DOI:10.13878/j.cnki.jnuist.2018.06.014

陈培培[1]　邵曦[1]

基于生成对抗网络的音乐标签自动标注

摘要

针对如何快速有效地对音乐信息进行查询、检索和组织的问题,提出了一种基于生成对抗网络模型的多标签音乐自动标注系统.通过音乐自动语义标注技术,可以提高音乐检索系统的性能.利用LDA方法对音乐标签进行聚类以获取主题类别,再通过生成对抗网络,找到音乐的音频特征与语义特征之间的映射关系.应用于CAL500数据集的5次交叉验证实验结果表明,该方法的综合性能指标与现有方法相比有较大的提升.

关键词

音乐自动标注;LDA模型;生成对抗网络

中图分类号 TN912

文献标志码 A

收稿日期 2018-04-20

资助项目 国家自然科学基金(70573025)

作者简介

陈培培,女,硕士生,主要研究方向为多媒体信息系统.chenpp904@163.com

邵曦(通信作者),男,博士,副教授,研究方向为多媒体信息系统与基于内容的音乐信息检索.shaoxi@njupt.edu.cn

1　南京邮电大学　通信与信息工程学院,南京,210003

0　引言

随着数字技术的飞速发展,人们会把海量的音乐资源上传到网上,因此音乐信息检索(MIR)系统得到的关注越来越多,但也给其处理音乐数据库带来了难度和挑战.目前,音乐检索系统的实现方法通常有两种,分别是基于内容的音乐分析及检索和基于文本的音乐分析及检索[1].前者主要是从音频文件中提取音频特征(如频谱、节奏、音色和音调等)并利用这些特征进行相似度匹配.类似于图像视觉特征和图像语义间存在的巨大鸿沟使得基于相似度的图像检索效果并不理想;由于可计算的音频特征与高层语义间也存在语义鸿沟,使得准确性通常不尽如人意,且系统的实现也较复杂.而后者由于是基于文本实现的,仅需采用文本信息(如音乐元数据、歌词和用户标签等)对音乐进行索引和检索,所以其过程与前者相比要简单得多.

随着Web2.0的发展,网络为多媒体信息提供了大量的用户标注的社会化标签,使得基于语义标签的音乐检索在许多应用场景中成为流行而实用的方法[2],例如基于标签的歌曲相似度计算、基于用户查询的相似歌曲列表推荐等,可以满足不同群体在不同环境中的需求.一些音乐推荐网站也将人工标签作为检索歌曲和导航的重要机制.由于网络用户标签的随意性和模糊性,音乐检索或推荐系统通常存在"冷启动"问题,因此利用统计学习算法进行高效的自动标注在当下显得尤为重要[3].

目前,主流的自动标注方法是通过学习歌曲的音乐内容来建立语义模型的.其中,有一类方法是基于判别模型的,如提升方法(boosting)[4]、隐马尔可夫模型(HMM)[5]和支持向量机(SVMs)[6],这类方法会学习如何根据音乐内容识别单个标签.然而,由于分配给每个标签模型的类标签不是均等表示的,这类方法会遇到不平衡的数据问题.另一类音乐自动标注方法是基于生成模型的.它们通过统计建模方法,可以从相关音频文件中学习到特定标签的特征分布,这些方法有高斯混合模型(GMMs)[7]、码字伯努利平均模型(CBAM)[8]以及狄利克雷混合模型(DMM)[9].基于这些标签模型,当标注未知音乐时,自动标注系统会生成关于音乐标签权重的向量.这个向量可被看作是一个多项式概率分布,用来表征每个标签与特定音频的相关性.然而,由于基于音乐内容的语义模型都是针对每个标签独立建模的,所以会造成标签间的联系产生的音乐上下文信息的丢失,这对于那些跟

音频特征关联性较大的流派类标签的自动标注效果较好,而对于诸如情绪等主观性较强的标签则很难学习和训练.

近年来,深度学习备受关注,并且在计算机视觉和自然语言处理领域已取得了很好的效果.深度学习是具有多层结构的机器学习算法,它能够有效地表征特征的潜在结构,其中深度置信网络和卷积神经网络就是两种典型的算法.在音乐信息检索领域,深度学习也被研究人员越来越多地应用于自动标注.如 Lee 等[10]提出了基于卷积深度置信网络的频谱学习和音乐分类算法;Sigtia 等[11]用深度神经网络进行音乐流派分类;Choi 等[12]提出了用完全卷积神经网络进行音乐自动标注.

自 Goodfellow 等[13]在 2014 年提出生成式对抗网络 GAN(Generative Adversarial Networks)后,各种基于 GAN 的衍生模型被提出,GAN 已经成为人工智能学界一个热门的研究方向.本文主要应用 LDA[14](Latent Dirichlet Allocation,潜在狄利克雷分配模型)和 GAN 两个模型实现音乐自动标注系统.研究重点主要在两个方面,一是利用 LDA 模型将音乐标签聚类以获取主题类别,用主题向量表示歌曲的语义特征;二是应用 GAN 的衍生模型 InfoGAN[15],通过训练这个网络,找到音乐的音频特征和语义特征之间的映射关系,从而实现对歌曲标签的标注.

1 基于生成对抗网络的音乐自动语义标注方法

本文所提的音乐自动标注系统的框架如图 1 所示.训练过程如下:首先,从音频文件中提取出歌曲的音频特征;其次,通过潜在语义建模,将上下文空间中的音乐标签建模为潜在概念空间;最后,将音乐的音频特征和语义特征通过生成对抗网络训练,找出它们之间的映射关系.在测试过程中,将未标注歌曲的音频特征通过生成对抗网络,由此得到该歌曲的预测标签.

1.1 基于 LDA 模型的语义建模

音乐标签矩阵包含了歌曲跟标签之间相关性大小的信息.如图 2 所示,矩阵 $\boldsymbol{A}$ 表示音乐标签矩阵,并且矩阵的每一列取得每一首歌曲的所有人工标注信息.通过把所有标注者针对每一个标签的标注值取平均,可以得到每首歌曲的标注向量.因此,可以推断如果有越多的标注者用词汇表中的特定词语来标注某首歌曲,那么在语义上描述该歌曲的那个词

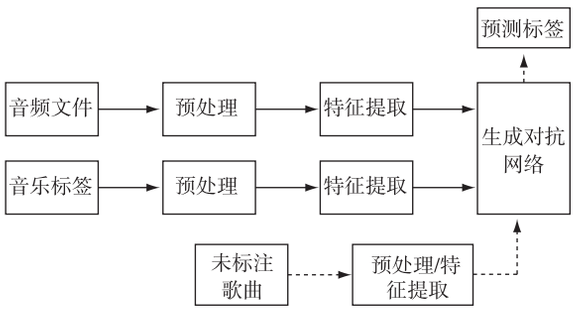

图 1 音乐自动标注系统框架

Fig. 1 The framework of music auto-tagging system

语就越重要.一些传统的语义模型已经被用来从音乐的社会标签中探索新的语义,比如基本矢量模型、潜在语义分析和 Aspect 模型.但是,这些传统模型在一些特定任务上表现得不是很好.本文提出用潜在狄利克雷分配模型(LDA)在社会标签中对语义进行建模.LDA 模型目前在文本挖掘领域包括文本主题识别、文本分类以及文本相似度计算方面都有应用,这里用 LDA 给音乐标签聚类,以获取其潜在的语义信息.

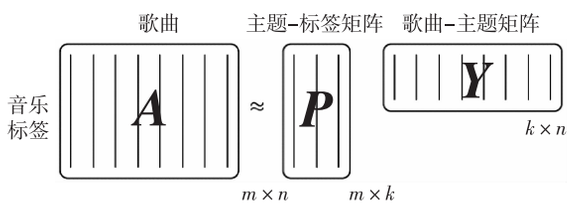

图 2 基于 LDA 模型分解的音乐标签矩阵

Fig. 2 LDA for music tags matrix

LDA 是一种文档生成模型.它认为一篇文章是有多个主题的,而这个主题又对应着不同的词.一篇文章的构造过程,首先是以一定的概率选择某个主题,然后再在这个主题下以一定的概率选出某一个词,这样就生成了这篇文章的第 1 个词.不断重复这个过程,就生成了整篇文章.LDA 的使用是上述文档生成的逆过程,它将根据得到的一篇文章,去寻找出这篇文章的主题,以及这些主题对应的词.LDA 模型的结构如图 3 所示.其中 α 和 β 都为 Dirichlet 分布的超参数,在实验中使用默认值 $1/k$(k 为隐含主题数);$\boldsymbol{\theta}$ 是一个主题向量,向量的每一列表示每个主题在文档出现的概率,该向量为非负归一化向量;$p(\boldsymbol{\theta})$ 是 $\boldsymbol{\theta}$ 的分布,具体为 Dirichlet 分布;N 表示要生成的文档的单词数;w_n 表示生成的第 n 个单词 w;z_n 表示选择的主题;$p(z \mid \boldsymbol{\theta})$ 表示给定 $\boldsymbol{\theta}$ 时主题 z 的概

率分布,具体为$\boldsymbol{\theta}$的值,即$p(z=i\mid\boldsymbol{\theta})=\boldsymbol{\theta}_i$;$p(w\mid z)$为主题$z$对应一个单词的概率分布.这种方法首先选定一个主题向量$\boldsymbol{\theta}$,确定每个主题被选择的概率,然后在生成每个单词的时候,从主题分布向量$\boldsymbol{\theta}$中选择一个主题z,按主题z的单词概率分布生成一个单词.由图3可知LDA的联合概率为

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

将此模型用于给音乐标签聚类,那么M是整个音乐集,w便是单个标签.目标便是得到每首歌曲的主题分布,以及各个主题下标签分布概率.把w当作观察变量,$\boldsymbol{\theta}$和z当作隐藏变量,就可以通过EM算法[16]学习出α和β,求解过程中遇到后验概率$p(\boldsymbol{\theta},z\mid w)$无法直接求解,需要找一个似然函数下界来近似求解,本文使用基于分解假设的变分法进行计算,用到了EM算法.每次E-step输入α和β,计算似然函数,M-step最大化这个似然函数,算出α和β,不断迭代直到收敛.

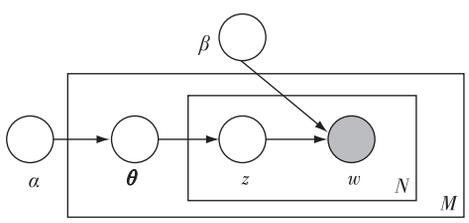

图3 LDA模型结构

Fig. 3 The structure of LDA model

如图2所示,可以使用LDA模型将音乐标签矩阵$\boldsymbol{A}(m\times n,n$首歌曲和$m$个标签)分解表示为$\boldsymbol{P}(m\times k)$和$\boldsymbol{Y}(k\times n)$的两个矩阵.矩阵$\boldsymbol{P}$是标签矩阵的语义主题,$\boldsymbol{P}$的每一列可以看作是特定主题的标签模式,而矩阵$\boldsymbol{Y}$可以看作是主题重要性指标矩阵,在$k$维的语义概念空间中,$\boldsymbol{Y}$的每个列向量$\boldsymbol{y}_i$可以被认为是特定歌曲$i$的关于主题分布的向量.

在音乐上下文建模后,矩阵$\boldsymbol{Y}$的列向量$\{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_i,\cdots,\boldsymbol{y}_n\}$可以看作是音乐的上下文信息的潜在表示.

1.2 基于生成对抗网络的音乐标签自动标注算法

生成对抗网络GAN是利用相互竞争游戏的一种深度生成模型.它的目标是学习生成器数据分布$P_g(x)$,使得该分布与真实的数据分布$P_{\text{data}}(x)$尽量接近.在原始的GAN中,D网络通过最大化判别真伪(G网络生成的伪造数据和真实数据)更新网络参数,G网络则是最大化欺骗D网络,提高数据造假的能力.为了生成一个样本,G使用了一个噪声变量z作为网络的输入.因为G网络的输入只有z,那么z就包含了生成一个样本所需的全部信息.原始GAN没有对生成器如何使用这个噪声做出约束,训练出来的生成器,对于z的每一个维度不能很好地对应到相关的语义特征.为了解决上述问题,本文使用GAN的衍生模型InfoGAN来实现音乐标签自动标注.

InfoGAN的输入噪声向量由两部分组成:1)z,可以看成是输入噪声向量;2)c,对应于语义向量.通过定义一系列的结构潜变量$c_1,c_2,\cdots$,这一系列潜变量相互独立,那么:

$$P(c_1,c_2,\cdots,c_L)=\prod_{i=1}^{L}P(c_i). \tag{2}$$

InfoGAN使用的是一种无监督的方法,让生成网络输入噪声变量z、潜变量c,即生成网络可以表示成:$G(z,c)$.然而,在标准的GAN中,如果直接这样作为网络的输入进行训练,那么生成器将忽略潜变量c的作用,即$P_G(x\mid c)=P_G(x)$,或者可以看成变量c与x相互独立、不相关.为了解决这个问题,InfoGAN模型中加入了信息正则化约束项:潜变量c与生成样本$G(z,c)$的互信息量应该较大,即$I(c;G(z,c))$应该较大.$I(x,y)$也可以看成X在给定Y与否的条件下不确定性的差值.如果X、Y相互独立,那么$I(X,Y)=0$,反之如果X和Y相关性较大,那么$I(X,Y)$也较大.因此对于$I(c;G(z,c))$来说,如果想要让它更大可以通过使$P_G(c\mid x)$更小来实现.为此在原始生成对抗网络GAN的损失函数$V(D,G)$的基础上$V(D,G)=E_{x\sim P_{\text{data}}}[\log D(x)]+E_{z\sim P_z}[\log(1-D(G(z)))]$,提出加入正则约束$I(c;G(z,c))$:

$$\min_G\max_D V_I(D,G)=V(D,G)-\lambda I(c;G(z,c)), \tag{3}$$

也就是在生成网络损失部分加入了互信息的惩罚.在实践中,如果直接最大化$I(c;G(z,c))$很难,因为需要求解后验概率$P(c\mid x)$.可以定义$Q(c\mid x)$来逼近$P(c\mid x)$,从而获得$p(c\mid x)$的变分下界.根据变分推断的理论,可以得到其下界函数如式(4)所示:

$$\begin{aligned}&I(c;G(z,c))=H(c)-H(c\mid G(z,c))=\\&E_{x\sim G(z,c)}[E_{c'\sim P(c\mid x)}[\log P(c'\mid x)]]+H(c)=\\&E_{x\sim G(z,c)}[\underbrace{D_{KL}(P(\cdot\mid x)\,\|\,Q(\cdot\mid x))}_{\geqslant 0}+\\&E_{c'\sim P(c\mid x)}[\log Q(c'\mid x)]]+H(c)\geqslant\\&E_{x\sim G(z,c)}[E_{c'\sim P(c\mid x)}[\log Q(c'\mid x)]]+H(c).\end{aligned} \tag{4}$$

假设$L_I(G,Q)=E_{c\sim P(c),x\sim G(z,c)}[\log Q(c\mid x)]+H(c)$,进一步化简式(4),可得:

$$L_I(G,Q)=E_{c\sim P(c),x\sim G(z,c)}[\log Q(c\mid x)]+H(c)=E_{x\sim G(z,c)}[E_{c'\sim P(c\mid x)}[\log Q(c'\mid x)]]+H(c), \quad (5)$$

$$I(c;G(z,c))\geqslant L_I(G,Q). \quad (6)$$

潜变量 C 的概率分布是人为设定的,$H(c)$ 不包含待优化的参数,因此 $H(c)$ 是一个常数,于是 InfoGAN 模型的损失函数,可以描述为

$$\min_{G,Q}\max_{D}V_{\text{infoGAN}}(D,G,Q)=V(D,G)-\lambda L_I(G,Q). \quad (7)$$

在实践中,用神经网络来参数化辅助分布 $Q(c\mid x)$,这也是一个判别模型网络,就是给定输入样本 x,判别对应的类别(c 对应类别标签).与 D 网络的差别在于,D 网络用于判别真伪;与 D 网络的共性是,Q 与 D 是参数共享网络,除了网络的最后一层分类层之外,因为 D 网络的最后一层是二分类,而 Q 网络则可能是其他多分类.具体网络结构如图 4 所示.具体的算法描述如下:

输入:随机噪声分布 $P_z(z)$,潜变量 c,真实样本,判别器 D 的迭代次数 r(默认为 1),学习率 l_1,生成器 G 的学习率 l_2,分类器 Q 的学习率 l_3,采样维度 t.

输出:D 的网络参数 w,G 的网络参数 u,Q 的网络参数 v.

Step0:初始化:w_0,u_0,v_0.

Step1:while u 未收敛 do.

Step2:D:For $j=0,1,2,\cdots,r$ do.

Step3:从随机噪声分布 $P_z(z)$ 和潜变量 c 中分别采样 t 个样本 $\{z^{(i)}\}_{i=1}^{t},\{c^{(i)}\}_{i=1}^{t}$.

Step4:从真实数据分布 P_{data} 中采样 t 个数据样本 $\{x^{(i)}\}_{i=1}^{t}$.

Step5:$d_w\leftarrow\nabla_w\frac{1}{t}\sum_{i=1}^{t}[\log D(x^{(i)})+\log(1-D(G(c^{(i)},z^{(i)})))]$.

Step6:$w\leftarrow w+l_1\cdot\text{SGD}(w,d_w)$.

Step7:end for.

Step8:G:从随机噪声分布 $P_z(z)$ 和潜变量 c 中分别采样 t 个样本 $\{z^{(i)}\}_{i=1}^{t},\{c^{(i)}\}_{i=1}^{t}$.

Step9:$g_u\leftarrow-\nabla_u\frac{1}{t}\sum_{i=1}^{t}\log(1-D(G(c^{(i)},z^{(i)})))$.

Step10:$u\leftarrow u-l_2\cdot\text{SGD}(u,g_u)$.

Step11:Q:将 Step8 中得到的假样本作为 Q 网络的输入.

Step12:$g_u,q_v\leftarrow\nabla\{E_{x\sim G(z,c)}[E_{c'\sim P(c\mid x)}[\log Q(c'\mid x)]]\}$.

Step13:$v\leftarrow v-l_3\cdot\text{SGD}(v,q_v),u\leftarrow u-l_3\cdot\text{SGD}(u,g_u)$.

Step14:end while.

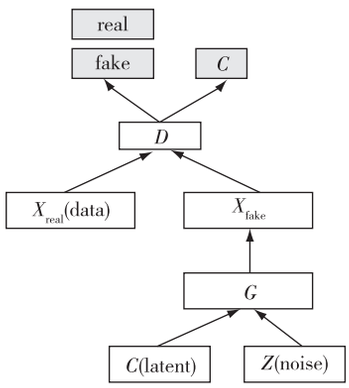

图 4 InfoGAN 的结构

Fig. 4 The structure of InfoGAN

2 实验和分析

2.1 数据集的选择

拟通过在 CAL500 数据集[7]上进行实验来评估本文所提出的自动音乐标记方法.CAL500 是西方流行音乐的 500 首音乐曲目的集合,每首音乐至少由 3 位注释人员手动注释.该数据集中有 174 个音乐相关的语义标签,包含了情绪、流派、乐器和声乐特点等多个不同的语义种类.在本实验中,为了与其他方法的结果进行比较,只考虑了 78 个标签,每一个标签至少标记过 50 首歌曲.实验采用 5 折交叉验证,每首歌曲在测试中出现一次.

CAL500 数据集提供了两种注释.一种是软注释,即对每首歌曲,把所有注释者针对每个标签的注释值取平均,这里使用软注释来实现上下文建模.另一个注释是基于“基本事实”的二元注释,即若值为 1 表示该标签标注了该歌曲,若值为 0 则表示该标签没有标注该歌曲.

2.2 实验及结果分析

本文采用 Matlab 进行程序开发,InfoGAN 的网络结构通过全连接层实现,具体参数设置如表 1 和表 2 所示.音频特征采用 Mel 倒谱系数来表示,本实验统一将音乐处理成采样率 16 kHz、wav 格式、单声道.对其进行预处理时,将音频通过 32 ms 的汉明窗,获得每帧音频信号的 512 个抽样点,提取 36 维 MFCC 系数.潜在语义 c 维数取决于标签主题数 k,噪声 z 维数这里选择 20 维.在本实验中,上文已得到的每首歌曲的主题向量 $\boldsymbol{y}_i$ 即为潜在语义 c. 在测试时,

表 1　网络参数设置

Table 1　Network parameter settings

判别网络 D/推断网络 Q	生成网络 G
输入:36 维的音频特征	输入:$c(k$ 维$)$,$z \in \mathbf{R}^{20}$ Unif$(-1,1)$
隐层:节点数 20,激活函数 relu	隐层:节点数 50,激活函数 relu
输出:D:节点数 1,激活函数 sigmoid Q:节点数 k,激活函数 softmax	输出:节点数 36,激活函数 sigmoid

音乐音频特征通过 Q 网络后得到的结果即是主题向量,用概率最高的几个主题对应概率最高的几个标签作为测试曲目的标签.

表 2　网络超参数设置

Table 2　Network hyper-parameter settings

参数	值
Optimizer	SGD
l_1	1×10^{-4}
l_2	1×10^{-3}
l_3	1×10^{-3}

本文用 3 个度量标准评估音乐标记实验结果,即准确率(P)、召回率(R)和 $F_{1\text{-measure}}$.准确率定义为歌曲被系统用标签 w 注释并且在数据集中也确实被 w 所标记.召回率定义为歌曲实际被标签 w 标记并且通过系统后也由 w 标记.$F_{1\text{-measure}}$ 是准确率和召回率的调和平均值,其表达式为:

$$F_{1\text{-measure}}=\frac{2PR}{P+R}. \tag{8}$$

本文共设计了 2 组实验并进行分析和统计,第 1 组是不同的标签潜在主题数 k 下,对音乐标记结果的影响;第 2 组是将本文所提方法与另 2 种算法得到的实验结果进行对比.

在第 1 组实验中,讨论上下文建模中潜在主题的维数 k 对音乐自动标记任务的影响.为了评估主题数 k 的影响,在相同条件下(相同的音乐内容分析设置),k 分别取值为 3,6,9,12 进行实验.图 5 显示了音乐标记的 3 个评估指标.由图 5 可以看出,当 k 取 6 时,音乐标记的 P,R 和 $F_{1\text{-measure}}$ 都表现最优.结果表明,音乐标签标注的最佳性能是通过捕获音乐社会标签的一些重要潜在主题来实现的.由于实验中设置的音乐标签并不包含过于丰富的语义,因此若潜在主题的数量太少,无法捕捉到音乐语境的变化,而潜在主题的数量过大,潜在主题的重载会过度驱散真正潜在话题的重要性.

在第 2 组实验中,本文选择与具有代表性的模型 HEM-GMM[7] 和 HEM-DTM[17] 进行对比.每种方法

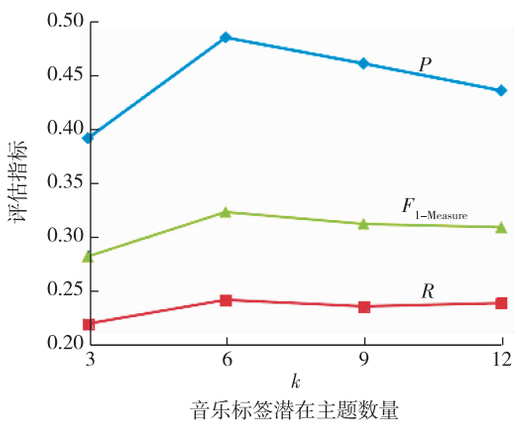

图 5　音乐标签潜在主题数量对音乐标记的影响

Fig. 5　The influence of the number of latent topics on music auto-tagging

的 3 种度量标准的实验结果如表 3 所示,可见本文方法的 3 种度量标准的结果较好,进一步证明了该解决方案的可行性.在最先进的自动标记系统中,由于所使用的协议不同,本文没有进一步实验与文中所提方法进行比较.

表 3　3 种方法的实验结果

Table 3　Experimental results for each method

方法	准确率 P	召回率 R	$F_{1\text{-measure}}$
HEM-GMM[7]	0.432	0.218	0.290
HEM-DTM[17]	0.458	0.234	0.310
本文方法	0.484	0.242	0.323

3　结束语

本文提出了基于生成对抗网络的音乐自动标记方法,首先通过 LDA 模型捕捉隐藏在音乐上下文背后的语义,继而利用生成对抗网络找到音乐音频特征与语义特征之间的映射关系.从实验结果可以看出,该算法在理论和实践上都具有现实意义.总体而言,与以往的解决方案相比,本文提出的解决方案取得较好的效果,有一定的可行性.但限于时间、数据、计算机性能等因素,本文并未通过实验来证明解决

方案的稳定性.未来可以选择更大规模的数据集进行测试,也可以选择更多的音频特征,使得歌曲表示更具有准确性,从而提高系统标注性能.

参考文献
References

[1] Schedl M, Gómez E, Goto M. Multimedia information retrieval: music and audio[C]// Proceedings of the 21st ACM International Conference on Multimedia. ACM, 2013:1117-1118

[2] Levy M, Sandler M. Music information retrieval using social tags and audio[J]. IEEE Transactions on Multimedia, 2009, 11(3):383-395

[3] 高天虹,马恩云.效率与成本是数据采集迎接挑战的关键[J].国外电子测量技术,2014,33(3):6-8
GAO Tianhong, MA Enyun. Efficiency and cost are key to meeting data acquisition challenges [J]. Foregin Electronic Measurement Technology, 2014, 33(3):6-8

[4] Bertin-Mahieux T, Eck D, Maillet F, et al. Autotagger: a model for predicting social tags from acoustic features on large music databases [J]. Journal of New Music Research, 2008, 37(2):115-135

[5] Coviello E, Lanckriet G R, Chan A B. The variational hierarchical EM algorithm for clustering hidden Markov models[C]// Advances in Neural Information Processing Systems. 2012:404-412

[6] Mandel M I, Ellis D P W. Multiple-Instance Learning for Music Information Retrieval[C]// Proceedings of ISMIR. 2008:577-582

[7] Turnbull D, Barrington L, Torres D, et al. Semantic annotation and retrieval of music and sound effects[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(2):467-476

[8] Hoffman M D, Blei D M, Cook P R. Easy as CBA: a simple probabilistic model for tagging music[C]// Proceedings of ISMIR, 2009, 9:369-374

[9] Miotto R, Lanckriet G. A generative context model for semantic music annotation and retrieval[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(4):1096-1108

[10] Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]// Advances in Neural Information Processing Systems. British Columbia, Canada: DBLP, 2009:1096-1104

[11] Sigtia S, Dixon S. Improved music feature learning with deep neural networks[C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. USA: IEEE, 2014:6959-6963

[12] Choi K, Fazekas G, Sandler M. Automatic tagging using deep convolutional neural networks[EB/OL]. [2016-05-10]. https://arvix.org/abs/1606.00298

[13] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]// Proceedings of the 2014 Conference on Advances in Neural Information Processing System. Montreal, Canada, 2014:2672-2680

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3:993-1022

[15] Chen X, Duan Y, Houthooft R, et al. Interpretable representation learning by information maximizing generative adversarial nets[C]// Proceedings of the 2016 Neural Information Processing Systems. Barcelona, Spain: Department of Information Technology IMEC, 2016:2172-2180

[16] Popescul A, Ungar L H, Pennock D, et al. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments[C]// Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. SanFrancisco: Morgan Kaufmann, 2001

[17] Coviello E, Chan A B, Lanckriet G. Time series models for semantic music annotation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(5):1343-1359

Music auto-tagging based on generative adversarial networks

CHEN Peipei[1] SHAO Xi[1]

1 College of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003

Abstract For the problem of how to query, retrieve, and organize music information quickly and efficiently, the performance of the music retrieval system can be improved through automatic music annotation technology. In this study, a multi-label music automatic annotation system based on generative adversarial networks (GANs) is proposed. The LDA model is used to cluster the music tags to obtain thematic categories, and then the mapping relationship between the audio features and the semantic features of the music is found by the generative adversarial network. For experimental verification, when the method proposed in this paper was applied to the CAL500 dataset in five cross-validation experiments, the comprehensive performance index of the method was greatly improved compared with existing methods.

Key words music automatic tagging; LDA model; generative adversarial network