

基于 EM 的隐马尔可夫过程随机迭代算法及其在生物序列启动子识别中的应用^{*}

罗泽举^{1,2} 朱思铭¹

(中山大学数学与计算科学学院 广州 510275)¹

(重庆工商大学计算机科学与信息工程学院 重庆 400067)²

摘要 随着后基因组时代的到来,如何去揭示序列背后隐藏的生命信息已成为当前生命科学探讨的主题。而控制基因表达的正是启动子序列,如何去识别和预测序列的启动子区域是基因研究的重点课题。隐马尔可夫模型是最近几年研究基因最主要的模型。本文首先探讨了 EM 算法并提出了随机迭代算法,在初始状态分布和散发矩阵都随机假设,而转移矩阵由序列计算出的条件下对人类启动子序列进行识别,平均识别率达到了 92.05%。改进了多分类问题中的“投票策略”,提出了“一票决定”算法,使算法次数由 $O(N^2)$ 降到了 $O(N)$,由此对多个 DNA 家族进行分类,正确率达 90.73%。从结果上看,在两类问题上,支持向量机比隐马尔可夫模型优越,但在处理多分类问题上隐马尔可夫模型却比支持向量机有更强的分类能力。

关键词 隐马尔可夫模型,随机迭代算法,“一票决定”算法,启动子的识别和分类

Random Iterated Algorithm for HMMS Based on EM and Selected Application in Biology Sequence Promoter Recognition

LUO Ze-Ju^{1,2} ZHU Si-Ming¹

(School of Mathematics Computational Science, Sun Yat-Sen University, Guangzhou 510275)¹

(School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067)²

Abstract With the coming of post-genomics era, how to find out the life information hidden at the back of sequences is current main subject in the study of life sciences. Promoter controls gene expression. How to recognize and predict the promoter regions is an important subject in gene study. Hidden Markov Models (HMMS) are the main method for researching gene over the last several years. Firstly, the EM algorithm is discussed in this paper and one kind of random iterated algorithm is proposed. It recognizes human promoters at the condition that the initial state distribution and the emitted matrix are randomly assumed and the state transition probability matrix is known by calculating, with mean rate of recognition up to 92.05%. And it improves on “strategy of voting” for multi-class classification, puts forward one kind of voting algorithm named “one ticket determines”, to make the order of calculation reduce from $O(N^2)$ to $O(N)$. From that DNA sequences are divided with, the rate of accuracy up to 90.73%. From the result, SVM has an advantage over HMMS for binary classification, but for multi-class classification, the HMMS is superior to SVM.

Keywords Hidden Markov models, Random iterated algorithm, “one ticket determines” algorithm, Promoter recognition and classification

1 引言

判断一门机器学习理论的发展是看它是否具备以下三方面的条件:首先它是否有丰富的数学内涵和理论基础;是否可以得到广泛的应用;是否在实际应用中效果良好。隐马尔可夫模型(Hidden Markov Models, HMMS)正是具备这三方面的条件而成为近年来十分引人关注的机器学习方向,特别是在生物序列特征的识别方面。

隐马尔可夫理论最初由 Baum 及其同事于上世纪 60 年代末、70 年代初提出^[1],开始用于语音识别^[2],该模型最早用于计算生物学是 80 年代末、90 年代初,目前已经用于计算机系统入侵检测、人脸识别、在线零售、手写识别^[3]等,特别是在生物信息检测方面,已用于 DNA 模型构建、蛋白质二级结构预测^[4]、基因预测^[5]、蛋白质家族的构建^[6]等。

启动子(promoter)是位于结构基因 5' 端上游的一段 DNA 序列,能够指导全酶(holoenzyme)同模板正确结合,活化 RNA 聚合酶,启动基因转录。本文将 HMMS 模型用于人类启动子识别,通过改进 EM 算法,将启动子序列重新编码,在随机假设初始分布和转移概率及计算出序列转移矩阵的条件下,能将人类启动子和其它生物启动子显著地区分开来,并且区分内含子(intron)和外显子(exon),其算法收敛速度比一般迭代算法快。

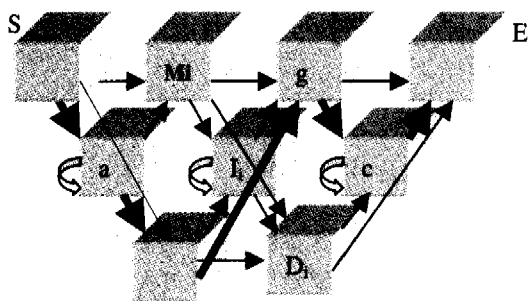
2 隐马尔可夫模型

2.1 隐马尔可夫模型的生物学描述

最初定义模型的时候,允许 HMMS 可以生成任意长度的序列。假设有一个开始状态,一系列的匹配状态 M_1, M_2, \dots, M_N , 插入态 I_1, I_2, \dots, I_{N+1} , 删除态 D_1, D_2, \dots, D_N 和结束

^{*}国家自然科学基金资助项目(No. 10371135)。罗泽举 博士研究生,主要研究方向:机器学习与模式识别、生物信息学。朱思铭 教授,博士生导师,主要研究方向:应用数学、常微分方程、计算机应用。

态。链从开始态开始,中间经过一系列的 M,I,D 状态,最后到结束态。观察符号序列是从 M,I 产生的,D 状态不产生观察符号(如图 1)。这个结构可以描述任何一个 DNA 或氨基酸 motifs 序列,而 N 是序列 motifs 的长度。



Di 代表删除态, Ii 代表插入态, Mi 代表匹配态

图 1 隐马尔可夫的描述

基因的进化就可以认为是这三个状态(删除态、插入态、匹配态)之间随机转移的结果。删除态代表基因序列中的空位和缺失,插入态代表基因的突变,匹配态代表某个特征序列。为了简化起见,如图 1,假设原始祖先序列是 TT,开始以某种转移概率插入了一个碱基 A,再以某随机概率转移到删除态,不生成任何符号,又以某种概率转移到匹配态 G,再随机转移到插入态插入碱基 C,然后结束。从而由特征序列 TT 最终形成了序列 TTAGC。当然这只是一条进化途径,由模型还可以形成其它许多序列,理论上讲,形成的 motifs 可以有无数多条,因为有无穷多种插入的可能。这些不同的祖先基因经过长时间的进化演变终于产生了大千世界丰富的动物界、植物界、微生物等等。

上述双重随机过程可以简化为图 2 的形式。

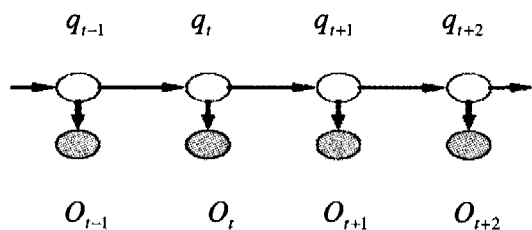


图 2 隐马尔可夫链双重随机过程的简单表示

其中 $q_1 \dots q_{t-1}, q_t, q_{t+1}, q_{t+2} \dots q_N$ 为隐藏的状态序列, $O_1 \dots O_{t-1}, O_t, O_{t+1}, O_{t+2} \dots O_M$ 为观察序列,问题是:如果我们只能看到观察序列,例如一段 DNA 序列 AGTCGTAGTCGATCGA,如何去揭示隐藏的状态序列呢?隐马尔可夫模型正是在这样的条件下被引入到计算生物学来的。

2.2 隐马尔可夫模型的定义

定义 1 一个模型 $\lambda = (S, \Sigma, A, B, \pi)$ 称为隐马尔可夫模型,其中:

- (1) $S = \{S_1, S_2, \dots, S_N\}$ 为状态集合, $N = |S|$ 是状态个数;
- (2) $\Sigma = \{O_1, O_2, \dots, O_M\}$ 是观察符号或观察向量的集合, $M = |\Sigma|$ 是观察符号或观察向量的个数;
- (3) $A = (a_{ij})$ 为状态转移概率矩阵,其元素 a_{ij} 表示从状态 S_i 转移到状态 S_j 的转移概率,记为

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N \quad (1)$$

满足:

$$a_{ij} \geq 0, \sum_{j=1}^N a_{ij} = 1; 1 \leq i, j \leq N \quad (2)$$

(4) $B = (b_j(k))$ 表示在状态 S_j 时产生观察符号 $v_k \in O$ 的离散概率值(v_k 为离散符号)或连续概率密度(v_k 是连续的观察符号)矩阵:

$$b_j(k) = P(v_k | q_t = S_j), 1 \leq j \leq N, 1 \leq k \leq M \quad (3)$$

(5) $\pi = (\pi_j)$ 是初始状态分布矩阵,其中: $\pi_j = P(q_1 = S_j)$, $1 \leq j \leq N$. 满足条件 $\pi_j \geq 0, \sum \pi_j = 1$.

上述定义中当观察符号 v_k 是离散符号时,叫离散马尔可夫模型;当 v_k 是连续符号时,叫连续马尔可夫模型。其中关键的模型参数是 A, B, π , 从而模型可以简记为 $\lambda = (A, B, \pi)$ 。

3 EM 算法

HMMS 模型的 EM (Expectation-Maximization) 算法最早由 Baum-Welch^[7,8] 提出,主要用来重新估计 HMMS 模型参数 A, B, π , 也就是 HMMS 的训练问题。它利用最优化方法,使得由模型生成序列的值 $P(O|\lambda)$ 达到最大。算法的思想是先得到一组参数值,再进行迭代运算,直到算法收敛到一个局部最优解。

重估算法如下:

设观察序列 $O = O_1^* O_2^* \dots O_n^*$, 相应的状态序列为: $Q = q_1 q_2 \dots q_n$; 定义变量:

$$\alpha_t(i) = P(O_1^* O_2^* \dots O_t^* | q_t = S_i, \lambda) \quad (4)$$

$$\beta_t(i) = P(O_{t+1}^* O_{t+2}^* \dots O_n^* | q_t = S_i, \lambda) \quad (5)$$

$$\eta^*(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}^*) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(O_{t+1}^*) \beta_{t+1}(j)} \quad (6)$$

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) = \sum_{j=1}^N \eta^*(i, j) \quad (7)$$

则状态 S_i 到 S_j 的期望值为:

$$n(i, j) = \sum_{t=1}^{n-1} \eta^*(i, j) \quad (8)$$

从状态 S_i 转出的期望值为:

$$n(i, +) = \sum_{t=1}^{n-1} \gamma_t(i) \quad (9)$$

状态为 S_i 且产生观察符号 v_k 的期望值为:

$$m(i, k) = \sum_{t=1}^n \gamma_t(i) \quad (10)$$

状态为 S_i 的期望值为:

$$m(i, +) = \sum_{t=1}^n \gamma_t(i) \quad (11)$$

则我们得到模型参数的三个重估公式:

$$\begin{aligned} \pi_i^* &= \gamma_1(i) \\ \alpha_{ij}^* &= n(i, j) / n(i, +) \\ b_j^*(k) &= m(i, k) / m(i, +) \end{aligned} \quad (12)$$

设当前模型为 $\lambda = (A, B, \pi)$, 重估后的模型为: $\lambda^* = (A^*, B^*, \pi^*)$ 。

Baum 等证明^[7,8] 得出,只要: $\max_{\lambda^*} Q(\lambda, \lambda^*) = \sum_Q P(Q|O, \lambda) \log\{P(O, Q|\lambda^*)\}$;

便有

$$P(O|\lambda^*) \geq P(O|\lambda) \quad (13)$$

且算法收敛到一个局部最大值。

4 随机迭代算法

本文提出一种新的随机迭代算法,其算法的步骤是:

Step1: 随机生成初始模型参数:

$\pi = (\pi_j), B = (b_j(k));$

Step2: 用人类启动子全序列计算出转移概率矩阵 $A = (a_{ij})$ 作为初始转移阵;

Step3: 用 EM 算法进行 HMMS 模型训练重新估算模型 3 个参数;

Step4: 训练出启动子序列的 HMMS 分数;

Step5: 计算出测试集的启动子序列的 HMMS 分数; 记录分数结果: 假阳性、假阴性及正确性的概率;

Step6: 如果假阳性、假阴性及正确性的概率不收敛, 则返回 Step1;

Step7: 用训练好的 HMMS 模型计算如下测试集的 4 个分数:

人类启动子序列 HMMS 分数;

人类内含子序列 HMMS 分数;

人类外显子序列 HMMS 分数;

其它生物启动子序列 HMMS 分数;

Step8: 算法终止。

该算法关键是 step1 和 step2, 也就是初始状态概率 π 和散发矩阵 B 都是随机假定的, 而不是一般算法中要事先计算出来; 而转移概率矩阵 A 不是随机假定的, 而是通过人类实际启动子全序列统计计算出来的; 结果显示, step1 的随机假定识别效果良好。

5 数据的获取和编码

5.1 数据的获取

从瑞士实验癌研究组织(The Swiss Institute for Experimental Cancer Research, ISREC)生物信息组支持的真核生物非冗余(non-redundant)启动子数据库(The Eukaryotic Promoter Database, EPD)提取数据。该数据库是基于欧洲分子生物实验室(European Molecular Biology Laboratory, EM-BL), 当前版本为 Release 83, 目前共收集有多个物种的 4809 个启动子序列(2005-8-10 前)(<http://www.epd.isb-sib.ch/>), 其中脊椎动物共有 2540 个(包括人类启动子有 1871 个)、植物质粒 198 个, 另有线虫(Nematode)、软体动物(Mollusc)、棘皮类动物(Echinoderm)等启动子序列。

从新加坡国立大学生物信息中心(NUS Bioinformatics Centre)的 Xpro 数据库下载真核生物(人类)的编码蛋白质的内含子和外显子序列(<http://origin.bic.nus.edu.sg/>), 格式有 FASTA 和 XML 两种, 文件基于 GenBank, 版本 139.0, 数据库共收集内含子和外显子序列 166555 个(2005-8-10 前)。

5.2 序列的编码

提取生物的特征来装配向量是目前普遍使用的办法。但是什么是一个 DNA 序列的特征? S. Kundsén 等研究过根据启动子词汇以序列 Motif 集作为启动子编码^[9], 也有人以基于序列相似性比作为编码规则的^[10], 但是这些编码模式并没有得到公共的论证, 当前一个标准的描述序列特征的规则并不存在。

笔者认为, 在如此长的 DNA 碱基对中(例如人类核基因组 DNA 约有 3×10^9 个碱基对(bp)), 有关序列的信息是直接隐藏在序列本身中的, 因此本文直接以序列作为装配对象, A、G、T、C 4 个碱基分别用 1、2、3、4 表示, 例如一个序列为:

GCGGAGGTTGCA

可以用向量

$X = (2, 4, 2, 2, 1, 2, 2, 3, 3, 2, 4, 1)$ 来表示。而且这样表示正好和隐马尔可夫过程一致, 表示刚好有 4 个观察符号的离散时间序列, 特别适合于训练。经过实验, 发现其效果好。

6 实验结果和讨论

6.1 两类识别模型的建立

我们用图 3 示出模型识别人类启动子序列和其它类序列。

以人类启动子 800 个序列作为训练集, 选择植物质粒 100 个、线虫(Nematode) 20 个序列、软体动物(Mollusc) 唯一的 3 个序列、棘皮类动物(Echinoderm) 30 个序列作为测试集; 人类内含子和外显子各选择其中 100 个也作为测试集。

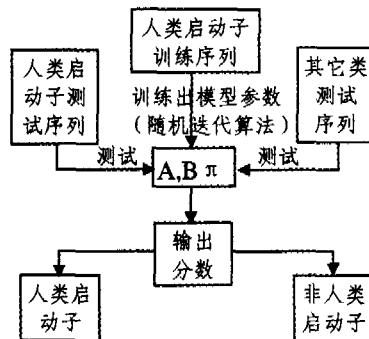


图 3 两类识别模型识别人类启动子和非人类启动子

所有序列都取统一长度为 300bps, 采用 300 个循环的迭代训练, 用算法计算它们各自的能生成该序列的最大可能的概率对数值, 得到表 1 所示的测试分数。

表 1 人类启动子和其它类的 L 值范围

类别	L 值范围
Man promoters	-373.56 ~ -383.33
Plant promoters	-296.73 ~ -309.61
Nematode promoters	-436.95 ~ -439.12
Mollusc promoters	-340.82 ~ -345.23
Echinoderm promoters	-450.39 ~ -468.23
Man intron	-496.16 ~ -518.83
Man exon	-440.31 ~ -453.53

为便于比较, 我们也把数据用支持向量机 SVM(support vector machine)训练测试, 发现表 2 所示的有趣的结果。从表中可以看到, 支持向量机在两类问题上的识别能力比隐马尔可夫模型要强些。但在随后的多分类问题上又可以看出, HMMS 又比 SVM 优越得多。

表 2 人类启动子和其它类正确识别率

类别	测试数	正确识别率	
		HMMS	SVM
Man promoters	100	93%	94%
Plant promoters	100	92%	93%
Nematode promoters	20	90%	92%
Mollusc promoters	3	100%	100%
Echinoderm promoters	30	93.3%	92.1%
Man intron	100	91%	95%
Man exon	100	92%	92%
平均		92.05%	93.38%

根据迭代规则,迭代过程中只要算法收敛就停止迭代,虽然我们选择了300个循环,但实际上只运行了63步就收敛了,说明随机迭代算法的收敛速度是很快的。

从表1可以看出人类启动子序列的训练L值和其它类别的启动子序列L值的显著区分情况,不同的类有不同的L值范围,因此我们的算法能明显识别人类启动子序列和其它类别的启动子及与内含子、外显子序列。

6.2 “一票决定”算法及多类DNA序列分类

通常用HMMS进行多类识别的方法是采用所谓“投票原则”。训练方法是先根据每一类单独训练出一组模型参数,也就是如果要区分N类,就先训练出N组模型参数,再用各自的测试集进行“投票表决”,算法要进行N次训练, N^2 次投票,共 $N+N^2$ 次,量级 $O(N^2)$,我们可以把这类投票模型叫做“多票决定制”,因为每一类在决定分类时都要投N次票后才能决定所属类,如图4所示。

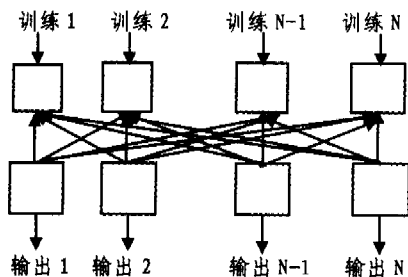


图4 传统多分类原则采用“多票决定制”

本文将传统投票模型进行改进,使得每一类在决定分类时只投“一次票”便可决定它的所属类别,我们把它叫做“一票决定”原则。其算法步骤是:

- (1)只训练出一类DNA序列;
- (2)其它类的数据一部分作为训练数据用(1)的模型计算L值,得出各类不同的L值范围;
- (3)其它类的测试数据再用(1)的模型计算L值,看L值落在什么范围(投票);
- (4)根据(3)识别和分类。

“一票决定”算法的HMMS多类辨别模型如图5所示。

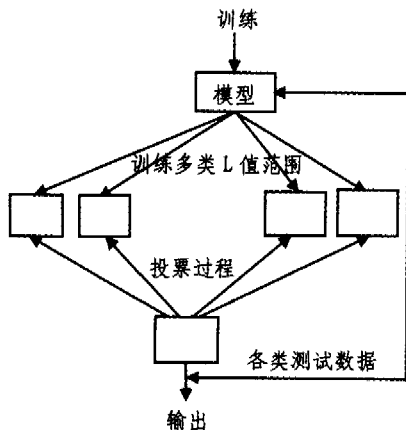


图5 新型的进行多分类的“一票决定”原则

可见,我们的模型训练 $1+N$ 次,投票 N 次,共 $2N+1$ 次,量级减少为 $O(N)$ 。

用上述新的投票方法进行多类“投票”,得到表3所示的分类结果,并和SVM训练多分类比较。

表3 7类DNA序列的正确识别率

类别	测试数	正确率	
		HMMS	SVM
Man promoters	100	91%	88%
Plant promoters	100	93%	89%
Nematode promoters	20	90%	85%
Mollusc promoters	3	100%	66.7%
Echinoderm promoters	30	86.7%	83.3%
Man intron	100	91%	92%
Man exon	100	89%	85%
平均		90.73%	88.08%

从表3可以看出,除了人类外显子的L值和棘皮类动物启动子L值有部分交叉(见表1),使得这两类的分辨率略低于其它类的分辨率外,整体分类效果是良好的,平均达到了90.73%。

我们从HMMS和SVM的分类结果比较上可以看出, HMMS的多分类处理能力明显比支持向量机要强得多。另外,实验中还发现HMMS计算处理的时间要比SVM少得多(表4)。

表4 HMMS与SVM的时间消耗对比(单位:s)

类型	HMMS	SVM
两类问题	50s	90s
多类问题	66s	125s

注:SVM选用多项式核函数

因为HMMS算法和支持向量机算法不同,SVM训练利用最优分类超平面把数据分为两类,多分类要进行多次处理;而HMMS由于各类有不同的L值范围,可以一次同时分出多类,这是隐马尔可夫模型在识别多类问题时所独有的优势。

近年来,SVM和HMMS都是模型识别的主流分支,SVM有它独有的小样本处理能力及基于核函数特征映射,使之得到迅速发展;而HMMS由于能对基因序列符号进行编解码,其插入、删除、匹配正好模拟了基因的进化过程,使它在生命科学领域备受欢迎。国际上已经建立的著名的基于隐马尔可夫模型蛋白质家族识别数据库(基于多重序列比对) Pfam [<http://pfam.wustl.edu/>]就是一例。

6.3 关于HMMS模型的局限性

虽然HMMS模型有它的优越性,但是和其它任何模型一样,隐马尔可夫模型也有它的局限性,主要体现在:

(1)线性模型的局限性。HMMS只是线性模型,无法描述如蛋白质的折叠关系等高阶相关性,这也是所有线性模型的局限性;

(2)马尔可夫假设的局限性。当前时刻的状态只依赖于前一时刻,然而有些问题,当前状态可能依赖于前面多个状态;

(3)独立性问题。HMMS假设观察符号之间是相互独立的,而事实上它们之间可能相互依赖;

(4)小样本训练的局限性。有的物种序列在数据库中非常少,因此训练的模型总存在偏差,例如L值范围的偏差;

(5)分类域值选取问题。过紧和过松的域值选取,会导致样本的少分或错分,因此要结合序列的其它性质来取域值可能会更好;

(6)HMMS模型得到的解是局部最优,如何得到全部最优也是将来要进一步研究的方向;

(7)序列的编码问题。特征提取直接影响识别效果,但一个序列的特征是什么?怎样提取一个序列的特征?这也是今后生物信息学要继续研究的方向。

但隐马尔可夫模型仍然具备机器学习的3大条件,已经成功进行了语音识别、手写数字识别、蛋白质分类、基因预测等重大问题。和其它机器学习方法(如SVM)一样,结合计算机计算能力的不断提高和计算生物学的发展,其应用还将越来越广泛,今后仍然是机器学习的重要方向。

参考文献

- 1 Baum L E, Egon J A. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull Amer Meteorol Soc., 1967, 73: 360~363
- 2 Baker J K. The dragon system-An overview. IEEE Trans. Acoust Speech Signal Processing, 1975, ASSP-23(1): 24~29
- 3 Schlapbach A, Bunke H. Using HMM based Recognizers for

writer identification and verification. In: Frontiers in Handwriting Recognition, 2004. 167~172

- 4 Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden Markov model. Computer Applications in the Biosciences(CABIOS), 1993, 9: 141~14
- 5 Krogh A. Two methods for improving performance of an HMM and their application for gene finding. In: Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology(ISMB), 1997. 179~186
- 6 Gough J, Chothia C. SUPERFAMILY: HM-MS representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research, 2002. 268~272
- 7 Baum L E, Sell G R. Growth functions for transformations on manifolds. Pac. J. Math., 1968, 27(2): 211~227
- 8 Baker J K. The dragon system-An overview. IEEE Acoust, Speech Signal Processing, 1975, ASSP-23(1): 24~29
- 9 Kundsén S. Promoter 2.0: for the Recognition of Promoter Sequences. Bioinformatics, 1999, 15: 356~361
- 10 Kasabov N, Pang Shaoning. Transductive Support Vector Machines and Applications in Bioinformatics for Promoter Recognition. In: IEEE Int Conf. neural Networks & signal Processing. Nanjing, China, Dec. 2003

(上接第194页)

输入信息,所以C1-C4的信息主要来自TimeWidget(在环境信息层专门负责处理时间信息的组件)。

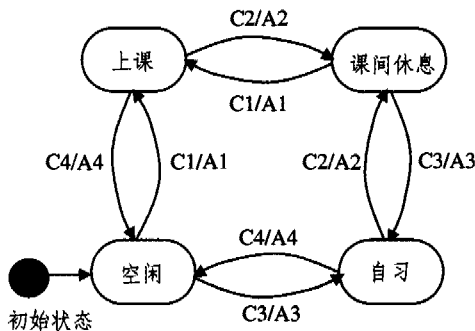


图3 SCR状态转换图

下面详细描述了系统在不同的状态下环境感知的功能。由于系统中提供了该教室的使用情况,则系统可以自动地在这几个状态之间进行切换。例如上午7:50时,教师手持PDA经教室门口安放的红外适配器进行身份识别后,系统将发出语音提示“某某老师已经来到,请大家做好上课准备”,并自动打开PowerPoint准备让教师使用。上午8:00时,系统将自动响铃以通知学生已经上课,同时将其内部状态切换为上课状态,自动加载上课相关的环境信息(如课程名称、学生人数的信息包括应到人数、实到人数等)并将信息显示到教师电脑桌面上。教师关闭系统消息显示后,系统启动了录音机准备记录教师的讲课录音,接着教师开始正常讲课。当到了下课时间时,系统再次响铃提示下课并提示下一节课的课程名称。教师结束讲课后,再次拿PDA对门外红外适配器照了一下通知系统离开的事件,系统检测到教师离开教室了以后则自动关掉电脑中打开的PPT、关掉录音设备等,同时将自身状态转为课间休息。当系统一旦进入课间休息的状态时,它将自动加载该状态下的环境信息(如学生喜欢听的音乐列表)并开始播放。而当到了上课时间时又能自动关掉音乐播放,再次执行响铃等操作。当到了自习的时间如晚上19:00时,系统将转入自习状态,此时将检测房间的亮度、温度等信息以便自动开启电灯、空调等设备,同时关注传感器传来的教室的音量信息,如果发现音量过大则自动发出提示音“现在是自习时间,请同学们保持安静”。当到了自习结束的时间以后,系统则自动关闭除电脑外的所有的设备以减少功耗,系统

转入空闲状态。

SCR原型系统的构造过程表明了本文提出的基于状态自动机的环境感知框架能较好地支持环境感知应用的开发,它加强了物理空间和信息空间的状态的联系,简化了系统分析和开发的过程,同时易于实现系统的修改和功能的扩充。

结束语 本文针对现有环境感知应用框架的不足,引入了信息空间环境状态的概念,加强了物理空间状态与信息空间状态的联系,同时提出了一个基于状态自动机的环境感知的改进框架,设计了相应的组件以进一步简化应用程序的开发并解决原有框架下较难处理的一些问题。实践证明,用基于状态自动机的环境感知应用框架开发应用程序是简单和快速的,同时很容易实现系统功能的扩展或重构。然而,该框架还存在些不足,如为了完成特定的功能,我们有可能定义过多的信息空间状态,而状态自动机引擎还没有对此进行优化处理。我们下一步的研究重点将集中在优化状态自动机引擎,使之能够自动优化多余的状态,争取用最少的状态来完成相同的功能。

参考文献

- 1 Mark W. The computer for the 21st Century. Scientific American, 1991, 265(3): 66~75
- 2 Dey A, Abowd G. Towards a Better Understanding of Context and Context-Awareness. In: Workshop on what, who, where, when and how of context-awareness at CHI 2000, April 2000
- 3 Dey A K, et al. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. HCI Jou, 2001, 16(2-4): 97~166
- 4 Roy, Want, Hopper A, Falcao V, Gibbons J. The Active Badge location system. ACM Transactions on Information Systems, 1992, 10(1): 91~102
- 5 Abowd G D, Atkinson C J, Hong J, Long S, Kooper R, Pinkerton M. Cyberguide: a mobile context-aware tour guide. Wireless Networks, 1997, 3(5): 421~433
- 6 Zhang T, Brüggé B. Empowering the User to Build Smart Home Applications, ICOST 2004
- 7 Humble J, et al. Playing with your bits: user-composition of ubiquitous domestic environments. In: Proc. UBIComp 2003, to appear
- 8 Schilit, Bill N, Norman I. Adams and Roy Want(1994). Context-aware computing applications. In: the Proc. of the 1st Intl. Workshop on Mobile Computing Systems and Applications, Santa Cruz, CA, IEEE. Dec. 1994. 85~90
- 9 Brown, Peter J, Bovey J D, Chen Xian. Context-aware applications: From the laboratory to the marketplace. IEEE Personal Communications, 1997, 4(5): 58~64
- 10 Davies, Nigel, Mitchell K, Cheverst K, Blair G. Developing a context-sensitive tour guide. In: the 1st Workshop on Human Computer Interaction for Mobile Devices, Glasgow, Scotland, May, 1998
- 11 Mealy G H. A Method for Synthesizing Sequential Circuits. Bell System Tech. J., 1955, 34: 1045~1079