文章编号:1001 - 2486(2002)06 - 0102 - 05

二维隐马氏模型的一种简化算法及其参数估计*

葛正坤,李　兵,王春玲
(国防科技大学理学院,湖南 长沙　410073)

摘　要:针对现有的二维隐马氏模型算法给出了一种简化算法及参数估计方法。该算法与现有的算法相比非常简单。基于此方法给出了相应的识别方法和参数估计,并且该估计与极大似然估计是等同的。

关键词:隐马氏模型;Viterbi 算法;EM 算法;参数估计;充分统计量;高斯混合
中图分类号:O211.62　　文献标识码:A

An Abridged Algorithm and Parameter Estimation of Two Dimension Hidden Markov Chain Model

GE Zheng-kun, LI Bing, WANG Chun-ling
(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: An abridged algorithm of 2D hidden Markov chain model and its parameter estimation method are made. This method is simplier compared with preceding ones. Based on this method the corresponding recognition method and parameter estimation are presented. And estimating parameters in this manner is equivalent to the maximizing of the likelihood.

Key words: hidden Markov chain model; Viterbi algorithm; EM algorithm; parameter estimation; sufficient statistics; Gaussian mixtures

二维隐马氏模型在解决文字识别中提供了很大的潜力,但是已经证明,一个连续的二维隐马氏模型将导致一个指数级复杂度的识别算法。本文采用 Viterbi 算法[1],将其看做是多维观测空间的一维隐马氏模型提供了一种简化的算法,从客观上简化了解码过程中的计算复杂度。本文中所用到的参数估计方法是一种与著名的隐马氏模型参数估计 Baum-Welch 算法相似的算法,这种新的算法使得每一状态的观测概率密度函数可以用不同的特征集来定义和估计。这种方法是基于充分统计量的,并且从理论上来说不会造成性能的损失。

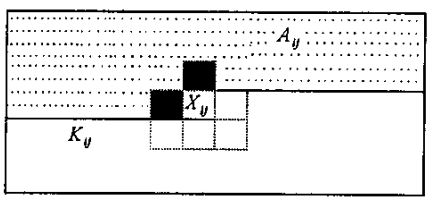

图 1　模型结构
Fig.1　Structure of the model

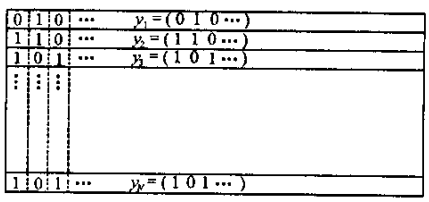

图 2　变换结构
Fig.2　Structure of transformation

1　二维隐马氏模型

该模型是在进行手写体汉字识别时所给出的,其结构如图 1 所示。隐马氏模型的数学表达为:$P(X_{ij}=x_{ij}\,|\,X\in A_{ij})=P(X_{ij}=x_{ij}\,|\,X\in K_{ij})$,其中:$x_{ij}$ 为随机变量 X_{ij} 的取值,$A_{ij}=\{(kl)\mid k<i \text{ or } k=i, l<j\}$表示点$(ij)$的"过去",$K_{ij}=\{(i,j-1),(i-1,j)\}$是点$(ij)$的近邻。根据完全二维隐马氏模型可以进行解码和识别。本文采用如图 2 的结构来解决。

* 收稿日期:2002 - 05 - 10
作者简介:葛正坤(1972—),男,讲师,硕士。

将观测阵列的每一行看作一个向量,即 $y_i = \{y_{i1}, y_{i2}, y_{i3}, \cdots, y_{iM}\}(1 \leqslant i \leqslant N)$ 其中每一个元素为 M 维向量,此时观测序列可以记为 $Y = (y_1, y_2, y_3, \cdots, y_N)$,状态序列可以记为 $X = (x_1, x_2, x_3, \cdots, x_N)$,其中每一个元素为 M 维向量,状态集合为 $\Theta = \{q_1, q_2, q_3, \cdots, q_L\}$。通过如上的简化之后,文字笔画之间的二维依赖关系简化为行与行之间的依赖关系,模型为:

观测为 M 维向量: $y_i = (y_{i1}, y_{i2}, y_{i3}, \cdots, y_{iM}), 1 \leqslant i \leqslant N$

状态序列为: $X = (x_1, x_2, x_3, \cdots, x_N)$

状态的个数为 L: $q_1, q_2, q_3, \cdots, q_L$

此模型可以看做是多观测空间的一维隐马氏模型,参数为 $\lambda = (\Pi, A, B)$,其中:

初始分布: $\Pi = \{\pi_i \mid \pi_i = P(x_1 = q_i), i = 1,2,3,\cdots,L\}$

转移概率: $A = \{(a_{ij})_{L\times L} \mid a_{ij} = P(x_{n+1} = q_j \mid x_n = q_i), 1 \leqslant i \leqslant L, 1 \leqslant j \leqslant L\}$

观测特征的条件分布: $B = \{(b_j(y_i))_{N\times L} \mid b_j(y_i) = P(y_i \mid x_i = q_j)\} 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant L\}$

并且 $$\sum_j^L \pi_j = 1, \quad \sum_j^L a_{ij} = 1, \quad i = 1,2,\cdots,L$$

给定模型下观测序列的联合概率密度函数为:

$$L(Y;\lambda) \triangleq P(Y;\lambda) = \sum_{\Theta} P(Y,\theta;\lambda) = \sum_{\Theta} \pi_{x_1} b_{x_1}(y_1;\lambda) \prod_{n=2}^{N} a_{x_{n-1}x_n} b_{x_n}(y_n;\lambda) \tag{1}$$

λ 的极大似然估计为: $$\hat{\lambda} = \mathrm{Arg}_{\lambda} \max L(Y;\lambda) \tag{2}$$

计算的过程中,为了防止数值下溢的情况,对出现的概率取对数。记

$P(X_1 = x_1, \cdots, X_n = x_n, Y_1 = y_1, \cdots, Y_n = y_n \mid \lambda)$ 为 $P(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n \mid \lambda)$

令 $$\delta_n(i) = \max_{x_1, x_2, \cdots, x_{n-1}} P(x_1, x_2, \cdots, x_{n-1}, x_n = q_i, y_1, y_2, \cdots, y_n \mid \lambda), 1 \leqslant i \leqslant L$$

则有 $$\delta_{n+1}(i) = \max_{x_1, \cdots, x_{n-1}, x_n} P(x_1, x_2, \cdots, x_n, x_{n+1} = q_i, y_1, \cdots, y_n, y_{n+1} \mid \lambda)$$

令 $\psi_n(i)$ 表示 $n-1$ 时刻时使得 $\delta_n(j) + \log(a_{ji})$ 达到最大的状态,

初始化 $$\delta_1(1) = \log(0.5) + b_1(y_1), \quad \delta_1(2) = \log(0.5) + b_2(y_1)$$
$$\delta_1(j) = -\infty (3 \leqslant j \leqslant L), \quad \psi_1(j) = 0 (1 \leqslant j \leqslant L)$$

迭代 $$\delta_n(j) = \max_{1\leqslant i\leqslant L} [\delta_{n-1}(i) + a_{ij}] + b_j(y_n), \psi_n(j) = \mathrm{Arg}\max_{1\leqslant i\leqslant N} [\delta_{n-1}(i) + a_{ij}]$$
$$(n = 2,3,\cdots,N-1, j = 1,2,\cdots,L)$$

终止 $$P^* = \max_{1\leqslant i\leqslant N} (\delta_N(i)) \qquad x_N^* = \mathrm{Arg}\max_{1\leqslant i\leqslant L} (\delta_N(i))$$

后推 $$x_n^* = \psi_{n+1}(x_{n+1}^*) \quad (n = N-1, N-2, \cdots, 3,2,1)$$

2 极大似然函数及参数冲估计的计算方法

定义 $$Z \triangleq \{[z_{1,0}, \cdots, z_{L,0}], [z_{1,1}, \cdots, z_{L,1}], \cdots, [z_{1,N}, \cdots, z_{L,N}]\}$$

其中 $$z_{j,n} = T_j(y_n) \qquad (j = 1,2,\cdots,L; \quad n = 1,2,\cdots,N)$$

变换后,原来的参数集替换为 $\lambda^z \triangleq [\{\pi_j\}, \{a_{i,j}\}, \{\mu_{jk}^z\}, \{U_{jk}^z\}, \{c_{jk}^z\}]$

其中 $\{\pi_j\}$, $\{a_{ij}\}$ 与参数组 λ 中的成分是一致的。相关状态的概率密度函数是高斯混合密度[2]:

$$p_j^z(z_j;\lambda^z) \triangleq \sum_{k=1}^{K} c_{jk}^z N(z_j, \mu_{jk}^z, U_{jk}^z) \tag{3}$$

$\sum_k c_{jk}^z = 1, z_j = (z_{j,1}, z_{j,2}, \cdots, z_{j,N}), N(z_j, \mu_{jk}^z, U_{jk}^z)$ 是联合高斯密度函数:

$$N(z_j, \mu^z, U^z) \triangleq (2\pi)^{-P_j/2} |U^z|^{-1/2} \cdot \exp\left[-\frac{1}{2}(z_j - \mu^z)'(U^z)^{-1}(z_j - \mu^z)\right]$$

p_j 是 z_j 的维数。

为了说明 λ, λ^z 之间的关系,需要定义一个对所有的状态都有‘相同作用’的状态 H_0。对上述计算过程中出现的状态 $q_1, q_2, \cdots, q_L$ 分别用 $H_1, H_2, H_3, \cdots, H_L$ 来代替。所选取的状态 H_0 满足下面两个条件:

1) $$p(Y \mid H_0) = \prod_{n=1}^{N} p(y_n \mid H_0) \tag{4}$$

2) $$p(Y \mid H_0) > 0, \text{对所有的 } Y \tag{5}$$

令 $\{z_n = T(y_n)\}(1 \leqslant n \leqslant N)$ 并且定义基于相关状态分布为 $\{p_j(z), 1 \leqslant j \leqslant L\}$ 的 HMM,这就是传统的隐马氏模型[3]。为使结果最优,Z 对于 L 个状态的分类必须满足充分性的要求,表达充分性要求的一个方法就是通过似然率,用它写成充分统计量的函数时,它是不变量[4,5],即

$$\frac{p(Y \mid H_j)}{p(Y \mid H_k)} = \frac{p(Z \mid H_j)}{p(Z \mid H_k)} \qquad (1 \leqslant j, k \leqslant L, j \neq k) \tag{6}$$

为使算法最优,要求 $Z_j = T_j(Y)$ 是 H_0 和 H_j 二元假设检验的充分统计量。若 Z_j 是充分统计量,则

$$\frac{p(y \mid H_j)}{p(y \mid H_0)} = \frac{p(z_j \mid H_j)}{p(z_j \mid H_0)} \qquad (1 \leqslant j \leqslant L) \tag{7}$$

即 Z_j 必须包含能从 H_0 中区分 H_j 的所有信息。与传统的方法比较 Z 中包含了所有的信息。对 Z 定义似然函数

$$L^z(Z; \lambda^z) \triangleq \sum_{\theta} \pi_{x_1} \left[\frac{p^z_{x_1}(z_{x_1,1}; \lambda^z)}{p(z_{x_1,1} \mid H_0)}\right] \cdot \prod_{n=2}^{N} \left[a_{x_{n-1},x_1} \frac{p^z_{x_n}(z_{x_n,n}; \lambda^z)}{p(z_{x_n,n} \mid H_0)}\right] \tag{8}$$

λ^z 的极大似然估计定义为： $$\hat{\lambda}^z \triangleq \operatorname{Arg}\max_{\lambda^z} L^z(Z; \lambda^Z) \tag{9}$$

要通过解决(9) 式来解决(2) 式,需要将变换后的参数组 λ^z 转换到参数组 λ 中去。

定理 1 设概率密度函数 $p_Y(y \mid H_0)$ 是定义在 Y(观测取值空间) 上,并且 $p_Y(y \mid H_0) > 0$,对任意的 $y \in Y$,令 Z 是通过多对一的特征变换 $z = T(y)$ 和 Y 有关的随机变量,这里 $T(y)$ 是 Y 的任意可测函数,当 Y 是来自于 $p_Y(y \mid H_0)$ 时,$p_Z(z \mid H_0)$ 是 Z 的概率密度函数,并且 $p_Z(z \mid H_0) > 0$,对所有的 $z \in Z$ 都成立。令 $f_Z(z)$ 是定义在 Z 上的任意的概率密度函数,则按如下定义 y 的函数是 Y 的概率密度函数： $$f_Y(y) = \frac{p_Y(y \mid H_0)}{p_Z(T(y) \mid H_0)} f_Z(T(y)) \tag{10}$$

证明略(参考变数定理[6] 的变化应用)。

定理 2 设 Y 的分布函数如(10) 式中定义,若 $Z = T(Y)$,则 Z 的概率密度函数是 $f_Z(z)$。证明略。

定理中虽然没有强调 Z 的充分性,但它所构造的概率密度函数 Z 是充分统计量。

推论 1 令 H_1 为某一定义在 Y 上的概率密度函数的任意假设,那么当 $T(Y)$ 是 H_1 相对于 H_0 的二元假设检验的充分统计量,那么当 $f_Z(z) \to p(z \mid H_1)$ 的时候,有 $f_Y(y) \to p(y \mid H_1)$。

推论 2 令 z^* 是 Z 中的一个点,那么对所有的满足 $T(y) = z^*$ 的 y,有：

$$\frac{f_Y(y)}{p_Y(y \mid H_0)} = \frac{f_Z(z^*)}{p_Z(z^* \mid H_0)}$$

这说明 $f_Y(y)$ 中所有满足 $T(y) = z^*$ 的点 y 与 z^* 有等同的性质。

假定给定一参数集 λ^z,它可以通过构造如下的概率密度函数来构造一个标准参数,即由 $\lambda^z \to \lambda$,写为 $\lambda = G(\lambda^z)$。与上述定理不同之处是下面的概率密度函数中含有参数。

$$p_j(y; G(\lambda^z)) \triangleq \left[\frac{p(y \mid H_0)}{p(T_j(y) \mid H_0)}\right] p^z_j(T_j(y); \lambda^z) \qquad 1 \leqslant j \leqslant L \tag{11}$$

在一般情况下,引用的假设可以是 j 的函数,记作 $H_{0,j}$。为了简单起见,选择 $H_{0,j} = H_0$,这样可以从定理 1 中看出 $p_j(y; G(\lambda^z))$ 是一个真正的概率密度函数。还可以从(1)、(4)、(8) 和(11) 式中看出

$$L^z(Z; \lambda^z) = \frac{L(Y; G(\lambda^z))}{p(Y \mid H_0)} \tag{12}$$

定义 $\hat{\lambda}^g \triangleq G(\hat{\lambda}^z)$,$\hat{\lambda}^g = \operatorname{Arg}\max_{\lambda^z} L(Y; G(\lambda^z))$,则 $\hat{\lambda}^g$ 对所有的 λ 使得 $L(X; \lambda)$ 达到了最大,且对某一 λ^z 满足 $\lambda = G(\lambda^z)$。当该统计量是充分统计量并且(7) 式成立时,由(7) 和(11) 式知：如果 $p^z_j(z_j; \lambda^z) \to p(z_j \mid H_j)$,那么 $p_j(y; G(\hat{\lambda}^z)) \to p(y \mid H_j)$。这样,通过对变换后的参数的估计便可以构造真正隐马氏

模型的参数。另外通过(12)式有

$$L^z(Z;\hat{\lambda}^z) \to \frac{L(Y;\lambda)}{p(Y \mid H_0)}$$

解决(2)式的一个迭代的算法是基于 EM 算法[7]。在每一次递推中参数 λ 的更新公式称之为重估计公式[3]。将(8)式写为

$$L^z(Z;\lambda^z) = \sum_{\Theta}\pi_{x1}\left[\sum_{k=1}^{K} c^z_{x_1 k} b^*_{x_1,k}(z_{x_1,1};\lambda^z)\right] \cdot \prod_{n=2}^{N}\left[a_{x_{n-1}x_n}\sum_{k=1}^{K} c^z_{x_n k} b^*_{x_n,k}(z_{x_n,n};\lambda^z)\right] \tag{13}$$

其中

$$b^*_{jk}(z_j;\lambda^z) \triangleq \frac{N(z_j,\mu^z_{jk},U^z_{jk})}{p(z_j \mid H_0)} \tag{14}$$

为了计算的方便,将上述式子变为[8]

$$L^z(Z;\lambda^z) = \sum_{\Theta}\sum_{K} p^*(Z,\Theta,K;\lambda^z) \tag{15}$$

其中

$$\sum_{K} \triangleq \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\cdots\sum_{k_N=1}^{K}$$

$$p^*(Z,\Theta,K;\lambda^z) \triangleq \pi_{x_1} b^*_{x_1,k_1}(z_{x_1,1};\lambda^z) c^z_{x_1,k_1} \cdot \prod_{n=2}^{N} a_{x_{n-1}x_n} b^*_{x_n,k_n}(z_{x_n,n};\lambda^z) c^z_{x_n,k_n} \tag{16}$$

给定一个参数值 λ^z,找到新的参数 $\lambda^{z'}$ 使 $L^z(Z;\lambda^{z'}) \geqslant L^z(Z;\lambda^z)$。

2.1 辅助函数

定理 3 定义 $Q(\lambda^z,\lambda^{z'}) \triangleq \sum_{\Theta}\sum_{K} p^*(Z,\Theta,K;\lambda^z)\log p^*(Z,\Theta,K;\lambda^{z'})$,如果 $Q(\lambda^z,\lambda^{z'}) \geqslant Q(\lambda^z,\lambda^z)$,那么 $L^z(Z;\lambda^{z'}) \geqslant L^z(Z;\lambda^z)$。根据 $\log x$ 当 $x>0$ 是严格凹函数,当且仅当 $p^*(Z,\Theta,K;\lambda^{z'}) = p^*(Z,\Theta,K;\lambda^z)$ 等式成立。

2.2 重估计算法

$$\begin{aligned} Q(\lambda^z,\lambda^{z'}) &= \sum_{\Theta}\sum_{K} p^*(Z,\Theta,K;\lambda^z)\log p^*(Z,\Theta,K;\lambda^{z'}) \\ &= \sum_{\Theta}\sum_{K}\left[\pi_{x_1} b^*_{x_1,k_1}(z_{x_1,1}) c^z_{x_1,k_1} \cdot \prod_{n=2}^{N} a_{x_{n-1}x_n} b^*_{x_n,k_n}(z_{x_n,n}) c^z_{x_n,k_n}\right] \cdot \\ &\quad \left[\log\mu'_{x_1} + \sum_{n=2}^{N}\log a'_{x_{n-1}x_n} + \sum_{n=1}^{N}\log b^*_{x_n,k_n}{}'(z_{x_n,n}) + \sum_{n=1}^{N}\log c^z_{x_n,k_n}{}'\right] \end{aligned}$$

$b^*_{jk}(Z;\lambda^z)$ 仅仅通过多变量高斯密度依赖于 $\lambda^{z'}$,可以认为数据 z_j 是固定的。(14)式中附加的条件,不依赖于 $\lambda^{z'}$ 的点,并不影响最大化的解决,只需要 $z_{j,n}$ 代替 y_n 的位置。

2.2.1 模型的前向过程

令

$$\alpha^c_n(i) \triangleq \frac{p(y_1,\cdots,y_n,x_n = q_i;\lambda^z)}{p(y_1,\cdots,y_n \mid H_0)}$$

初始化:

$$\alpha^c_1(i) = \pi_i \frac{p^z_i(z_{i,1};\lambda^z)}{p(z_{i,1} \mid H_0)} \quad (1 \leqslant i \leqslant N)$$

归纳:

$$\alpha^c_{n+1}(j) = \left[\sum_{i=1}^{L}\alpha^c_n(i)a_{ij}\right]\frac{p^z_j(z_{j,n+1};\lambda^z)}{p(z_{j,n+1} \mid H_0)} \quad (1 \leqslant n \leqslant N-1, 1 \leqslant j \leqslant L)$$

终止:

$$L^z(Z;\lambda^z) = \frac{L(Y;G(\lambda^z))}{p(Y \mid H_0)} = \sum_{i=1}^{L}\alpha^c_N(i) \tag{17}$$

2.2.2 模型的后向过程

后向参数 $\beta^c_n(i)$ 类似地可以定义:

初始化:

$$\beta^c_N(i) = 1$$

归纳:

$$\beta^c_n(i) = \sum_{j=1}^{L} a_{ij}\frac{p^z_j(z_{j,n+1};\lambda^z)}{p(z_{j,n+1} \mid H_0)}\beta^c_{n+1}(j) \quad (t = N-1,\cdots,1, 1 \leqslant i \leqslant L)$$

终止:

$$L^z(Z;\lambda^z) = \sum_{i=1}^{L}\beta^c_1(i)P^z_i(y_1)$$

2.3 重估计公式

记
$$\gamma_n(j) = P(x_n = q_j \mid Y) = \frac{\alpha_n^c(j)\beta_n^c(j)}{\sum_{i=1}^{L}\alpha_n^c(i)\beta_n^c(i)}$$

$$\xi_n(i,j) = P(x_n = i, x_{n+1} = j \mid Y) = \frac{\alpha_n^c(i)a_{ij}(p_j^z(z_{j,n+1};\lambda^z)/p(z_{j,n+1} \mid H_0))\beta_{n+1}^c(j)}{\sum_{k=1}^{L}\sum_{m=1}^{L}\alpha_n^c(k)a_{km}(p_m^z(z_{m,n+1};\lambda^z)/p(z_{m,n+1} \mid H_0))\beta_{n+1}^c(m)}$$

更新的状态为
$$\hat{\mu}_i = \gamma_1(i)$$

更新状态的转移矩阵为：
$$\hat{a}_{ij} = \frac{\sum_{n=1}^{N-1}\xi_n(i,j)}{\sum_{n=1}^{N-1}\gamma_n(i)}$$

2.4 高斯混合重估计公式

令
$$\gamma_n^c(j,k) \triangleq \gamma_n(j)\left[\frac{c_{jk}^2 N(z_{j,n}, \boldsymbol{u}_{jk}^z, \boldsymbol{U}_{jk}^z)}{p_j^z(z_{j,n};\lambda^z)}\right]$$

则
$$\hat{c}_{jk}^z = \frac{\sum_{n=1}^{N}\gamma_n^c(j,k)}{\sum_{n=1}^{N}\sum_{r=1}^{K}\gamma_n^c(j,r)}, \quad \hat{\boldsymbol{u}}_{jk}^z = \frac{\sum_{n=1}^{N}\gamma_n^c(j,k)z_{j,n}}{\sum_{n=1}^{N}\gamma_n^c(j,k)}$$

$$\hat{\boldsymbol{U}}_{jk}^z = \frac{\sum_{n=1}^{N}\gamma_n^c(j,k)(z_{j,n} - \boldsymbol{u}_{jk}^z)(z_{j,n} - \boldsymbol{u}_{jk}^z)'}{\sum_{n=1}^{N}\gamma_n^c(j,k)}$$

3 结论

本文运用二维隐马氏模型对文字识别给出了一种简单的算法，并在此基础上给出了相应的参数估计方法，对于二维隐马氏模型通过改变模型参数对似然函数进行了估计，但是对于真正的二维隐马氏模型非简化的算法和较为简洁的参数估计还没有得到较好的解决。

参考文献：

[1] 邓明华. 数学模型补充讲义(未定稿)[M]. 1999,10.

[2] Baggenstoss Paul M. A Modified Baum-Welch Algorithm for Hidden Markov Models with Multiple Observation Spaces [J]. IEEE Transactions on Speech and Audio Processing, 2001,9(4).

[3] Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [J]. Proc. IEEE, 1989,77:257 – 286.

[4] Lehmann E H. Testing Statistical Hypotheses [M]. New York: Wiley, 1959.

[5] Kendall M, Stuart A. The Advanced Theory of Statistics [M]. London: Charles Griffin, 1979.

[6] Royden H L. Real Analysis, 3rd ed [M]. Englewood Cleffs, NJ: Prentice-Hall, 1988.

[7] Baum L E. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains [J]. Ann. Math. Statist. 1970,41:162 – 171.

[8] Juang B H. Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains [J]. AT&T Tech. J. 1985,64(6):1235 – 1249.