基于 CART 技术的汉语韵律短语分析^{*}

A CART-based Prosodic Phrasing Method for Chinese Text-to-Speech

谯卫军 林福宗 李建民 张 钺

(清华大学计算机科学与技术系 智能技术与系统国家重点实验室 北京100084)

Abstract The prediction of prosodic phrase boundaries is an important step for a TTS system. This paper proposes a new method of phrasing which uses Classification and Regression Trees (CART) techniques to generate decision tree automatically. First, the principle of CART is introduced and two improvements are proposed. Second we collect 559 sentences (of approximately 78 min length) and built a speech corpus. The prosodic boundaries are manually marked and word identification, POS tagging and syntactic analysis are also done on the text. And 22 features for modeling prosody are proposed. Lastly, a decision tree is trained on 371 sentences, and tested on 188. The success rate is 92.9%.

Keywords Text-to-speech, Prosodic phrasing, CART

韵律短语分析是文语转换(Text-to-Speech, TTS)系统中的一项重要内容。它负责在输入文本的不同位置上,标记上不同级别的符号形式的韵律短语边界,从而把一个句子分隔成不同层次的短语。韵律短语分析的结果直接影响着 TTS 系统的后继步骤,如韵律参数的生成。杨玉芳^[1]研究了不同级别的韵律短语边界与附近音节的时长、停顿、基频等韵律参数之间的关系,考察了这些参数随边界等级的系统变化。此外,在口语理解等应用系统中,韵律短语分析也有可用之处。因为口语理解的一个难点在于自然语言存在着歧义现象:即便是完全相同的一个句子,如果在表述时断句的位置不同,就可能产生完全不同的含义,所以有的研究者就利用句子的韵律边界信息来排除歧义^[2]。

传统的韵律短语分析方法主要是规则法,也就是通过研究句子的语法结构与韵律结构之间的关系,手工制定一些启发式规则来预测韵律边界的位置和级别^[3,4]。这种方法具有较好的解释性,但缺点在于:规则的制定、修改和维护比较困难,而且难以适应新的语言和领域。因此,近些年来的研究主要集中在基于统计的方法,也就是说,首先构造一个带有韵律短语标记的语料库,并用它来训练各种统计模型,然后用训练好的模型来预测韵律短语边界。常用的模型包括 Markov 模型^[5], 神经网络^[6]和分类与回归树 CART^[7~9]。它们克服了规则法的缺点,而且正确率较高。尤其是 CART 方法,它以决策树作为输出结果,解释性较好,因此近年来受到较大的关注。

对于汉语 TTS 系统而言,由于韵律短语分析问题与句子的语法信息密切相关,而汉语的复杂性又使得精确的语法分析难以实现,因此现有的系统大都回避了这个问题,或者只是在各种标点符号处插入不同长短的停顿。已有的一些工作包括:Chou^[10]利用词性(Part-of-Speech, POS)组概率来预测韵律短语边界,统计了构成韵律短语的各种 POS 组模式及其概率。谯卫军^[11]利用二元 POS 组概率和短语长度概率来预测韵律边界。

1 CART 技术

1.1 基本原理

CART 技术是 Breiman 等人提出的一种基于二叉树的统计模型^[12],它能根据训练样本的特征变量和预测变量的取值情况,自动地生成决策树或回归树。CART 的一个优点在于预测过程的直观性,对于生成的树,人们可以通过分析各个非终结点上所用的问题,来评估各个特征变量所起的作用,从而对需要解决的模式分类或回归估计问题有一个更深入的了解。就韵律短语分析而言,由于短语边界的级别个数是有限的,因而也可以把它看成是一个模式分类问题。使用 CART 方法来构造一个树形分类器的算法流程总结如下:

不妨假设训练样本集 $L = \{(x_n, y_n), n = 1, \dots, N\}$, 其中 x_n 是第 n 个训练样本的特征向量, y_n 是类别标记, N 是训练样本的个数。特征向量 x 中的每一个成分 x_i 都是实数型变量或有穷离散符号型变量。

第一步,创建根结点 t_1 ,它包含所有训练样本。

第二步,根据每个特征变量 x_i 的类型,定义一组二值型问题。对于实数型变量,问题形如“ $x_i \leq c$?”;对于有穷离散符号型变量,假设它的定义域为 S ,问题形如“ $x_i \in S_1$?”,其中 S_1 是 S 的子集。然后,对每个问题定义一个分裂操作,即从根结点 t_1 生长出两个儿子结点,并对 t_1 中的每个样本,如果它对问题的回答为“是”,则把它复制到左结点;如果回答“否”,则把它复制到右结点。

第三步,根据一个评判准则,从这组分裂操作中选择最优的一个,对 t_1 进行分裂。就分类问题而言,其目标是尽可能减少预测误差,由此提出了杂度函数(impurity function)的概念,用以描述一个结点内各个类的概率分布情况。常用的如 Gini 函数: $\Phi(t) = \sum_{i,j} p(i/t)p(j/t)$, 其中 $p(i/t)$ 表示结点 t 当中类别为 i 的概率。假设一个分裂操作把 t 分裂为 t_L 和 t_R ,那么定义杂度的下降值为: $\Delta\Phi = \Phi(t) - G_L * \Phi(t_L) - G_R * \Phi(t_R)$, 其中 G_L 和 G_R 是左右结点的样本所占的比例。因此,所谓的最优分裂就是使 $\Delta\Phi$ 最大的那个分裂操作。

第四步,使用贪心算法,对新生成的子结点重复上述步骤,不断进行分裂,直到每个结点中的所有样本都属于同一类或特征向量完全相同。这样便得到了一棵过大的决策树 T_{max} 。

^{*} 国家自然科学基金项目(69823001)和973项目(G1998030509),谯卫军 博士生,研究方向为文语转换,林福宗 副教授,研究方向为多媒体,李建民 博士生,研究方向为文语转换,张 钺 教授,博士生导师,研究领域为人工智能理论,神经网络、遗传算法理论等。

然后使用 cost-complexity 算法对 $T_{\max}$ 进行剪枝, 得到它的一系列子树 $T_0 = T_{\max}, T_1, \dots, T_k, \dots, T_K = \{t_i\}$ 。接下来的问题就是从它们当中选择一棵最优子树, 使得对未知的数据具有最小的预测误差。这通常有两种方法, 一是使用一个独立的测试样本集, 逐个计算每棵子树的测试误差 $R^v(T_k)$, 然后从中选择误差最小的。二是使用交叉验证 (cross-validation) 算法, 对同一个样本集进行训练和测试。如图1所示。

$T_0 = T_{\max}$	$T_0^{(1)} = T_{\max}^{(1)}$	$T_0^{(2)} = T_{\max}^{(2)}$	...	$T_0^{(V)} = T_{\max}^{(V)}$
T_1	$T_1^{(1)}$	$T_1^{(2)}$	...	$T_1^{(V)}$
...	...	...	...	...
T_k	$T_k^{(1)}$	$T_k^{(2)}$	...	$T_k^{(V)}$
...	...	...	...	...
T_K	$T_K^{(1)}$	$T_K^{(2)}$	...	$T_K^{(V)}$

图1 交叉验证算法

算法把样本集 L 随机分为 V ($V=10$) 个大小相同的子集 $L_v, v=1, \dots, V$, 并分别用 $L-L_v$ 作为训练集, 得到 V 棵辅助树: $T_{\max}^{(1)}, T_{\max}^{(2)}, \dots, T_{\max}^{(V)}$, 然后同样对它们进行剪枝。算法的思路就是当 V 比较大时, 可以把 $T_k^{(1)}, T_k^{(2)}, \dots, T_k^{(V)}$ 都近似地看成 T_k , 而由于 L_v 未曾用于 $T_k^{(v)}$ 的生成, 因此可以用 L_v 去测试 $T_k^{(v)}, v=1, 2, \dots, V$, 并用它们的误差之和来作为 T_k 的交叉验证误差 $R^v(T_k)$ 。然后, 从 T_k 当中选择一棵最优子树。既可以选择误差最小的, 也可以在误差与树的规模之间做一个折衷。

第五步, 对生成的决策树的每个叶子结点, 根据多数原则标上相应的类别标记。在测试新样本时, 从根结点开始, 根据每个非终结点上的问题来决定向左还是向右, 最终落到某个叶子结点, 然后用它的类别标记来作为预测的结果。

1.2 改进方法之一

常的决策树方法生成的是多叉树, 允许多个儿子结点, 每个结点代表一个属性值, 而 CART 生成的是二叉树, 只允许两个儿子结点。这样做的优点是减少结点的个数, 提高预测的精度。但问题是: 假设 x_q 是离散的符号变量, 它的属性值个数为 M , 那么由 x_q 生成的问题个数为 $2^{M-1}-1$, 当 M 较大时, 要想从中选择一个最优分裂, 在计算量上是不可实现的。因此, 针对二分类问题, 这里提出一种在多项式时间内寻找最优分裂的方法。

假设当前结点中的样本为 $L = \{(x_n, y_n), n=1, \dots, N\}$, 其中 $x_n = (x_{n1}, x_{n2}, \dots, x_{nT}), T$ 表示特征变量的个数, $y_n \in \{0, 1\}$ 。不失一般性, 这里只考虑第 q 个特征变量 x_q 的情形, 不妨令 $x_q \in S = \{s_1, s_2, \dots, s_M\}$ 。此外, 对于 $i=1, 2, \dots, M$, 定义 $p_i = P(y=1 | x_q = s_i)$ 。那么把 p_i 按照从小到大的顺序进行排序, 得到 P'_i , 其相应的属性值为 s'_i 。然后定义 $M-1$ 个 S 的子集 $S_1 = \{s'_1\}, S_2 = \{s'_1, s'_2\}, \dots, S_{M-1} = \{s'_1, \dots, s'_{M-1}\}$, 显然每个子集都对应一个分裂操作。假设 S_b 是其中的最优分裂, 那么可以证明它也是特征变量 x_q 的最优分裂。Breiman 曾经证明过相似的问题^[12], 但过程较为冗杂, 这里对其简化如下。

首先, 对于 $s \in S$, 定义 $g_s = P(x_q = s), p_s = P(y=1 | x_q = s)$ 。假设一个分裂操作把 L 划分为 L_L 和 L_R 两部分, 其对应的 x_q 的属性值集合为 S_L 和 S_R , 那么对于 $s \in S_L$, 定义 $G_L = \sum g_s, P_L = (1/G_L) \sum (p_s g_s)$; 同样定义 G_R 和 P_R , 显然 $G_L + G_R = 1$ 。令 $\Psi(L_L) = G_L \Phi(P_L) + G_R \Phi(P_R)$, 其中 $\Phi(p) = p(1-p)$, 显然有 $\Psi(L_L) = \Psi(L_R)$ 。再定义 $\Psi(L) = \Phi(G_L P_L + G_R P_R)$, 由于 Φ 是上凸函数, 因此有 $\Psi(L_L) = G_L \Phi(P_L) + G_R \Phi(P_R) \leq \Phi(G_L P_L + G_R P_R) = \Psi(L)$, 因此, 最优分裂就是使 $\Psi(L_L)$ 最小的分裂操作。

其次, 如果 L_L 是最优的, 那么可以假定 $\Psi(L_L) \neq \Psi(L)$ (否则任何一个分裂操作都是最优的)。同时假定 $P_L < P_R$ (若 $P_L > P_R$, 则把 L_L 和 L_R 交换; 若相等, 则有 $\Psi(L_L) = \Psi(L)$)。令 $y_L = \max\{p_i, s \in S_L\}, y_R = \min\{p_i, s \in S_R\}$ 。如果 $y_L < y_R$, 那么 S_L 即为某个 $S_i (i=1, \dots, M-1)$; 否则, 令 $L_{L1} = \{x \in L_L: x_q = s \text{ 且 } p_i < y_L\}, L_{L2} = \{x \in L_L: x_q = s \text{ 且 } p_i = y_L\}, L_{R1} = \{x \in L_R: x_q = s \text{ 且 } p_i > y_R\}, L_{R2} = \{x \in L_R: x_q = s \text{ 且 } p_i = y_R\}$, 那么根据 Breiman 的定理 9.6^[12], 必有:

$$\Psi(L_L) \geq \lambda \Psi(L_{L1}) + (1-\lambda) \Psi(L_{R2}) \quad (1)$$

其中 $\lambda = G_{R1} / (G_{L2} + G_{R1}), G_{L2} = |L_{L2}| / N, G_{R1} = |L_{R1}| / N$ 。这样, 若 L_L 最优, 则 $\Psi(L_{R2}) \geq \Psi(L_L)$, 代入(1)式, 可知 $\Psi(L_{L1}) \leq \Psi(L_L)$, 即 L_{L1} 也是最优的。然后, 令 $L_L = L_L - L_{L2}, L_R = L_R + L_{L2}$, 重复上述步骤, 直到 $y_L < y_R$ 。而此时 L_L 必不为空 (否则 $\Psi(L_L) = \Psi(L)$)。

显然, 改进后的时间复杂性仅为 $O(M \log M)$ 。

1.3 改进方法之二

CART 方法的另一个缺点是局部最优性, 每次选择的只是相对于当前结点的最优分裂。事实上, 从全局的角度来看, 这样选出的最优分裂并不一定最好。例如, 对于特征变量 x_q 而言, 如果属性值为 s 的样本在当前结点里只出现了一次, 那么 $p_i = 0\%$ 或 $p_i = 100\%$, 但这个概率显然不太可靠, 所以引入了一个惩罚因子 ϵ , 把 p_i 修正为 p'_i 。

$$\epsilon = 1 / (Q(Q-1) + 2) \quad \text{其中 } Q = N * g_i$$

$$p'_i = p_i + \text{sgn}(0.5 - p_i) * \epsilon$$

2 基于 CART 的韵律短语分析

韵律短语分析问题可以形式地定义为: 假设输入的句子由一序列的词语组成: $(w_1, \dots, w_i, w_{i+1}, \dots, w_N)$, 在任意两个相邻的词之间定义词边界 (w_i, w_{i+1}) , 并定义几种边界类型, 如 $\{0, 1\}$: 0 表示不是韵律边界, 1 表示是韵律边界。此外, 每个词边界还可以获得一些特征信息。韵律短语分析的任务就是对句子中的每一个词语边界, 根据它的特征信息来判断它的类型。

2.1 语料库

为了进行汉语的韵律短语分析, 我们首先从中央电视台的新闻联播节目收集了 559 个不同类型的句子 (约 78 分钟), 构造了一个语音语料库。考虑到句子中的标点符号等信息很难从语音中听出来, 而这些信息又可能对短语分析有影响, 我们从报纸、互联网等媒介找到了每个句子的文字稿, 以确保语音与文字的完全一致。其次, 我们对文字语料库进行了分词和词性标注处理, 并手工校正了其中的错误。这里用的词性标记以“八·五”《关于汉语语料库词性标记的规定》中的标记集为基础, 并做了一些修改以适应韵律短语分析的特点, 共有 72 个词性。然后, 手工标注了每个句子的语法信息, 包括主语、谓语、宾语、定语、状语和补语等句子成分信息和名词短语、介词短语等短语成分信息。最后, 通过听音实验和分析句子的波形与基频曲线, 标记了韵律短语边界和重音的位置和级别。

2.2 特征集

在分类器的设计中, 特征的选择是很重要的。我们首先采用了文[7, 8]中提出的 19 个特征变量, 包括 POS 信息、语法信息、时长信息和重音信息。

- Pos{1-4}: 词边界左右各两个词的词性。
- Sbs: 语法树中包含了 w_i 和 w_{i+1} 的最小语法成分。
- Sll: 包含 w_i 但不包含 w_{i+1} 的最大成分。

- Srl : 包含 w_{i+1} 但不包含 w_i 的最大成分。
- Ltw, Lic : 句子的长度(以词和字为单位)。
- Lsw, Lsc : 从句首到当前位置的距离。
- Lew, Lec : 从当前位置到句尾的距离。
- $Lllw, Lllc$: 语法成分 Sll 的大小。
- $Lrlw, Lrlc, Srl$ 的大小。
- $Acc(1-2)$: w_i 和 w_{i+1} 是否是重音。

此外,我们还提出了3个新的特征:

• $Gbipos$: 韵律边界在当前二元 POS 组中出现的概率。它的计算方法是:在训练集中统计每一个二元 POS 组出现的次数 N_1 和其中出现的边界个数 N_2 ,再相除^[11]。

• $Lrrw, Lrrc$: 在 Sbs 中,位于 Sll 右边的所有成分的大小。

3 实验结果

我们随机选择了371个句子(8170个词边界)来训练,然后用剩下的188个句子(4084个词边界)来测试。结果如表1所示。为了比较,同时列出了文[7~9]中有关英语、西班牙语和韩语的结果。

表1 几个 CART 应用的实验结果

	N	P_0 (%)	$ T_k $	$R^v(T_k)$ (%)	$R^u(T_k)$ (%)
Old-1	8170	77.6	37	7.8	7.9
Old-2	8170	77.6	37	7.5	7.6
New-1	8170	77.6	22	6.9	7.2
New-2	8170	77.6	24	6.8	7.1
Wang	3677	80.9	6	10	—
Hirsch	19473	83.7	14	5.8	—
Lee	2286	60.6	8	13.9	15.2

在表1中, N 表示训练样本的个数, P_0 表示非韵律边界在其中所占的比例,它常用来作为一个基准值。即假定所有的样本都预测为非韵律边界时的正确率。 $|T_k|$ 表示最优子树的叶子结点个数, $R^v(T_k)$ 表示训练集的交叉验证误差, $R^u(T_k)$ 表示测试误差。New 和 Old 表示新旧特征集,后面的数字表示方法1还是方法2。

从实验结果可以看出,在引入新的特征变量后,各方面的指标都有明显的提高,而方法2也略优于方法1。此外,Hirschberg^[8]的结果最好,交叉验证的正确率为94.2%。不过值得指出的是:(1)可能由于不同语言的原因,该训练集的基准值也最大。(2)与 Wang^[7]的实验一样,该训练集中包含了句子之间的边界点,而句子之间一般都是韵律边界(在该实验中的比例是98.1%)。如果去除这些样本的话,该训练集的基准值将达到92.7%。另外 Lee^[9]的正确率最低,原因可能是:(1)该训练集的基准值很小,非边界样本与边界样本的比例相

差不大。(2)训练集的规模较小,而对于基于统计的方法而言,样本库的规模非常重要。(3)所选的特征不够充分,例如没有考虑重音信息等。

结束语 文章提出了用 CART 技术来解决汉语 TTS 系统中韵律短语分析问题,并针对 CART 自身的缺点提出了两种改进的方法。今后的研究工作包括:(1)对决策树方法的进一步改进,尤其是在选择每个结点的最优分裂时如何纳入更多的全局信息。(2)对特征集的改进,现有的特征是否已经充分够用,或者是否存在冗余。(3)现有的工作是基于手工标注好的样本库,如何自动地获取所需的信息并把该方法集成到汉语 TTS 系统中去。

参考文献

- 1 杨玉芳.句法边界的韵律学表现[J].声学学报,1997,22(5):414~421
- 2 Ostendorf M, Wightman C, Veilleux N. Parse scoring with prosodic information: an analysis/synthesis approach [J]. Computer Speech and Language, 1993, 7: 193~210
- 3 Bachenko J, Fitzpatrick E A. Computational grammar of discourse-neutral prosodic phrasing in English [J]. Computational Linguistics, 1990, 16: 155~170
- 4 Quene H, Kager R. The derivation of prosody for text-to-speech from prosodic sentence structure [J]. Computer Speech and Language, 1992, 6: 77~98
- 5 Taylor P, Black A W. Assigning phrase breaks from part-of-speech sequences [J]. Computer Speech and Language, 1998, 12: 99~117
- 6 Muller A F, Zimmermann H G, Neuneier R. Robust generation of symbolic prosody by a neural classifier based on autoassociators [A]. In: Proc. of Intl. Conf. on Acoust, Speech and Signal Processing, 1996. 1285~1288
- 7 Wang M Q, Hirschberg J. Automatic classification of intonational phrase boundaries [J]. Computer Speech and Language, 1992, 6: 175~196
- 8 Hirschberg J, Prieto P. Training intonational phrasing rules automatically for English and Spanish text-to-speech [J]. Speech Communication, 1996, 18: 281~290
- 9 Lee S, Oh Y H. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems [J]. Speech Communication, 1999, 28: 283~300
- 10 Chou F C, Tseng C Y, Chen K J, et al. Automatic generation of prosodic structure for high quality mandarin speech synthesis [A]. Proc ICASSP, 1997. 1624~1627
- 11 谌卫军,林福宗,李建民,等.基于概率统计的韵律短语分析[J].计算机工程与应用,2001,37(3): 10~11
- 12 Breiman L, Friedman J, Olshen R, et al. Classification and Regression Trees [M]. Belmont, CA: Wadsworth, 1984