基于凝聚信息瓶颈的音频事件聚类方法

李艳雄,王 琴,张 雪,邹 领
(华南理工大学电子与信息学院,广东广州 510640)

摘 要: 为了进一步提高音频事件聚类算法性能,本文基于凝聚信息瓶颈理论提出一种音频事件聚类方法.首先,论述信息瓶颈原理及其推导过程;然后,详细论述一种基于凝聚信息瓶颈的音频事件聚类方法,包括源变量、相关变量和目标变量的定义,聚类的具体步骤,算法主要计算量分析等.采用取自两个数据库的音频事件样本进行测试,实验结果表明:与目前文献报道的方法相比,本文方法在多种实验条件下都获得了更高的 K 值(平均类纯度和平均音频纯度的几何平均值),而且运算速度更快.

关键词: 凝聚信息瓶颈; 音频事件聚类; 音频内容分析
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2017)05-1064-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.05.006

Audio Events Clustering Based on Agglomerative Information Bottleneck

LI Yan-xiong, WANG Qin, ZHANG Xue, ZOU Ling
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong* 510640, *China*)

Abstract: In order to further improve the performance of methods for audio events clustering, this paper proposes a method for audio events clustering based on the theory of agglomerative information bottleneck. First, the principles and derivations of information bottleneck are briefly introduced. Then, the proposed method is described in detail, including the definitions of source variables, relevance variables and destination variables, the steps of the proposed method and the analyses of main computational loads of all methods. The proposed method and two kinds of previous methods (including the method based on spectral clustering, and the method based on both Bayesian information criterion and agglomerative hierarchical clustering) are evaluated on the experimental data extracted from two different corpora of audio events. The experimental results show that the proposed method obtains higher K values (geometric mean of average clustering purity and average audio purity) and runs faster than the previous methods under several experimental conditions.

Key words: agglomerative information bottleneck; audio events clustering; audio content analysis

1 引言

随着多媒体技术的发展,记录有各种音频事件的音频文档(影视剧音轨、智能手机录制的音频等),正迅猛增长.如何有效检测、辨识音频文档中的各类音频事件,受到越来越多的关注[1].目前主要采用两种处理方法:(1)有监督识别;(2)无监督聚类.前者首先从各个音频事件中提取特征参数,再通过训练好的分类器,例如隐马尔科夫模型(Hidden Markov Model, HMM)、高斯混合模型(Gaussian Mixture Model, GMM)、支持向量机(Support Vector Machine, SVM)、深度神经网络(Deep Neural Network, DNN)等,将各音频样本辨识为预先定义的类别.后者首先提取特征参数,但无需事先训练分类器,而是采用某种聚类算法将相同类别的音频段合并在一起,并分配一个标签给各类别.

目前监督式识别音频事件的研究报道比较多.这些方法所采用的特征基本相同,例如梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)、感知线性预测(Perceptual Linear Prediction)、过零率(Zero Crossing Rate)、基频(Pitch)等,或者上述特征的组合.它们的差异主要是采用不同分类器.采用 DNN 作为分类器的有 Ian McLoughlin[2]、Oguzhan[3]等人提出的方

收稿日期:2015-10-28;修回日期:2015-11-20;责任编辑:马兰英
基金项目:国家自然科学基金(No.61101160);中央高校基本科研业务费专项资金重点项目(No.2015ZZ102);广州市珠江科技新星专项(No.2013J2200070)

法.采用随机森林回归框架(Random Forest Regression Framework)作为分类器的有 Huy Phan 等人[4]提出的方法.采用 SVM 作为分类器的有 Kucukbay[5]、Lu[6]、Huy[7]、Li[8]、Jose[9]、Peng[10]等人提出的方法.采用 GMM 作为分类器的有 Zhang 等人[11]提出的方法.采用 HMM 作为分类器的有 Jose[9]、Maria[12]、Cai[13]、贺前华[14]等人提出的方法.而罗森林等人[15]融合 GMM 和 SVM 识别特定音频事件.

监督式音频事件分类方法的研究报道很多,而无监督音频事件聚类的研究工作却非常少.Lu 等人[16]提出一种无监督的音频事件聚类方法.采用谱聚类方法对 5 种音频事件进行聚类,将相同类别音频段合并在一起从而得到音频类别个数及各个音频类别所包含音频段.谱聚类方法需要计算两两类别之间的亲和矩阵(Affinity Matrix)并对矩阵进行谱分解,计算量比较大.除了谱聚类之外,目前常见音频聚类算法还有以贝叶斯信息准则(Bayesian Information Criterion,BIC)为收敛准则的凝聚分层聚类(Agglomerative Hierarchical Clustering,AHC)算法,即基于 AHC + BIC 的聚类算法[17].该方法目前被用于说话人聚类,而没被用于音频事件聚类,本文也将其作为一个音频事件聚类基准方法.该方法需要根据不同数据集调整 BIC 惩罚系数,而且计算两个样本之间的 BIC 距离需要较大计算量.

复杂音频文档所包含的音频事件类型及类别数一般都是未知的.如何从中找出未知类别的音频事件,是目前音频文档内容分析的难点.监督式音频事件分类方法需要事先知道音频事件的类型及类别数,并为它们训练一个模型(HMM,SVM,GMM,DNN 等);而无监督的音频事件聚类方法不需要上述先验知识,且无需事先训练分类器.因此,在复杂音频内容分析中,无监督音频事件处理方法更具普适性.为了进一步提高无监督音频事件聚类方法性能,本文提出一种基于凝聚信息瓶颈(Agglomerative Information Bottleneck,AIB)的音频事件聚类方法.通过查阅国内外文献可知,目前还没有信息瓶颈理论在音频事件聚类方面的研究报道.因此,本文主要贡献是:提出一种无监督音频事件处理方法——基于 AIB 的音频事件聚类方法,讨论源变量、相关变量和目标变量的定义、聚类流程、算法主要计算量等,并与文献报道的无监督方法进行比较.

2　信息瓶颈理论简介

信息瓶颈方法是在率失真(Rate Distortion)理论[18,19]基础上发展而来的概率分布聚类方法,采用联合概率分布表示数据,以互信息(Mutual Information)作为度量手段,刻画样本和样本属性的相关性,不需要对样本之间的距离函数做任何假设[20,21].它可以描述为:给定源变量 X 和相关变量 Y 的联合概率分布 $P(X,Y)$,把源变量 X(待聚类样本)所包含信息压缩到目标变量 C(聚类结果)时,最大化保留目标变量 C 与相关变量 Y 之间的互信息 $I(Y,C)$ 即尽可能保留相关结构,同时尽量压缩源变量 X 与目标变量 C 之间的互信息 $I(C,X)$ 即尽可能压缩数据.在压缩数据和保留相关结构的过程中,目标变量 C 就相当于源变量 X 和相关变量 Y 之间的瓶颈,如图 1 所示.

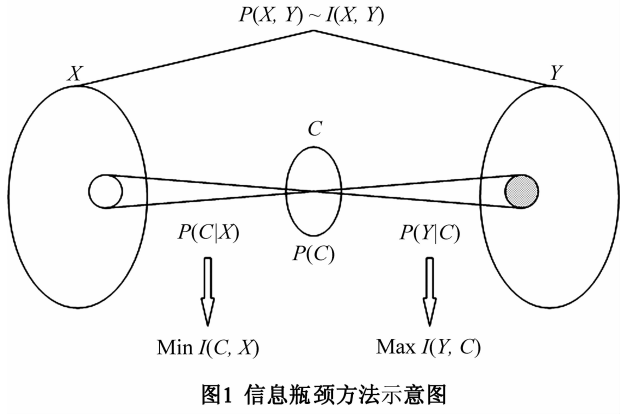

图1 信息瓶颈方法示意图

信息瓶颈方法试图寻找关于相关变量 Y 的信息最大压缩与最大保留之间的折衷,相当于最大化下述目标函数:

$$F = I(Y,C) - \frac{1}{\beta} I(C,X) \tag{1}$$

其中 β 是拉格朗日乘子,用来平衡互信息 $I(Y,C)$ 和互信息 $I(C,X)$ 使得目标函数 F 达到最大.$I(Y,C)$ 和 $I(C,X)$ 分别定义为:

$$I(Y,C) = \sum_{y\in Y,\ c\in C} p(c)p(y\mid c)\log\frac{p(y\mid c)}{p(y)} \tag{2}$$

$$I(C,X) = \sum_{x\in X,\ c\in C} p(x)p(c\mid x)\log\frac{p(c\mid x)}{p(c)} \tag{3}$$

随机变量 $X\sim p(x)$ 的熵 $H(X)$ 定义为:

$$H(X) = H[p(x)] = -\sum_{x\in X} p(x)\log p(x) \tag{4}$$

二维随机变量 $(X,C)\sim p(x,c)$ 的二维联合熵 $H(X,C)$ 定义为:

$$\begin{aligned} H(X,C) &= H[p(x,c)] \\ &= -\sum_{x\in X}\sum_{c\in C} p(x,c)\log p(x,c) \end{aligned} \tag{5}$$

在给定 X 时,关于 X 的条件熵 $H(C\mid X)$ 定义为:

$$\begin{aligned} H(C\mid X) &= -\sum_{x\in X}\sum_{c\in C} p(x,c)\log p(c\mid x) \\ &= -\sum_{x\in X} p(x)\sum_{c\in C} p(c\mid x)\log p(c\mid x) \end{aligned} \tag{6}$$

互信息 $I(X,C)$ 与熵的关系:

$$I(X,C) = H(C) - H(C\mid X) = H(X) - H(X\mid C) \tag{7}$$

目标函数 F 的解空间(详见文献[19]):

$$\begin{cases} p(c\mid x) = \dfrac{p(c)}{Z(x,\beta)}\exp(-\beta D_{KL}[p(y\mid x)\parallel p(y\mid c)]) \\ p(y\mid c) = \sum\limits_{x\in X} p(y\mid x)p(c\mid x)\dfrac{p(x)}{p(c)} \\ p(c) = \sum\limits_{x\in X} p(c\mid x)p(x) \end{cases} \tag{8}$$

其中,$Z(x,\beta)$是一个概率归一化函数,$D_{KL}[p(y|x)\parallel p(y|c)]$表示 $p(y|x)$与 $p(y|c)$之间的 Kullback-Liebler 散度. $Z(x,\beta)$和 $D_{KL}[p(y|x)\parallel p(y|c)]$分别表示为:

$$Z(x,\beta) = \sum_{c\in C} p(c)\exp(-\beta D_{KL}[p(y\mid x)\parallel p(y\mid c)]) \tag{9}$$

$$D_{KL}[p(y\mid x)\parallel p(y\mid c)] = \sum_{y\in Y} p(y\mid x)\log\frac{p(y\mid x)}{p(y\mid c)} \tag{10}$$

从式(8)可知,当β趋于无穷大时,即$\beta\to\infty$,随机映射 $p(c|x)$变成对源变量 X 的硬判决:$p(c|x)$只取 0 和 1 这两个值. 从上述解空间的定义可以看出,目标函数 F 具有一个形式解,可以借助具体的信息瓶颈算法(例如 AIB 等)得到具体解.

凝聚信息瓶颈算法是一种迭代合并的方法:将每个样本分配给一个类,即初始类别数等于样本个数;再每次合并使目标函数变化量 $\Delta F(c_{r_1},c_{r_2})$最小的两个类 c_{r_1}和 c_{r_2}. $\Delta F(c_{r_1},c_{r_2})$定义为:

$$\Delta F(c_{r_1},c_{r_2}) = (p(c_{r_1})+p(c_{r_2}))d_{r_1r_2} \tag{11}$$

其中,$d_{r_1r_2}$是两个 Jensen-Shannon 散度之组合:

$$d_{r_1r_2} = JS[p(y|c_{r_1}),p(y|c_{r_2})] - \frac{1}{\beta}JS[p(c_{r_1}|x),p(c_{r_2}|x)] \tag{12}$$

其中,JS 表示两个概率分布之间的 Jensen-Shannon 散度,定义为:

$$JS[p(y|c_{r_1}),p(y|c_{r_2})] = \pi_{r_1}D_{KL}[p(y|c_{r_1})\parallel q_Y(y)] + \pi_{r_2}D_{KL}[p(y|c_{r_2})\parallel q_Y(y)] \tag{13}$$

$$JS[p(c_{r_1}|x),p(c_{r_2}|x)] = \pi_{r_1}D_{KL}[p(c_{r_1}|x)\parallel q_X(x)] + \pi_{r_2}D_{KL}[p(c_{r_2}|x)\parallel q_X(x)] \tag{14}$$

其中,
$$\begin{cases} q_Y(y) = \pi_{r_1}p(y|c_{r_1}) + \pi_{r_2}p(y|c_{r_2}) \\ q_X(x) = \pi_{r_1}p(c_{r_1}|x) + \pi_{r_2}p(c_{r_2}|x) \end{cases} \tag{15}$$

$$\begin{cases} \pi_{r_1} = \dfrac{p(c_{r_1})}{p(c_{r_1})+p(c_{r_2})} \\ \pi_{r_2} = \dfrac{p(c_{r_2})}{p(c_{r_1})+p(c_{r_2})} \end{cases} \tag{16}$$

目标函数 F 随着类别数的减少而单调递减[19],不断合并使 $\Delta F(c_{r_1},c_{r_2})$最小的两个类,直到预定类别个数为止. 合并 c_{r_1}和 c_{r_2}得到的新类 c_r 表示为:

$$\begin{cases} p(c_r) = p(c_{r_1}) + p(c_{r_2}) \\ p(y|c_r) = \dfrac{p(y|c_{r_1})p(c_{r_1}) + p(y|c_{r_2})p(c_{r_2})}{p(c_r)} \\ p(c_r|x) = 1, \quad \forall x\in c_{r_1},c_{r_2} \end{cases} \tag{17}$$

3 音频事件聚类方法

为了将信息瓶颈方法应用于音频事件聚类,本节首先定义源变量 $X=\{x_j\}$、相关变量 $Y=\{y_j\}$、目标变量 $C=\{c_r\}$和条件概率 $p(y_j|x_j)$;然后给出聚类具体步骤,并对算法主要计算量进行分析. 其中,$1\leqslant j\leqslant J$,J 表示聚类样本总个数;$1\leqslant r\leqslant N_c$,$N_c$ 表示类别数;$N_{\max}$表示聚类后的最大可能类数.

3.1 变量的定义

从每个样本中提取特征参数(例如 MFCC),并作为源变量 x_j. 由于高斯混合模型 GMM 被广泛用于表示各种概率分布,GMM 作为相关变量 y_j. GMM 个数 L 等于待聚类样本个数 J,即每个样本用一个 GMM 描述. 条件概率 $p(y_j|x_j)$表示第 j 个高斯混合模型 GMM_j 对第 j 个样本的相关性,定义为:

$$p(y_j\mid x_j) = \frac{\xi_j\left(\sum\limits_{m=1}^{M}\omega_m^j b_m^j(x_j)\right)}{\sum\limits_{l=1}^{L}\xi_l\left(\sum\limits_{m=1}^{M}\omega_m^l b_m^l(x_j)\right)} \tag{18}$$

其中,$1\leqslant j\leqslant L$,L 表示高斯混合模型个数;M 表示高斯混合模型的高斯元个数;ξ_j 表示权重系数,等于第 j 个样本总帧数与所有样本总帧数的比值;ω_m^j 表示高斯混合模型 GMM_j 中的第 m 个高斯元的混合权重系数;$b_m^j(x_j)$表示高斯混合模型 GMM_j 中的第 m 个高斯元对特征 x_j 的高斯概率. $b_m^j(x_j)$定义如下:

$$b_m^j(x_j) = \frac{1}{(2\pi)^{\frac{Q}{2}}|\boldsymbol{\Sigma}_m^j|^{\frac{1}{2}}}\exp\left[-\frac{(x_j-\boldsymbol{u}_m^j)(x_j-\boldsymbol{u}_m^j)^{\mathrm{T}}}{2\boldsymbol{\Sigma}_m^j}\right] \tag{19}$$

其中,T 表示矩阵的转置运算;Q 表示特征 x_j 的维数;u_m^j 表示 GMM_j 的第 m 个高斯元的均值矢量;$\boldsymbol{\Sigma}_m^j$ 表示 GMM_j 的第 m 个高斯元的协方差矩阵.

通过估计条件概率 $p(y_j|x_j)$,将源变量 x_j 映射到了相关变量 y_j 所表示的空间. 将源变量 x_j 对应的条件概率 $p(y_j|x_j)$作为聚类的输入,从而将相同类别的音频事件合并到同一个类,使源变量所包含的信息压缩到目标变量 c_r. 聚类后各个音频事件类别就是目标变量 c_r.

3.2 聚类方法步骤

本文聚类方法步骤如下:

(1)对各个样本进行分帧、加汉明窗、并提取特征参数 x_j,$1\leqslant j\leqslant J$.

(2)采用所有特征参数生成一个通用背景模型(U-

niversal Background Model,UBM). 再采用第 j 个样本的特征参数 x_j 对 UBM 进行自适应更新,得到刻画第 j 个样本的高斯混合模型 GMM_j. 根据文献[22]的最大后验概率算法(Algorithm of Maximum A Posteriori)更新模型参数. 根据式(18)计算条件概率 $p(y_j|x_j)$.

(3)将 J 个待聚类音频特征参数分成 J 类(初始类别数 N_c 等于 J),每类只包含一个样本特征参数,并计算对应条件概率:

$$\begin{cases} c_j = x_j \\ p(c_j) = p(x_j) \\ p(y_j|c_j) = p(y_j|x_j), \forall y_j \in Y \\ p(c_r|x_j) = 1, \text{if } r = j; \text{otherwise}, p(c_r|x_j) = 0 \end{cases} \tag{20}$$

(4)计算当前目标变量 C 与相关变量 Y 之间的互信息 $I(Y,C)$,源变量 X 与相关变量 Y 之间的互信息 $I(X,Y)$. 如果两者的比值(当前聚类结果互信息与原始互信息的比值)大于预设门限 T_I(由调参数据集确定):

$$\frac{I(Y,C)}{I(X,Y)} > T_I \tag{21}$$

则跳到第(5)步;否则,跳到第(6)步.

(5)根据式(11)计算任意两个类的目标函数变化量 $\Delta F(\cdot,\cdot)$,得到 $0.5N_c(N_c-1)$ 个值. 假设 c_{r_1} 和 c_{r_2} 这两个类的目标函数变化量 $\Delta F(c_{r_1},c_{r_2})$ 最小,则将 c_{r_1} 和 c_{r_2} 合并为一个新类:$c_r = \{c_{r_1}, c_{r_2}\}$,根据式(17)计算合并后的类的概率参数,$N_c = N_c - 1$(类数减一),跳到第(4)步继续迭代.

(6)如果 $N_c \leqslant N_{\max}$,即聚类后的类别数已小于等于设定的最大可能类数,则 N_c 作为最后类别数,跳到第(7)步;否则跳到第(5)步继续迭代.

(7)如果条件概率被聚类合并在同一个类别中,则与之对应的样本被判为同类音频事件.

3.3　主要计算量分析

完全量化给出各方法计算复杂度(加减乘除次数)是比较困难的,本文通过分析它们的主要计算量定性比较其计算复杂度. 从聚类流程看,本文方法与基于 AHC + BIC 方法[17]基本相同,属于层次聚类算法,主要区别在于两个样本(或类)之间的距离计算. 基于谱聚类方法不属于层次聚类算法,它首先构造三个矩阵(亲和矩阵 $\boldsymbol{A}$、对角矩阵 $\boldsymbol{D}$ 和归一化的亲和矩阵 $\boldsymbol{L}$),再对矩阵 L 进行谱分解,并采用 N_c 个最大特征值所对应的特征向量构造出一个 $J \times N_c$ 的矩阵 $\boldsymbol{V}$[16]. 矩阵 $\boldsymbol{V}$ 的每一行当作一个聚类对象,即对 J 个 N_c 维行向量进行 K-means 聚类[16]. 假设样本总数为 J;最终类别数为 N_c,聚类过程的中间类别数为 N'_c,$N_c \leqslant N'_c \leqslant J$;各方法采用相同特征参数(维数为 Q);第 i 和第 j 个样本(或类)的特征矩阵分别为 x_i 和 x_j,帧数分别为 N_i 和 N_j. 从第 3.2 节的聚类方法步骤可知,本文方法的主要计算量是目标函数变化量 $\Delta F(c_{r_1},c_{r_2})$ 的迭代计算:从初始化到最后收敛需要迭代 $(J-N_c)$ 次,每次迭代需要计算 $0.5N'_c(N'_c-1)$ 次 $\Delta F(c_{r_1},c_{r_2})$,其中 N'_c 初始值为 J,逐次减一,直到等于 N_c. 由公式(11)至(14)可知,$\Delta F(c_{r_1},c_{r_2})$ 的主要计算量是 4 个 KL 散度 $D_{KL}[\cdot \| \cdot]$ 的计算. 令 KL 散度 $D_{KL}[\cdot \| \cdot]$ 的计算量为 $f(D_{KL}[\cdot \| \cdot])$,则本文方法主要计算量为:$(J-N_c) \times 0.5N'_c(N'_c-1) \times 4f(D_{KL}[\cdot \| \cdot])$.

基于 AHC + BIC 方法主要计算量是两个样本(或类)之间 BIC 距离 ΔBIC 的计算:

$$\begin{aligned} \Delta\mathrm{BIC} = {} & N_i \times \ln(|\det(\mathrm{cov}(x_i))|) \\ & + N_j \times \ln(|\det(\mathrm{cov}(x_j))|) \\ & + \lambda \times \ln(N_{ij}) \times (Q + 0.5Q \times (Q+1)) \\ & - N_{ij} \times \ln(|\det(\mathrm{cov}(x_{ij}))|) \end{aligned} \tag{22}$$

其中,$\mathrm{cov}(\cdot)$ 表示特征矩阵的协方差矩阵;$\det(\cdot)$ 表示行列式值;$\ln(\cdot)$ 表示自然对数;x_{ij} 由 x_i 和 x_j 合并而成,$N_{ij} = N_i + N_j$ 是特征矩阵 x_{ij} 的帧数. 从公式(22)可知,ΔBIC 的主要计算量是计算 3 个 $\ln(|\det(\mathrm{cov}(\cdot))|)$. 令 $\ln(|\det(\mathrm{cov}(\cdot))|)$ 的计算量为 $f(\ln(|\det(\mathrm{cov}(\cdot))|))$,且聚类后的类别数与本文方法的相同,则基于 AHC + BIC 方法的主要计算量为:$(J-N_c) \times 0.5N'_c(N'_c-1) \times 3f(\ln(|\det(\mathrm{cov}(\cdot))|))$.

基于谱聚类方法主要计算量是:矩阵 $\boldsymbol{A}$ 各元素的计算、矩阵 $\boldsymbol{L}$ 的谱分解和特征向量的 K-means 聚类. 矩阵 $\boldsymbol{A}$ 是一个 $J \times J$ 对称矩阵,需要计算 $0.5J(J+1)$ 个不同元素 a_{ij},元素 a_{ij} 表示第 i 个与第 j 个样本之间的距离. 如果 a_{ij} 是特征距离,则对应的方法是基于特征距离的谱聚类方法;如果 a_{ij} 是模型距离,则对应的方法是基于模型距离的谱聚类方法[23]. 逐帧计算 x_i 和 x_j 之间的欧氏距离,得到距离矩阵 $\boldsymbol{E}$:

$$\boldsymbol{E} = \begin{bmatrix} e_{11}, \cdots\cdots, e_{1N_j} \\ \cdots, \quad e_{uv}, \cdots \\ e_{N_i1}, \cdots\cdots, e_{N_iN_j} \end{bmatrix}_{N_i \times N_j} \tag{23}$$

其中 e_{uv} 表示 x_i 的第 u 帧与 x_j 的第 v 帧之间的欧氏距离. 计算矩阵 $\boldsymbol{E}$ 各元素的平均值 e:

$$e = \frac{1}{N_i \times N_j} \sum_{u=1}^{N_i} \sum_{v=1}^{N_j} e_{uv} \tag{24}$$

第 i 个与第 j 个样本之间的特征距离 $d_1(x_i,x_j)$ 定义为:

$$d_1(x_i,x_j) = \exp\left(-\frac{e^2}{2\sigma^2}\right) \tag{25}$$

其中,σ 为常数. 令特征距离 $d_1(x_i,x_j)$ 的计算量为 $f(d_1(x_i,x_j))$. 采用所有特征参数生成一个 UBM. 再采用 x_i 和 x_j 分别对 UBM 进行自适应更新[22],分别得到第 i 和 j 个样本的高斯混合模型 GMM_i 和 GMM_j. 第 i 与第 j 个样本之间的模型距离(Bhattacharyya 距离)$d_2(x_i,x_j)$

定义为:

$$d_2(x_i,x_j)=\frac{1}{8}(u_i-u_j)^{\mathrm{T}}\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(u_i-u_j)+\frac{1}{2}\ln\left(\frac{\det\left(\frac{\Sigma_i+\Sigma_j}{2}\right)}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right) \tag{26}$$

其中,u_i(或 u_j)和 Σ_i(或 Σ_j)分别是 GMM_i(或 GMM_j)的均值矢量和协方差矩阵. 令模型距离 $d_2(x_i,x_j)$ 的计算量为 $f(d_2(x_i,x_j))$. 因此,由特征距离和模型距离构成的亲和矩阵 $\boldsymbol{A}$ 的主要计算量分别为:$0.5J(J+1)\times f(d_1(x_i,x_j))$ 和 $0.5J(J+1)\times f(d_2(x_i,x_j))$. 由于矩阵 $\boldsymbol{E}$ 的计算量非常大(N_i 和 N_j 越大,计算量越大),$f(d_1(x_i,x_j))$ 远大于 $f(d_2(x_i,x_j))$.

矩阵 $\boldsymbol{L}$ 定义为:

$$\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{27}$$

矩阵 $\boldsymbol{D}$ 是一个对角矩阵,分别将矩阵 $\boldsymbol{A}$ 中每一行元素相加,依次放在矩阵 $\boldsymbol{D}$ 的主对角线上. 令矩阵 $\boldsymbol{L}$ 谱分解计算量为 $f(Dec(\boldsymbol{L}))$. 矩阵 $\boldsymbol{V}$ 的 J 个 N_c 维行向量的 K-means 聚类的计算复杂度为 $O(J^{QN_c+1})$[24],令其计算量为 $f(Clu(\boldsymbol{V}))$. 基于特征距离和基于模型距离谱聚类方法的主要计算量分别为:$0.5J(J+1)\times f(d_1(x_i,x_j))+f(Dec(\boldsymbol{L}))+f(Clu(\boldsymbol{V}))$ 和 $0.5J(J+1)\times f(d_2(x_i,x_j))+f(Dec(\boldsymbol{L}))+f(Clu(\boldsymbol{V}))$.

综上所述,各方法主要计算量如表 1 所示.

从表 1 可知,本文方法主要计算量小于 AHC + BIC 方法的主要计算量,因为两个概率分布的 KL 散度的计算量 $f(D_{KL}[\cdot\parallel\cdot])$ 小于特征参数协方差矩阵的行列式值的自然对数的计算量 $f(\ln(|\det(\mathrm{cov}(\cdot))|))$;模型距离谱聚类方法的计算量小于特征距离谱聚类方法的计算量,因为 $f(d_2(x_i,x_j))$ 小于 $f(d_1(x_i,x_j))$. 通过表 1 列出的主要计算量,不能直接判定本文方法和 AHC + BIC 方法的计算量与谱聚类方法计算量之间的关系,将在实验部分通过实际运行时间来确定.

表 1 各方法的主要计算量

方法	主要计算量
本文方法	$(J-N_c)\times 0.5N'_c(N'_c-1)\times 4f(D_{KL}[\cdot\parallel\cdot])$
AHC + BIC	$(J-N_c)\times 0.5N'_c(N'_c-1)\times 3f(\ln(\lvert\det(\mathrm{cov}(\cdot))\rvert))$
特征距离谱聚类	$0.5J(J+1)\times f(d_1(x_i,x_j))+f(Dec(\boldsymbol{L}))+f(Clu(\boldsymbol{V}))$
模型距离谱聚类	$0.5J(J+1)\times f(d_2(x_i,x_j))+f(Dec(\boldsymbol{L}))+f(Clu(\boldsymbol{V}))$

4 实验及结果分析

4.1 实验数据及设置

实验数据取自两个数据库:Digital Juice Sound FX Library[25] 和 BBC Sound Effects Library[26],共有 11 类音频事件:掌声、哭声、低频乐器声、鸟叫声、鼓声、女声、枪声、男声、雨声、流水声、风声. 各音频事件样本时长范围:1 ~ 6 秒,16KHz 采样、16bits 量化、单通道 WAV 格式. 实验数据详细情况如表 2 所示. 表 2 列出每类音频事件样本个数是为了让读者知道所用实验数据情况,以及方便评估聚类方法性能,而聚类方法并不知道音频事件类别个数以及样本个数.

表 2 实验数据中各类音频事件样本个数

数据类型	时长	掌声	哭声	低频乐器	鸟叫	鼓声	女声	枪声	男声	雨声	流水	风声
调参集	1 ~ 6s	75	80	78	96	85	89	75	87	90	87	88
测试集	≤3s	100	100	100	100	100	100	100	100	100	100	100
	>3s	50	50	50	50	50	50	50	50	50	50	50
	1 ~ 6s	145	146	140	146	143	142	147	143	143	150	142

实验平台:Intel(R) Core(TM) i5-2400,3.10GHz CPU,4GB RAM,C/C++ 编程. 帧长为 32 ms,帧移为 16 ms. 特征参数为 MFCC 及其一阶差分,共 24 维,$Q=24$,所有方法都采用相同特征参数. 高斯混合模型的高斯元个数 M 设置为 16. 对比的方法是谱聚类方法[16,23] 和基于 AHC + BIC 的方法[17]. n_{ik} 表示在第 i 类中第 k 种音频事件样本帧数;N_s 表示音频事件类型总数;N_c 表示聚类类别总数;N 表示音频样本总帧数;$n_{\cdot k}$ 表示第 k 种音频事件样本总帧数. $n_{i\cdot}$ 表示第 i 类所包含的音频事件样本总帧数. 第 i 类的纯度,$\pi_{i\cdot}$:

$$\pi_{i\cdot}=\sum_{k=1}^{N_s}\frac{n_{ik}^2}{n_{i\cdot}^2} \tag{28}$$

平均类纯度 ACP(Average Clustering Purity):

$$\mathrm{ACP}=\frac{1}{N}\sum_{i=1}^{N_c}\pi_{i\cdot}n_{i\cdot} \tag{29}$$

第 k 种音频事件的纯度,$\pi_{\cdot k}$:

$$\pi_{\cdot k}=\sum_{i=1}^{N_c}\frac{n_{ik}^2}{n_{\cdot k}^2} \tag{30}$$

平均音频纯度 AAP(Average Audio Purity):

$$\mathrm{AAP}=\frac{1}{N}\sum_{k=1}^{N_s}\pi_{\cdot k}n_{\cdot k} \tag{31}$$

最后,采用 K 值评价算法整体性能:

$$K=\sqrt{\mathrm{ACP}\times\mathrm{AAP}} \tag{32}$$

K 值越大,算法性能越好.

4.2 实验结果

采用表 2 调参数据集确定各方法参数最优值,测试数据集用于评估各方法性能. 基于特征距离和基于模型距离的谱聚类方法的尺度因子 σ 分别为 2 和 15. 基于 AHC + BIC 方法的 BIC 惩罚系数为 2.

4.2.1 本文方法参数的确定

拉格朗日乘子 β 用来平衡聚类过程中信息的保留与压缩程度,其取值影响聚类结果. K 与 β 的关系如图 2 所示. 随着 β 的变化,K 值跟着变化,当 $\beta=11$ 时,获得最高的 K 值.

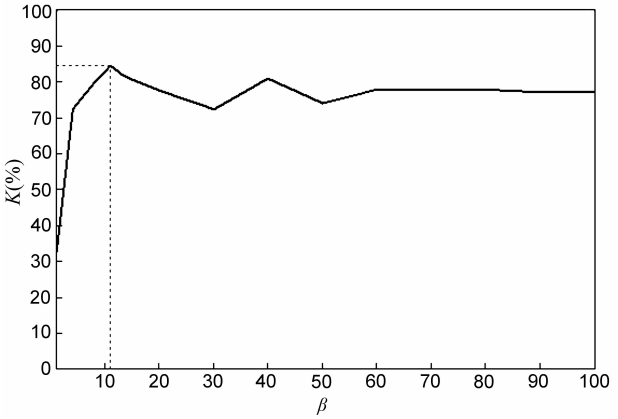

图2　K值与β的关系

图 3 给出了 β 取不同值时,K 值与聚类类别数 N_c 之间的关系. 当 $\beta=1$ 时,K 值基本上不受 N_c 的影响. 根据公式(1)有:$I(C,X)=\beta(F+I(Y,C))$,β 取较小值时,互信息 $I(C,X)$ 会变小. 又由于凝聚信息瓶颈方法中,$p(c\mid x)\in\{0,1\}$,条件熵:$H(C\mid X)=-\sum_{x\in X}p(x)\sum_{c\in C}p(c\mid x)\log p(c\mid x)=0$,互信息 $I(C,X)=H(C)-H(C\mid X)=H(C)$. 因此,当 β 取值较小时,聚类后目标变量 C 的熵 $H(C)$ 也会变小. 导致聚类后的样本分布非常不均衡,大部分样本都被聚到一个类中(即 $H(C)\approx 0$). 即会出现一个大的类和一些小的类,K 值几乎保持不变(不受 N_c 的影响). 相反地,如果 β 取值较大时(例如 $\beta=1000$),聚类后的样本分布会趋于均衡,K 值随着类别数 N_c 的变化而出现较大变化. 当 β 取中间值时(例如 $\beta=11$),K 值随着 N_c 的变化也会出现较大变化. 在 N_c 取值为[12 20],K 值很高,且当 $N_c=14$ 时,K 值达到最大. 最优聚类类别数 N_c 为 14. 图 4 给出了 K 值与门限 T_I 的关系. 当 $T_I=0.63$ 时,K 值达到最大. 因此,本文方法的参数设置为:拉格朗日乘子 $\beta=11$,互信息比值门限 $T_I=0.63$,聚类后最大的类别数 $N_{\max}=20$. $N_{\max}$的取值不小于最优聚类类别数($N_c=14$).

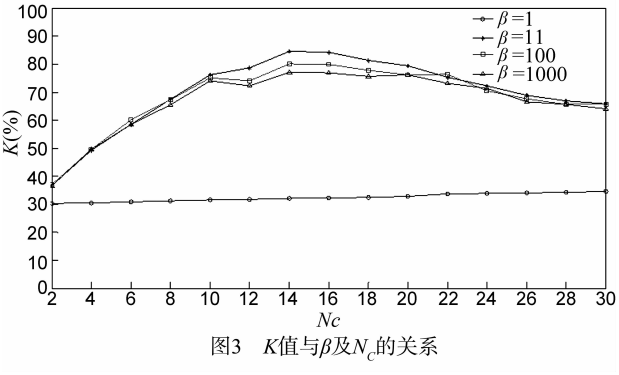

图3　K值与β及N_C的关系

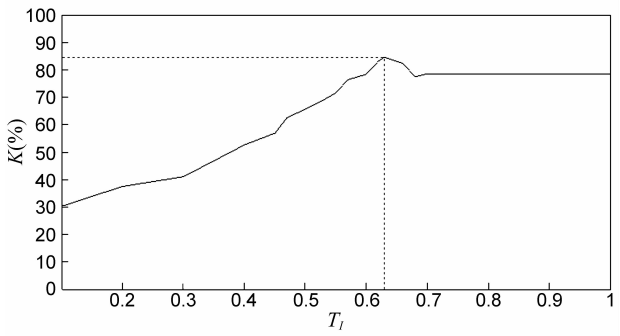

图4　K值与T_I的关系

4.2.2 各种方法的性能比较

各方法的实验结果如表 3 所示.

表 3　各方法的性能比较

样本时长	聚类方法	耗时(s)	ACP(%)	AAP(%)	K(%)
≤3s	特征距离谱聚类	1632	58.28	66.57	62.29
	模型距离谱聚类	657	68.60	71.75	70.16
	AHC + BIC	529	67.51	68.74	68.12
	本文方法	179	82.95	72.27	77.43
>3s	特征距离谱聚类	1098	74.49	80.08	77.24
	模型距离谱聚类	315	73.93	80.36	77.08
	AHC + BIC	248	78.59	82.37	80.46
	本文方法	89	91.07	84.25	87.59
1~6s	特征距离谱聚类	4398	63.85	69.50	66.62
	模型距离谱聚类	1265	68.72	72.78	70.72
	AHC + BIC	1011	73.42	74.23	73.82
	本文方法	326	87.92	74.09	80.71

从表 3 可看出,在不同样本时长条件下(聚类样本个数不同),本文方法运算速度和 K 值都优于其他方法,特别是在样本时长较短时(≤3s),本文方法与其他方法的 K 值差异更大. 这表明:本文方法取得了更优聚类性能,对样本时长不敏感,具有更好普适性. 从运算时间来看,完成聚类所需时间从小到大依次是:本文方法、基于 AHC + BIC 的方法、基于模型距离和基于特征距离谱聚类方法. 从 K 值来看,与其他方法相比,本文方法获得了更高 K 值,表明其性能更优. 本文方法采用概率分布的 Jensen-Shannon 散度度量类之间相似度,通过最大化目标函数获得最优聚类结果. 基于 AHC + BIC

方法采用 BIC 距离度量类之间相似度,当样本时长较大时(> 3s)取得了较高 K 值(高于谱聚类方法的 K 值),但在样本时长≤3s 时,其 K 值低于基于模型距离谱聚类方法的 K 值. 谱聚类方法是基于谱图理论的聚类方法,通过构建样本集矩阵并计算该矩阵特征值和特征向量,最后采用 K-means 聚类算法对特征向量进行聚类. 在样本时长≤3s 时,基于模型距离谱聚类方法明显优于基于特征距离谱聚类方法,稍优于基于 AHC + BIC 方法;但在样本时长 > 3s 时,基于模型距离谱聚类方法的 K 值最低. 将 1 ~ 6s 的音频事件样本作为实验数据时,本文方法也获得了最高 K 值,其次是基于 AHC + BIC、基于模型距离和基于特征距离的谱聚类方法. 为了更进一步了解本文方法的聚类情况,表 4 给出了它在对测试集中所有时长(1 ~ 6s)样本进行聚类时的混淆矩阵. 真实类别数为 11,而聚类后的实际类别数是 14,多了 3 个虚假类别. 样本数最少的 3 个类别分别是第 4 类(49 个样本)、第 12 类(68 个样本)和第 6 类(70 个样本). 平均类纯度等于 1 的类分别是:第 2 类、第 3 类、第 4 类、第 8 类、第 10 类、第 11 类、第 12 类和第 14 类. 平均音频纯度最高的是女声(等于 1),其次是哭声、男声、掌声. 平均音频纯度较低的是风声、鸟叫声和雨声,它们与其他音频事件混淆程度较高,较难区分.

表 4 本文方法聚类 1 ~ 6s 的样本时的混淆矩阵

类别	掌声	哭声	低频乐器	鸟叫	鼓声	女声	枪声	男声	雨声	流水	风声
1	141/145						9/140				
2									75/143		
3			99/140								
4											49/142
5				52/146				142/143			
6			28/140		16/143		6/140				20/142
7	4/145	145/146			2/143			1/143			
8							132/140				
9			13/140	1/146						22/150	73/142
10				93/146							
11					125/143						
12									68/143		
13		1/146				142/142					
14										128/150	

5 结论

本文提出一种基于凝聚信息瓶颈的音频事件聚类方法:在聚类过程中,通过最大化保留目标变量与相关变量之间的互信息且尽可能压缩源变量与目标变量之间的互信息,从而得到最优聚类结果. 该方法采用音频事件特征参数作为源变量、高斯混合模型作为相关变量、聚类后的类作为目标变量、采用 Jensen-Shannon 散度(两个 Kullback-Liebler 散度之和)度量类之间的距离. 与其他方法相比,在聚类各种时长音频事件样本时,本文方法都取得了更高 K 值,且速度更快.

参考文献

[1] Dan Stowell, et al. Detection and classification of acoustic scenes and events[J]. IEEE Trans. on Multimedia, 2015, 17(10):1733 – 1746.

[2] Ian McLoughlin, Zhang Haomin, Xie Zhipeng, et al. Robust sound event classification using deep neural networks[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2015, 23(3):540 – 552.

[3] Oguzhan Gencoglu, Tuomas Virtanen, Heikki Huttunen. Recognition of acoustic events using deep neural networks [A]. The 22nd European Conference on Signal Processing [C]. Lisbon: EURASIP, 2014. 506 – 510.

[4] Huy Phan, Marco Maaβ, Radoslaw Mazur, et al. Random regression forests for acoustic event detection and classification[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2015, 23(1):20 – 31.

[5] Selver Ezgi Kucukbay, et al. Audio-based event detection in office live environments using optimized mfcc-svm approach[A]. The IEEE 9th International Conference on Semantic Computing[C]. New York: IEEE, 2015. 475 – 480.

[6] Lu Xugang, et al. Sparse representation based on a bag of spectral exemplars for acoustic event detection[A]. Inter-

national Conference on Acoustics, Speech and Signal Processing[C]. New York:IEEE,2014. 6255-6259.
[7] Huy Dat Tran,et al. Sound event recognition with probabilistic distance SVMs[J]. IEEE Trans. on Audio, Speech, and Language Processing,2011,19(6):1556 – 1568.
[8] Lu Li,et al. A SVM-based audio event detection system [A]. International Conference on Electrical and Control Engineering[C]. New York:IEEE,2010. 292 – 295.
[9] Jose Portelo,et al. Non-speech audio event detection[A]. International Conference on Acoustics, Speech and Signal Processing[C]. New York:IEEE,2009. 1973 – 1976.
[10] Peng Ya-Ti, Lin Ching-Yung, Sun Ming-Ting. Audio event classification using binary hierarchical classifiers with feature selection for healthcare applications[A]. IEEE International Symposium on Circuits and Systems [C]. New York:IEEE,2008. 3238 – 3241.
[11] Zhang Xueyuan,He Qianhua,Feng Xiaohui. Acoustic feature extraction by tensor-based sparse representation for sound effects classification[A]. International Conference on Acoustics, Speech and Signal Processing[C]. New York:IEEE,2015. 166 – 170.
[12] Maria E. Niessen,Tim L. M. Van Kasteren,Andreas Merentitis. Hierarchical modeling using automated sub-clustering for sound event recognition[A]. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics [C]. New York:IEEE,2013. 1 – 4.
[13] Cai Rui,et al. A flexible framework for key audio effects detection and auditory context inference[J]. IEEE Trans. on Audio, Speech, and Language Processing, 2006, 14(3):1026 – 1039.
[14] 贺前华,等. 基于两步判决的口语中非文字音频事件检测方法[J]. 华南理工大学学报(自然科学版),2011,39(2):20 – 25.
He Qianhua,et al. Two-stage decision based detection of non-lexical audio events in spontaneous vocalization[J]. Journal of South China University of Technology (Natural Science),2011,39(2):20 – 25. (in Chinese)
[15] 罗森林,等. 融合 GMM 及 SVM 的特定音频事件高精度识别方法[J]. 北京理工大学学报,2014,34(7):716 – 722.
Luo Senlin,et al. High-precision specific audio event recognition method combining SVM and GMM[J]. Transactions of Beijing Institute of technology,2014,34(7):716 – 722. (in Chinese)
[16] Lu Lie,Alan Hanjalic. Audio keywords discovery for text-like audio content analysis and retrieval[J]. IEEE Trans. on Multimedia,2008,10(1):74 – 85.
[17] Margarita K,Vassiliki M,Constantine K. Speaker segmentation and clustering[J]. Signal Processing,2008,88(5):1091-1124.
[18] Cover T M,Thomas J A. Elements of Information Theory [M]. New York:Wiley,1991.
[19] N Tishby,F Pereira,W Bialek. The information bottleneck method[A]. The 37th Allerton Conference on Communication,Control and Computing[C]. Monticello:Springer, 1999. 368 – 377.
[20] N Tishby,F Pereira,W Bialek. The information bottleneck method[R]. NEC Research Institute TR,1998.
[21] N Slonim. The information bottleneck:theory and applications[D]. Ph. D. dissertation,Hebrew Univ. of Jerusalem, Jerusalem,Israel,2002.
[22] C H Lee,C H Lin,B H Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models[J]. IEEE Trans. on Signal Processing,1991, 39(4):806 – 814.
[23] 陈芬. 无监督说话人聚类方法研究及实现[D]. 广州:华南理工大学,2012.
Chen Fen. Research on unsupervised speaker clustering and its implementation[D]. Guangzhou:South China University of Technology,2012. (in Chinese)
[24] M Inaba, N Katoh, H Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering[A]. The 10th ACM Symposium on Computational Geometry[C]. New York:ACM,1994. 332 – 339.
[25] Digital Juice, Inc. , The Digital Juice Sound FX Library [DB/OL]. http://www. digitaljuice. com, accessed March 2009.
[26] British Broadcasting Corporation (BBC), BBC Sound Effects Library [DB/OL]. http://www. sound-ideas. com/bbc. html,accessed May 2010.

作者简介

李艳雄(通信作者)　男,1980 年出生,湖南嘉禾人,现为华南理工大学电子与信息学院先上岗副教授,硕士生导师,主要从事语音及音频处理、模式识别方面的研究工作.
E-mail:eeyxli@ scut. edu. cn

王　琴　女,1990 年出生,湖北京山人,现为华南理工大学通信与信息系统专业硕士研究生,主要从事语音及音频处理方面的研究工作.