

文章编号:1001-9081(2005)12-2792-03

## 基于 MBIC 的决策树聚类算法在连续语音识别中的应用

陈国平<sup>1,2</sup>, 杜利民<sup>2</sup>, 付跃文<sup>3</sup>, 王劲林<sup>1,2</sup>

(1. 中国科学院 声学研究所, 北京 100080; 2. 中国科学院 研究生院, 北京 100080;

3. 南京工业大学 信息科学与工程学院, 江苏 南京 210009)

(chengguoping97@tsinghua.org.cn)

**摘 要:**提出了一种采用最小贝叶斯信息准则(Minimum Bayesian Information Criterion, MBIC)来最优化控制决策树结点分裂程度的算法。首先在理论上证明了 MBIC 能够较好地解决模型参数复杂度与训练数据集规模之间的权衡问题,然后给出了基于 MBIC 的决策树分裂停止准则的计算公式。汉语连续语音全音节识别实验表明:与传统的最大似然准则(Maximum Likelihood Criterion, MLC)相比,MBIC 对声学模型参数和训练数据集的变化具有更好的适应能力。

**关键词:**连续语音识别;决策树聚类;最小贝叶斯信息准则;分裂停止准则

**中图分类号:** TP391.42 **文献标识码:** A

## Clustering algorithm based on the MBIC decision-tree for CSR

CHEN Guo-ping<sup>1,2</sup>, DU Li-min<sup>1,2</sup>, FU Yue-wen<sup>3</sup>, WANG Jin-lin<sup>1,2</sup>

(1. Speech Interaction Technology Research, Institute of Acoustic, CAS, Beijing 100080, China;

2. Graduate School of Chinese Academy Sciences, Beijing 100080, China;

3. College of Information Science and Engineering, Nanjing University of Technology, Nanjing Jiangsu 210009, China)

**Abstract:** an algorithm based on Minimum Bayesian Information Criterion(MBIC) was proposed to help optimize the node-splitting degree in a decision tree. First, it was proved in theory that MBIC can find a good balance between the complexity of model parameters and the scale of the training sets. Then, a formula was proposed to describe MBIC decision tree splitting and stopping criterion. Finally, the experiment on Chinese all-syllable recognition shows that MBIC has much better adaptive ability to variable acoustic model parameters and training sets than the classical Maximum Likelihood Criterion method.

**Key words:** Continuous Speech Recognition(CSR); clustering based on decision-tree; Minimum Bayesian Information Criterion(MBIC); splitting and stopping criterion

## 0 引言

近来主流连续语音识别系统都采用连续密度的 HMM 模型和上下文相关的声学模型对语音数据进行建模。在连续语音中,协同发音现象十分严重,采用上下文相关单元是很有必要的。在实际情况下,由于上下文单元数目通常非常庞大,训练数据就会显得相对不足,一般会有有一半以上的上下文单元没有对应的训练数据,通过共享不同模型状态可以有效地解决数据稀疏问题。

模型状态共享策略大致可以分为两类:一类是基于数据驱动的,另一类是基于决策树的。基于决策树的状态共享可以得到与基于数据驱动相似的聚类性能,此外这种聚类方法还为训练数据集中没有包含但实际语流中又可能会出现语音单元提供一个较为可靠的参数估计。

基于最大似然准则(Maximum Likelihood Criterion, MLC)的决策树状态共享<sup>[1]</sup>已在连续语音识别的模型状态共享中得到了广泛应用,但 MLC 本身并不能有效地控制决策树结点的分裂程度。在大部分情况下,随着分裂数目增多,其似然值几乎一直在增大,最后的叶结点数目通常和参与共享的状态数目一样多,无法解决数据稀疏问题。通过人工选取适当的

阈值,可以有效地控制结点的分裂程度,从而改善决策树的聚类性能。然而最优的阈值会随着声学模型和训练数据集的改变而改变。本文提出的最小贝叶斯信息准则(Minimum Bayesian Information Criterion, MBIC)可以在模型复杂度和训练数据规模之间找到一个合理的平衡,从而最优化地控制决策树结点的分裂程度。

## 1 决策树状态共享策略

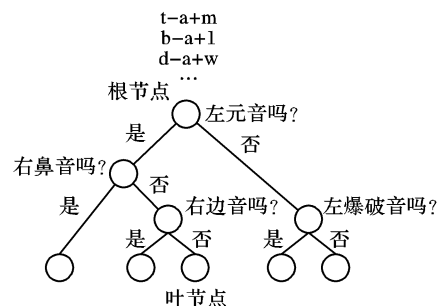


图1 决策树的结构

如图1,基于决策树的状态共享是一种自顶向下的聚类过程。假设上下文相关模型的同一个中心基元的同一个状态

收稿日期:2005-06-22;修订日期:2005-08-30

**作者简介:**陈国平(1979-),男,江苏宜兴人,博士研究生,主要研究方向:语音识别、语音合成;杜利民(1957-),男,四川人,研究员,博士生导师,主要研究方向:语音信号与信息处理技术;付跃文(1968-),男,山西孝义人,博士,主要研究方向:信号处理与模式识别;王劲林(1964-),男,北京人,研究员,主要研究方向:多媒体通信。

汉语中,标准的声韵母共有 59 个,其中声母 21 个,韵母 38 个。除了零声母外,汉语的每个音节都是由声母和韵母构成。根据汉语的这个特点,本文新增加 6 个零声母{aa, ee, ii, oo, uu, vv}的定义,构成包含 65 个基本单元的汉语声韵母扩

展集(见表1),在保证每个音节的声韵母结构的一致性的同时,还有如下优点:

1)当考虑上下文声韵母模型时,由声韵母扩展集得到的上下文基元数目不到3万,而由标准声韵母得到的上下文基元数目将超过10万。

2)引入零声母可以大大减少识别中零声母音节的插入错误。

表1 汉语声韵母扩展集

声母基元(27个)	韵母基元(38个)
b, p, m, f, d, t, n, l, g,	a, ai, an, ang, ao, e, ei, en, eng,
k, h, j, q, x, zh, ch, sh,	er, o, ong, ou, i, il, i2, ia, ian,
z, c, s, r, aa, ee, ii, oo,	iang, iao, ie, in, ing, iong, iou, u,
uu, vv	ua, uai, uan, uang, uei, uen, ueng,
	uo, v, van, ve, vn

本文使用从左向右无跳转的连续单高斯 HMM 模型来建模,对于声母和静音基元使用3个状态描述,对于韵母基元使用5个状态描述。

训练时先利用K均值模型参数估计算法和 Baum-Welch 模型参数估计算法对上下文无关的声韵母模型进行训练(最大迭代次数为20次,收敛域值为0.0001),接着将上下文无关的声韵母模型转换成上下文相关的声韵母模型(TriIF),接着利用嵌入式 Baum-Welch 算法对 TriIF 模型进行参数估计训练(3次迭代),接着利用决策树聚类将 TriIF 模型进行聚类得到共享的上下文相关的声韵母模型(Tied-TriIF),最后利用嵌入式 Baum-Welch 算法对 Tied-TriIF 模型进行参数估计训练(2次迭代)。实验利用 HTK 工具 V3.2.1<sup>[3]</sup>进行模型训练和音节识别,增加了 MBIC 计算的代码。

### 3.2 问题集的设计

决策树是一个二叉分类树,每个结点的分裂都是在一系列回答“Yes/No”的问题下进行的。这些问题就构成了决策树的问题集。问题集的好坏会直接影响到上下文模型状态的聚类性能。本文使用的问题集是基于语言和语音学知识的<sup>[4,5]</sup>。根据这些先验知识,中心基元的上下文(即中心基元左右两边相邻的基元)被划分成若干类,每一类作为一个问题。本文设计问题集是针对声韵母基元的。现举例如下:

作为问题的声母基元类有:

双唇音: {b, p, m} 舌根音: {g, k, h}

作为问题的韵母基元有:

麻韵: {a, ia, ua} 前元音: {i, v}

### 3.3 实验条件及方法

实验中使用的数据是“863 汉语语音数据库”中的女声数据,数据库共有2573个不同的句子,划分成A、B、C、D四组。实验时将A、B、C组作为训练集,D组作为测试集。

实验中,特征参数使用 MFCC(包括静态参数、一阶差分和二阶差分参数以及  $C_0 + \Delta C_0 + \Delta^2 C_0$ ,本文以下用  $S$  表示静态参数及  $C_0$ ,  $\Delta$  表示一阶差分系数及  $\Delta C_0$ ,  $\Delta^2$  表示二阶差分系数及  $\Delta^2 C_0$ );识别网络为困惑度为408的全音节网络。决策树状态聚类的性能用连续语音识别的音节准确率来进行评价。

### 3.4 实验结果及分析

实验分为两部分进行。

第一部分实验:固定声学模型的特征维数(39维)不变,改变训练数据集的规模(60人和30人)。目的是考察 MBIC 对训练数据集规模变化的适应能力(见表2和表3)。

训练集60人时,MLC下的最优阈值为1300,最优的音节识别率为72.60%;训练集30人时,MLC下的最优阈值变为

1100,最优的音节识别率为73.16%。当训练数据集规模下降50%时,最优阈值下降了15%,这与式(9)描述的与对数占有数成正比的关系基本相符。训练集30人时,MLC下阈值为1300的音节识别率为72.89%,而 MBIC 下音节识别率为73.12%;训练集60人时,MLC下阈值为1100的音节识别率为72.21%,而 MBIC 下音节识别率为72.56%;所以与 MLC 相比,MBIC 对训练数据集规模变化具有更好的适应能力。

第二部分实验:固定训练数据集的规模(60人)不变,改变特征参数的维数(39维和26维)。目的是考察 MBIC 对特征参数变化的适应能力(见表2和表4)。

表2 训练集60人,39维( $S + \Delta + \Delta^2$ )

准则	阈值	聚类后状态数目	音节准确率(%)
MBIC	-	4087	72.56
	200	26283	70.59
	1000	6112	71.98
	1100	5403	72.21
MLC	1200	5004	72.42
	1300	4860	72.60
	1400	4606	72.47
	3000	2697	71.84

表3 训练集30人,39维( $S + \Delta + \Delta^2$ )

准则	阈值	聚类后状态数目	音节准确率(%)
MBIC	-	3412	73.12
	200	20081	71.38
	900	4791	72.97
	1000	4369	73.10
MLC	1100	4057	73.16
	1200	3801	72.99
	1300	3591	72.89
	3000	2034	71.73

表4 训练集60人,26维( $S + \Delta$ )

准则	阈值	聚类后状态数目	音节准确率
MBIC	-	5226	69.41
	200	19222	67.63
	800	5407	69.33
	900	4931	69.44
MLC	1000	4529	69.39
	1100	4231	69.35
	1200	3913	69.21
	1300	3751	69.07
	3000	2119	68.29

特征参数维数39时,MLC下的最优阈值为1300,最优得音节识别率为72.60%;特征参数维数26时,MLC下的最优阈值变为1000,最优的音节识别率为69.44%。当特征维数下降33%,阈值下降了30%,这与式(9)描述的阈值大小与特征维数成正比的关系基本相符。特征参数维数26时,MLC下阈值为1300的音节识别率为69.07%,而 MBIC 下音节识别率为69.41%;特征参数维数39时,MLC下阈值为1000的音节识别率为71.98%,而 MBIC 下音节识别率为72.56%;所以与 MLC 相比,MBIC 对特征参数变化具有更好的适应能力。

## 4 结语

分裂停止准则对基于 MLC 决策树的聚类性能有重大影响。经典的分裂停止准则通常依赖于训练数据集规模和声学

文章编号:1001-9081(2005)12-2795-03

## 在线废料建模在特定领域语音识别中的应用

辛璐璐, 谢莎莎, 孙甲松, 王作英  
(清华大学 电子工程系, 北京 100084)  
(xll03@mails.tsinghua.edu.cn)

**摘要:**严格按照语法规则模型指导声学层识别的特定领域语音识别系统,难以处理未经规则描述的插入语或语气词等语言现象。针对这一问题,将在线废料建模方法应用于该系统,详细讨论了此方法中模型参数  $N$  的选择策略,分析验证了语料的信噪比  $SNR$  值与参数  $N$  之间的相关性,提出了基于此相关性的模型参数优化方法,使得系统的句子识别率和槽识别率相对基线系统分别提高了 18.34% 和 11.47%。

**关键词:**语音识别;废料;在线废料建模;词图搜索

**中图分类号:** TP391.42 **文献标识码:** A

## Application of on-line filler modeling in specific domain speech recognition system

XING Lu-lu, XIE Sha-sha, SUN Jia-song, WANG Zuo-ying  
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

**Abstract:** It is difficult for a speech recognition system in specific domain to deal with parentheses or sentence particles which is not depicted in rule since it instructs acoustic layer recognition in strict syntax rules model. To solve this problem, an on-line filler modeling method based on wordgraph decoder was applied to specific domain speech recognition system. The choice of the parameter  $N$  in this method was discussed in detail, and the correlation between materials'  $SNR$  and the parameter  $N$  was also analyzed and validated. Moreover, a model parameter optimization method, which benefits from the correlation, was proposed. Computer simulation validates that the proposed method increases recognition rate by 18.34% for sentence and 11.47% for slot relatively.

**Key words:** speech recognition; filler; on-line filler modeling; wordgraph decoder

## 0 引言

语音识别是人与机器进行口语交流的关键技术,在智能家电、智能办公以及机器人技术等方面有着广阔的应用前景。而语言模型是高性能语音识别系统不可或缺的重要组成部分,它包括统计模型和规则模型两种。目前常用的统计模型为  $N$ -gram 模型;而规则模型则是指传统的规则文法。对于特定领域的语音识别任务,如果采用传统的  $N$ -gram 统计语言模型,由于很难获得足够的文本语料,常出现严重的数据稀疏现象,往往不能达到令人满意的效果;与之相对,由于特定领域的语言规则比较简单,语法和句型相对统一,因此一般采用规则模型效果会更好。基于图搜索的特定领域的语音识别系统<sup>[1]</sup>即直接使用规则模型指导声学层识别。

但是在日常生活中,人们说话时常含有如“请问”,“是

吗”这类的插入语或语气词,它们不表示实际意义,不影响整句话的语义理解,通常称之为废料。语法规则难于表述这一现象,所以直接按照语法规则进行识别搜索的系统,对于这种含有废料的情况难以得到正确的识别结果,这将导致整个系统性能的降低。

为了处理这种情况,多采用离线废料建模的方法<sup>[2]</sup>,它预先对废料建立模型。在数据量较小时,离线废料模型能比较精细地刻画废料特性。但是,由于废料非常广泛,使得模型的设计和训练相当困难,不同应用环境下的模型需要重新训练。同时,Hever 等人<sup>[3]</sup>也提出了在线废料建模的方法,此算法简单,易于实现。扩展的模板匹配方法<sup>[4]</sup>把在线废料建模方法应用到模板匹配中,增强了系统的稳健性,取得了较好的识别效果。

本文将在线废料建模方法应用于特定领域的语音识别,

收稿日期:2005-06-24 基金项目:国家 863 计划资助项目(2001AA114071)

作者简介:辛璐璐(1981-),女,四川乐山人,硕士研究生,主要研究方向:特定领域中的语音识别; 谢莎莎(1982-),女,山西运城人,硕士研究生,主要研究方向:语音识别中的搜索和剪枝算法; 孙甲松(1962-),男,山东潍坊人,副教授,主要研究方向:语言模型,中文信息处理; 王作英(1935-),男,江西赣县人,教授,博士生导师,主要研究方向:非特定人连续语音识别。

特征参数。本文提出的基于 MBIC 的决策树聚类方法,对训练数据集规模和声学特征参数维数的变化则具有较好的适应能力,该方法可以应用于声学模型的自适应训练。另外,该方法对声学特征参数的类型变化也有较好的适应能力。

## 参考文献:

- [1] YOUNG S J, ODELL J J, WOODLAND P C. Tree-Based State Tying for High Accuracy Acoustic Modelling[A]. Proceedings ARPA Workshop Human Language Technology[C]. Berlin, 1994. 286-291.

- [2] 龚光鲁, 钱敏平. 应用随机过程教程及在算法和智能计算中的随机模型[M]. 北京: 清华大学出版社, 2004.  
[3] YOUNG S, EVERMANN G, HAIN T, et al. The HTK Book (for HTK Version 3.2)[M]. Cambridge University, 2002.  
[4] 吴宗济, 林茂灿, 等. 实验语音学概要[M]. 北京: 高等教育出版社, 1989.  
[5] 黄伯荣, 廖序东. 现代汉语(增订三版). 上册[M]. 北京: 高等教育出版社, 2002.