

文章编号:1006-3080(2017)04-0540-06

DOI:10.14135/j.cnki.1006-3080.2017.04.013

基于混合判别受限波兹曼机的音乐自动标注算法

王诗俊, 陈宁

(华东理工大学信息科学与工程学院, 上海 200237)

摘要:对于音乐自动标注任务,在很多情况下,未标注的歌曲量远远超过已标注的歌曲数据,从而导致训练结果不理想。生成模型能够在某种程度上适应少量数据集的情况,得出较为满意的结果,然而,在有充分数据集的情况下生成模型的效果却劣于判别模型。本文提出了一种结合生成模型与判别模型两者优势的面向音乐自动标注的混合判别波兹曼机模型,该模型可明显提升音乐自动标注的准确率。实验结果表明,混合波兹曼机的效果不仅好于传统的机器学习模型,同时,模型在拥有足够训练数据量的情况下与判别模型效果相当,且在训练集较少的情况下效果也好于判别模型。另外,为了防止模型过拟合,还引入了 Dropout 规则化方法以进一步加强模型的性能。

关键词:音乐自动标注;混合判别受限波兹曼机;机器学习;人工智能

中图分类号:TP391

文献标志码:A

Annotating Music with Hybrid Discriminative Restricted Boltzmann Machines

WANG Shi-jun, CHEN Ning

(School of Information Science and Engineering, East China University of
Science and Technology, Shanghai 200237, China)

Abstract: For the music annotation, the amount of unlabeled music data is often much more than the labeled ones such that the training results are usually unsatisfying. Although generation model can be suitable for the smaller training data case to some extent and get higher quality results, it may be inferior to the discriminative model in the case of sufficient training data. By combining the advantages of the generation model and the discriminative model, this paper presents a hybrid discriminative restricted Boltzmann machines. The proposed hybrid model can improve the accuracy of the music annotation tasks. The experiment results show that the hybrid model is much better than the traditional machine learning models. Moreover, it is also better than the single discriminative Boltzmann machines for the case that the amount of training data is small and can attain the similar performance to the discriminative model in the case that the amount of training data is sufficient. Besides, the Dropout method is introduced in this paper to improve the model and prevent the overfitting for the smaller training data.

Key words: annotating music; hybrid discriminative restricted Boltzmann machines; machine learning; artificial intelligence

收稿日期:2016-09-28

基金项目:国家自然科学基金(61271349)

作者简介:王诗俊(1991-),男,硕士生,研究方向为音频信号处理。

通信联系人:陈宁, E-mail:chenning_750210@163.com

随着数字音乐的发展,在线歌曲的数量呈指数级增长。让用户快速找到自己感兴趣的歌曲,是目前互联网服务必须解决的一个问题。一个可行的方法是利用歌曲标签,即用户可以通过搜索标签的方式找到自己喜欢的歌曲。

Tingle 等^[1]利用两种新的特征——Echo Nest Timbre (ENT)和 Echo Nest Songs (ENS),以声学标签 (Acoustic tags) 和流派标签 (Genre tags) 作为标注为音频进行自动标注。Sordo 等^[2]利用基于音频内容相似性的模型对具有近似风格或感情的音乐进行了自动标注。

除上述标签外,一种最容易获得的标签是社群标签 (Social tags)^[3],社群标签的获得不需要专业人士的参与,极大地节约了人力和时间成本。所有用户都可以为歌曲进行标注,例如“女声”、“舒缓”、“好听”等语义级别描述。文献[4]提出了利用 MFCC (Mel Frequency Cepstrum Coefficient) 特征来训练 AdaBoost 模型,该模型能够根据音频特征和社群标签进行自标注。文献[5]利用社群标签进行情感音乐分类,使用聚类的方式将音乐分为“高兴”、“悲伤”、“愤怒”和“温和”4 种情绪。文献[6]利用文本挖掘和信息检索技术对庞大的语义级社群标签进行文本分析,从而最终对音乐进行情感分类。文献[7]结合利用用户信息、社群标签以及音频特征为用户推荐音乐。

但社群标签存在一些问题。第一,并不是所有的歌曲都拥有足够的社群标签,只有少量的热门歌曲拥有足够多的标签,新歌或是冷门歌曲的标签都很缺乏。第二,用户可能会使用同义词(如 Favorite, Favorites)、无用词(如 Awesome, Own it)、歧义词(如 Love 可以代表用户喜爱的歌曲也可代表是爱情歌曲)来标注音乐。第三,由于社群标签是语义级特征,需要对其进行语义分析。因此,可以参与模型进行训练的歌曲量其实并不多。正因为如此,在少量训练数据集的情况下对音乐进行准确的自动标注非常重要。

目前,拥有高准确率的深度学习模型被用于音乐自标注研究中。文献[8]利用深度置信网络 DBN (Deep Belief Network) 来自动提取输入音频的特征以进行流派分类,不过该模型需要足够的训练集,并且流派类别非常少,属于某个流派的歌曲数量依旧十分庞大,无法使用户快速找到自己喜欢的歌曲,本文不再对该方法进行比较。

生成模型 (Generative model) 可以解决训练数据集数量较少的问题,在少量数据集情况下的效果

要好于分类问题中常使用的判别模型 (Discriminative model)^[9],但在拥有大量训练数据的情况下,判别模型却要优于生成模型。为了兼顾两者的优势,本文提出了一种将生成波兹曼机和判别波兹曼机混合组成的混合判别波兹曼机 (Hybrid Discriminative Restricted Boltzmann Machines, HDRBM)。实验证明该模型即使在少量带标签音乐训练集的情况下也能达到精准标注的结果。

另外在训练模型时,可能因为训练数据集太少会导致过拟合的情况。为了防止过拟合,本文在混合判别波兹曼机中加入了 Dropout^[10] 规则化方法。Dropout 能够提高模型的准确率^[11],本文实验证实了 Dropout 使混合波兹曼机在小规模数据集下也能保持准确性。

混合判别波兹曼机模型在音乐数据训练集较为充分时,标注准确率能达到与单一的判别波兹曼机相仿的效果,在小规模训练集上训练时,混合判别波兹曼机的准确度相比判别波兹曼机或是其他机器学习模型更为优秀。同时,本文的实验证实了 Dropout 方法确实能够防止因为少量音乐数据集而导致的过拟合现象。

1 受限波兹曼机

1.1 生成波兹曼机

如图 1 所示,传统波兹曼机是一种无向生成模型,拥有一层输入层和一层隐层,层与层之间互连接,层之内无连接。通过训练,隐层可以学习到输入层的概率分布模型,也可以理解为隐层学习到了输入层的特征。

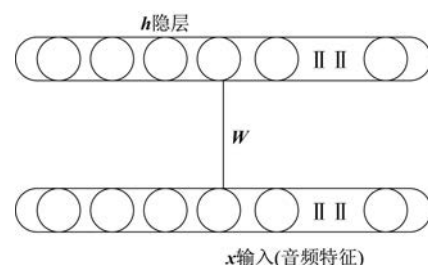


图 1 生成受限波兹曼机

Fig. 1 Generation restricted Boltzmann machines

传统受限波兹曼机是无监督学习模型,若将音频放入模型中学习,隐层将能够学习到音频特征,并重构原始的输入音频信息。正因为这种特性,传统波兹曼机常常被当作多层深度学习模型中的一层,用来提取上一层输入的特征。不过,将原来的

输入(音频)联合其类别(社群标签)一同输入波兹曼机,就能使波兹曼机学习到输入和标签的联合分布,即生成波兹曼机,如图2所示。

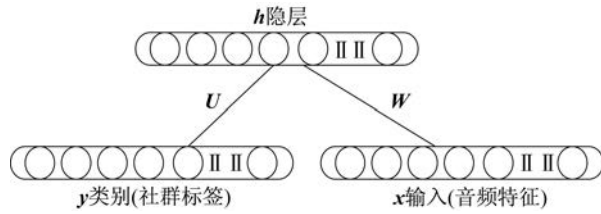


图2 判别受限波兹曼机

Fig. 2 Discriminative restricted Boltzmann machines

在音乐自标注任务中,假设生成波兹曼机有 n 个隐节点,输入数据为 d 维音频特征,有 c 种音乐标签,由此形成的概率分布为

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h})} \quad (1)$$

其中: $\mathbf{x} = (x_1, \dots, x_d)$ 为输入数据; $\mathbf{y} = (y_1, \dots, y_c)$ 为标签; $\mathbf{h} = (h_1, \dots, h_n)$ 为隐层节点; $Z = \sum_{\mathbf{x}, \mathbf{y}, \mathbf{h}} e^{-E(\mathbf{x}, \mathbf{y}, \mathbf{h})}$ 为归一化系数。

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{b}^T \mathbf{x} - \mathbf{c}^T \mathbf{h} - \mathbf{d}^T \mathbf{y} - \mathbf{h}^T \mathbf{U} \mathbf{y} \quad (2)$$

式(2)为波兹曼机的能量公式,其中,参数 \mathbf{W} 为隐层和输入层之间的连接权值系数; \mathbf{U} 为隐层和标签之间的连接权重系数; $\mathbf{b}, \mathbf{c}, \mathbf{d}$ 分别为输入层、隐层和标签的偏置系数。

假设有训练集 D_{train} ,以最小化负对数似然为目标,即最小化目标函数:

$$L_g(D_{\text{train}}) = - \sum_{i=1}^{|D_{\text{train}}|} \lg p(\mathbf{x}_i, \mathbf{y}_i) \quad (3)$$

其中, $p(\mathbf{x}_i, \mathbf{y}_i)$ 为一个样本 \mathbf{x} 和一个标签 \mathbf{y} 的联合分布。

为了最小化式(3),使用随机梯度下降法,得到由两个期望组成的梯度公式:

$$\frac{\partial \lg p(\mathbf{x}_i, \mathbf{y}_i)}{\partial \theta} = -E_{\mathbf{h}|\mathbf{x}_i, \mathbf{y}_i} \left[\frac{\partial}{\partial \theta} E(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \right] + E_{\mathbf{x}, \mathbf{y}, \mathbf{h}} \left[\frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{y}, \mathbf{h}) \right] \quad (4)$$

式(4)的第1项可以通过给定的 $\mathbf{x}_i, \mathbf{y}_i$ 计算得到;第2项却因为需要所有的 \mathbf{x} 和 \mathbf{y} 求和,不能通过直接的计算求得,可以通过逼近的方法得到。文献[12]提出了一种对比散度方法(Contrastive Divergence, CD),通过以 $(\mathbf{x}_i, \mathbf{y}_i)$ 为起始点的有限步长吉布斯采样来逼近所要求的期望。

1.2 判别波兹曼机

生成波兹曼机能得到音频特征与社群标签之

间的联合概率分布,但本文关注的是社群标签的自动标注,因此,我们的目标是对输入音乐的精准标注,而不是概率分布,即将目标函数 $p(\mathbf{x}, \mathbf{y})$ 替换为 $p(\mathbf{y}|\mathbf{x})$ 。继续考虑负对数似然,目标函数从式(4)的联合分布变为条件分布:

$$L_d(D_{\text{train}}) = - \sum_{i=1}^{|D_{\text{train}}|} \lg p(\mathbf{y}_i | \mathbf{x}_i) \quad (5)$$

通过与式(4)相似的式(6)计算梯度下降:

$$\frac{\partial \lg p(\mathbf{y}_i | \mathbf{x}_i)}{\partial \theta} = -E_{\mathbf{h}|\mathbf{x}_i, \mathbf{y}_i} \left[\frac{\partial}{\partial \theta} E(\mathbf{x}_i, \mathbf{y}_i, \mathbf{h}) \right] + E_{\mathbf{y}, \mathbf{h}|\mathbf{x}_i} \left[\frac{\partial}{\partial \theta} E(\mathbf{x}_i, \mathbf{y}, \mathbf{h}) \right] \quad (6)$$

同生成波兹曼机一样,在训练判别波兹曼机时,同样可以利用对比散度方法计算梯度,最大的区别是在计算式(6)的第2项时,由于是计算后验概率,不需要考虑输入 \mathbf{x}_i 的重构,因此将 \mathbf{x}_i 固定即可。

2 基于混合判别波兹曼机的音乐自标注模型

相比于生成波兹曼机,判别波兹曼机的优势体现在训练时数据集的量上,大量的训练集将使判别波兹曼机发挥更出色,反之,生成波兹曼机在少量的数据集上更有优势。

因此,为了既能适应本文音乐数据集较少的情况,又能在日后数据集足够充足的情况下不影响性能,本文提出了一种结合两者优势的基于混合判别波兹曼机模型的音乐自标注模型,使得音乐标注任务在不同数量的音乐数据集上的标注结果达到比单一模型更好的效果,同时标注的准确度好于传统的机器学习模型。

混合判别波兹曼机的目标函数由式(3)和式(5)共同组成:

$$L_h(D_{\text{train}}) = L_d(D_{\text{train}}) + \alpha L_g(D_{\text{train}}) \quad (7)$$

其中,可调参数 α 表示生成模型对于整个模型的影响比重。若 α 较大,则更偏重于生成波兹曼机,即少量音乐数据集的情况;反之,则偏重于判别波兹曼机。在实际运用中,可以利用交叉验证的方法调节 α ,以达到在不同音乐训练集的情况下都能为新的音乐标注上准确的社群标签。事实上,也可以将式(7)的后一项看成判别波兹曼机的正则化项。

另一方面,在机器学习训练中,有时由于训练数据集不充足,模型会产生过拟合问题,最终训练得到的模型对在训练时“未见过”数据的分类效果很差。为此,本文引入了近些年在深度模型中广泛

使用的 Dropout 规则化方法。

在传统的神经网络模型训练过程中,层与层之间的权值和偏置都会在每一次的训练过程里强制更新,因此可能会导致一些系数过分地共适应(co-adaptation),也就是说,一些权重系数可能会收敛于某个值,而这个值可能会过分依赖于另一些权重系数所收敛的值。

在训练过程中,Dropout 将一层单元中的一部分随机置零,也就相当于舍弃被置零的单元与下一层的链接,这样便会阻止神经元间的过分的共适应。文献[10]证实了带有 Dropout 的受限波兹曼机将优于传统受限波兹曼机。

本文将 Dropout 应用于混合判别波兹曼机中,使输入音频特征的一部分置零,以达到防止模型过拟合的效果。同时,实验结果也证实了 Dropout 确

实能使模型的标注准确率更高。

图 3 为基于混合判别波兹曼机的音乐自标注模型的训练流程图。首先从原始音乐中提取出色特征和旋律特征,然后送入混合波兹曼模型中进行训练,以最小化负对数似然为目标函数,其中混合模型使用了 Dropout 方法防止过拟合。通过交叉验证方法,调整模型中的参数使模型达到最优的效果。

与流派标注或其他标注问题不同的是,在为音乐标注社群标签时,一首歌曲一般有多个标签,如一首音乐可能被同时标记为“男声”、“摇滚”等,因此,在使用模型为测试集音乐标注时,将混合波兹曼机由 softmax 函数得到的一组概率值向量 $p(y|x)$,即表征模型将一首歌曲标记为某个标签的把握程度,取其概率最大的一部分标记与真实标记进行比较来计算模型的准确率。

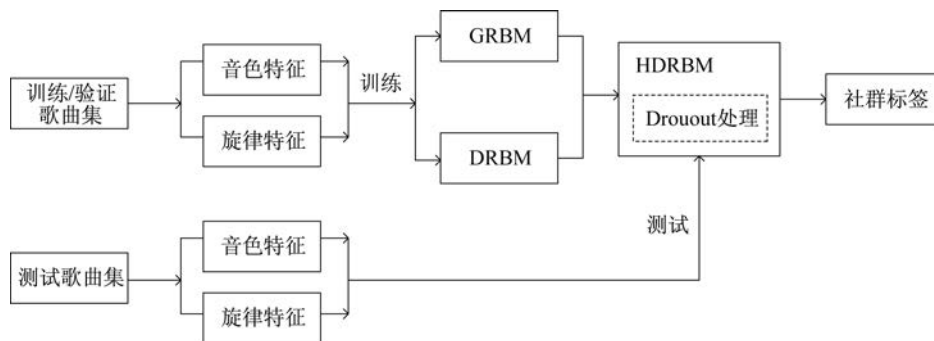


图 3 带 Dropout 的混合判别波兹曼机的训练流程图

Fig. 3 Training flowchart of the Hybrid discriminative restricted Boltzmann machines with Dropout

3 实验结果与分析

3.1 数据集

本文实验采用混合判别波兹曼机模型为音乐自动标注。社群标签数据集采用 Amazon.com 的 Mechanical Turk 数据集^[13]。Mturk 让用户任意描述一段 10 s 的歌曲片段,用户可以随意去标记片段的流派、情感和乐器类型等多种多样的标签(例如“孤独”、“快节奏”等语义标签)。本实验一共选取 915 首 10 s 片段,同时摘取最常用且有意义的 25 种社群标签。

模型输入特征采用文献[14]中的音色与旋律特征。音色特征是音频梅尔倒谱系数(MFCC)协方差的均值与方差,这种特征能捕捉到音频的乐器信息。旋律特征由 4 个频带得到,能够提取到鼓乐器成分,还能够分离具有强烈节奏的乐曲与其他乐曲,如舞曲与摇滚民谣。上述特征都进行了归一化预处理,具有零均值与单位方差。音色特征有 189

维,旋律特征有 200 维,将两者合并,输入为 389 维特征。

3.2 实验模型

本文实验采用了判别波兹曼机、混合判别波兹曼机、带有 Dropout 处理的混合判别波兹曼机、经典分类器——支持向量机(SVM),以及多层感知机 MLP 模型。

所有的波兹曼机在建立模型时都引入了高斯单元^[15]以适应连续音频特征 x 和二值标签序列 y ,具体来说,在训练时,通过隐层生成的可视层具有正态分布(连续值序列),通过隐层生成的标签为二值序列。

实验还使用了线性核支持向量机(SVM)来进行比较,由于是多分类问题,故采用一对多分类器。

3.3 实验

3.3.1 实验 1 实验 1 比较了在训练数据集与测试数据集数量不同的情况下,判别波兹曼机、混合判别波兹曼机、SVM 与多层感知机 MLP 结果的差异。实验结果如图 4 所示。

训练过程中,采用 10 折交叉验证(10-fold validation)设置使模型最优的参数。数据集被分为 10 折,模型将被训练 10 次,每一次的训练都采用 9 折的样本集,剩下的 1 折样本子集作为测试集。这种方法能最大限度地利用整个数据集样本,并且由于验证集基本涵盖了整个数据集,能最大限度考察模型的泛化能力并依据验证集设置模型参数。最终,本实验的波兹曼机与多层感知机的模型系数如表 1 所示,其中,将式(8)中的 α 设为 0.005,以达到较为显著的效果。

从图 4 中可以看到,当训练数据集足够多时,混合判别波兹曼机和判别波兹曼机都能达到较为满意的准确率,且两者的结果非常接近且都优于 SVM 与 MLP。当训练数据集开始减少时,所有的分类器的准确率都开始下降,但混合判别波兹曼机的优势开始体现出来,比判别波兹曼机准确率高 2% 左右。当训练数据集进一步减少时,混合判别波兹曼机的准确率与判别波兹曼的准确率之间的差距又有了进一步的拉大。

表 1 实验 1 中波兹曼机与多层感知机的模型参数

Table 1 Model parameters of the Boltzmann machines and the MLP in experiment 1

分类器	学习速率	隐节点数	CD 次数
DRBM	0.01	1 000	10
HDRBM	0.01	1 000	10
MLP	0.01	128 (2 层)	

表 2 实验 2 中的模型参数

Table 2 Model parameters in experiment 2

分类器	学习速率	隐节点数	CD 次数	α	Droupout 率
SVM	0.01				
MLP	0.01	250 (2 层)			
DRBM	0.01	50	10		
HDRBM (未规则化)	0.01	50	10	0.05	
HDRBM (规则化)	0.01	50	10	0.05	0.75

表 3 实验 2 中 5 种分类器的 AUC 准确率结果

Table 3 AUC accuracy of five classifiers in experiment 2

分类器	AUC 准确率/%
MLP	65.6
SVM	66.0
DRBM	68.4
HDRBM (未规则化)	68.4
HDRBM (规则化)	68.9

更多的实验数据表明,在有足够多的训练数据学习时,混合判别波兹曼机与判别波兹曼的效果非常近似,当不断减少训练数据量时,混合判别波兹曼机的优势逐渐体现出来,最终的差距维持在 3% 左右。

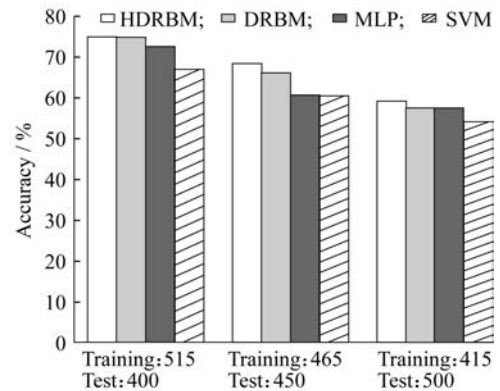


图 4 不同数据集样本数量下 4 种分类器的准确率

Fig. 4 Accuracy of four classifiers in different amounts of datasets

3.3.2 实验 2 实验 2 把经 Dropout 处理后的混合判别波兹曼机与文献[16]中的判别波兹曼机、多层感知机(MLP)、SVM 进行比较,观察 Dropout 规则化处理方法的效果。训练采用 5 折交叉验证来设置参数与比较模型,其中 3 折是训练集,1 折是验证集,1 折是测试集,即 549 个训练样本,验证集和测试集分别为 189。同文献[16]一样,也采用 AUC 判别标准^[17]。模型参数如表 2 所示,实验结果如表 3 所示。

实验结果表明,Dropout 确实能够提升混合判别波兹曼机的性能,防止因为训练样本过少而产生的过拟合情况。在同样的环境参数下,经过 Dropout 的混合判别波兹曼机比未经过 Dropout 的单一判别波兹曼机或混合判别波兹曼机的准确率提高了 0.5% 左右。

4 结束语

本文通过实验论证了混合判别波兹曼机在音乐标注上的效果不仅好于传统分类器,且在训练数据集较少的情况下,效果也能优于判别波兹曼机。这一优势在带标注数据较少的情况下,即带标注数据少,而未标注数据量极大的情况下很有意义。同时,为了进一步应对实验中少量训练集的情况,引入了 Dropout 方法,以解决混合判别波兹曼机模型过拟合问题。

在音乐推荐系统中,往往需要计算得到符合用户需求的音乐,从而能够将优秀的推荐结果呈现给用户,而通过人工对音乐进行标注、分类等工作需要耗费大量的人力成本以及时间成本。本文研究的模型能够在一定程度上为音乐进行自动标注,从而能为后续的推荐系统算法,包括基于内容或基于协同过滤等推荐算法提供便利。

目前的深度学习模型中,卷积神经网络(Convolutional neural network)或是深度置信网络(Deep belief network)的分类准确率很高,我们将把深度学习模型与本文研究的混合判别波兹曼机相结合,即将深度学习模型最后的线性回归(Logistic regression)层替换为混叻判别波兹曼机,同时加上 Dropout 等规则化方法,试图能够在少量样本数量的情况下得到更好的标注结果。

参考文献:

- [1] TINGLE D, KIM Y E, TURNBULL D. Exploring automatic music annotation with acoustically-objective tags [C]// Proceedings of the International Conference on Multimedia Information Retrieval. Philadelphia, PA, USA: ACM, 2010: 55-62.
- [2] SORDO M, LAURIER C, CELMA O. Annotating music collections: How content-based similarity helps to propagate labels [C]// Proceedings of the 8th International Conference on Music Information Retrieval. Vienna, Austria: DBLP, 2007: 531-534.
- [3] LAMERE P. Social tagging and music information retrieval [J]. Journal of New Music Research, 2008, 37(2): 101-114.
- [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] LAURIER C, SORDO M, SERRA J, *et al.* Music mood representations from social tags [C]// Proceedings of the 10th International Society for Music Information Retrieval Conference. Kobe, Japan: DBLP, 2009: 381-386.
- [6] LEVY M, SANDLER M. A semantic space for music derived from social tags [C]// Proceedings of the 8th International Conference on Music Information Retrieval. Vienna, Austria: DBLP, 2007: 411-416.
- [7] SYMEONIDIS P, RUXANDA M M, NANOPOULOS A, *et al.* Ternary semantic analysis of social tags for personalized music recommendation [C]// 9th International Conference on Music Information Retrieval. Philadelphia, USA: DBLP, 2008: 219-224.
- [8] HAMEL P, ECK D. Learning features from music audio with deep belief networks [C]// Proceedings of the 11th International Society for Music Information Retrieval Conference. Utrecht, Netherlands: DBLP, 2010: 339-344.
- [9] NG A, JORDAN M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes [C]// Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2002: 841-848.
- [10] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4): 212-223.
- [11] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, *et al.* Dropout: A simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
- [12] HINTON G E. Training products of experts by minimizing contrastive divergence [J]. Neural Computation, 2002, 14(8): 1771-1800.
- [13] MANDEL M I, ECK D, BENGIO Y. Learning tags that vary within a song [C]// Proceedings of the 11th International Society for Music Information Retrieval Conference. Utrecht, Netherlands: DBLP, 2010: 399-404.
- [14] MANDEL M I, ELLIS D P W. A web-based game for collecting music metadata [J]. Journal of New Music Research, 2008, 37(2): 151-165.
- [15] WELLING M, ROSEN-ZVI M, HINTON G E. Exponential family harmoniums with an application to information retrieval [C]// Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2004: 1481-1488.
- [16] MANDEL M, PASCANU R, LAROCHELLE H. Autotagging music with conditional restricted Boltzmann machines [C]// Asian Conference on Information and Database Systems. USA: IEEE, 2012: 284-293.
- [17] CORTES C, MOHRI M. AUC optimization vs. error rate minimization [J]. Advances in Neural Information Processing Systems, 2004, 16(16): 313-320.