# 基于主动学习的环境音分类研究

张 雁[1],吕丹桔[1],王红崧[2]
(1. 西南林业大学 计算机与信息学院,云南 昆明 650224;
2. 西南林业大学 生态旅游学院,云南 昆明 650224)

**摘 要**:环境音分类是当前语音识别领域的研究热点。主动学习是利用未标记数据,在少量标记数据代价下提高监督学习算法的分类性能的方法。文中提出了熵优先采样(Entropy Priority Sampling,EPS)方法和简单不一致采样(Simple Disagreement Sampling,SDS)方法作为主动学习选择样本的策略。针对环境音数据,提取 11 维的 CELP 音频特征,采用单一分类器与 EPS,SDS 方法对不同标记训练样本比例下的分类实验结果进行了比较分析。结果表明,主动学习方法在标记样本数较少的情况下,能取得较好的分类效果,并且 EPS 方法的性能优于 SDS 方法。

**关键词**:主动学习;环境音分类;采样;熵优先采样;简单不一致采样
**中图分类号**:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)06-0110-04
doi:10.3969/j.issn.1673-629X.2014.06.028

# Research on Environmental Audio Classification Based on Active Learning

ZHANG Yan[1],LÜ Dan-ju[1],WANG Hong-song[2]
(1. School of Computer and Information,Southwest Forestry University,Kunming 650224,China;
2. School of Ecological Tourism,Southwest Forestry University,Kunming 650224,China)

**Abstract**:Environmental audio classification has been the focus in the field of speech recognition. Active learning enhances the performance of supervised learning classification under the case of few labeled data. Propose EPS (Entropy Priority Sampling) and SDS (Simple Disagreement Sampling) methods as the selecting sampling strategies in active learning. For the given environmental audio data, the CELP features in 11 dimensions are extracted. The experiments with the single classifier, EPS and SDS on the environmental audio are carried out in order to illustrate the results of the proposed methods and compare their performance under different percent training sample. The experimental results show that active learning can effectively improve the performance of environmental audio data classification, even under the fewer number of the training examples. The EPS method outperforms the SDS.

**Key words**:active learning;environmental audio classification;sampling;EPS;SDS

## 0 引 言

音频分类是提取音频结构和内容的途径之一,也是音频检索和分析的基础[1]。环境音的分类越来越吸引广大的研究者[2]。已有的音频分类的技术有最小距离分类器、神经网络、支持向量机[3]、决策树和隐马尔可夫模型[4]等。分类器的最优化是与具体分类问题密切相关的,要找一个普适的最优分类器是分类研究的一个难点。

在实际应用中,环境音数据的分类需要大量的训练数据,获取这些标记数据需要大量的人力和物力,而且还很费时。这就要求分类的方法在少量的标记样本和大量的未标记数据情况下,获得较高的分类正确率[5-6]。传统的监督分类(即被动学习)构建正确率满足要求的分类器将十分困难。因此,主动学习方法[7-8]被提出以有效地处理这类问题。在主动学习中,学习器主动地选择包含信息量大的未标记样本将其交由专家进行标记,作为新增加的标记样本成为训练数据,以保证在训练集比较小的情况下,不断补充训练集,获得较高的分类正确率,有效地降低建立高性能分类器的代价[9]。

**收稿日期**:2013-08-22 **修回日期**:2013-11-28 **网络出版时间**:2014-02-24
**基金项目**:云南省教育科研基金(2012C098)
**作者简介**:张 雁(1973-),女,副教授,CCF 会员,研究方向为机器学习。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0915.037.html

## 1　主动学习

主动学习可以自主选择对学习过程最有用的未标记样本来请求用户标记,并将这些样本加入到已有的训练数据集中,使用分类器在此选择未标记样本,这样不断地选择扩大标记样本训练集合,能够最大程度地提高监督学习对未标记样本分类的准确性。图 1 描述了主动学习的基本流程。从已标记样本中建立训练样本,训练监督分类器,在每次学习过程中,分类器主动根据采样策略从未标记数据中选择最富有信息的样本,提交专家(Oracle)标记后,补充到训练集中进行下一次迭代。

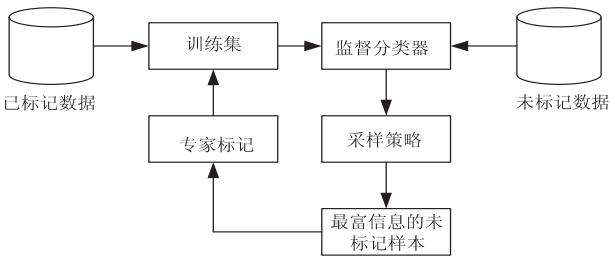


图 1　主动学习的基本流程

主动学习算法是一个迭代的过程,分类器使用迭代时反馈的样本进行训练,不断提升分类效率。目前,采样策略是主动学习中的核心,也是区分不同主动学习算法的关键[10]。根据不同的采样方法,常用于分类问题的主动学习算法主要有三种形式:

(1)基于委员会的查询方法(QBC)[11];

(2)基于边缘的查询方法(MS)[12];

(3)基于后验概率的查询方法(PP)[13]。

基于委员会投票的主动学习方法首先根据已有的类标签数据建立两个或多个分类器,组成“委员会”,利用这个委员会对预测样本进行分类投票,然后选择投票最不一致的样本作为候选样本。这种方法能够选择对学习过程最为有用的样本,从这些样本中获得丰富的分类信息,加快学习过程,因此它能使用很少的已标记样本达到给定的分类精度,降低了人工标注的代价。

### 1.1　主动学习 EPS 方法

主动学习中最常用的采样技术是基于委员会的采样(Query By Committee,QBC)[11],基于委员会建立多样性的集成分类器,并应用于未标记数据。选择各个集成分类器最小置信度的数据作为信息量最大的样本,请专家进行标记后用于扩充训练数据集。

设基于 QBC 中的分类器有 $k(k\geqslant 3)$ 个,选择最富有信息的数据进行标记的选择策略是:每个未标记的样本点都被输入 $k$ 个分类器中,而使得各个分量分类器的判决结果最不一致的数据点作为最富有信息的样本。不一致性的度量采用每个分量分类器对分类结果的投票熵 Entropy$(x)$ 来确定。

$$\text{Entropy}\ (x) = -\sum_{i=1}^{C} p_i \log p_i \tag{1}$$

$$p_i = \frac{V(i)}{k} \tag{2}$$

式中,$V\ (i)$ 是类别 $i$ 的投票数;$C$ 是总的分类类别数。

优先选取投票熵值高的。这个数据采样的过程命名为 Entropy Priority Sampling (EPS)。算法描述如下:

输入:初始训练集 $L$,未标记数据集 $U$,学习算法 Learn$_k$($k\geqslant 3$);

输出:最终的分类器 $H_{\text{out}}$。

```
begin
Train k classifiers H_k on L with Learn_k(k≥3);
Repeat N times
L_a←ø;
for each x_i ∈ U
H_k (x_i) (k = 1,2,…);
计算 Entropy(x_i);
end for
L_a←{x|Entropy(x) is highest}
U←U-L_a;      //从 U 中移除 L_a 样本
Labeling(L_a);      //标记 L_a;
L ← L ∪ L_a ;
Train k classifiers H_k on L with Learn_k(k≥3);
end Repeat
H_out←Ensemble Method(H_k) (k=1,2,…)
end
```

### 1.2　主动学习 SDS 方法

EPS 方法中,参与选择样本的委员会分类器是 $k$ ($k\geqslant 3$),$k$ 越大,选择最富信息样本点需要的时间代价就越高。因此,可以简化为分类器 $k=2$,选择的策略为两个分类器预测结果不一致的点,也称为争议点(Contention Points,CPs),作为选择的对象。此采样方法命名为 SDS(Simple Disagreement Sampling)。SDS 的过程描述如下:

输入:初始训练集 $L$,未标记数据集 $U$,学习算法 Learn$_k$($k=2$);

输出:最终的分类器 $H_{\text{out}}$。

```
begin
for i=1 to 2
S_i←Bootstrap(L);
for t=1 to N      //N iterations
H_1←learn_1(S_1);H_2←learn_2(S_2);CPs←ø;
for each x_i ∈ U
CPs←CPsU {x_i|x_i ∈ U and H_1(x_i) ≠ H_2(x_i);
U ←U-{CPs} ;      //将争议点从 U 中移除
NewL←Label(CPs);    //标记争议点
for i=1 to 2
S_i←S_i || NewL;
```

end for

$H_{out}$←Ensemble Method($H_1$,$H_2$)

end

## 2 实验数据和方法

实验的数据是网络与实地采集的数据,抽样率 8 k,位比特 16 bit,单声道,共收集 5 类自然界环境音频数据(鸟类叫声、蛙叫声、风声、雨声和雷声),语音数据时长为 10 min。音频前期处理是将静音和无关噪声的滤除。

### 2.1 环境音特征提取

在 Matlab7.1 平台中将音频数据经 G.723.1 数据编码,形成比特流,接收端将比特流解包为帧比特信息,将每帧的 0 ~ 23 位($LPC_0$ ~ $LPC_2$)提取 LPC 的 10 阶系数,构成 10 维的参数特征。将第 24 ~ 30,32 ~ 38 位($ACL_0$,$ACL_2$)信息解码提取基音周期延迟,构成 1 维的参数特征,合成 11 维的 CELP(Code Excited Linear Predictive)音频特征。特征的提取流程如图 2 所示。

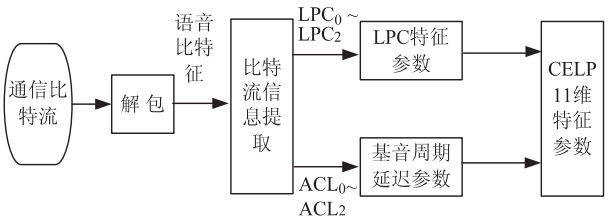


图 2 基于环境音编码流音频特征合成

### 2.2 实验的方法

实验的环境是基于 Weka[14] 的二次开发,通过 Matlab 程序将环境音待分类音频特征参数和训练特征参数转换为 CSV 格式的文件后再转换得到 ARFF 格式的文件。用 Weka 二次开发的分类模块对 ARFF 文件进行分类,结果为 ARFF 文件。单分类器采用了决策树 J48,朴素贝叶斯网络 NB 和 RBF(Radial Basis Function),主动学习采用了 SDS 和 EPS 算法,最终的分类器是通过最后投票决策(Ensemble Method)获得。图 3 为环境音分类的流程。

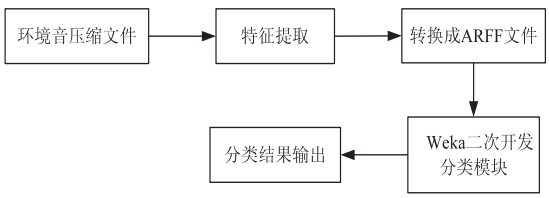


图 3 环境音分类的流程

## 3 实验结果与分析

实验以帧为单位作为统计数据依据,正确分类也以帧为单位进行统计。为了避免训练阶段由于数据量过大而产生溢出,实验中按每类数据量总帧数的约 1/3 进行的量抽样作为每个类别总的数据量,共抽样 10 次,每类数据按 75% 作为训练样本,25% 作为测试样本。最后统计 10 次分类的结果,取平均值作为最终分类识别结果。5 类的音频信号分帧情况和分类样本如表 1 所示。

表 1 实验所用的数据集

| 类别 | 总数据帧 | 每次抽样帧数 | training | test |
|---|---|---|---|---|
| 鸟叫 | 3 900 | 3 000 | 2 250 | 750 |
| 风声 | 6 434 | 2 000 | 1 500 | 500 |
| 雨声 | 6 200 | 2 000 | 1 500 | 500 |
| 蛙叫 | 3 334 | 1 200 | 900 | 300 |
| 雷声 | 1 634 | 1 000 | 750 | 250 |
| 合计 | | | 6 900 | 2 300 |

对每个数据集,25% 的数据作为测试数据,75% 的是训练集。在训练集中,标记数据各自取训练集的 10%,20%,40%,60% 和 80%。实验中比较了传统监督分类和主动学习方法,监督分类器采用决策树 J48,朴素贝叶斯 NB 和径向基函数,主动学习采用 EPS 和 SDS 方法。其中,EPS 算法中 $k=3$,使用三种不同的分类算法,即 J48 决策树,NB(Naïve Bayesian)和 RBF(Radial Basis Function)来选择最富有信息的样本点。主动学习的 SDS 方法,采用了决策树 J48 和 RBF 分类算法来选择样本。实验的环境是基于 Weka 的二次开发。每组标记比例的数据是随机选取的,实验结果取每种方法运行测试 10 次的平均值。表 2 显示的是在每种训练样本比例下,各类数据集分类的错误率。

表 2 不同比例训练样本的三类方法的平均分类错误率

| 训练样本 | Single-classifier | | | Active Learning | |
|---|---|---|---|---|---|
| | J48 | NB | RBF | EPS | SDS |
| 10% | 0.215 9 | 0.227 1 | 0.182 3 | 0.170 9 | 0.181 0 |
| 20% | 0.198 4 | 0.221 6 | 0.181 2 | 0.159 4 | 0.167 0 |
| 40% | 0.181 0 | 0.211 1 | 0.186 6 | 0.149 5 | 0.163 8 |
| 60% | 0.165 3 | 0.206 2 | 0.185 1 | 0.147 4 | 0.162 1 |
| 80% | 0.165 9 | 0.208 6 | 0.187 5 | 0.141 7 | 0.166 6 |

从实验结果数据可看出,主动学习方法比传统的监督分类如决策树 J48,朴素贝叶斯 NB 和径向基函数分类的正确率高。对于主动学习而言,EPS 性能明显优于 SDS,在 EPS 中,用了三个不同的学习算法来选择样本,而 SDS 是采用了两个不同的学习算法。在很少的训练样本下,10%,20% 和 40% 的标记数据时,SDS 相对单分类器 J48 的性能分别提高了 20.84%,19.68% 和 17.42%。而 SDS 相对单一分类器 J48 性能提高率分别是 16.19%,15.81% 和 9.49%。它们都充分利用了未标记样本的信息来降低对训练标记样本的数量要求,能更有效地提高几乎所有的标记比例的训

练数据的分类性能。

图 4 是环境音数据对应各方法分类性能的比较。总体来看,EPS 和 SDS 都比单一分类器性能好,而且曲线比较平滑和稳定(除了 SDS 在标记样本比例增大到 60% ~80% 外)。在标记数据占总训练样本数量 10% ~50% 时,主动学习能得到比较理想和稳定的分类状态,性能的提高比较明显。因为在学习的过程中,未标记数据可能会被误标记,以致在扩大的训练标记样本中引入了噪音数据,在下一次学习训练迭代时,包含了错误标记的训练数据,引入了训练误差。尤其 SDS 算法,虽然训练的时间减少了,但是当已标记样本比例增加时,性能反而有所降低。

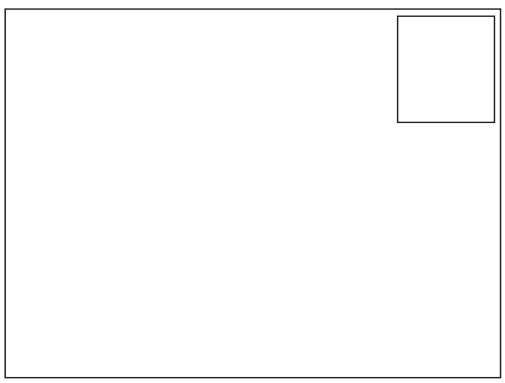

图 4　数据集各方法的性能

## 4　结束语

在训练标记数据较少,存在大量未标记数据的情况下,主动学习的分类效果明显优于传统的监督分类,因而提供了一条解决分类性能的有效途径。文中采用了 EPS 和 SDS 主动学习对实验中环境音数据进行分类取得了较好的预测效果。

如何提高选择样本的多样性,研究采样策略仍然是主动学习的研究热点。一方面将集成学习引入到主动学习中以保证选择样本的多样性;另一方面,采样策略算法应结合实际研究数据对象环境音的音频结构和提取的音频特征,更有效地选择具有高信息量的样本,减少计算的复杂度,提高分类的准确度和分类器的泛化能力。

**参考文献:**


[1] 白　亮,老松杨,陈剑赟,等. 基于支持向量机的音频分类与分割[J]. 计算机科学,2005,32(4):87-90.

[2] 张小梅,杨鼎才. 基于支持向量机模型的环境音分类研究[J]. 电子测量技术,2008,31(9):121-123.

[3] 余清清,李　应,李　勇. 基于 SVM 模型的自然环境声音的分类[J]. 计算机与数字工程,2010,38(7):1-5.

[4] 郑怡文. 典型的音频分类算法[J]. 计算机与现代化,2007(8):59-63.

[5] Zhou Zhihua, Wang Yu. Machine learning and application[M]. Beijing:Tsinghua University Press,2007.

[6] Seeger M. Learning with labeled and unlabeled data[R]. Edinburgh,UK:University of Edinburgh,2001.

[7] Zhu X. Semi-supervised learning literature survey[R]. Wisconsin:University of Wisconsin at Madison,WI,2008.

[8] 龙　军,殷建平,祝　恩,等. 主动学习研究综述[J]. 计算机研究与发展,2008,45(Sup):300-304.

[9] 刘　康,钱　旭,王自强. 主动学习算法综述[J]. 计算机工程与应用,2012,48(34):1-4.

[10] 吴伟宁,刘　扬,郭茂祖,等. 基于采样策略的主动学习算法研究进展[J]. 计算机研究与发展,2012,49(6):1162-1173.

[11] Seuong H S, Opper M, Sompolinski H. Query by committee[C]//Proceedings of the 5th annual workshop on computational learning theory. New York,NY,USA:ACM,1992:287-294.

[12] Campbell C, Cristianini N, Smola A J. Query learning with large margin classifiers[C]//Proc of ICML. San Francisco, CA,USA:Morgan Kaufmann Publishers Inc,2000:111-118.

[13] Roy N, McCallum A. Toward optimal active learning through sampling estimation of error reduction[C]//Proc of ICML. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001:441-448.

[14] Witten I H, Frank E, Hall M A. Data mining: practical machine learning tools and techniques[M]. 3rd ed. San Francisco,CA,USA:Morgan Kaufmann Publishers Inc,2011.


(上接第 109 页)


始化方法[J]. 微电子学与计算机,2013,30(6):80-83.

[9] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of ACM,1975,18(11):613-620.

[10] 林春实,方　燕,全吉成. 汉语文献自动分词与标引技术发展浅析[J]. 情报学报,1997(S1):45-49.

[11] 化柏林. 知识抽取中的停用词处理技术[J]. 现代图书情报技术,2007(8):48-51.

[12] 李　健. 聚类分析及其在文本挖掘中的应用[D]. 西安:西安电子科技大学,2005.

[13] 李永森,杨善林,马溪骏,等. 空间聚类算法中的 K 值优化问题研究[J]. 系统仿真学报,2006,18(3):573-576.

[14] Shekhar S, Chawla S. 空间数据库[M]. 谢昆青,译. 北京:机械工业出版社,2004.

# 基于主动学习的环境音分类研究



作者： 张雁， 吕丹桔， 王红崧， ZHANG Yan， L Dan-ju， WANG Hong-song
作者单位： 张雁,吕丹桔,ZHANG Yan,L Dan-ju(西南林业大学 计算机与信息学院,云南 昆明,650224)
， 王红崧,WANG Hong-song(西南林业大学 生态旅游学院,云南 昆明,650224)

刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(6)