一些MFCC所不具备的频谱信息。
3. OSC使用K-L变换得到最终的倒谱系数而MFCC则使用的是DC T变换。

与MFCC相比，OSC反映的频率范围更贴近音乐的表示，同时其反映了子

频带上的频谱变化，弥补了MFCC所丢失的某些音乐上的结构特性。



3) Chromagram feature(Pitch Class Profile PCP)

Chorma特征也是目前广泛使用的用于表示短时音频窗口的结构化内容(包括音调(key)、和弦(chord))而提取的特征。该特征基于现在音乐理论的十二音阶体系，计算给定频带范围的频谱能量并映射到12个音级类别((C, C#, D, D#, E,E#, F, F#, G,G #,A, A#, B, B#)，每个类中包含了所有八度范围的能量。要理解音级与八度的概念，我们以钢琴为例，它可以弹出88个音高(pitch)，我们可以发现这些音高都是以7个白键((do, re, mi... ti)及其中间交错的5个黑键为一组循环出现的，例如某组中re与下一组的re为一个八度的关系，如果忽略这种组间的关系，那么这十二个音将构成上述12个音级。本文中使用sonic-annotator中的Queen Mary plugin set提取chorma一个典型的Chorma提取过程主要分为如下几个步骤

1.将时域信号转换到频域信号中。由于在时域信号中通常无法直接提取与音频相关的旋律信息，而人的听觉系统对于频度震级((frequency magnitude)比相位信息(phase information)更加敏感，因此提取chromagram特征的第一步往往使用短时傅里叶变换(Short Time Fourier Transform STFT)将时域信号转换到频度谱中。

2.然后我们需要对信号进行降噪处理。显而易见在频谱图中并非所有信号都对音乐旋律的提取有所帮助，我们将这些无用的频谱信号通过滤波器移除，并使用调音算法将音乐音调间的频率间隔调至标准间隔A=440Hz。

3.将信号根据所选窗口的长度转换为帧，根据每个音高(pitch)的显著度计算其在每一帧中的能量值，并将其记录下来得到一个音高直方图。

4.在3中音高直方图的基础上，将同一时间同一音级中不同八度的音符能量值叠加到相应音级所对应的特征上，得到最终的chroma特征向量。与MFCC, OSC等低阶特征相比，Chorma作为基于内容的更高阶的语义特征，描述了音乐音阶的分布，反映了更多音乐结构上的特点，对音乐情感的表述能起到一定的作用。

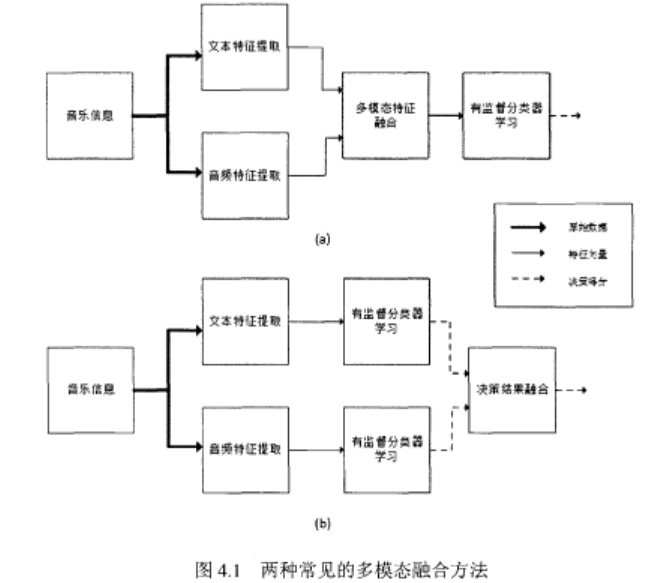
由于MFCC和OSC特征与Chroma在不同的级别提取音乐特征，因此他们的窗口大小并不相同，为了将这三种特征融合在一起，我们将窗口大小相同的MFCC与OSC特征先拼接在一起，然后将Chromagram在其窗口范围内每个小窗口上复制并与上述特征拼接。此外，为了与歌词特征同步，我们使用了两种不同级别的表示方式:

1.为了与句子级别的歌词特征同步，我们根据每个歌词句子的起止时间将音频信号分割成多个部分，对于每个部分，我们有音频特征集{v1, v2, v3, ...vn}（vi属于Rm)，计算该集合中所有特征向量的均值和方差，将此均值向量与方差向量拼接起来得到最终的句子级别音频特征。

2.为了与文件级别的歌词特征同步，对于每首歌曲m我们直接计算该m中所有的特征向量的均值与方差，并将其拼接得到最终唯一的音频特征表示。

**多模态音乐情感分类方法**

**---基于多模态的Hough森林的音乐情感分类方法**



利用分层投票框架将句子级别的音频与歌词特征在Hough投票空间中进行融合，并计算他们的音乐情感类别属性在时间上的相关性用作相应投票的权重，原则上，相关性越高，相应时间片段的音乐特征在情感类别上更加明确，因此具有更高的投票权重。通过聚合音乐不同时间片段的多模态情感类别投票，我们最终获得音乐的整体情感类别情感。

**针对句子层次音乐表示的Hough森林分类模型**

Hough森林有一组二叉分类决策树组成，每颗决策树接收输入向量并将根据某些分裂准则进行树的构造，并在输出其在每个可能类上的投票结果，最终将每棵树的结果整合在一起得到整个森林最终的输出结果。在每个决策树的非叶节点上将根据输入向量的某个特征进行计算并与阈值都欧式训练过程中得到的。Hough森林在随机森林的基础上，更进一步的使用的广义Hough变换对每个由样本所产生的相关参数在Hough空间上进行投票，这种特性使得Hough森林对于很多任务更具有多样性以及普适性。

第四章多模态音乐情感分类方法

息、)在Hough空间上进行投票，这种特性使得Hough森林对于很多任务更具多

样性以及普适性。

在我们的工作中，我们使用Hough森林模型学习从每首歌曲句子级别的特

征到他们可能所属的音乐情感类别的概率的映射。根据来自不同情感类别的训

练歌曲句子级别的特征集合，我们的训练过程如下:

1.根据所需参数(例如:树的最大深度、节点的最小样本个数)初始化含有L棵决策树的集合T1, T2,...,TL

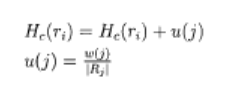
2.将歌曲级别的情感标签ri((ri[1 ...K]),K表示情感类别的个数)、歌曲的编号n，以及句子编号t，赋予每个句子级别的特征vi.

3.将所有的{<vi,ri,ni,ti>}作为训练集训练特征随机森林集合{Ti, T2,T3...TL}，一旦训练完成，每个决策树的叶子节点上将包含若干训练歌曲的句子级别特征以及他们在对应的情感类别上的概率投票值。

**基于特征融合的多模态Hough投票机制**

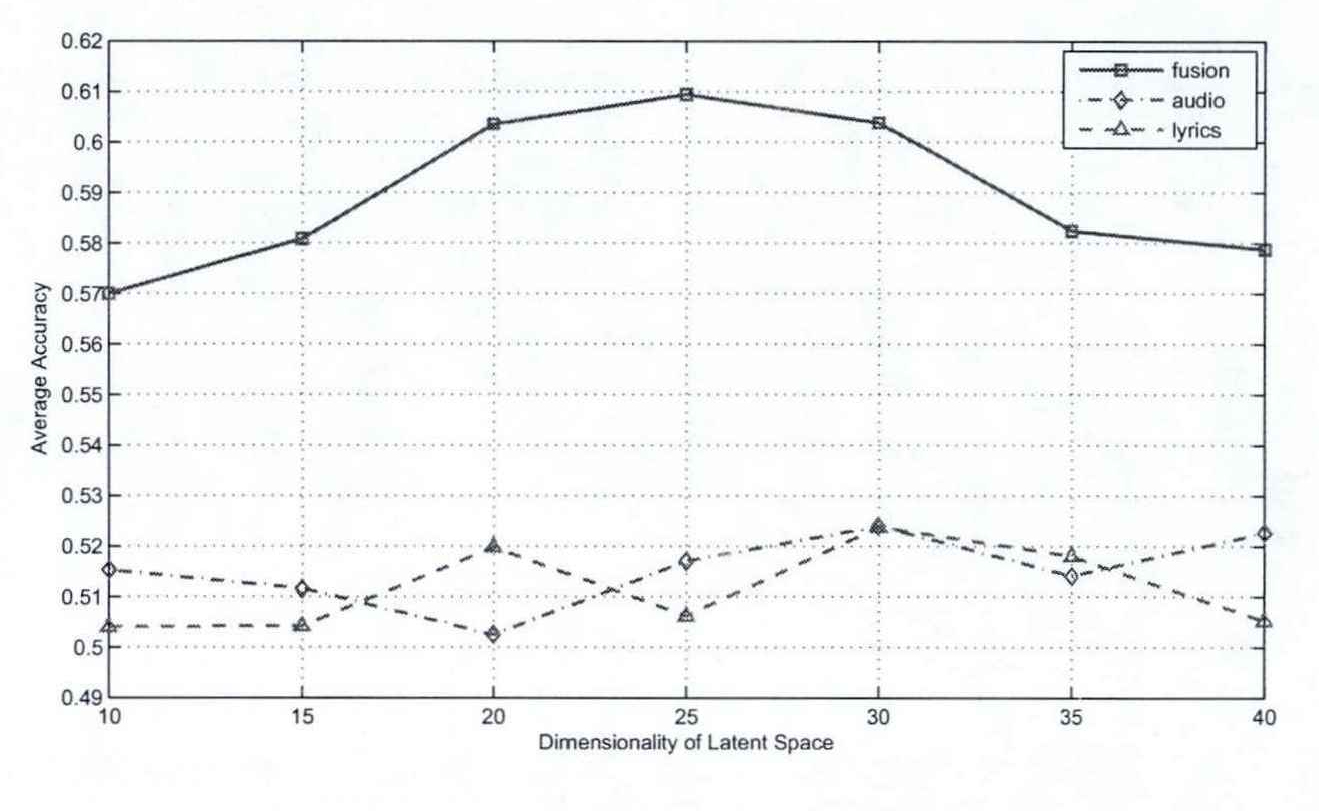
在这种投票机制中，同一个句子的两个模态的特征首先被拼接成一个复合特征。我们将该复合特征作为训练过程的输入向量。然后给定一个未知标签的音乐s，我们将歌曲中每一段句子所对应的复合特征sj传入森林中，并将sj，所到达的叶子节点中包含的训练样本集Rj及其在对应情感类别的贡献度收集起来。

然后我们将所有音乐片段sj，所对应的训练样本的在对应情感分类上的投票概率值通过广义Hough变换聚集起来。具体来说，我们初始化一个一维Hough矩阵He(k)(k = 1..K)，矩阵中的每个元素代表对应的第k个情感类别的累计投票值。对于每个音乐片段sj，我们假设凡中的每个训练样本{<vi,ri,ni,ti>}对于sj属于情感类别ri均具有一定的贡献值，我们根据该贡献值将该Hough矩阵对应的位置以如下方式进行累加:



在这里，w(j)表示Rj中的训练样本所对应的权重，用于表示他对于投票的重要程度，该值是在Hough森林训练过程中得到的，|Rj|表示测试句子sj对应的训练样本集的大小，在所有片段进行如上处理之后，He(k)(k=1,…,K)中拥有最大投票值的类即被标记为该测试歌曲的类标签。

实验结果：



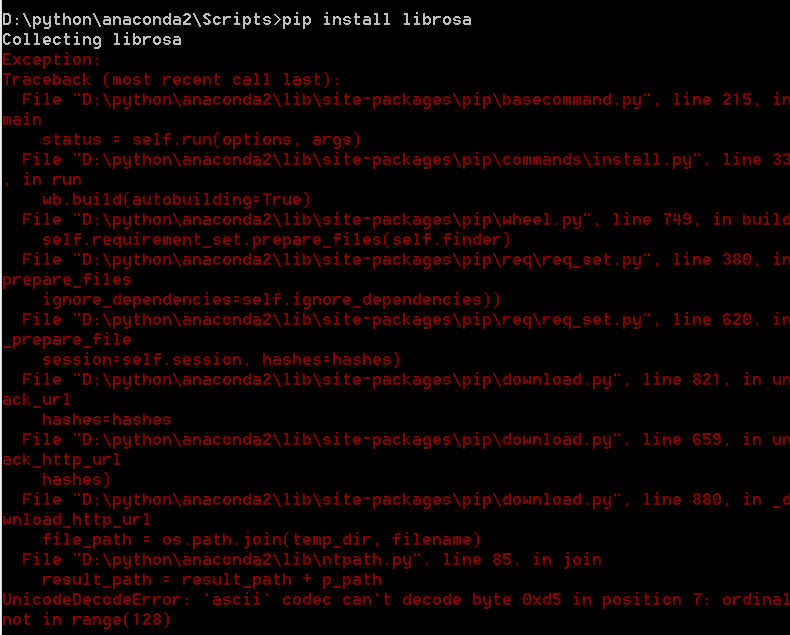
《神经网络在音乐分类中的应用研究》

本文提出一种基于BP神经网络的音乐分类方法，首先采用倒谱系数特征提取法对音乐特征进行提取，以感知特征和Mel倒谱系数组成特征向量；在音乐分类方面，以BP神经网络模型作为分类器，对同一音频抽取若干样本，对样本识别结果采用投票法判定该音频的音乐类别。仿真实验结果表明该方法拥有更好的抗干扰能力和正确率。

BP神经网络分类算法

BP 神经网络是一种多层前馈神经网络，该网络是主要特点是信号传递前向传递，误差反向传播。在前向传递中，输入信号从输入层经隐含层逐层处理，直至输出层。第一层的神经元状态只影响下一层神经元状态。如果输出层得不到期望输出，则转入反向传播，根据预测误差调整网络权值和阈值，从而使 BP 神经网络预测输出不断逼近期望输出。

一直尝试在anaconda2上安装librosa，但总是无法成功



后来上网查到解决方法：找到anaconda文件夹下的lib中的mimetypes.py中加入

import sys

if sys.getdefaultencoding() != 'gbk':

reload(sys)

sys.setdefaultencoding('gbk')

保存后再次输入pip install librosa就安装成功了