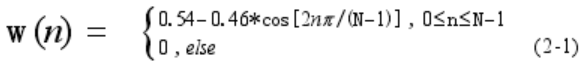
本周我继续阅读了基于深层神经网络的音频特征提取及场景识别研究

进一步对基于MFCC 和KNN的场景识别基线系统进一步了解

该论文使用的基线系统是基于MFCC和 KNN构建的，大致流程如下：首先是音频信号的预处理工作，然后对预处理得到的中间结果进行特征分析，最后在得到的特征向量上进行音频场景分类。音频信号预处理工作主要是对由音频设备采集到的原始模拟信号进行处理，常规操作有：预滤波、模数转换、预加重、分帧、加窗等；特征提取的目的是提取出相对来说最能体现出当前音频信号的内容数据，我们采用的特征是MFCC,所以涉 及到 的工作有： 快速傅里叶变换 (Fast Fourier Transformation,FFT)、三角滤波、离散余弦变换(Discrete Cosine Transformation, DCT)、倒谱均值减、差分计算等；最后一步是对特征数据进行分类，我们选择的 KNN 分类器无须进行训练，可以直接对特征向量进行分类操作。

分帧、加窗

声卡设备采集到的信号是模拟信号，经过数字化处理会得到对应的数字信号，但是随着时间变化它的幅度值也在变化且变化较剧烈。现在成熟的信号处理技术往往分析的都是时不变信号即平稳信号，所以假设音频信号在短时间内具有平稳性，来适应传统的分析方法。为了得到上面假设的平稳信号，首先得对原始音频信号进行短时切分操作，即用窗函数平滑地在时间轴上滑动将整个信号分成以帧为单位的短时信号；不同的窗函数形式或者不同的窗长都会影响最终的结果。实验中采用 hamming 窗，计算公式如(2-1)所示，帧叠我们设置为窗长的一半。N 为窗长。



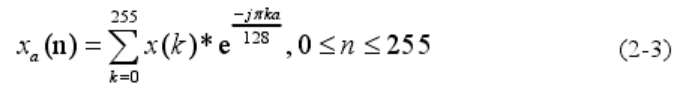
特征提取

模式识别任务中，非常重要的一个环节就是提取训练样本数据的特征。特征数据的好坏会直接决定后面分类器分类结果的准确程度。MFCC具有其它声学特征不具备的性质，即它根据人耳的听觉机理对原始频率（Hz）进行了非线性映射，是在目前识别任务中最常用的音频特征计算公式如(2-2)所示。

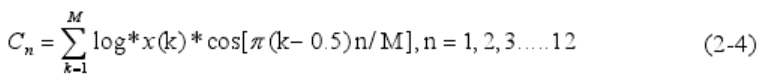


上式中，f 为原始频率。MFCC系数计算过程如下：首先经过上面介绍的预处理操作（分帧、加窗处理）得到中间数据。然后对上面得到的中间数据进行 FFT, FFT 是把音频信号从时域向频域进行转换，于是得到频谱数据，计算公式如(2-3)所示。上式子中

x(k) 为中间数据，a为帧序号。



对上面得到的频谱数据进行平方操作我们就得到能量谱，然后用M 个梅尔带通滤波器对能量谱进行滤波操作，最后将所有经过滤波得到的子带能量进行叠加求和得到功率谱。将上面得到的功率谱取对数，然后进行离散余弦变换,最后就得到 L 个我们要求的 MFCC系数。一般L在12到16这一区间进行取值，实验中我们取L值为12，计算公式如(2-4)所示。

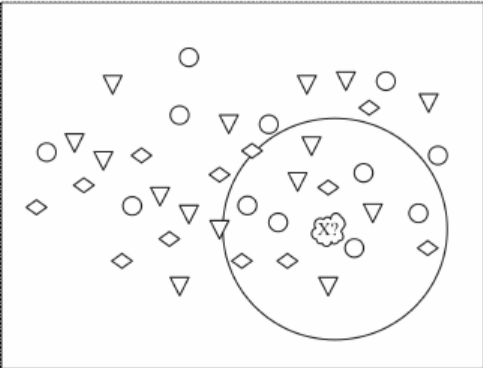


上面得到的结果只是静态的音频特征，如果在静态音频特征基础上级联静态特征的差分参数对识别性能的提升会有很大的帮助，因为差分参数中含有动

态音频特性。所以，实验中我们加入了一阶差分参数，最终得到的是 24 维的MFCC特征向量。

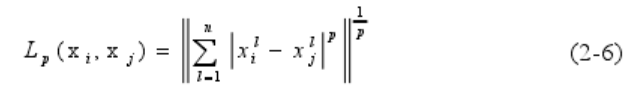
KNN 模型及基本要素

如果给定**训练样本集合**、**样本之间的距离度量方式**、**分类决策规则**还有 **k值**，那么对于任意一个给定的测试样本实例，它的类别也就随之确定。决策过程如图 2-3 所示，在训练样本xi对应的S(xi ,k )中，我们根据投票原则找出包含样本数量最多的类别为圆圈，所以 xi的类别为圆圈。



**距离度量方式**

训练样本特征向量之间的距离反映了样本之间的相似程度，不同的距离度量方式最终所确定的近邻点是不同的。常用的距离度量方式有欧式距离，曼哈顿距离等，我们可以用统一的闵可夫斯基距离又称为Lp范数(norm)来定义，如公式(2-6)所示。式中，当 p =2时，为欧氏距离，当 p =1 时，为曼哈顿距离, 我们选择欧式距离作为样本特征向量之间的距离度量方式。



**k 值的选择**

实际应用中，k 一般选择较小的数值且选择的 k 值一般不大于样本数量的平方根。

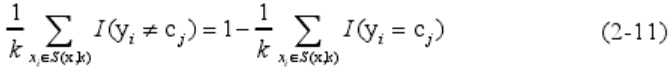
**分类决策规则**

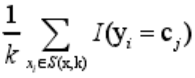
k 近邻算法中的分类决策规则一般是投票原则。即由 S( xi, k)中包含样本点多的类别来作为测试样本的类别。如果假设分类的损失函数为 0-1 损失函数的话，分类函数如式(2-9)所示，那么投票表决规则就等同于经验风险最小化。原因如下：



首先误分类的概率如式(2-10)所示。



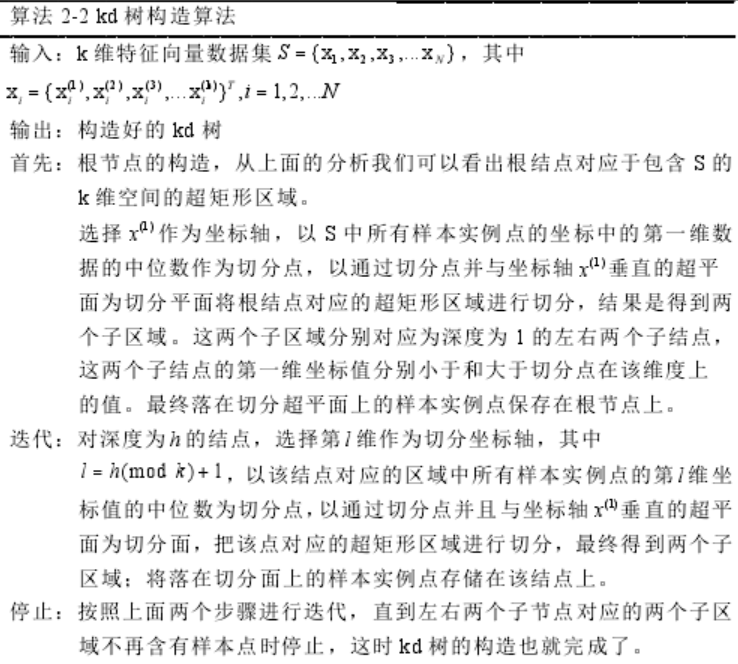
对于给定的测试样本实例 X，如果涵盖 S(x,k)的区域的类别是cj ，则误分类率计算公式如(2-11)所示。

如果我们要使误分类数量最小，即经验风险最小，就要让最大，所以投票表决规则等价于经验风险最小化。

k 近邻的 kd-tree 实现算法

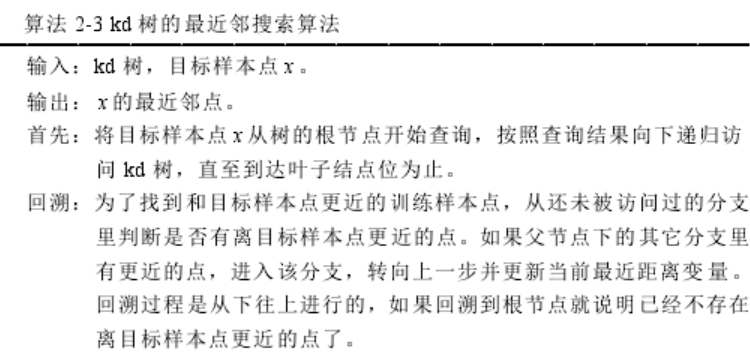
我们在实现 k 近邻算法的过程中，最重要的环节就是找出k 个近邻样本点，这就需要计算样本向量之间的距离。kd 树是一种二叉树，它可以被看做是对 k 维空间的一个线性划分，这样我们就可以用它来进行存储k 维样本数据点kd 树的构造过程是递归进行的，我们可以简单地描述其为递归的对 k 维空间进行切分，进而生成子结点，如果子结点内不再含有样本点就停止切分。递归过程中，我们会将相应的样本点保存在生成的子结点中

Kd树的构造算法如下图所示



优势：K 近邻搜索是在上面构造完成的 kd 树上进行的，其相对于遍历搜索来说，可以省去大部分样本特征向量之间的距离计算，大大减少了计算量。

给定一个待测试样本点，要搜索其最近邻点。首先我们必须找到包含该实例点的叶子结点，然后以此叶子结点为出发点逐步向父结点回溯，在回溯过程中我们要持续查找与目标样本点距离最近的结点，并做到随时更新当前的最近邻点；如果确定不可能存在更近的结点时停止搜索。如算法 2-3 为 kd 树的最近邻搜索算法。



上面的算法中判断未被访问过的分支中是否有离目标样本点更近的点时

涉及到以下两个操作：

（1）如果该子区域中含有的实例点比当前最近距离小的话，那么我们就把它更新为最近邻点。

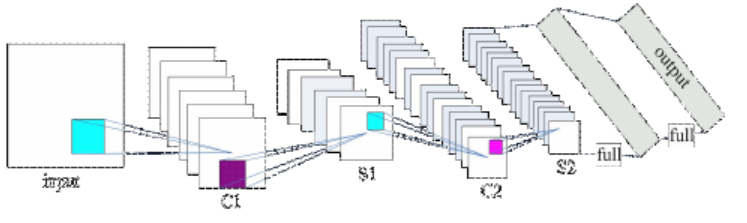
（2）当前最近邻点肯定内含于某个结点的其中一个子结点对应的矩形区域中，我们就检查其父结点的其它分支结点对应的矩形区域是否含有更近的点，如果没有就继续向上查找。

这篇论文还介绍了**基于 CNN卷积神经网络的音频特征提取及场景识别**

由于MFCC场景识别存在一定缺陷：1不能够完整的对音频场景的信息进行刻画。2，海滩场景信号一般主要由浪花拍打沙滩的声音和海鸥的叫声等关键声学事件组成，其中两种声学事件发生的绝对时间和相对时间都不确定，如果用 MFCC 长时统计值的话,相同的场景得到的统计值结果差别可能会很大，造成误识；

如图 3-1所示为一个卷积神经网络的拓扑结构示意图，除了输入输出层外，每层由多个二维平面组成，每个平面由多个相互独立的神经单元构成。从下面的拓扑结构示意图可以看出，

CNN 从左由输入层开始向右依次连接卷积层和下采样层，其中 C 表示卷积层 S 表示下采样层。卷积层和下采样层可以设置多个且它们之间是间隔设置的，最后按照全连接方式连接几个普通网络层。CNN 的输入数据为二维矩阵，在本实验系统中，输入数据就是语谱图。



在这一部分介绍了给定一个卷积神经网络模型，如何进行特征图的求解计算问题，即卷积神经网络的前向传播过程。前向传播过程中唯一的一个常量是输入矩阵，其它像卷积滤波器对应的矩阵系数、偏置系数以及最后的全连接层中的矩阵系数都是随机初始化的较小值，它们都是在后面的学习过程中待学习的参数变量。

在CNN方法使用该过程中，卷积滤波器的设计，特征提取及分类，参数调整同要重要

最后论文又介绍了基于解卷积神经网络的音频特征提取及场景识别，下周会尝试第一种方法的代码实现部分