**viterbi算法**

G. **DAVID** FORNEY, JR.

***Invited Paper***

**摘要**

viterbi算法(VA)是估计无记忆噪声信道中离散时间的有限状态马尔可夫过程状态序列的最佳回溯算法。在许多像数字通信领域中的很多问题都可以用这种算法解决。本文给出了一个指导说明来实现和测试。可以预见这种算法在很多领域中会得到越来越多的使用。

1 简介

viterbi算法(VA)于1967年作为卷积码译码方法被提出。从那时起，它就被公认为是一个用于数字信号估值的有吸引力的解决方案，就像卡尔曼滤波器被用作许多类似的估值问题一样。就像卡尔曼滤波器一样，VA使用回溯法跟踪状态随机过程，这在一定意义上是最佳的，并且便于进行试验和分析。然而这个潜在的过程是假定有限状态的马尔可夫过程而非高斯过程，这会导致在结构上有所不同。

本文的目的主要是指导介绍VA的结构和测试。这也意味着在写作时(1972年夏)或多或少的尽可能多次回顾以前的工作。我们的信念是将该算法应用到越来越多的领域。我们希望我们能为读者加速这个过程。

2 问题的现状

在一般形式中，VA可能被视为一个计算在无记忆噪声信道中的有限状态离散时间马尔可夫过程状态序列最大后验概率(MAP)的解决办法。在这部分里我们提出这个问题的一般性，然后通过不同的问题举例论证使其更加适合这样的模型。对这个一般的方法也简单作说明。

隐马尔科夫过程特征如下。时间是离散的。时刻k的状态xk是有限数目M中的状态m，1≤m≤M，即状态空间X就是简单的{ 1, 2, . . , M }，首先我们假定此过程只从时刻0运行到时刻K，并且第一个和最后一个状态x0和xk是已知的；则这个状态序列用有限向量x = (x0, . . ,xk) 代表。我们看到后来延长到无限序列是微不足道的。这个过程是马尔可夫过程，在这种意义下，若时刻k及以前概率给定，则概率仅取决于k时刻的状态xk：P(xk+1|x0,x1, . . . ,xk)=P(xk+1|xk)

转移概率P(xk+1|xk)可能随时间变化，但是我们在表达式中没明确表示这一点。我们将时刻k的转移状态ξk定义为状态对 (xk+l,xk) ： 

我们令Ξ(可能随时间变化)为P(xk+1|xk)≠0条件下的ξk=(xk+1,xk)的转换过程。

图 最通用的模型

很明显|Ξ|≤M2。显而易见的是状态序列x与转换序列ξ=(ξ0, . . . ,ξk-1)有一一对应的关系。(我们记作。)

这个过程是假定信号在无记忆噪声信道；也就是说，序列z的观测值zk一定概率取决于ξk在时刻k的转移概率：

基于这种理论，z可以描述为无记忆信道的输出，这种信道的输入序列是的ξ(见 图1)。其次，尽管我们没有明确的说，信道仍可能是时变的，在这种意义上P(zk| ξk)可能是关于自变量k的一个函数。这个公式可以用下面几种特殊情况概括：

(1)zk仅取决于状态xk的情况：

(2).这种情况下zk在一定概率下取决于k时刻的yk，此时yk是一个关于变换ξk或状态xk的确定过程。

例如：下面的模型经常出现在数字通信中。有一个输入序列**u**=**(u0,u1,** . . .)，它的每一个uk都是按照分布律**P(uk)**按一定概率单独生成的，并且序列的状态值是有限的，比如m。有一个无噪声信号序列y无法观测，每个yk都是由当前输入值和前v个输入值确定的函数值：**yk** = **f (uk,** . . . ,**uk-v)。**观察序列z是序列y在无记忆信道中的输出。我们把这样的一个过程称为寄存器移位过程(如图2所示)，它可以模拟为一个以uk为输入的长度为v的移位寄存器。(同样的这是一个m序列马尔可夫过程。) 为了实现我们通用模型的通信我们定义：

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形2.wmf

图 移位寄存器模型

(1)状态：

(2)变换：

则状态数为|X|=mv，变换的个数|Ξ|=mv+1。如果输入序列开始于0时刻，终止于k-v时刻，也就是说u= ( . . . 0, u0, u1, . . . , uk-v ,0, 0, . . . )，那么移位寄存器会有效地处理，从时刻0开始，在时刻k以x0=xK=(0, 0, . . . , 0)结束。

最后，我们用VA来说明如何解决这个问题。如果给定一个观察序列z。它是离散时间有限状态的经过无记忆噪声的马尔可夫过程，我们为状态序列x找到最大的后验概率P(x|z)。同样的，找到转移序列ξ的最大(由于)后验概率P(ξ|z)。因为，所以在移位寄存器模型中这个过程相当于找到最有可能的输入序列u；或者，找到最有可能的信号序列y。众所周知MAP(最大后验概率)减小了检测全部序列(数据分组，信息或可能的错误字符)的错误可能性，从而在这种意义上达到整体最优化。我们会看到它在必然可以有效地应用。应用实例：我们现在举一些不同领域的例子说明上述论点，包括卷积码，码间串扰，连续相位的FSK(频移键控)，文本识别。

1. 卷积码

一个码率为1/n的二进制卷积码编码器是一个像图2那样的移位寄存器电路，

uk为输入信息比特，yk为输出的n比特分组，yk=(p1k, . . . , pnk)，每个都同样的加上了v+1个信息位**(uk,** . . . ,**uk-v)**的监督位(模2加)。当被编码的序列y被送进一个无记忆信道后，就表示为图2所示的模型。图3显示了一个特定码率为，且v=2的卷积码的电路。

更一般的卷积码编码器：码率为可能为k/n，输入可能不是二进制序列，编码器可能存在反馈。无论如何，码要由寄存器移位过程生成。

我们可能也会注意到其他类型的传输码可以模型化为有限状态机因此成为我们通用设备。

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形3.wmf

图 一个卷积码编码器

B. 码间串扰

在模拟信道中进行数字传输时，我们经常会遇到下列情况。输入序列为u，它的时间和取值都是离散的，被用来调制一些连续的波形，这些波形通过信道传输后进行采样。理想情况下，采样值zk与其对应的uk相同，或是它的某些函数。然而事实上，采样值zk会被噪声和uk中相临的输入码元干扰。后者的干扰作用被称为码间串扰。有时为了使频谱成型要有意引入码间串扰，这就是所谓的部分响应系统。

在这种情况下输出抽样信号模型化为：zk=yk+nk。yk是一个有限输入信号的确定性函数，即：yk = f (uk, . . . , uk-v)，nk是一个高斯白噪声序列。图2正是这样。

详细一点说，在脉冲幅度调制中(PAM) 信号序列y可以看作输入序列u与时间离散的信道中的冲击响应序列(h0,h1, . . .)的卷积 : 

如果在i>v时hi=0(有限长脉冲响应)，然后我们就得到了我们的移位寄存器模型。图4是对这样的模型在三个时间跨度单位中(v=2)的码间串扰的说明。

在[29]中说明了当实际中时间是连续的时候的情况，即接收信号r(t)的形式为：。对于一些脉冲响应h(t)，以码元间隔为T发送信号，在实现高斯噪声的作用下不会丧失最优性的情况下可以减少为上述的离散时间形式。

C. 连续相位FSK

引用这个例子不是因为它的实际的重要性，而是因为，首先，它会形成一个简单的模型，我们后面会在一个例子中使用，其次，它展示了甚至是在传统情况下VA是怎样给予我们开辟了新视野的。

在FSK中，输入的数字序列u在每个发送信号间隔T中在m中频率中选择一个；也就是说，传输信号η(t)为：η(t)=cos[ω(uk)t+θk], kT≤t<(k+1)T 。其中ω(uk)是由uk选择的频率，θk是某一相位角度。考虑到频谱成型和变换间隔处相位连续的简单调制器，这是可取的，即：ω(uk-1)kT+θk-1≡ω(uk)kT+θk(module 2π) 。

这称为连续相位FSK。

相位连续在寄存器中加入调制过程；即它使得传输的信号在第k个间隔取决于以前的信号。拿可能是最简单的情况(偏移率=1/2)来说，令输入序列u为二进制，选择w(0)和w(1)使得w(0)在T秒内经历一个整数倍周期，w(1)经历半整数的奇数倍周期；即w(0)T≡0和w(1)T=π(以2π为模)。如果，根据u0为0或1判断θ0=0, θ1=0或π，类似的根据传输的1的个数的奇偶判断θk=0或π。

这里我们有两状态过程X= {1, π}。传输的信号yk是关于当前输入uk和状态xk的一个函数：yk=cos[w(uk)t+xk]=cosxk\*cosw(uk)t , kT≤t<(k+1)T。由于变换ξk=(xk+1,xk)是关于当前状态xk和输入uk的一对一的单值函数，我们可以认为yk由ξk决定。

如果我们把和作为信号空间的一组

基底，我们可以写作yk=y0kη0(t)+y1kη1(t)，给出如下坐标(y0k,y1k)：

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形4.wmf

图 存在码间串扰和高斯白噪声的PAM系统模型



图 偏移率1/2的有高斯白噪声和相干检测的二进制连续相位FSK



最后，如果接收信号ξ(t)是η(t)叠加高斯白噪声v(t)，将接收信号与分解信号η0(t)和ηl(t)在每个码元间隔做相关运算(相干检测)，我们才可能不丢失信息的接收时间离散的输出信号：zk=(z0k,z1k)=(y0k,ylk)+(n0k,n1k)，这里n0和n1是独立等方差的高斯白噪声序列。图5展示了该模型，信号产生器根据上述规则生成(y0k,y1k)。

D．文本识别

我们举这个例子来说明VA的使用并不局限于数字通信。在光学文字辨识(OCR)识别器中，每个文字经过扫描，隔离显著特征，并进行判决决定是什么字母或字符。当一部分字符以自然语言的文本出现时，语境曾长时期被认为可以帮助读者解决语义模糊问题。

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形6.wmf

图6 利用VA提高根据上下文语境识别字符能力

模型化上下文语境的一种方法是将像英语这样的自然语言看成一种离散时间的马尔可夫过程。例如，我们可以假定每个字母出现的概率取决于前v个字母，并且估计这(v+1)个字母组合的概率。然而这样的模型不能充分描述产生的自然语言(例如二字母和三字母的英文，见Shannon [45]、[46])，他们说明了语言的统计独立性并且容易被掌握。

使用这样的模型后，英文字母被看做一个mv个状态的马尔可夫过程，m是不同字符的数目，比如27(26个字母和一个空格)。如果再假设OCR的输出zk只取决于相应的输入yk，对于我们应用VA算法根据语境输出的序列，OCR识别器是一个无记忆信道；见图6。这里的“OCR输出”可以是由任何原始数据产生的。这些数据可能是一系列0和1，可以由识别器经过判决在没有上下文语境的情况下产生。一般来说，数据越原始VA算法越有用。

E．其他

在发现磁记录设备与有码间串扰的信道有相似之处后，Kobayashi [50] 提出了将VA应用到数字磁记录系统。Timor [51] 已经将该算法应用到序列排序系统中。该算法已经被提议在信源编码中使用，最后Preparata和Ray [53] 提议将该算法使用到语法模式识别中的“语义映射”上。这些就是被发现的该算法的应用。

3. VA算法

我们现在说明上述MAP序列估计问题在形式上与寻找某个图上的最短路径是完全相同的。因此VA作为一种自然回溯的解决方法出现了。

对于像图3或图4(在这个例子中是一个de Bruijn图[54])那样的4状态的寄存器移位过程，我们习惯将离散时间有限状态的马尔可夫过程和图7那样的状态转移图联系起来，这里的节点代表状态，树枝代表状态转移，在一个时间段内通过状态转移图追溯一定的路径，从一个状态转移到另一个状态。

在图7(b)中我们介绍了一个相同情况的更详细的描述，在这里每个节点都对应某个时刻的一个不同的节点，每个树枝代表在下个时刻转移到某个新状态。网格图开始时刻是x0，终止时刻是xK。最重要特性是对于每个可能的状态序列x，篱笆图中都会相应有一条单独的路径，反之亦然。

现在我们展示怎样实现。如果给定一个观察序列I，每个路径都可能被指定一个与-ln P(x,z)成正比的“长度” ，这里x是表示那条路径的状态序列。由于ln P(x,z)是P(x,z)的单调函数而且在路径和信号序列之间有一条一对一的对应关系，这样就使得我们通过寻找长度-ln P(x,z)最小的路径，能够解决寻找P(x|z)的值最大的的状态序列的问题，或者说P(x,z) =P(x|z)P(z)最大的。由于马尔可夫过程的无记忆性，我们只是作简单地观察，P(x,z)因子如下：



因而如果我们为每个树枝(转移)分配一个“长度” ：



那么对应于某个序列x的路径的总长度的定义为：



在一个图上寻找最短路径是运筹学中的一个古老的问题。最简单的办法是由Minty在1957年的一个信件中给出的，我们几乎全部引用如下：

最短路径问题的解决方法可以很简单，如下：在旅行网中建立一个线模型，其中节点代表城市，串的长度代表距离(或开销)。把节点“洛杉矶”抓在你的左手里，把“波士顿”抓在你的右手里并且将它们分开。如果模型变得纠缠在一起，让助手解开结再重新打结直到不纠缠为止。最后一个或多个路径会绷紧，选择出最短路径。由于模型使用上一二次后会产生混淆，因此为节点打上标签会很好。

不幸的是，Minty算法不能很好的适应现代计算机的计算方式，助手也没从前那样顺从。因此有必要继续研究在运筹学[56]中众所周知的VA算法来进行更多的观测。

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形7.wmf

图7 (a) 4状态寄存器移位过程的状态图 (b) 4状态寄存器移位过程的网格图

我们用**x0k**表示一个由时刻k的状态序列**x**=**(x0,x1,** . . . , **xk)** 组成的段**(x0, x1,** . . . , **xk)。**在网格图中**x0k**对应于一个从节点**x0**开始，在 **xk**结束的路径段。对于任何时刻k的节点**xk**，一般都会有几个这样的路径段，每个路径段长度：



这样的最短路径段地被称为对应于节点**xk**的幸存路径，记作，对于任意时刻k>0，总共会有M条幸存路径，每个都对应一个**xk**。观测过程是这样的：最短的完全路径必须从这些幸存路径开始。(如果没有，但是在时刻k经过状态**xk**，那么我们能够将它的初始段替换为来得到更短的路径度量。)

这样在任何时刻k我们只需要记住M条幸存路径和它们的长度。为了到达第k+1时刻，我们只需要将所有第k时刻的

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形8.wmf

图8 (a)标以树枝长度的网格图:M=4,K=5 ;(b)VA算法最短路径回溯判决

幸存路线向后延伸一个时间单位，并计算延伸长度路径段的长度，对于每个节点**xk+1**选择最短的延伸路径段，它终止在相应于第(k+1)时刻的幸存路径**xk+1**。如果幸存路径数目超过M，回溯过程就会无限持续下去。

许多读者会认为这个算法是前向动态规划[57],[58]的简单版本。这个问题就是计算最短路径。

我们在图8中对算法作了说明，是一个简单的5个时间单位4状态网格图。图8(a)是全部网格图，每个树枝标以一个长度。(在实际应用中，这个长度会是接收数据的函数。) 图8(b)是5个回溯长度，由算法确定从第一个码元到最后一个码元的最短路径。每个时刻在时间长度内只有4(或者更少)的幸存路径。

该算法的一个正式表示如下：

维特比算法

存储：



初始化：



回溯：对于所有的ξ=(xk+1,xk)计算



对于每个**xk+l**寻找；储存**Γ(xk+l)**和相应的幸存路径。

重复将k设为k+1直到k=K。

对于有限状态序列x，算法终止在时刻K作为幸存路径存储的最短的完全路径 。

在实际中某些微小的修改时必须的，当状态序列很长或无限时，将幸存路线截短为可控制长度δ是必要的。换句话说该算法必须在时刻k时在时刻k-δ之前进行一定的判决。注意图8(b)中所有的时刻4的幸存路径都在时刻2及之前都经过相同节点。一般而言，如果截短深度δ选择的足够大，很有可能所有的时刻k的幸存路径在时刻k-δ及之前都会经过相同的节点，因此最大可能路径的初始时间段取决于时刻k-δ并且可以作为算法的硬判决输出；在这种情况下截短不会有影响。在少数幸存路径不相同的情况下，算法在时刻(k-δ)的任何合理的判决将会生效[20]，[21]：选择时刻(k-δ)的任意一个节点，或者对应于最短幸存路径的节点，或者是对应于最大度量的节点，等等。如果δ足够大的话他们在实现时是基本相同的。

同样，如果k变得很大，则很有必要通过在每个时刻将所有**Γ(m)**减少一个常量来重新归一化长度。

最后该算法需要重新开始，此时不需要知道初试状态**x0**。在这种情况下它可以以任何设置合理的初始节点长度进行初始化，如对于所有的m有**Γ(m)**=0，或者如果每个状态的先验概率已知则有**Γ(m)**= -ln。通常在初始化完成的瞬间，很有可能所有的幸存路径都将会合并成为正确的路径。这样算法能在无特殊流程的情况下进行自身的同步。

该算法的复杂度是容易进行估计的。首先，存储：该算法需要M个存储空间，每个存储空间对应一个状态并且能够存储一个“长度” **Γ(m)**和一个δ个符号的截短幸存路径表，其次，计算：该算法在每个时间单位必有**|Ξ|**次加法和M次在**|Ξ|**个结果中比较，每个加法对应一个状态转移。因此存储的数量与状态数量和计算状态转移数量的次数成正比，在一个寄存器移位过程中，M=**mv**且**|Ξ|**=**mv+1**，因此算法的复杂性随移位寄存器长度v呈指数级增加。

在前面几段里，我们忽略了包括生成的增量长度λ(ξk)的复杂度。在一个寄存器移位过程中，一般来说**P(xk+1|xk)**的值为1/m还是0取决于**xk**后面的**xk+1**是否是被允许的；这样所有被允许的状态转移与-ln **P(xk+1|xk)**是等价的，这部分λ(ξk)可以忽略。更值得注意的是在一般情况下**P(xk+1**|**xk)**是事先知道的：于是这部分肯能要重新计算。-lnP(zk|ξk)部分只取决于数据；而且通常对于许多ξk会有相同的输出**yk**，因此对给定zk需要计算-lnP(zk|yk)的值或者查询所有这些ξk。(当噪声是高斯白噪声时，-ln P(zk|yk)仅仅与(zk-yk)2成比例。) 最后，所有这些都可以在中心回溯单元之外进行。

此外我们注意到一旦计算出λ(ξk)，观察序列**zk**不再需要存储，这在即时应用中是个吸引人的特点。

对寄存器移位过程的网格图进行深入研究后，可以发现更多可以利用的细节。对于一个二进制移位寄存器，在任何时间单位的状态转移都可以分成**2v-1**个互不相关的组，每个都由一对普通的状态产生且终止于另一对普通状态。如图9所示是一个这种典型单元，时刻k的状态标以x**’0** 和x’l ，时刻k+1的状态标以**x’0** and **x’1**，其中**x’**代表一个在任何单元中存在的v-1比特位的序列。例如，图8中网格图的每个时间单元都有两个这样的单元构成。我们注意到在一次回溯中只有相同单元的不同运算结果会相互影响。图10显示的是一个实现计算的基本逻辑单元。一个高速并行算法可以用**2v-1**个相同的这样的逻辑单元实现；一个低速并行算法可以用1个这样的单元进行分时实现；软件版本的用相应的子程序实现。

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\图形9.wmf

图 9 一般的寄存器移位单元网格图

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\PICTURE\图形11.wmf

图 10 二进制寄存器移位的基本VA逻辑单元

许多读者会注意到这里的网格图，这些网格图会让他们想起快速傅里叶变换(FFT)的计算流图。事实上除了在长度上，它们是相同的，而且一般的FFT确实也由单元结构组成。虽然VA的加-比运算不包含在FFT中，但是可以认为一些由FFT发展起来的记忆结构会同样有用。

由于高度平行化的结构并且只进行加、比、选的操作，VA很适合高速应用。一个v=6 (**M=64, |Ξ|**=**128**) 码的卷积码译码器是由356个转移逻辑单元电路构成，它能以可能代表现代工艺的最高**2** Mbits/s的速度运算[22]。即使一个简单码的软件译码器也能以大约**1000** bits/s的速度在一个小型计算机上运行。这样不太复杂的电路使VA应用在很多信号处理系统中。

实现的详情请参阅[22]、[23]中的相关参考。

4.性能分析

就像VA的实现一样，它的性能分析也很重要。在很多情况下，产生错误的概率可以在一定紧边界的上下区间上下界波动。事实上即使当VA没实现时，对它的性能的估计可以反映出它的一个简单方案的性能与理想的差距，而且能找到简单的很接近最佳性能的次佳方案。

在性能分析中的重要观点对错误结果的分析。令x为实际状态序列，令为由VA选择的实际序列。经过一个很长的时间后x和一般会多次分离和合并，如图11所示。每个不同的分离称为一个错误事件。如果x是无限的，那么错误事件可能通常是无限长度的，但是一个无限长误差事件的概率通常是0。

错误事件的重要性在于在一定概率下它们之间是相互独立，用概率论的语言描述是可再生的。此外，它们允许我们在每个时间单位内计算错误率，这是必要的，因为通常当一个长度K趋近于无穷时的MAP分组错误率估算值趋近于1。相反的我们从合理的给定时间开始计算错误事件概率,即错误事件在那个时刻不再继续。

对于图7(b)中的网格图，给定适当的路径x，开始于一定时刻k的所有可能的错误事件集合εk是树状结构的网格图，它开始于xk，在xk时树枝都终止在适当的路径，如图12所示。在编码理论中称之为不正确子集(时刻k)。

任何特定的错误事件概率是容易计算的；观察结果可能仅仅是这样，在x不等于的时间跨度内，的可能性比x大。如果错误事件长度为τ，就仅仅是关于两个长度为τ的序列一个双假设判决问题。并且一般会有标准的解决方法。

εk中的任何错误事件概率P(εk)会变成上界，通常是紧边界，通过一个联合边界，即通过所有εk中错误事件概率的和。

当这个和是无限大时，它通常决定于一个或几个代表可能错误事件的较大时间段，其和非常接近P(εk)。实际的上界经常由流图技术取得[4]、[5]、[29]。



图 篱笆图中的正确路径x(粗线) 和 估计路径x(细线)，显示了3个错误事件

另一方面，错误事件概率的一个紧边界区间下边界。取εk中有最大可能性的不同错误事件。假如实际的状态序列是两种可能性之一：对应于错误事件的实际的正确路径或错误路径。即使有这方面的信息，如果错误路径与给定的z很接近时仍然会出错，因此错误概率不会比特定的错误事件概率低。即使你有一种策略，在给定信息情况下，在缺少天赋的情况下，你的错误率会更高。总之，任何一定的错误事件的概率是P(εk)的下界。

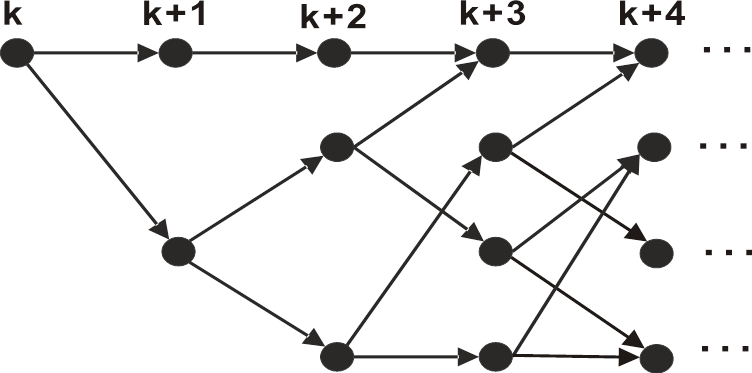


图 12 图7(b)网格图中的正确路径x(粗线)和时刻k的错误路径子集

一个重要的观察方面是这个下边界适用于任何判决方案,不仅仅是VA。简单扩展码给出了数值相关的错误率的的下界。如果VA的性能接近这些边界，它的相关结果可能是最佳且最有效的[30]。

总之，开始于时刻k的错误事件概率的上下边界如下：

max P(错误事件) ≤P(εk)≤max P(错误事件) + 其他项目

幸运的话这些区间是紧边界的。

例子

更具体一些，由于结果是有指导意义的，我们对前面定义过的简单连续相位FSK进行计算。图13是这个过程的二状态网格图，图14是一般的错误子集的第一部分。最少可能错误事件长度为2，由对{cos **w(1)t,** -cos **w(1)t}**的判决构成，而不是{cos **w(1)t,** cos **w(0)t**}。在我们的坐标标记中，{**(0, l), (0,** -1)}优于{**(1, 0), (1, 0)**}被选择。这是一个四维信号空间双假设判决问题[59]。在方差为σ**2**的高斯白噪声情况下，只有两个信号之间的欧氏距离会产生影响；在这种情况下d=，因此这个错误事件的概率是：**Q(d/2**σ**)=Q(l/**σ**)**，其中Q(x)是高斯概率误差函数，定义为：



C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\PICTURE\图形14.wmf

图 13 连续相位FSK网格图

C:\Documents and Settings\Administrator.CHINA-CCE6C076A\桌面\PICTURE\图形15.wmf

图 14 典型错误子集。 粗线：正确路径。细线：错误路径。

通过图14的测验我们发现在正确路径中，长度为3，4，. . . 的错误事件对应于距离， , . . . ；因此可以取得P(εk)的上下界：



Q(x)会随着x快速下降，这意味着对于任何合理的方差为σ2的噪声，P(εk)可以准确的估计为 。这个结果独立于时刻k的正确路径x，这一点从图3中很容易验证。

在这个结果中有趣的是，对于正交信号来说在一个码元间隔内的相干检测最好可以达到Q(1/2σ)，增加信号的信噪比3dB。因此可以认为连续相位FSK比偏移率非连续的好3dB，和反相相位变换键控性能相当。(虽然我们只证明了比率为的情况，但是它适用于任何比率。) 即使VA是一个很简单的2状态网格图，只有一种类型的错误事件对允许更简单的次优方案达到同样性能有一些意义[43][44]。

5.应用

我们对前面的结果做出总结，无论在理论还是实践中VA算法都得到了应用。

1. 卷积码

这个算法首先是由卷积码发展起来的，自然会在此有很大作用。

最主要的理论成果在维特比原稿中[1][5]-[7]。结果表明，对于一个合理的网

格编码和MAP解码，对于所有码率R小于信道容量的情况下，长度限制为v的错误率可以以指数方式减少。而且，我们认为衰减的速度比复杂度相当的分组码解码速度快。(尽管和相同解码延时的分组码相同)。在我们看来，分组码和卷积码的最明显的比较如下，适当的截短一个卷积(网格)码,可以生成一个任意长度和码率的有效的最佳分组码，但是这个分组码的码率较低[7]。事实上一个研究表明，卷积码的性能范围是紧边界，如果卷积码性能较好，理论上就可能截短成更好的分组码。

对于对称无记忆信道中的非渐进长度的固定二进制卷积码，一般会得到这样的结果[6]，P(εk)近似可以P(εk)=Nd2-dD由给出，其中d是自由距离，即错误子集εk中的任何正确路径的最小汉明距离。Nd是这些路径数目；D是巴特查里亚距离：，其中它们的和包括信道中输出空间Z中的所有输出z。

在高斯信道中P(εk)≈Ndexp(-dREb/N0)，其中Eb/N0是每信息位的信噪比。边界密度可以用模拟的方法确定[8]，[9]，[22]。比如对于一个v=6，d=10，R=的码，误码率10-3，10-5，和10-7对应于Eb/N0为3.0， 4.3，和5.5 dB，一般来说它的边界在几十分贝之内[22]。

空间通信的可用信道被准确地模型化为高斯白噪声信道。将VA应用到这样的信道是非常吸引人的，因为它的性能优于所有其它编码方案，在复杂度不是很高的情况下能速度很快，并且在随参信道参数变化时能稳定工作。大量系统原型已经被开发出来并进行了测试[21]-[26]，有些十分原始，很可能维特比译码器将会在空间通信系统中得到普遍应用。

最后，维特比译码器已经作为超高性能联合编码方案中作为基本单元使用[10]-[12]、[27]并且在有码间串扰的卷积码译码中也在使用[35]、[60]。

1. 码间串扰

VA在码间串扰中的应用是进年来才有的，主要的成就已经成为一套理论。

在高斯白噪声中PAM的主要结论是P(εk)可以以下面的紧边界式子描述：

KLQ( dmin/2σ)≤P(εk) ≤KUQ(dmin/2σ)

其中KL 和 KU都是较小的常量，Q(x)是前面定义过的高斯误差概率函数，σ2是噪声方差，dmin是两个不同信号之间的最小欧氏距离[29]。这个结果说明在大多数信道中码间串扰并不会带来明显的性能衰减，这是相当令人惊讶的。

例如，在最常见的部分响应系统中，VA恢复了相当于在全响应系统3dB的信息损失[29]、[32]。简单的次优处理器几乎可以进行处理。

一些工作人员针对于未知时变信道提出自适应算法版本[34]，[35]，[42]。Ungerboeck[41]，[42]和Mackechnie[35]已经得到结论，PAM中只需要一个匹配滤波器而不是白化滤波器。

似乎最佳路径很可能是在数字调制系统中VA产生最大效果是揭示了一些例子，在这些例子中常规检测技术由于性能较低而失败，所以建议采用次优的方式缩小距离。PAM信道在没有由于传输带宽接近0时而产生过多的噪声的情况下不是线性均衡的，这种信道最可能是这种非线性技术。

1. 文本识别

前面进行的试验用于修正由噪声模拟文字识别器产生的乱码文本的图表统

计[47]。结果与那些Raviv(使用统计图表)相似[48]，尽管算法很简单而且只有硬判决而不是置信度。看起来这个算法可能是一个用来解决复杂的文字识别系统有用的辅助系统，它用于有不同文字置信水平产生歧义时。

6 总结

VA对我们在一些问题上的理解产生了巨大影响，尤其是在卷积码理论和码间串扰理论中。它也开始对空间通信线路工程产生了实际性的影响。在码间串扰区的工作量显示了它可以进行更近一步的应用。由于该模型在应用中的普遍性和直接性，它在理论和实际中的分析和应用在未来的几年里将会越来越多。

附录

相关算法

在这个附录中我们提到一些VA相关的处理结构，主要是顺序解码和最小比特错误概率算法。我们也提到了生成可靠信息的扩展算法。

当网格图变大时，自然会放弃彻底搜索用于序列反复搜索的VA算法，检查出可能是最短路径的那些路径。在编码文献中，这些算法统称为顺序解码[13]-[16]。最简单的解释是“堆栈”算法[15]、[16]，其中维持一个了最短的部分路径的列表，列表最顶端的路径是被延长的，它的后继重新排列在表中直到最终节点，否则会无限制减少。(有些路径最终会这样，这保证了在编码应用中减少了一种倾向，这种倾向认为当所有错误路径长度都增加时正确路径的长度会减少。) 研究一个有限长网格是从什么地方开始和结束也是十分有用的。

在编码应用中，序列解码和维特比译码有很多相同的特点，包括相同的错误概率。它允许更长的也更高效码的译码，它的代价是输出数据z所需缓冲存储运算量的变化。如果不检查其它路径时找到最佳路径，它取决于在译码器识别能力，因此很可能在编码外用处不大。因此需要一个有限网格或可能的错误事件

之间的距离较大。

在码间串扰的文献中，许多早期的尝试是为了寻找使用误比特率作为最佳标准的的最佳非线性算法。这个过程的马尔可夫性会产生可控制算法，但是它逊色于维特比算法[36]-[42]。

这些算法的一般原理如下。首先，我们对网格图中每个状态xk计算联合概率P(xk, z)，或者对每个状态转移ξk计算P(ξk, z)。可以观察到：

P(xk, z)=P(xk, z0k-1)P(zkK|xk, z0k-1)=P(xk, z0k-1)P(zkK|xk)

因此对于给定的xk，从时刻k到时刻K的输出zkK是独立于从时刻0到时刻k-1的输出z0k-1。同样的：

P(ξk, z)=P(xk, xk+1, z)

=P(xk, z0k-1)P(xk+1,zk|xk,z0k-1)P(zk+1K|xk+1, xk, z0k)

=P(xk, z0k-1)P(xk+1,zk|xk)P(zk+1K|xk+1)

现在我们给出递推公式：



这使我们能够根据M个概率值P(xk-1,z0k-2)使用|Ξ|次次乘法和加法计算M个概率值P(xk, z0k-1)，使用长度的指数形式：e-λ(ξk-1)=P(xk|xk-1)P(zk-1|ξk-1)。同样我们可以向后回溯：

它有相同的复杂度。对所有节点的前后回溯完成后，要对所有节点计算P(xk, z) 和/或 P(ξk, z)。详细一点说，让我们考虑寄存器左移过程，设S(uk)为所有开始部分为uk的所有状态xk+1。那样：



由于P(uk, z)=P(uk|z)P(z)，uk 的MAP估计简化为寻找它的最大值。同样的，如果我们要找yk的MAP估计，也就是说设S(yk)为产生yk的所有ξk并计算：



在估计一些数值时可以使用一种简单的步骤，一种确定性的状态函数或转移函数。

在前后回溯和因此所需要存储所有的数据时，除了需要乘法，这个算法不比VA算法更有优势。以下的修正算法[39]、[48]消除了后者的不良特性，但这是以次佳性能和额外成本为代价的。让我们将自己局限于输入序列为u的寄存器左移过程，在估计uk时允许时刻k+δ及之前的观测值，其中δ≥v- 1。我们就有：



在和中的mδ+1个值可以重复判决，通过：



当只有前向回溯时，我们必须存储mδ+1个值而不是mv。如果δ较大，这样做就没有意义；如果δ接近于v-1，则估计值毫无疑问是次佳的。

因此这些变化的算法看上去很明显不如VA。然而，像这样的算法可能需要在特定情况下与VA混合使用，比如跟踪有限状态信道中的有限状态原始码，其中只有原始码中的状态序列是有用的。

最后，我们考虑VA的输出增量。有一个判断算法有多优秀的通用好方法，就是计算所有路径在合并处的深度；这可以应用于建立广播通信信道，同步等等[11]、[28]。一个更有选择性的用来指示每个部分的可靠性的指标，是最佳路径和次优路径在在合并时间点时长度上的差距。这个可靠性的指标可以量化成可去除的输出。最后，该算法能被变成存储L条最佳路径，而不是一条简单的路径，如每个回溯期中的最佳路径，这样最后生成了一个L条可能路径序列列表。

参考文献

卷积码

[1] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” IEEE Trans. Inform. Theory, vol. IT-13, pp. 260-269, Apr. 1967.

[2] G. D. Forney, Jr., “Convolutional codes I: Algebraic structure,” IEEE Trans. Inform. Theory, vol. IT-16, pp. 720-738, Nov. 1970.

[3] --------, “Review of random tree codes,” NASA Ames Res. Cen., Moffett Field, Calif., Contract NAS2-3637, NASA CR 73176, Final Rep., Dec. 1967, appendix A.

[4] G. C. Clark, R. C. Davis, J. C. Herndon, and D. D. McRae, “Interim report on convolution coding research,” Advanced System Operation, Radiation Inc., Melbourne, Fla., Memo Rep. 38, Sept. 1969.

[5] A. J. Viterbi and J. P. Odenwalder, “Further results on optimal de coding of convolutional codes,” IEEE Trans. Inform. Theory (Corresp.), vol. IT-15, pp. 732-734, Nov. 1969.

[6] A. J. Viterbi, “Convolutional codes and their performance in communicationsystems,” IEEE Trans. Commun. Technol.vol. COM-19, ,pp. 751-772, Oct, 1971.

[7] G. D. Forney, Jr., “Convolutional codes II: Maximum likelihood decoding,’’ Stanford Electronics Labs., Stanford, Calif., Tech. Rep. 7004-1, June 1972.

[8] J. A. Heller, “Short constraint length convolutional codes,” in Space Program Summary 37-54, vol. III. Jet Propulsion Lab., Calif. Inst. Technol., pp. 171-177, Oct.-Nov. 1968.

[9] --------, “Improved performance of short constraint length convolutional codes,” in Space Program Summary37-56,vol. III. Jet Propulsion Lab., Calif. Inst. Technol., pp. 83-84, Feb.-Mar. 1969.

[10] J. P. Odenwalder, “Optimal decoding of convolutional codes,” Ph.D. dissertation, Dep. Syst. Sci., Sch. Eng. Appl. Sci., Univ. of California, Los Angeles, 1970.

[11] J. L. Ramsey, “Cascaded tree codes,” MIT Res. Lab. Electron., Cambridge, Mass., Tech. Rep. 478, Sept. 1970.

[12] G. W. Zeoli, “Coupled decoding of block-convolutional concatenated codes,” Ph.D. dissertation, Dep. Elec. Eng., Univ. of California, Los Angeles, 1971.

[13] J. M. Wozencraft, “Sequential decoding for reliable communication,” in 1957 IRE Nat. Conv. Rec., vol. 5, pt. 2, pp. 11-25.

[14] R. M. Fano, “A heuristic discussion of probabilistic decoding,” IEEE Trans. Inform. Theory, vol. IT-9, pp. 64-74, Apr. 1963.

[15] K. S. Zigangirov, “Some sequential decoding procedures,” Probl. Pered. Inform., vol. 2, pp. 13-25, 1966.

[161 F. Jelinek, “Fast sequential decoding algorithm using a stack,” IBM J. Res. Develop., vol. 13, pp. 675-685, Nov. 1969.

[171 L. R. Bahl, C. D. Cullum, W. D. Frazer, and F. Jelinek, ”An efficient algorithm for computing free distance,” IEEE Trans. Inform.

[18] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of Theory (Corresp.), vol. IT-18, pp. 437-439, May 1972. linear codes for minimizing symbol error rate’ (Abstract), in 1972 Int. Symp. Information Theory (Pacific Grove, Calif., Jan. 1972), p. 90.

[19] P. L. McAdam, L. R. Welch, and C. L. Weber, “M.A.P. bit decoding of convolutional codes,” in 1972 Inf. Symp. Informafion Theory (Pacific Grove, Calif., Jan. 1972), p. 91.

空间应用

[20] Linkabit Corp., “Coding systems study for high data rate telemetry links, ’Final Rep. on N4SA Ames Res. Cen., Moffett Field, Calif., Contract NASZ-6024, Rep. CR-114278, 1970.

[21] G. C. Clark, “Implementation of maximum likelihood decoders for convolutional codes,” in Proc. Int. Telcmetering Conf. (Washington, D. C., 1971).

[22] J. A. Heller and I. M. Jacobs, “Viterbi decoding for satellite and space communication,” IEEE Trans. Commun. Technol., vol. COM-19, pp. 835-847, Oct. 1971.

[23] G. C. Clark, Jr., and R. C. Davis, “Two recent applications of errorcorrection coding to communications system design,” IEEE Trans. Commun. Technol., vol. COM-19, pp. 856-863, Oct. 1971.

[24] I. M. Jacobs and R. J. Sims, “Configuring a TDMA satellite communication system with coding, in Proc. 5th Hawaii Int. Conf. Systems Science. Honolulu, Hawaii: Western Periodicals, 1972, pp. 443-446.

[25] A. R. Cohen, J. A. Heller, and A. J. Viterbi, “A new coding technique for asynchronous multiple access communication,” IEEE Trans. Commun. Technol., vol. COM-19, pp. 849-855, Oct. 1971.

[26] D. Quagliato, “Error correcting codes applied to satellite channels,” in 1972 IEEE Int. Conf. Communications (Philadelphia, Pa.), pp. 15/13-18.

[27] Linkabit Corp., “Hybrid coding system study,” Final Rep. on Contract NAS2-6722, NASA Ames Res. Cen., Moffett Field, Calif., NASA Rep. CR114486, Sept. 1972.

[281 G. C. Clark, Jr., and R. C. Davis, “Reliability-of-decodingindicators Systems Science. Honolulu, Hawaii: Western Periodicals, 1927, pp. 447-450.

Intersymbol Interference

[29] G. D. Forney, Jr., “Maximum-likelihood sequence estimation of digital sequences in thep resence of intersynlbol interference,”IEEE Trans. Inform. Theory, vol. IT-18, pp. 363-378, May 1972.

[30] --------, “Lower bounds on error probability in the presence of large intersymbol interference,” IEEE Trans. Commun. Technol. (Corresp.), vol. COM-IO, pp. 76-77, Feb. 1972.

[31] J. K. Omura, ”On optimum receivers for channels with intersymbol interference,” abstract presented at the IEEE Int. Symp. Information Theory, Noordwijk, The Netherlands, June 1970.

[32] H. Kobayashi, ”Correlative level coding and maximum-likelihood decoding, IEEE Trans. Inform. Theory, vol. IT-17, PP. 586-594, Sept. 1971.

[33] M. J. Ferguson, ”Optimal reception for binary partial response channels,” Bell Syst. Tech. J., vol. 51, pp. 493-505, Feb. 1972.

[34] F. R. Magee, Jr., and J. G. Proakis, “Adaptive maximum-likelihood sequence estimation for digital signaling in the presence of intersymbo1 interference,” IEEE Trans. Inform. Theory (Corresp.), vol. IT-19, pp. 120-124, Jan. 1973.

[35] L. K. Mackechnie, ”Maximum likelihood receivers for channels having memory,’ Ph.D. dissertation, Dep. Elec. Eng., Univ. of Notre Dame, Notre Dame, Ind., Jan. 1973.

[36] R. W. Chang andJ . C. Hancock, “Onr eceiver structures for channels having memory,’ IEEE Trans. Inform. Theory, vol. IT-12, pp. 463- 468, Oct. 1966.

[37] K. Abend, T. J. Harley, Jr., B. D. Fritchman, and C. Gumacos, “On optimum receivers for channels having memory,” IEEE Trans. Inform. Theory (Corresp.), vol. IT-14, pp. 819-820, Nov. 1968.

[38] R. R. ?wen, “Bayesian decision procedures for interfering digital signals, IEEE Trans. Inform. Theory (Corresp.), vol. IT-15, pp. 506-507, July 1969.

[39] K. Abend and B. D. Fritchman, “Statistical detection for communication channels with intersymbol interference,” Proc. IEEE, vol. 58, pp. 779-785, May 1970.

[40] C. G. Hilborn, Jr., “Applications of unsupervised learning to problems of digital communication,” in Proc. 9th IEEE Symp. Adaptive Processes, Decision, and Control (Dec. 7-9, 1970).

C. G. Hilborn, Jr., and D. G. Lainiotis, “Optimal unsupervised learning multicategory dependent hypotheses pattern recognition,” IEEE Trans. Inform. Theory, vol. IT-14, pp. 468-470, May 1968.

[41] G. Ungerboeck, “Nonlinear equalization of binary signals in Gaussian noise,”I EEE Trans. Commun. Technol.v, ol. COM-19, pp. 1128-1137, Dec. 1971.

[42] ---------, ”Adaptive maximum-likelihood receiver for carrier-modulated data transmission systems,” in preparation.

连续相位FSK

[43] M. G. Pelchat, R. C. Davis, and M . B. Luntz, “Coherent demodulation of continuous-phase binary FSK signals,” iPnroc. Inl. Telemetry Conf. (Washington, D. C., 1971).

[44] R. de Buda, “Coherent demodulation of frequency-shift keying with low deviation ratio,” IEEE Trans. Commun. Technol., vol. COM-20, pp. 429-435, June 1972.

文本识别

[45] C. E. Shannon and W. Weaver, The MUthemdica! Theory of Communication. Urbana, Ill.: Univ. of Ill. Press, 1949.

[46] C. E. Shannon, “Prediction and entropy of printed English,” Bell Syst. Tech. J . , vol. 30, pp. 50-64, Jan. 1951.

[47] D. L. Neuhoff, “The Viterbi algorithm asa n aid in text recognition,’ Stanford Electronic Labs., Stanford, Calif., unpublished.

[48] J. Raviv, “Decision making in Markov chains applied to the problem of pattern recognition,” IEEE Trans. Inform. Theory, vol. IT-13, pp. 536-551, Oct. 1967.

杂项

[49] H. Kobayashi, “A survey of coding schemes for transmission or recording of digital data,” IEEE Trans. Commun. Technol., vol. COM-19, pp. 1087-1100, Dec. 1971.

[50] --------, “Application of probabilistic decoding to digital magnetic recording systems,” IBM J. Res. Develop., vol. 15, pp. 64-74, Jan. 1971.

[51] U. Tirnor, ‘Sequential ranging with the Viterbi algorithm,” JePt ropulsion Lab., Pasadena, Calif., JPL Tech. Rep. 32-1526, vol. 11, pp. 75-79, Jan. 1971.

[52] J. K. Omura, “On the Viterbi algorithm for source coding” (Abstract), in 1972 ZEEE Znt. Symp. Information Theory (Pacific Grove, Calif., Jan. 1972), p. 21.

[53] F. P. Preparata and S. R. Ray, “An approach to artificial nonsym-Calif., Jan. 1972), p. 21. bolic cognition,” Inform. Sci., vol. 4, pp. 65-86, Jan. 1972.

[54] S. W. Golomb, Shift Regisln Sequcrrccs. San Franasm, Calif.: Holden-Day, 1967, pp..13-17.

[55] G. J. Minty, ‘A comment on the shortest-route problem,” Opa. Res., vol. 5, p. 724, Oct. 1957.

[56] M. Pollack and W. Wiebenson, “Solutions of the shortest-route problem-A review,” Opa. Res., vol. 8, pp. 224-230, Mar. 1960.

[57] R. Busacker and T. Saaty, Finite Graphs and Networks: An Intro - duction with Applications. New York: McGraw-Hill, 1965.

[58] J. K. Omura, “On the Viterbi decoding algorithm,” ZEEE Trans. Inform. Thmy, vol. IT-15, pp. 177-179, Jan. 1969.

[59] J. M. Wozencraft and I. M. Jacobs, Principres of Cmnunicaliorr Engineering. New York: Wiley, 1965, ch. 4.

[60]. K. Omura, ‘Optimal receiver design for convolutional codes and Sci., vol. 3, pp. 24S-266, July 1971. channels with memory via control theoretical concepts,” Inform. Sci., vol. 3,pp. 243-266, July 1971.