用语义信息方法求检验、估计和混合模型的最大互信息和最大似然度

鲁晨光

lcguang@foxmail.com

摘要：不固定 Shannon 信道或产生假设的规则，求解最大互信息和最大似然度(等价于最大平均对数似然度)，是非常困难的，我们不得不使用迭代方法。根据鲁晨光提出的语义信息测度和 $R(G)$ 函数(信息率失真函数 $R(D)$的推广; G 是语义互信息的下限)可以得到一种新的迭代方法，用于检验、估计、混合模型。该语义信息测度被解释为平均对数标准(normalized)似然度, 而似然度由真值函数通过语义贝叶斯推理产生。一组真值函数构成一个语义信道，语义信道和 Shannon 信道相互匹配和迭代就能得到产生最大互信息和最大平均对数似然度的 Shannon 信道。该算法可谓信道匹配算法，简称 CM 算法。迭代的收敛可以通过 $R(G)$函数得到直观解释和证明。CM 算法用于检验、估计和混合模型的几个例子显示运算简单(可以用 Excel 文件演示)，收敛快速(随机选择的例子，收敛需要的迭代次数大多接近 5 次)。对于混合模型，CM 算法和 EM 算法类似； 但是和标准 EM 算法比，CM 算法有更好的收敛性和更多潜在应用。

关键词：Shannon 信道；语义信道；语义信息；似然度；混合模型；EM 算法; 机器学习。

演示迭代的 Excel 文件下载：http://survivor99.com/lcg/CM-iteration.zip

1 引言

用最大互信息或最大似然度准则优化检验，估计， 预测， 分类， 及机器学习是非常重要的。互信息意味节省的平均码长并给予小概率事件的正确预测以更高评价，在信源不是等概率的时候应该是很好的准则。然而 Shannon 信息论[1]使用失真准则而不是互信息准则优化检验和估计，这是因为优化需要改变 Shannon 信道，但是不固定 Shannon 信道， 就不能计算互信息。

“最大似然度”(ML)出现在很多文章中，它有两个不同含义。 只有少数意指本文所说的最大似然度。考虑信息传递

$$\text{事物 } X \text{—>观察特征 } Z\epsilon C \text{—>假设 } Y=f(Z)$$

其中 X, Y, Z 是离散随机变量， $f(Z)$是决策函数， 它确定 C 的一个划分和一个 Shannon 信道。最优划分提供最大互信息(MMI). 对于给定一个 Shannon 信道，存在一个最大似然度；对于所有可能的 Shannon 信道(可能受到一定限制， 比如条件概率是高斯分布的限制), 存在另一种最大似然度。它等价于最大平均 log 似然度(MALL)。本文重点讨论后者， 即 MALL。后面最大似然度意指 MALL。

最大互信息和最大似然度是如此相关，以至于我们在寻求最大互信息分类的时候，我们可能需要用似然度方法；在我们需要寻求最大似然估计(其中所有不同估计的概率分布是不确定的)的时候， 我们可能需要使用信息和熵的概念。最近几十年，信息测度[1]和似然度[2]之间的关系已经引起越来越多的关注[3][4][5]。

我们是否可以把似然度放进信息公式中，更有效地结合 Shannon 信息论和似然度方法？Akaike [3]指出:最大似然准则等价于最小 Kullback-Leibler (KL)距离(divergence)准则[6]。这是一个重要发现。然而，这一距离并不意味传递的信息。虽然一些研究者使用的 log(相对似然度)[5]或 log(标准似然度)[6])很像是信息测度，但是它们不能直接用 KL 公式或 Shannon 互信息公式表示， 原因是样本分布一般不同于似然函数。

已有不少求解最大互信息和最大似然方法，包括 Newton 方法[7]，EM 算法[8]，最小最大方法[9]。但是我们仍然想要更高效率，收敛理由更清楚， 更广应用的不同迭代方法。

非常不同的是，鲁晨光于 1993 [10][11][12]直接用平均标准(normalized)似然度定义语义信息测度，得到广义 KL 公式和广义互信息公式。虽然他没有提到“似然度”， 但是他使用“预测的概率”， 这实际上就是似然度。所得信息公式之所以被称之为语义信息公式，是因为其中似然函数是通过假设 Y 的真值函数, 加上 X 的先验概率分布产生的。鲁晨光还提出 $R(G)$函数，它是 Shannon 信息率失真函数 $R(D)$[13]的推广，其中 G 是语义互信息或平均 log 标准似然度的下限，R 是给定 G 时 Shannon 互信息的最小值。现在我们发现，用鲁晨光的信息测度和 $R(G)$函数可以更方便地求解最大互信息和最大似然度。

为了介绍鲁晨光的方法并显示它和流行的似然方法和流行的信息论方法兼容，我们先说明鲁氏真值函数和似然函数之间的联系。一个真值函数等于一个标准似然函数加上一个系数，这个系数使得真值函数的最大值是 1，并取值于实数区间[0,1]。给定样本分布或信源, 通过语义贝叶斯推理，真值函数和似然函数可以相互确定。一组真值函数构成一个语义信道，表示接收者理解的一组假设的语义；而 Shannon 信道表示发送者使用假设的规则。

这一发现具体说来，就是让语义信道和 Shannon 信道相互匹配和迭代可以求出检验，估计和混合模型的最大互信息和最大似然度(其中混合模型通常用 EM 算法[8]求解)。特别是，通过 $R(G)$函数，我们可以容易解释和证明新的迭代算法的收敛性。我们且称这种算法为信道匹配算法，或 CM 算法。

下面我们首先联系似然方法简要介绍语义信道，语义信息测度和 $R(G)$函数——用尽可能和流行的似然方法兼容的方式；然后举例说明如何将 CM 算法用于检验，估计和混合模型；最后比较 CM 算法和 EM 算法从而显示 CM 算法的优点和意义。

2 语义信道和语义贝叶斯推理

语义信道由一组模型的真值函数表示。下面我们从 Shannon 信道谈起。

2.1 Shannon 信道和转移概率函数

设 X 是表示信源(或数据，样本)的随机变量，取值于集合 $A=\{x_1, x_2, \ldots, x_m\}$; Y 是表示信宿(或假设)的随机变量，取值于集合 $B=\{y_1, y_2, \ldots, y_n\}$; Z 是表示观察条件的变量(或矢量)的随机变量，取值于集合 $C=\{z_1, z_2, \ldots, z_w\}$. 我们根据 Z，选择 Y, 预测 X。比如对于天气预报，X 是日降雨量，Y 是预报语句(比如“明天有小到中雨”)，Z 是预测依据的气象数据；对于医学检验，X 是真有病或真没病， Y 是阳性或阴性，Z 是化验数据。

我们用 $P(X)$表示 X 的概率分布，$P(X)$又叫信源；用 $P(Y)$表示 Y 的概率分布，$P(Y)$又叫信宿。Shannon 把概率 $P(y_j|X)$函数(y_j 不变 X 变)称作 y_j 的转移概率函数，那么一个 Shannon 信道就是一组转移概率函数，$P(y_j|X), j=1, 2, \ldots, y_n$ 构成的：

$$P(Y|X) \Leftrightarrow \begin{bmatrix} P(y_1|x_1) & P(y_1|x_2) & \ldots & P(y_1|x_m) \\ P(y_2|x_1) & P(y_2|x_2) & \ldots & P(y_2|x_m) \\ \ldots & \ldots & \ldots & \ldots \\ P(y_n|x_1) & P(y_n|x_2) & \ldots & P(y_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} P(y_j|X) \\ P(y_j|X) \\ \ldots \\ P(y_n|X) \end{bmatrix} \tag{2.1}$$

其中双向箭头表示等价。转移概率函数有如下性质：

1)和条件概率函数 $P(Y|x_i)$和 $P(X|y_j)$不同，它们是归一化的， 而转移概率函数 $P(y_j|X)$不是归一化的， 即一般情况下，$\sum_i P(y_1|e_i) \neq 1$.

2)可以用 $P(y_j|X)$和 $P(X)$做贝叶斯推理得到 X 的后验概率分布 $P(X|y_j)$(客观的统计结果，不是似然函数), 而且 $P(y_j|X)$ 乘上一个系数 k， 推理不变, 即

$$\frac{P(X)kP(y_j|X)}{\sum_i P(x_i)kP(y_j|x_i)} = \frac{P(X)P(y_j|X)}{\sum_i P(x_i)P(y_j|x_i)} = P(X|y_j) \tag{2.2}$$

2.2 语义通信模型和语义信道

从假设-检验的角度看，X 是证据或样本，Y 是假设; 我们用样本概率分布 $P(X|.)$检验和评价一个假设。现在我们用 Θ 表示模型变量，用 θ_j 表示 Θ 的一个取值，并且 $Y=y_j$ 时，$\Theta=\theta_j$。我们用 $y_j(X)$表示一个谓词, 其语义被真值函数 $T(\theta_j|X) \in [0,1]$所定义。流行方法中的模型 θ 就是这里的 Θ。这里的似然函数 $P(X|\theta_j)$就是流行方法中的似然函数 $P(X|y_j, \theta)$。一个模型 θ_j 包含一个或几个模型参数， 所以我们也可以像在流行的似然度方法中那样，把 θ_j 理解为预测模型的参数。不同的是，一个预测模型在这里可以独立地用真值函数表示出来。我们也可以认为 $T(\theta_j|X)$是用标准似然函数定义的，即 $T(\theta_j|X)=k\,P(\theta_j|X)/P(\theta_j)=k\,P(X|\theta_j)/P(X)$, k 使得 $T(\theta_j|X$ 的最大值是 1. $T(\theta_j|X)$就是假设 y_j 的或谓词 $y_j(X)$真值函数，所以我们有时也可以把 $T(\theta_j|X)$写成 $T(y_j|X)$。当 $X=x_i$ 时，就有命题 $y_j(x_i)$，命题的真值是 $T(\theta_j|x_i)$。如果 $T(\theta_j|X) \in \{0,1\}$，它就是 A 上使 θ_j 为真的子集的特征函数。所以我们可以把 θ_j 理解为一个模糊集合的特征函数或隶属函数[14]。

对比流行的似然度方法，上述方法使用子模型 θ_1 , θ_2, …, θ_n 而不是一个模型 θ or Θ, 而一个子模型是从一个似然函数 $P(X|\theta_j)$中分离出来的，是通过真值函数 $T(\theta_j|X)$定义的。$P(X|\theta_j)$等价于流行的似然度方法中的 $P(X|y_j, \theta)$。一个检验 y_j 的样本也是一个子样本，或条件样本。这些改变将使新的似然度方法(即语义信息方法)更加灵活，更加兼容 Shannon 信息论。

当 $X=x_i$ 时，$y_j(X)$变成命题 $y_j(x_i)$, 其真值是 $T(\theta_j|x_i)$. 于是，一个语义信道由若干真值或真值函数构成：

$$T(\Theta|X) \Leftrightarrow \begin{bmatrix} T(\theta_1|x_1) & T(\theta_1|x_2) & \ldots & T(\theta_1|x_m) \\ T(\theta_2|x_1) & T(\theta_2|x_2) & \ldots & T(\theta_2|x_m) \\ \ldots & \ldots & \ldots & \ldots \\ T(\theta_n|x_1) & T(\theta_n|x_2) & \ldots & T(\theta_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1|X) \\ T(\theta_2|X) \\ \ldots \\ T(\theta_n|X) \end{bmatrix} \tag{2.3}$$

真值函数 $T(\theta_j|X)$也不是归一化的，其最大值是 1。其意义是：作为预测模型，它和 $P(y_j|X)$及 $P(\theta_j|X)$类似，可以做贝叶斯推理——语义贝叶斯推理(或者说集合贝叶斯推理[10])， 产生似然函数：

$$P(X \mid \theta_j) = P(X)T(\theta_j \mid X) / T(\theta_j), \qquad T(\theta_j) = \sum_i P(x_i)T(\theta_j \mid x_i) \tag{2.4}$$

其中 $T(\theta_j)$就是 y_j 的逻辑概率，我们也可以把它写成 $T(y_j)$。由公式(2.2)可知， 当真值函数和转移概率函数成正比的时候，语义贝叶斯推理和贝叶斯推理等价。上述公式曾被 Thomas 提出[15]。

y_j 的逻辑概率 $T(\theta_j)$和被选择的概率 $P(y_j)$非常不同。$T(\Theta)$也不是归一化的， 一般有 $T(\theta_1)+T(\theta_2)\ldots+T(\theta n)>1$。设 y_1=“小雨”(“明天有小雨”的简写， 后面同理)， y_2=”中雨”， y_3=”小到中雨”。一定有 $T(\theta_3)\approx T(\theta_1 \cup \theta_2)>T(\theta_1)$, 但是可能 $P(y_3)<P(y_1)$或 $P(y_3)=0$。一个永真句，比如“有雨或无雨”，的逻辑概率是 1， 但是它被选择的概率几乎是 0。

$P(X|\theta_j)$和 $P(X|y_j)$不同，$P(X|y_j)$是样本的概率分布(注意样本也是有条件的)。实际上样本和 y_j 只有时间上的先后关系而没有必然因果关系，即 $P(X|y_j)$只表示 y_j 被选择后 X 发生的概率。而 $P(X|\theta_j)$则来自根据 y_j 的语义的推理。和流行的统计方法不同， 本文假设样本较大，我们并不直接使用时间序列中的样本，而是统计出这些样本在集合 A 中元素上的概率分布。

一个语义信道后面总有一个 Shannon 信道。以天气预报为例，转移概率函数 $P(y_j|X)$反映预报语句 y_j 的选择规律，因预报员而异——有人预报对的多，有人预报错的多。而 $T(\theta_j|X)$反映 θ_j 的语义，主要取决于语言的定义， 但是也受过去的预报规则的影响(后面详谈)。不同的人理解的语义 $T(\theta_j|X)$是大体相同的。

2.3 理解 GPS 定位——似然函数还是真值函数？

考虑全球定位系统(GPS)显示屏上的定位(小圆圈)的语义。它表示实际位置大概在某处。即 y_j=”$X\approx x_j$”。 一个时钟，一个秤，一个温度表，含义类似。具有这样含义的 y_j 通常被称为无偏估计。设 $Z\epsilon C_j$ 时，$y_j=f(Z|\ Z\epsilon C_j)$, 则 $P(y_j|X)=P(C_j|X), j=1, 2, \ldots, n$, 构成一个 Shannon 信道，反映估计的选择规则。而一个无偏估计的语义信道可以表示为：

$$T(\theta_j|X)=\exp[-|X-x_j|^2/(2d^2)], j=1, 2, \ldots, n \tag{2.6}$$

其中 d 是标准差。考虑 GPS 定位的特殊环境如图 2 所示。其中定位指示小车在高楼上，楼的左边是高速公路，右边是普通公路。请问小车在哪里可能性最大？

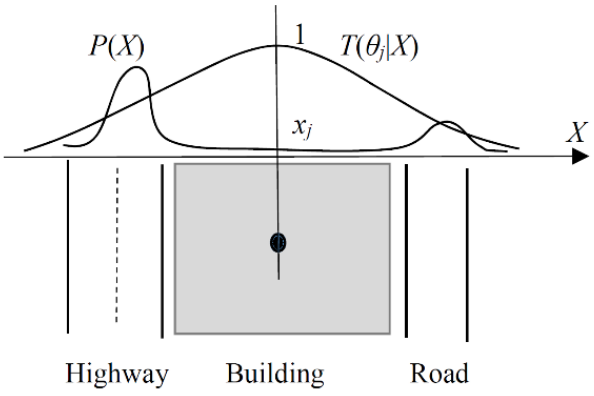


图 1 GPS 定位图解。先验概率分布不均且不断变化时，如何根据定位预测小车实际位置？用山形分布的真值函数表示定位语义，可以得到和常识一致的概率预测，而用似然函数表示定位的语义不行

根据常识， 小车在高速公路上概率最大。如果直接用似然函数预测小车位置，小车在楼顶上的概率最大——这是不对的。用转移概率函数表示定位的语义，这个想法很好。但是转移概率函数是难以得到的， 特别是在定位有系统误差的时候。用猜想的转移概率函数， 并且忽略不必要的系数，那它就变成式(2.5)的真值函数了。根据语义贝叶斯公式，预测的小车的概率分布或似然函数是

$$P(X\mid\theta_j)=\frac{P(X)\exp[-(X-x_j)^2/(2d^2)]}{\sum_i P(X)\exp[-(X-x_j)^2/(2d^2)]} \tag{2.7}$$

这样就能利用先验知识，和贝叶斯推理兼容，得到和人脑推理一致的结论，不会认为小车在楼顶上概率最大。其中分子就是 $T(\theta_j)$，很像是热力学和统计学中的配分函数。

从 GPS 的例子可以看出，语义信道比 Shannon 信道简单。后面将说明：医学检验的语义信道也比 Shannon 信道更简单，更易于理解。

3 语义信息测度及语义信道(预测模型)优化

下面提供最大语义信息估计，它在本质上就是最大标准似然度估计，它将兼容最大似然估计和最大似然比估计，但是适合信源可变的场合。

3.1 用真值函数和标准似然度度量语义信息

在 Shannon 信息论中，只有统计概率，没有逻辑概率， 也没有预测的概率(似然度)。鲁晨光曾提供的语义信息测度同时用到这三种概率[10]。其中 y_j 提供关于 x_i 的信息量就是标准似然度的对数：

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.1}$$

其中用到语义贝叶斯公式，并假设先验似然度等于先验概率。对于无偏估计，真值函数和信息之间的关系如图 2 所示。

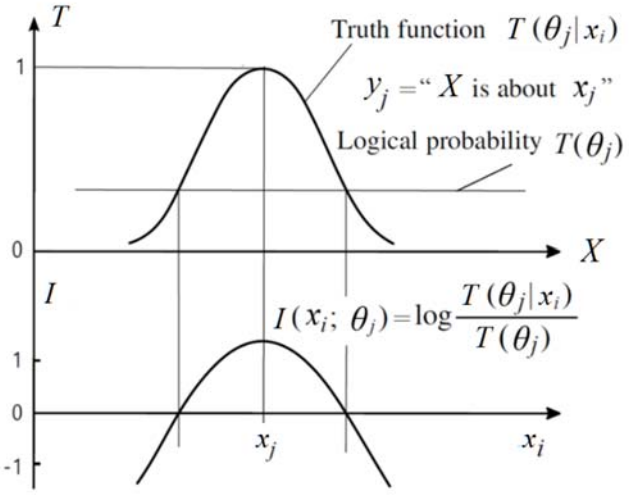


图 2 语义信息图解。偏差越大， 信息越少；逻辑概率越小，信息量越大；错误预测提供的信息是负的。

对 $I(x_i; \theta_j)$求平均，就得到广义或语义 Kullback-Leibler(KL)公式(后面简称语义 KL 公式)：

$$I(X;\theta_j)=\sum_i P(x_i \mid h_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{3.2}$$

对数左边的是统计概率 $P(x_i|y_j)$，$i=1, 2, \ldots$，它们构成样本概率分布 $P(X|y_j)$(注意：样本分布也是有条件的), 是用以检验 θ_j 的。如果 $y_j=f(Z|Z\epsilon C_j)$, 那么就有 $P(X|y_j)=P(X|Z\epsilon C_j)$(后面简记为 $P(X|C_j)$)。

虽然 Akaike 揭示了似然度和 KL 信息之间的联系[4] ，但是广义 KL 信息和似然度之间的关于更加简单。KL 信息可以写成相对熵 $H(P(X)||P(X|y_j))$， 意味用 $P(X)$取代 $P(X|y_j)$失去的信息，广义 KL 信息意味失去信息的减量(即增加的信息)；随 likelihood 增大而增大。而根据 Akaike 解释，KL 信息随 likelihood 增大而减小。所以广义 KL 信息更符合我们的信息观念。它可能是负的， 正好表示错误预测或谎言会减少语义信息或似然度。

对 $I(X;\theta_j)$求平均，就得到广义或语义互信息公式：

$$\begin{aligned} I(X;\Theta)&=\sum_j P(y_j)\sum_i P(x_i \mid h_j)\log\frac{P(x_i|\theta_j)}{P(x_i)} \\ &=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}=H(X)-H(X \mid \Theta) \\ H(X \mid \Theta)&=-\sum_j\sum_i P(x_i,y_j)\log P(x_i|\theta_j) \end{aligned} \tag{3.3}$$

其中 $H(X)$是 X 的 Shannon 熵，Θ 是一组模型 $\theta_1, \theta_2, \ldots, \theta_n$ 中一个。$H(X|\Theta)$是 X 的广义后验熵. 如果 $P(X)$是根据预测得到的，记为 $Q(X)$(后面用到), 则相对熵或 KL 距离

$$\begin{aligned} H(Q \| P)&=\sum_i P(x_i)\log[P(x_i)/Q(x_i)=H_\Theta(X)-H(X) \\ H_\Theta(X)&=-\sum_i P(x_i)\log Q(x_i) \end{aligned} \tag{3.4}$$

其中 $H_\Theta(X)$是 X 的广义熵。容易证明，每一种广义熵都大于或等于它所对应的 Shannon 熵。仅在预测和统计一致时，两者相等。

假设相应 y_j 有 N_j 个样本，它们来自 N_j 个独立同分布随机变量，其中 N_{ij} 个是 x_i, 当 N_j 无穷大时，就有 $P(X|y_j)= N_{ij}/N_j$。因此就有 log(标准似然度)：

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(e_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_j I(X;\theta_j) \tag{3.5}$$

对不同的 $y_j, j=1, 2, \ldots, n$ 求平均，就得到平均 log(标准似然度)，它和语义互信息的关系：

$$N\sum_j \frac{N_j}{N}\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}} = N\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}$$
$$= NI(X;\Theta) = NH(X) - NH(X\mid\Theta) \tag{3.6}$$

可见最大似然准则等价于最小广义后验熵准则。因为优化模型 θ_j 或 Θ 时，$P(X)$不变，所以最大语义信息准则也等价于最大似然准则。容易证明，Shannon 互信息是语义互信息在似然函数和样本分布符合时的特例，后者兼容前者。

3.2 语义信道优化

优化一个模型 Θ 等价于优化一个语义信道。给定 y_j 时优化 θ_j，也就是优化 $T(\theta_j|X)$, 于是有

$$T^*(\theta_j\mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) \tag{3.7}$$

$I(X;\theta_j)$可以写成两个 KL 距离的差，

$$I(X;\theta_j) = \sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i)} - \sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i\mid\theta_j)} \tag{3.8}$$

因为 KL 大于或等于 0，所以当

$$P(X|\theta_j)=P(X|y_j) \tag{3.9}$$

时，$I(X;\theta_j)$最大，等于 KL 信息 $I(X;y_j)$。 两边除以 $P(X)$得到

$$\frac{T(\theta_j\mid X)}{T(\theta_j)} = \frac{P(y_j\mid X)}{P(y_j)} \text{ 或 } T(\theta_j|X)\propto P(y_j|X) \tag{3.10}$$

令 $T(\theta_j|X)$的最大值等于 1，可以得到

$$T^*(\theta_j|X)=P(y_j|X)/P(y_j|x_j^*) \tag{3.11}$$

其中 x_j^*是使 $P(y_j|X)$达最大的 x_i。$P(y_j|X)$通常很难求解，但是有了样本分布 $P(X|y_j)$和 $P(X)$，求解 $T(\theta_j|X)$反而比求解 $P(y_j|X)$容易，因为：

$$T^*(\theta_j|X)= [P(X|y_j)/P(X)]/[P(x_j^*|y_j)/P(x_j^*)] \tag{3.12}$$

其中 x_j^*也是使 $P(X|y_j)/P(X)$达最大的 x_i.

在式(3.3)中，我们可以固定 $P(Y|X)$，通过改变 $T(X|\Theta)$最大化 $I(X;\Theta)$。这一过程可谓“语义信道匹配 Shannon 信道”。在 $P(X|\theta_j)=P(X|y_j)$或 $T(X|\Theta)\propto P(Y|X)$对所有 j 成立时，语义互信息 $I(X;\Theta)$达到其最大值， 并且等于 Shannon 互信息 $I(X;Y)$. 这时语义信道匹配 Shannon 信道。但是反过来，令 $P(y_j|X)\propto T(\theta_j|X)$(对所有 j)未必能增加 Shannon 互信息或语义互信息。给定语义信道，可能存在更好的 Shannon 信道， 传递更多的语义信息。后面谈及。

像最大后验估计——简称 MAP(Maximum A Prior)估计，MSI 方法也考虑先验。 不同的是 MAP 考虑 Θ 的先验，MSI 考虑的是 X 的先验。但是和 MAP 相比，MSI 估计更和贝叶斯推理兼容。上述模型优化方法可谓最大语义信息(MSI)方法。式(3.7)适合参数估计;，而等式(3.11)和(3.12)适合于大样本非参数估计，

4 Shannon 互信息和语义互信息的匹配函数 *R*(*G*)

R(*G*) 函数是信息论失真函数即 *R*(*D*)函数的推广。它最初是用来解决图像压缩问题的——如何根据视觉分辨率压缩图像。现在发现它可以用来解释两种信道相互匹配和 CM 算法。

4.1 从 *R*(*D*)函数到 *R*(*G*)函数

R(*D*)函数中 *R* 是信息率， *D* 是平均失真上限， *R*(*D*)表示给定失真上限时 Shannon 互信息的最小值。*R*(*G*)函数类似，*G* 是语义信息或平均 log(标准似然度)的下限，*R*(*G*)表示给定 *G* 时 Shannon 互信息的最小值。信息率失真函数 *R*(*D*)的参数形式是[15] :

$$\begin{aligned} D(s) &= \sum_i \sum_j d_{ij} P(x_i) P(y_j) \exp(s d_{ij}) / \lambda_i \\ R(s) &= sD(s) - \sum_i P(x_i) \ln \lambda_i \end{aligned} \tag{4.1}$$

其中 $\lambda_i=\sum_j P(y_j)\exp(sd_{ij})$是配分函数，类似于语义贝叶斯公式中的逻辑概率。

现在我们用 $I_{ij}=I(x_i;\theta_j)=\log[T(\theta_j|x_i)/T(\theta_j)]$ 代替 d_{ij}。给定信源 *P*(*X*)和语义信道 *T*(*Θ*|*X*), 令 *G* 是语义平均信息 *I*(*X*; *Θ*)的下限，求 Shannon 互信息的最小值 *R*=*R*(*G*), 它被定义为：

$$R(G) = \min_{P(Y|X):I(E;\Theta)\geq G} I(X;Y) \tag{4.2}$$

其中 *I*(*X*; *Y*)是 Shannon 互信息。仿照带参数 *s* 的 *R*(*D*)函数的推导过程[16]，那么我们得到 *R*(*G*)函数的参数表示[11][12]:

$$\begin{aligned} G(s) &= \sum_i \sum_j I_{ij} P(x_i) P(y_j) 2^{sI_{ij}}) / \lambda_i = \sum_i \sum_j I_{ij} P(x_i) P(y_j) m_{ij}^s / \lambda_i \\ R(s) &= sG(s) - \sum_i P(x_i) \ln \lambda_i \end{aligned} \tag{4.3}$$

其中 $m_{ij}=T(\theta_j|x_i)/T(\theta_j)=P(x_i|\ \theta_j)/P(x_i)$是标准似然度，$\lambda_i.=\sum_j P(y_i)\ m_{ij}^s$。我们也可以用 $m_{ij}= P(x_i|\ \theta_j)$，它导致相同的 m_{ij}^s/λ_i. *R*(*G*)函数曲线的形状是碗形的，如图 3 所示.

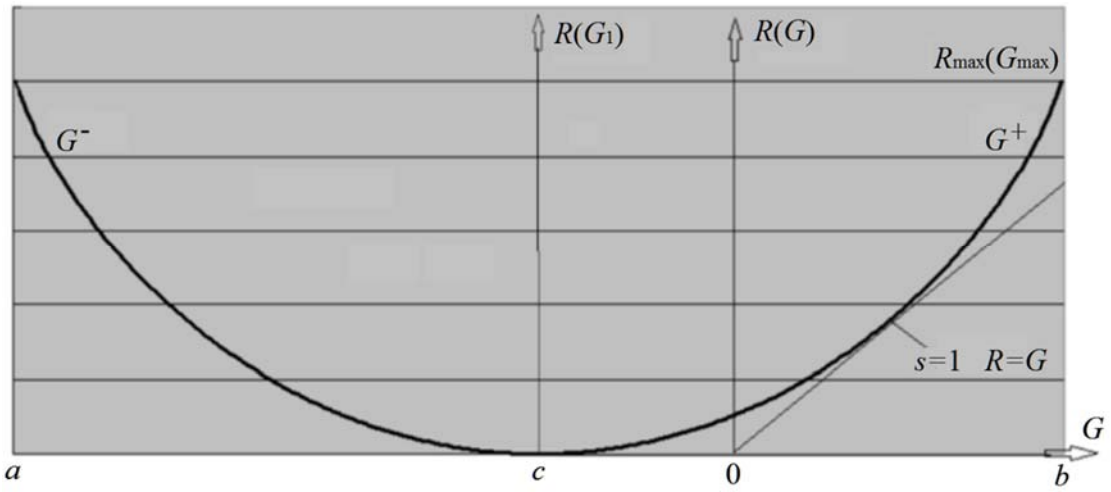


图 3 二元信源的 $R(G)$函数。$P(x_0)=P(x_1)=0.5$。$s=1$ 时 $R=G$，意味着语义信道匹配 Shannon 信道; 右上角高点意味着 Shannon 信道匹配语义信道，R 和 G 同时达最大。

和 $R(D)$函数不同， 给定 R 时， G 有最大值， 也有最小值。最小值是负的，意味着要造成敌人最大信息损失， 也必须有相应的客观信息 R。当 $R=0$ 时， G 是负的， 意味着，听信随机预测， 会减少我们已有的信息。

在信息论失真理论中，$\mathrm{d}R(s)/\mathrm{d}D(s)=s(s\leqslant 0)$。容易证明， 也有 $\mathrm{d}R(s)/\mathrm{d}G(s)=s$。但是 s 可以小于 0， 也可以大于 0。s 增大提高预测模型的正确性和精度，等于减小条件概率或转移概率分布的标准偏差。

s 由正的 s_1 变位-s_1 时， $R(-s_1)=R(s_1)$且 G 从语义互信息的上限 G^+变为下限 G^-(见图 3)。

当 $s=1$ 时，$\lambda_i=1$，$R=G$；意味着语义信道匹配 Shannon 信道，两种信息相等。

当 $s=0$ 时，$R=0$， $G(s=0)=\sum_i\sum_j I_{ij}P(e_i)$。我们记 $c=G(s=0)$。

我们以二元信源为例说明 $R(G)$函数。假设 $P(x_0)=P(x_1)=0.5$ 且

$$I_{ij}=\begin{cases} b>0, & i=j \\ a<0, & i\neq j \end{cases}$$

仿照二元信源信息率失真函数推导过程[13,16]， 可得

$$\begin{aligned} R(G)&=\frac{b-G}{b-a}\log(b-G)+\frac{G-a}{b-a}\log(G-a)-\log(b-a)+H(X) \\ &=\frac{h-G_1}{2h}\log(h-G_1)+\frac{h+G_1}{2h}\log(h+G_1)-\log h=H(X)-H(\frac{h-G_1}{2h}) \end{aligned} \tag{4.4}$$

其中 $h=(b-a)/2$, $c=(a+b)/2$, $G_1=G-c$。$R(G)$函数图形如图 3 所示, 其中假定 $T(\Theta_1|x_1)=T(\Theta_0|x_0)=1$ 且 $T(\Theta_1|x_0)=T(\Theta_0|x_1)=0.2$。于是 $b=0.737$ 比特, $a=-1.585$ 比特, $c=-0.424$ 比特。

我们可以把 $r=G/R$ 定义为信息效率，在 $s=1$ 即 $P(X|\theta_j)=P(X|y_j)$，$j=1, 2, \ldots, n$ 时，r 最大值是 1。

实际上，在信息论失真理论，如果允许有意制造更大的失真 D，那么 $R(D)$函数形状也是碗状曲线。这意味着，我们可以借助 $R(D)$函数优化假情报，造成更大失真从而迷惑敌人。

4.2 从 $R(G)$函数看医学检验的最大互信息和最大平均似然比

对于医学检验(参看图 4)，这时 $A=\{x_0, x_1\}$， $B=\{y_0, y_1\}$。其中 x_0 是真没病者， x_1 是真有病者；y_0=检验呈阴性，y_1=检验呈阳性。

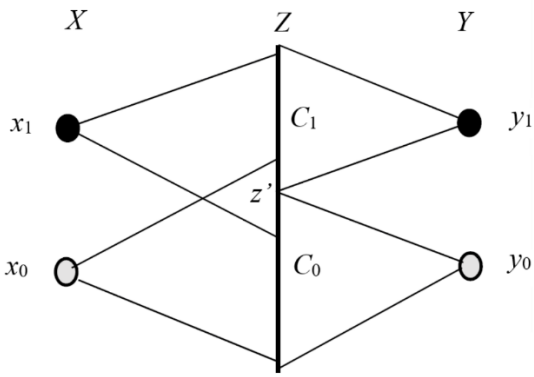


图 4 医学检验图解(二元 Shannon 有噪声信道，互信息且随判决分界点 z'改变)

医学检验中把 x_1 检验为阳性的概率叫做敏感性(sensitivity), 把 x_0 检验为阴性的概率叫做特异性(specificity) [17]. 检验的敏感性和特异性构成 Shannon 信道，如表 1 所示。

表 1 医学检验的敏感性和特异性构成 Shannon 信道 $P(Y|X)$

	真有病 x_1	真没病 x_0
检验是阳性 y_1	$P(y_1\|x_1)$=敏感性	$P(y_1\|x_0)$=1-特异性
检验是阴性 y_0	$P(y_0\|x_1)$=1-敏感性	$P(y_0\|x_0)$=特异性

如果我们相信阳性表示绝对有病，阴性表示绝对无病，那么就有 $T(y_1|x_1)= T(y_0|x_0)=1$, $T(y_1|e_0)= T(y_0|x_1)=0$。但是采用这样的语义信道，有一个反例存在，就会有负无穷大信息。为此，我们需要考虑预测和检验的置信度(confidence level)——用 b 表示，并且用 b'=1-|b|表示不信度(no-confidence –level,即显著性水平)。y_j 的其真值函数可定义为(假设 b>0)：

$$T(\theta_j|X)= b' + bT(y_j|X) \tag{4.5}$$

设阳性 y_1 的置信度 b_1,不信度是 b_1'; 阴性 y_0 的置信度是 b_0, 不信度是 b_0'.则医学检验的语义信道如表 2 所示。

表 2 医学检验的语义信道——含有两个不信度或显著性水平 b_1' 和 b_0'

	真有病 x_1	真没病 x_0
检验是阳性 y_1	$T(y_1\|x_1)$=1	$T(y_1\|x_0)= b_1$'
检验是阴性 y_0	$T(y_0\|x_1)= b_0$'	$T(y_0\|x_0)$=1

根据式(3.9), 两个优化的不信度是

$$b_1'^*= P(y_1|x_0)/P(y_1|x_1); \quad b_0'^*= P(y_0|x_1)/P(y_0|x_0) \tag{4.6}$$

医学界用似然比 LR^+信道 LR^-表示检验有多好. 公式(4.6)来自最大语义信息检验， 它和最大似然比检验[17]是兼容的，因为：

$$\mathrm{LR}^{+}=P(y_1|x_1)/P(y_1|x_0)=1/b_1'^{*}=1/(1-b_1^{*});\quad \mathrm{LR}^{-}=P(y_0|x_0)/P(y_0|x_1)=1/b_0'^{*}=1/(1-b_0^{*}) \tag{4.7}$$

其中 b_1^*是优化的 y_1 的置信度，它和 LR^+是一一对应的。但是 b_1^*在 0 和 1 之间变化，更加易于理解。b_0^*同理。

Thornbury 等人曾提出用 LR 做贝叶斯推理[17]. 不信度用于语义贝叶斯推理更方便。比如[1]，对于一种 HIV 检验，敏感性是 $P(y_1|x_1)=0.917$, 特异性是 $P(y_0|x_0)=0.999$, 则优化的 b_1'是 $b_1'^{*}=0.0011$(根据式(4.6))。假设受试者来自普通人群的随机抽查，$P(x_1)=0.002$, 则根据语义贝叶斯公式有

$$P(x_1|y_1)=0.002/(0.002+0.0011*0.998)=0.65;$$

假设受试者来自同性恋人群，$P(x_1)=0.1$， 则

$$P(x_1|y_1)=0.1/(0.1+0.0011*0.99)=0.991。$$

但是如果直接用阳性的似然函数 $P(X|\Theta_1)$预测 HIV 感染率，在 $P(x_1)$变化时，过去得到的似然函数就不再适合了。

最大似然比被用作检验优化。考虑 y_1 和 y_0 被选择的概率不同， 分别是 $P(C_1)$和 $P(C_0)$，则总的似然比是：

$$r_L=\left[\prod_{i=0}^{1}\left(\frac{P(x_i\,|\,\theta_1)}{P(x_i\,|\,\theta_0)}\right)^{P(x_i|C_1)}\right]^{NP(C_1)}\left[\prod_{i=0}^{1}\left(\frac{P(x_i\,|\,\theta_0)}{P(x_i\,|\,\theta_1)}\right)^{P(x_i|C_0)}\right]^{NP(C_0)} \tag{4.8}$$

这里没有考虑显著性水平限制。根据式 $R(G)$函数和语义互信息公式(3.3), $\max(\log r_L)=N(G^+-G^-)$(参看图 3 中的 G^+和 G^-)。因为 R 和 G^+ 确定好时，s 和 G^-也就确定了，所以最大似然比准则和最大似然准则(或最大语义信息准则)是等价的。

一个二元 Shannon 信道可能是无噪声的， 使得 R 的最大值是 $R_{\max}=H(X)$。后面检验的例子中， 噪声是不可避免的，为了求 R 最大值 $R_{\max}$——一般小于 $H(X)$，只能使用迭代方法。

5 信道匹配(CM)算法用于检验和估计

本节将介绍 CM 算法用于检验和估计，用 $R(G)$函数解释迭代收敛，并提供几个例子显示迭代过程和收敛速度。

5.1 用 $R(G)$函数解释信道匹配和迭代收敛过程

匹配 I (Right-step): 语义信道匹配 Shannon 信道

在 Shannon 信道不变时，我们改变语义信道使之匹配 Shannon 信道，即对每个 y_j, 使 $P(X|\theta_j)=P(X|y_j)$ 或 $T(\theta_j|X)\propto P(y_j|X)$，以至于语义互信息 $I(X;\Theta)$达到其上限 Shannon 互信息 $I(X;Y)$。目的是使 $G=R$，即使坐标点(R, G)位于 $R(G)$曲线和直线 $R=G$ 相切的地方($s=1$)。细节见图 5.

匹配 II (Left-step): Shannon 信道匹配语义信道

[1] https://arxiv.org/abs/1609.07827

语义信道 $T(\Theta|X)$不变时，我们改变 Shannon 信道 $P(Y|X)$使之匹配语义信道，从而使 Shannon 互信息 $I(X;Y)$和语义互信息 $I(X;\Theta)$同时达最大(达到 $R(G)$函数的右上方高点)。在此过程中，$P(Y|X)$匹配 $T(\Theta|X)$. $R(G)$函数提醒我们，可以通过增大参数 s 提高 R 和 G。匹配时(G,R)位于图 5 中一个 $R(G)$函数曲线的右上角。

匹配 III: 语义信道和 Shannon 信道相互匹配和迭代

依次采用匹配 I(R-step)和匹配 II (Left-step)，即迭代，使 R 和 G 同时达到最大(参看图 5)。用似然度的语言来说，就是预测模型 Θ 和 Y 的选择规则依次相互匹配，实现最大平均对数似然度。可以说，一个语义信道来自一个旧的 Shannon 信道，并且可以成为一个阶梯，让我们得到更好的 Shannon 信道； 通过新的更好的 Shannon 信道又可以得到一个更好的语义信道。这个过程可以不断重复。

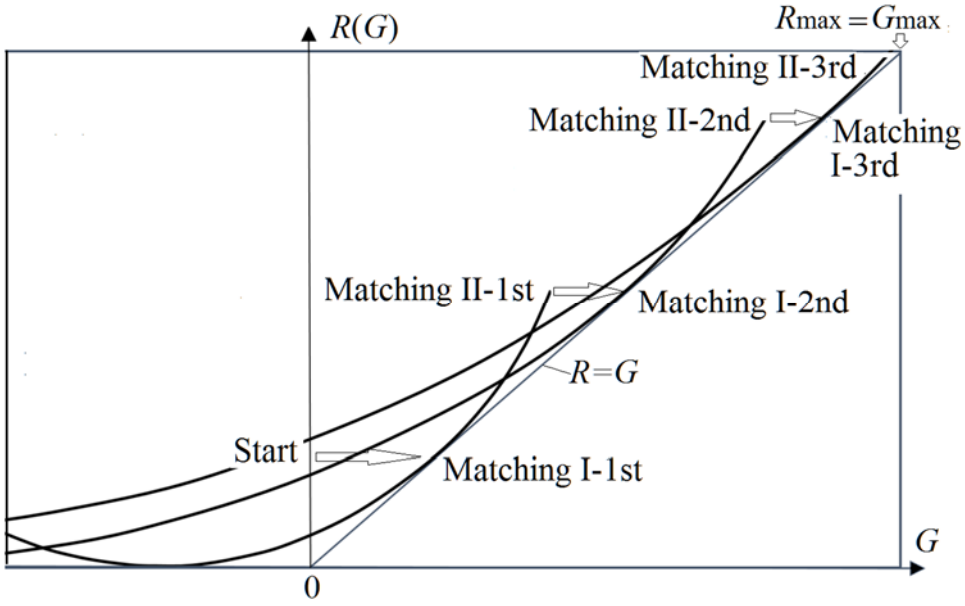


图 5 语义信道和 Shannon 信道相互匹配实现最大互信息和最大似然度。Matching I 找到 $R=G$ 的点， 而 Matchng II 找到 $R(G)$函数的最高点; 重复匹配可以得到最大的 R 和 G。

5.2 检验和估计的三个迭代实例

CM 算法用于检验和估计，比用于混合模型简单。迭代收敛可以从 $R(G)$函数看出，不再证明。下面通过具体例子说明收敛。对于图 5 所示检验，优化 Shannon 信道就是优化判决分界点 z', 当 Z>z'时，我们判断 $Y=y_1$=阳性, 否则判断 $Y=y_0$=阴性。

举例说，假定 $Z\in C=\{1, 2, \ldots, 100\}$。给定 x_1 和 x_0 时，$P(Z|X)$是高斯分布：

$$P(Z|x_1)=K_1\exp[-(Z-c_1)^2/(2d_1^2)], \quad P(Z|x_0)=K_0\exp[-(Z-c_0)^2/(2d_0^2)]$$

其中 K_1 和 K_0 是归一化常数。从 $P(X)$和 $P(Z|X)$, 可以算出样本分布 $P(X|Z)$。假定开始的划分点是 z', 比方说 $z'=50$, 做下面迭代运算：

右步骤(Matching I)：计算下面各项：

1)构成 Shannon 信道的 4 个转移概率：

$$P(y_0 \mid x_0) = \sum_{z_k=1}^{z'} P(z_k \mid x_0), \qquad P(y_1 \mid x_0) = 1 - P(y_0 \mid x_0)$$

$$P(y_1 \mid x_1) = \sum_{z_k=z'+1}^{100} P(z_k \mid x_1), \qquad P(y_0 \mid x_1) = 1 - P(y_1 \mid x_1)$$

2) 两个不信度 b_1'* 和 b_0'*(根据式(4.6));

3) 逻辑概率 $T(\theta_1)=P(x_1)+b_1$'$*P(x_0)$ 和 $T(\theta_0)=P(x_0)+b_0$'$*P(x_1)$;

4) $I_{ij}=I(x_i; \theta_j)$，对于 i=0, 1 和 j=0, 1;

5) 给定不同 Z 时的平均语义信息 $I(X; \theta_1|Z)$ 和 $I(X; \theta_0|Z)$ (显示为两条曲线):

$$I(X;\theta_j|z_k) = \sum_i P(x_i \mid z_k) I_{ij}\text{，} \ j=0, 1; \ k=1, 2, \ldots, 100 \tag{5.1}$$

左步骤(Matching II)：比较上述两根信息曲线， 可以发现两个曲线的交叉点， 用这一点做新的 z'. 如果它和上个 z'相同，则令最优分界点 z*=z', 迭代结束；否则转到**右步骤。**

下面是三个计算实例的报告，第二和第三个实例有两个分界点 z_1'和 z_2', 迭代原理相同。

迭代实例 1(2×2 Shannon 信道)

输入数据: $P(x_0)$=0.8; c_0=30, c_1=70; d_0=15, d_1=10。分界起点 z'=50。

迭代过程: 匹配 II-1 得到 z'=53; 匹配 II-2 得到 z'=54; 匹配 II-3 得到 z*=54. 用 $I(x_0;Z)$和 $I(x_1;Z)$的交叉点做起点，结果是一样的。

比较：$P(X)$=0.72 bit; $I(X; Z)$=0.55 bit; $I(X; Y)$=0.47 bit.

分析：如果使用最小误差准则， 则最优划分点是 57，而上面 z*=54。可以看出，和最小误差准则相比，最大互信息准则准则更注重小概率事件的正确预测， 允许更多的假阳性和较少的假阴性.

迭代实例 2 (2×3 Shannon 信道)

为了减少误判造成的信息损失，我们在上个例子中增加一个输出 y_2, 其语义是“检验无效”。当 z_1'< $Z \le z_2$'(误判较多)时,令 $Y=y_2$.

输入数据: $P(x_0)$=0.8; c_0=30, c_1=70; d_0=15, d_1=10。开始分界点是 z'_1=50 ， z_2'=60.

迭代过程：匹配 II-1 得到 z_1'=46，z_2'=57; 匹配 II-2 得到 z_1'=47，z_2'=59; 匹配 II-3 得到 z_1*=47，z_2*=59.

比较：$H(X)$=0.72 比特; $I(X; Z)$=0.55 比特; $I(X; \Theta)=I(X; Y)$=0.52 比特。然而在实例 1 中 $I(X; Y)$=0.47 比特。所以此 2×3 信道传递信息比上述 2×2 信道多。

这个例子也说明了，增加中性假设可以替代显著性水平限制。

迭代实例 3 (3×3 Shannon 信道)

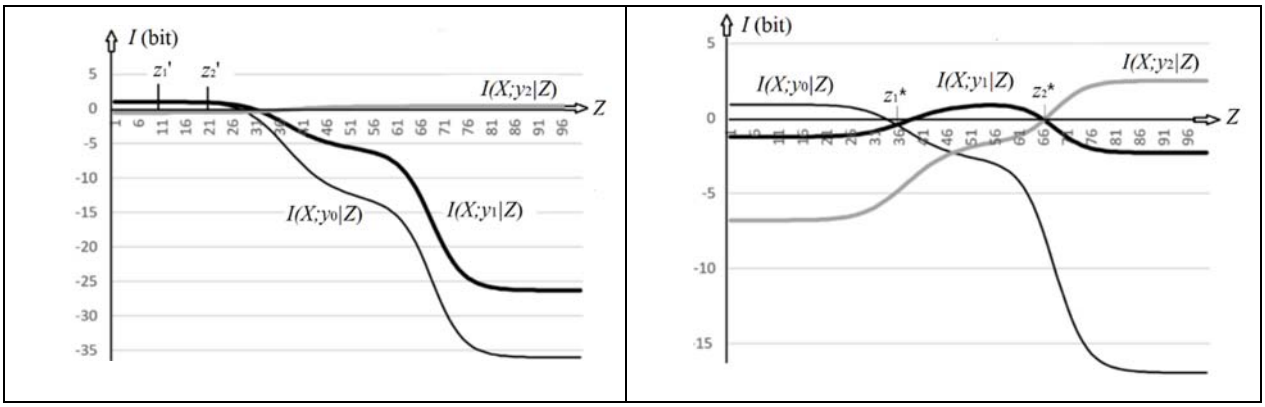
该实验用一个简化的估计测试迭代起点对速度和收敛的影响。分别一对较好的迭代起点和一对很坏的迭代起点做比较.

首先输入数据：$P(x_0)$=0.5, $P(x_1)$=0.35, $P(x_2)$=0.15; c_0=20, c_1=50, c_2=80; d_0=15, d_0=15, d_1=10, d_2=10。

迭代结果：

用好的分界起点 z_1'=50 和 z_2'=60, 迭代次数是 4; z_1*=35，z_2*=66。

用很差的分界起点 z_1'=9 和 z_2'=20, 迭代次数是 11; 也有 z_1*=35，z_2*=66。图 6 显示了迭代前后的信息曲线(正的区域大就好)。



(a) (b)

图 6 分界起点很差时的迭代。(a)显示迭代开始时，三条信息曲线正的部分很小；(b)显示了迭代收敛时三条信息曲线正的部分较大。该图说明迭代收敛是稳定的。

5.3 分析和讨论

以上三个例子显示 CM 算法用于检验和估计时，收敛是快速且必然的。一般迭代 3 到 5 次就收敛。上面例子中，Z 是一维的， 如果是多维空间中的矢量，运算会复杂些， 但是收敛速度应该类似。

在上面例子中没有用到优化 Shannon 信道的 $R(G)$函数中的参数 s, 是因为最优划分点已经包含了优化的 s 信息。不过 $R(G)$函数的参数形式提醒我们可以使用下面模糊决策(或分类)函数(考虑到信息效率 $G(s)/R(s)$:

$$P(y_j \mid z_k) = \frac{P(y_j)[\exp(I(X;\theta_j \mid z_k))]^s}{\sum_{j'} P(y_j)[\exp(I(X;\theta_{j'} \mid z_k))]^s} \quad , \quad j=1, 2, \ldots, n \tag{5.2}$$

当 $s\to\infty$, $P(y_j|Z)$就变成 C_j 的特征函数(在 C 被最优划分时)， 并且告诉我们最优划分点。即使 Z 是多维空间矢量，上式同样成立， 从而可以省去寻找边界的麻烦。

我们可以把 CM 算法用到一般预测， 比如天气预报， 和用于估计类似。只是真值函数更加多样化。这时候 CM 算法就可以解释语义进化。Shannon 信道反映语言用法，而语义信道反映听众理解方式. 语义信道匹配 Shannon 信道(Right-step)就是理解匹配用法；Shannon 信道匹配语义信道((Left-step)就是用法匹配理解。语义信道和 Shannon 信道相互匹配(迭代)，就是演讲者用法和听众理解相互匹配，相互促进。自然语言应该就是通过这种方式进化的。

如果样本分布不连续或不规则，两条信息曲线 $I(X;\theta_j|z_k)$和 $I(X;\theta_{j+1}|z_k)$会不会不止一个交点，会不会导致局部收敛？这需要进一步讨论。

6 CM 算法用于混合模型

本节将用 CM 算法求解混合模型， 借助于 $R(G)$函数证明迭代收敛， 用两个例子说明迭代过程和收敛速度。然后比较 CM 算法和 EM 算法。

6.1 用 $R(G)$函数解释迭代过程

假设样本分布 $P(X)$是某种条件概率函数(比如高斯分布)$P^*(X|Y)$产生的。我们只知道模型构件 Y 是 n 个，但是不知道真的 $P(Y)$， 记为 $P^*(Y)$. 要求的是接近 $P^*(Y)$的 $P(Y)$和接近 $P^*(X|Y)$的 $P^*(X|\Theta)$。和检验不同，检验中的 Z 和 X 现在合并成 X. 预测不再有对错，但是要求预测的样本分布——记为 $Q(X)$——和 $P(X)$尽可能接近， 即相对熵 $H(Q||P)$尽可能小。

我们再分别用 $P^*(Y|X)$和 $R^*=I^*(X;Y)$表示相应的 Shannon 信道 $P(Y|X)$和 Shannon 互信息 $I(X;Y)$。当 $Q(X)=P(X)$时，应有 $P(X|\Theta)=P^*(X|Y)$，$G=G^*=R^*$。求解最大似然混合模型和求解最大似然检验和估计不同，在 Shannon 信道匹配语义信道(即 Left-steps)的时候，我们并不求最大互信息 $I(X;\Theta)$，而是求尽可能和 $P^*(X|Y)$一致的 $P(X|\Theta)$(这是图 7 中 Left-step a 的目的), 以及尽可能和 $P^*(Y)$一致的 $P(Y)$(这是图 7 中 Left-step b 的目的)，也即寻求和 R^*接近的 R。而流行的 EM 算法及其变种，总是追求和证明某个似然度(比如 $\log P(X^N,Y|\Theta)$)在两种步骤中都只增不减。

仅当最优模型求出后， 如果我们需要根据 X 选择 Y(判决或分类)的时候， 我们才求产生 $R_{max}(G_{max})$的 Shannon 信道(见图 7 中 Left-step c)。

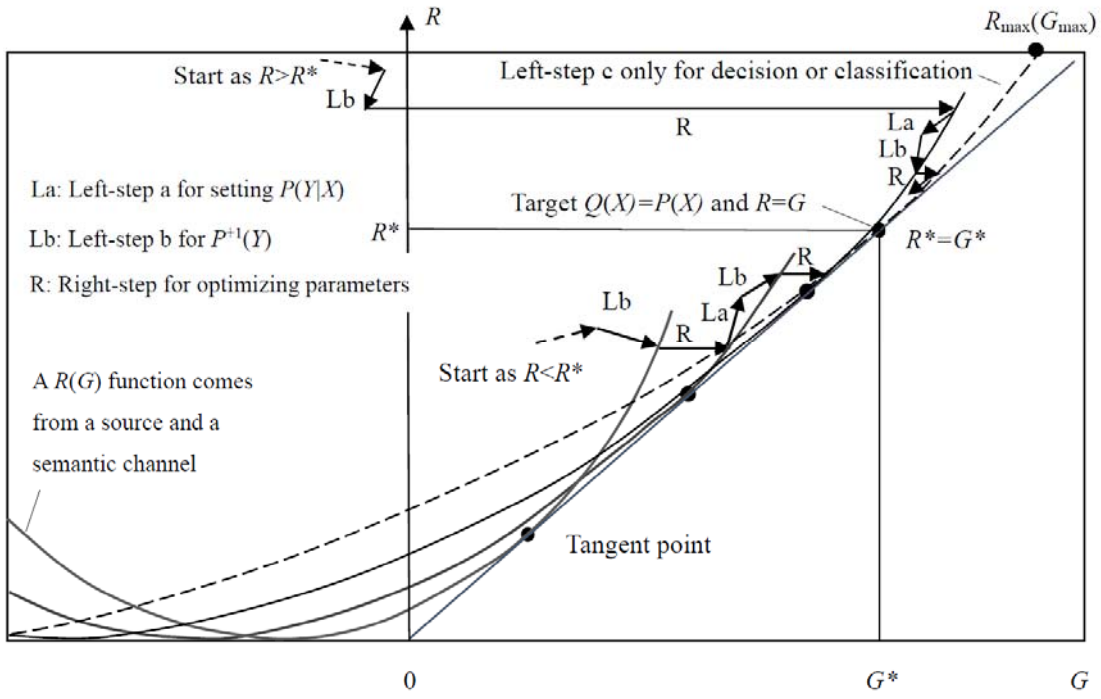


图 7 用 CM 算法求解混合模型。其中显示了两个迭代例子(分别是 $R<R^*$和 $R>R^*$时的例子)。Left-step a 和 Left-step b 使得 R 接近 R^*(意味(G, R)纵向改善)， 而 Right-step 增大 G， 使得(G, R)接近直线 $R=G$(意味(G, R)横向改善).

如果我们猜测 $P(X)$是 n 个高斯分布函数产生的， 则似然函数是:

$$P(X|\theta_j)= k_j \exp[-(X-c_j)^2/(2d_j)^2], j=1,2,\ldots, n$$

设 n=2，则参数是 c_1, c_2, d_1, d_2。我们不妨设 $P(y_1)=P(y_2)=0.5$。匹配从 Left-step a 开始。

Left-step a: 令转移概率函数是通过似然函数 $P(X|\Theta)$和 $P(Y)$产生的：

$$P(y_j \mid X) = P(y_j)P(X \mid \theta_j) / Q(X), \qquad Q(X) = \sum_j P(y_j)(X \mid \theta_j)，\ j=1, 2, \ldots, n \tag{6.1}$$

EM 算法[8]已经使用了这一公式。这一公式也是 $R(D)$和 $R(G)$函数推导过程中用到的[16]。

于是语义互信息是：

$$I(X;\Theta) = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)} \tag{6.2}$$

Left-step b：求新的 $P(Y)$, 即重复用下式迭代求

$$P(y_j) \Leftarrow \sum_i P(x_i)P(y_j \mid x_i) = \sum_i P(x_i) \frac{P(x_i \mid \theta_j)}{\sum_k P(y_k)P(x_i \mid \theta_k)} P(y_j)，\quad j=1, 2, \ldots, n \tag{6.3}$$

求得的 $P(Y)$记为 $P^{+1}(Y)$。这是因为式(6.1)中的 $P(Y|X)$并不是合格的 Shannon 信道，它不能保证$\Sigma_i P(x_i)P(y_j|x_i)=P(y_j)$。上面迭代也是为了使 $P(Y)$和 $P(X)$匹配更好。Byrne 已经将这一迭代用于改进的 EM 算法[18].

n=2 时，c_1 和 c_2 的选择要避免 $P(y_1)$和 $P(y_2)$两个同时小于或大于 $P(X)$分布的平均值(免得出现 $P(y_1)$=0 或 $P(y_2)$=0，再调大困难)。n>2 时，不妨在迭代开始几次避免每个 $P(y_j)$太小，比如小于 $0.1/n$ 时就停止迭代。只要有一次找到 $P^{+1}(y_j)=P(y_j)\neq 0$(对所有 j), 我们就应找到(G, R)收敛到(G^*, R^*)的途径。以后应允许某个 $P(y_j)$=0。

如果 $H(Q||P)$小于一个很小的数， 比如 0.001 比特，则迭代结束；否则继续。

Right-step: 优化模型参数。即改变 log 右边似然函数中的参数. 使 $I(X; \Theta)$最大。转到 Left Step a.

关于 Left-step c 模型 $P(X|\Theta)$优化后，或许我需要根据 X 选择 Y(决策或分类)， $R(G)$函数中的参数 s 提醒我们可以采用下面模糊决策(或分类)函数

$$P(y_j \mid X) = P(y_j)[P(X \mid \theta_j)]^s / Q(X), \qquad Q(X) = \sum_j P(y_j)[P(X \mid \theta_j)]^s，\ j=1, 2, \ldots, n \tag{6.4}$$

其中 $P(y_j|X)$类似于 logistic 函数。当 $s\rightarrow+\infty$时，模糊决策就变为清晰决策。和 MAP(Maximum A Prior)决策不同，上述决策仍然坚持最大似然准则。也即最大互信息准则。R 和 G 增加如图 7 中 Left-step c 所示。

6.2 用两个例子说明迭代过程

6.2.1 迭代实例 1 ($R<R^*$)

表 3 中含有产生样本分布 $P(X)$的 $P^*(X|Y)$中的真实参数和产生 $Q(X)$的 $P(X|\Theta)$中的参数. 从开始的(G, R)收敛到(G^*, R^*)的过程如如图 7 中 $R<R^*$的迭代所示，迭代过程如图 8 所示，迭代结果如图 9 和表 3 所示。

表 3 $R<R^*$时的模型参数和迭代结果(Right-steps 5 次)

Parameters	Real $P^*(X\|Y)$ and $P^*(Y)$			Starting paremeters $H(Q\|\|P)$=0.410 bit			Ending parameters $H(Q\|\|P)$=0.00088 bit		
	c	d	$P^*(Y)$	c	d	$P(Y)$	c	d	$P(Y)$
y_1	35	8	0.7	30	15	0.5	35.4	8.3	0.720
y_2	65	12	0.3	70	10	0.5	66.2	11.4	0.280

*The number of iterations is 5 (5 Right-steps and 6 Left-step b's)

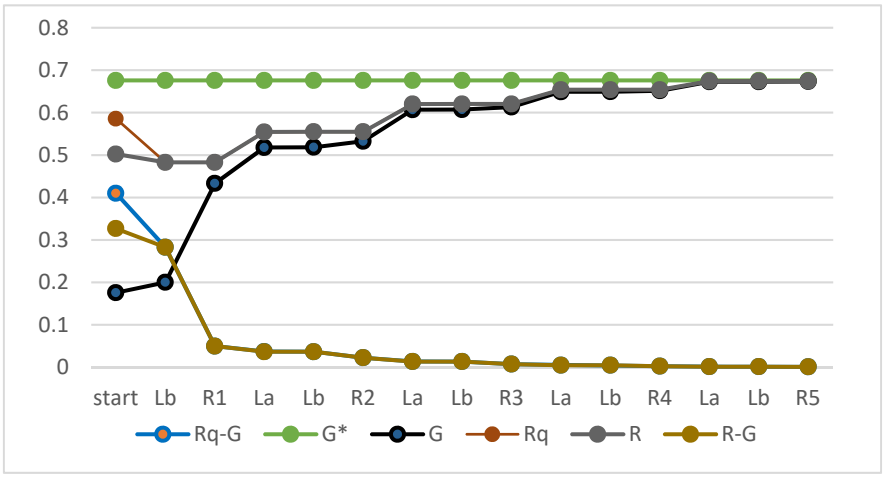


图 8 $R<R^*$时的迭代过程。其中 Rq 是式（6.4）中的 R_Q. G 是单调增加的，除了第一个 left-step b, R 也都是单调增加的。 R 和 G 不断增加并逐渐接近 $R^*=G^*$使得 $H(Q||P)=R_Q$-G 接近 0。

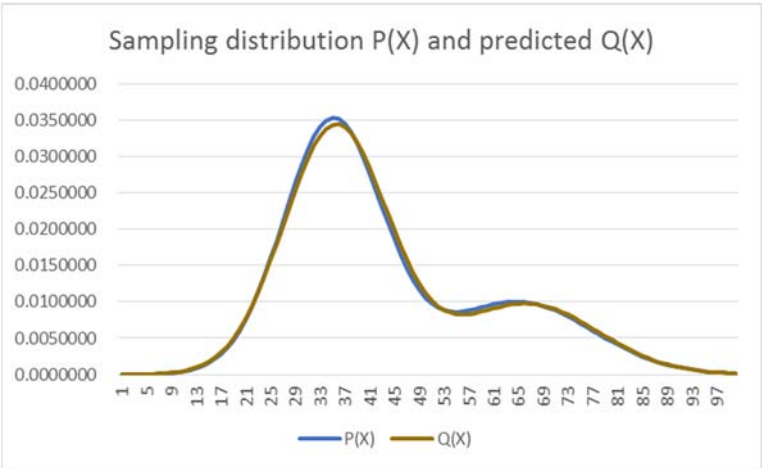


图 9 样本分布 $P(X)$和模型预测的分布 $Q(X)$比较(5 次迭代后)

分析：5 次 Right-step 之后，Θ 中的参数就已经接近真实参数，所产生的分布 $Q(X)$和真值分布比，差别也很小。但是这个过程很容易让人认为：上述三种步骤，每一步之后，G 都会增加，

最大化 G 会最小化 $H(Q||P)$——这是我们的目的。然而，这种想法是错误的，因为 Left a 和 Left b 并不一定增加 G。实际上还有大量反例存在。 幸运的是，这些反例最终也会收敛——迭代实例 2 将说明。

6.2.2 迭代实例 2 ($R>R^*$)

参数见表 4，迭代过程见图 10。

表 4 $R>R^*$时的模型参数和迭代结果(Right-step 5 次)

Parameters	Real $P^*(X\|Y)$ and $P^*(Y)$			Starting paremeters $H(Q\|\|P)$=0.680 bit			Ending parameters $H(Q\|\|P)$=0.00092 bit		
	c	d	$P(Y)$	c	d	$P(Y)$	c	d	$P(Y)$
y_1	35	8	0.1	30	8	0.5	38	9.3	0.134
y_2	65	12	0.9	70	8	0.5	65.8	11.5	0.866

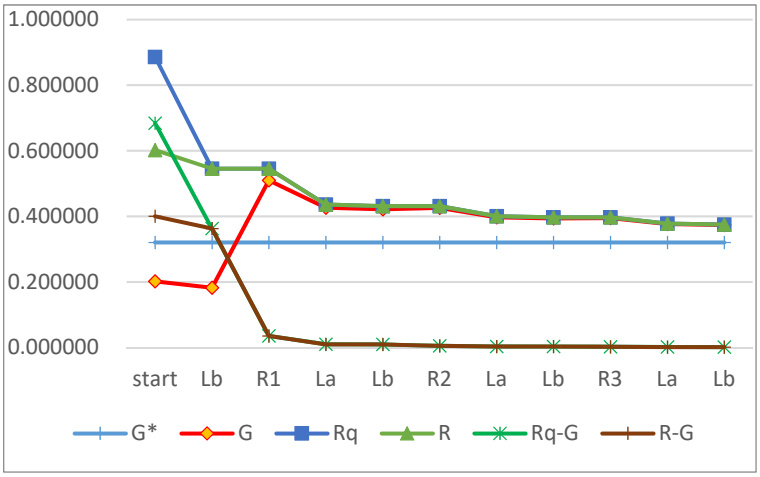


图 10 $R>R^*$时的迭代过程。其中 Rq 是式（6.4）中的 R_Q. $H(Q||P)=R_Q-G$ 在所有步骤中都是减小的。 R 单调减小，逐步接近 R*；在第一个 Right-step 后，G 在 Left-steps 减小，在 Right-steps 微量增加。

分析: G 不是单调增大的也不是单调减小的， 它在所有 Left-step 减小， 在所有 Right step 增大。这个例子是对所有证明 EM 算法或其变种收敛的作者是一个挑战。

6.3 CM 算法收敛证明和解释

证明 要证明 CM 算法收敛，我们需要证明 $H(Q||P)$在三种步骤中的每一步都是减小或非增的。

考虑 Right-step。设(6.1)所示 Shannon 信道传递关于 $Q(X)$的 Shannon 互信息是 R_Q ，关于 $P(X)$的 Shannon 的 Shannon 互信息是 R， 我们有

$$R_Q = I_Q(X;Y) = \sum_i \sum_j P(x_i)\frac{P(x_i|\theta_j)}{Q(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{Q(x_i)} \tag{6.5}$$

$$R = I(X;Y) = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)}$$
$$= \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{Q(x_i)} P(y_j) \log \frac{P(y_j \mid x_i)}{P^{+1}(y_j)} = R_Q - H(Y \| Y^{+1}) \tag{6.6}$$
$$H(Y \| Y^{+1}) = \sum_j P^{+1}(y_j) \log[P^{+1}(y_j) / P(y_j)]$$

比较关于 G 的式(6.2)和关于 R_Q 的式(6.4)可见

$$H(Q \| P) = R_Q - G = R + H(Y \| Y^{+1}) - G \tag{6.7}$$

在 Right-step, Shannon 信道和 R_Q 不变，语义信道或其参数变，G 的增量就等于 $H(Q||P)$的减量。如果 G 增大，则 $H(Q||P)$减小。

考虑 Left-step a。这时 $Q(X)$变为 $Q^{+1}(X)$。因为 $Q^{+1}(X)$是在 $P(Y)$不变，而似然函数 $P(X|\Theta)$改善后产生的，比 $Q(X)$更接近 $P(X)$, 所以 $H(Q^{+1}||P) \le H(Q||P)$，即 $H(Q||P)$是减小的。

考虑 Left-step b。这时我们通过迭代求 $P^{+1}(Y)$。结果是(G, R)移到 $P(X)$和 $P(X|\Theta)$(即语义信道)确定的 $R(G)$函数曲线上——从 $R(D)$函数[16]和 $R(G)$函数[12]的求解过程看出。这些过程也用类似迭代方法。在这个函数曲线上，$R(G)$是给定 G 时 R 的最小值，所以 R-G=R_Q-G 会减小。所以 $H(Q^{+1}||P) < H(Q||P)$。

在三种步骤中，$H(Q||P)$都会减小， 所以迭代收敛。 **证毕**。

下面我们通过图 7 进一步解释（不严格的证明）为什么迭代会使(G, R)收敛到(G^*, R^*)(全局收敛， 而不是局部收敛)。

要使(G, R)收敛到(G^*, R^*)(R^*=G^*), 我们需要横向改善——使(G, R)接近直线 R=G， 也需要纵向改善——R 接近 R^*。

在 Right-step 中，(G, R)向右移动，使得(G, R)接近 R=G 直线(而不是 G 接近 G^*)，实现横向改善同时纵向不变。

在 Left-step a 中，构成 $P(y_j|X)$的 $Q(X)$变为 $Q^{+1}(X)$，广义 KL 信息是

$$H(Q^{+1} / Q \mid P) = \sum_i P(x_i) \log \frac{Q^{+1}(x_i)}{Q(x_i)} = \sum_i [\sum_j P(x_i) P^*(x_i \mid y_j)] \log \frac{\sum_j P(y_j) P(x_i \mid \theta_j)}{\sum_j P(y_j) P^{-1}(x_i \mid \theta_j)} > 0 \tag{6.8}$$

比较不同 y_j 时的广义 KL 信息的平均

$$\Delta R_a = \sum_i \sum_j P(x_i) P^*(x_i \mid y_j) \log \frac{P(x_i \mid \theta_j)}{P^{-1}(x_i \mid \theta_j)} \tag{6.9}$$

可见 ΔR_a 和 $H(Q^{+1}/Q||P)$接近，所以也有 $\Delta R_a > 0$, 即 $P(X|\Theta)$比 $P^{-1}(X|\Theta)$更接近 $P^*(X|Y)$。所以 Left-step a 之后，R 更接近 R^*，实现(G, R)纵向改善。这一步 G 也可能减小。如果减小， 那是因为 G 接近 R，从而也接近 G^*。

在 Left-step b 中，迭代求 $P^{+1}(Y)$会使 $H(P_Y||P_Y^{+1})$变为 0。$R>R^*$时， R_Q 向 R 靠拢； $R<R^*$时，R_Q 向 R 靠拢，所以在两种情况下，R 都更接近 R^*。因为(G, R) 移到右边的 $R(G)$曲线上，更靠近直线 $R=G$。 所以$(G，R)$离$(G^*，R^*)$更近。纵向和横向都得到改善。G 在第一个 Left-step b 中例外——远离 G^*。这可能是因为 Start-step 没有改善 $Q(X)$。对此例外需要进一步研究。

综上所述，迭代会使(G, R)全局收敛于(G^*, R^*)。

6.4 CM 算法和 EM 算法比较

和已有的算法比，CM 算法和 EM 算法最为相似，比较两者可以加深我们对两者的理解。

6.4.1 两种算法的差别

在提出标准 EM 算法的 Dempaste， Laird 和 Rubin 的文章[8]和提供改进的收敛证明的 WU 的文章[18]中，混合模型的似然度被写成 $\log P(X^N|\Theta)\geq L=Q-H$，用上述语义信息方法叙述就是：

$$\begin{aligned}
&\log P(X^N \mid \Theta)=N\sum_i P(x_i)\log P(x_i \mid \Theta) =N\sum_i P(x_i)\log Q(x_i)\\
&\geq L = N\sum_i\sum_j P(x_i)P(y_i \mid x_i)\log\frac{P(x_i,y_j \mid \theta_j)}{P(y_i \mid x_i)}\\
&= N\sum_i\sum_j P(x_i)P(y_i \mid x_i)\log P(x_i,y_j \mid \theta_j) - N\sum_i\sum_j P(x_i)P(y_i \mid x_i)\log P(y_i \mid x_i)\\
&= Q - H
\end{aligned} \tag{6.10}$$

如果把 Q 中 $P(Y)$或 $P(Y|\Theta)$移到 H 中，Q 就变成$-NH(X|\Theta)$， H 就变成$-NR_Q$。不等式两边再加上样本分布的 Shannon 熵 $H(X)$, 那么就得到 $H(Q||P) \leq R_Q-G$, 它类似于式(6.7). 容易证明

$$Q=-NH(X,Y|\Theta)=NG-NP(X)-NH(Y) \tag{6.11}$$

其中 $H(Y)=-\sum_j P^{+1}(y_j)\log P(y_j)$是广义熵。

E-M 算法中的 E-step 和 CM 算法中的 Left-step a 相同(见式(6.1))。 而 EM 算法中的 M-step 将最大化负的广义联合熵$-H(X,Y|\Theta)= G-H(X)-H(Y)$(其中 $H(Y)$是广义熵)，也就是同时增大 G 并减小 $H(Y)$. 我们可以粗略地认为 M-step 是把 CM 算法中的 Left-step b 和 Right-step 合并成一步。即

E-step of EM =Left-step a of CM

M-step of EM≈Left-step b +Right-step of CM

Neal 和 Hinton 的变种 EM 算法[19]用 $F(P_Y, \Theta)=-H(X,Y|\Theta)+H(Y)$ (其中 $H(Y)$是 Shannon 熵而不是广义熵) 取代$-H(X,Y|\Theta)$作为优化的目标函数。$F(P_Y, \Theta)$和 CM 方法中的 $G=I(X; \Theta)$相似。但是这一 EM 算法在 E-step 和 CM 算法中的 Left-step 不同，它仍然最大化 $F(P_Y, \Theta)$，而 CM 算法在 left-step b 仅仅优化 $P(Y)$， 并不最大化 G. 比如在实例 2 中相反——在第一个 Right-step 之后 G 总是减小的。

还有其它改进的 EM 算法[7, 20-24], 都具有某些优点。但是像 CM 算法这样，考虑 R 收敛到 R^*和 $R-G$ 收敛到 0，是没有的。

6.4.2 EM 算法的收敛证明存在的问题

标准 EM 算法在 M-step 同时做 Lefte-step b 和 Right-step. 而 Left-step b 和 Right-step 目的完全不同。在有些情况下, 使 G 接近 G^*或 $P(X|Y, \Theta)$接近 $P^*(X|Y)$是不容易的。 比如在实例 2 中，$R^*=G^*$比较小，G 在第一个 Right-step 之后应该是下降的。因为 G 只有下降才能接近 G^*. 而对于实例 1， $H(Y)$应该是上升的， 因为 $H(Y)$只有上升才能接近 $H^*(Y)$. 而 M-step 最大化 Q 等于一律要求 G 上升 $H(Y)$下降。如果 EM 算法中的 M-step 像 CM 算法中 Left-b 那样先优化 $P(Y)$，再优化似然函数中的参数 Θ，那么它就和 CM 算法等价。 这也是为什么 EM 算法可以收敛的原因。

根据[8][19]作者的证明，M-step 增加 Q，E-step 也不降低 Q。Byrne[18]已经指出, 这些作者对“E-step 也不降低 Q”证明是有问题的。本文则得出结论：在 $R>R^*$时, “E-step 也不降低 Q”不利于收敛。实例 2 中真实的 $Q^*=-NH^*(X,Y)=-6.031N$。而第一次参数优化后，$Q=-6.011N>Q^*$。如果该证明对于实例 2 是真的，则不断最大化 Q 时，Q 将不会接近较小的 Q^*。如果 E-step 降低 Q，则他们的收敛证明不能成立。

Neal 和 Hinton 的变种 EM 算法[18]把优化 $P(Y)$从 M-step 移到 E-step，可以加快收敛速度。但是作者肯定 $F(P_Y, \Theta)$(类似于语义互信息 G)是一直上升的。这是有问题的。 对于实例 2，在 E-step 最大化 $F(P_Y, \Theta)$不利于 G 接近 G^*。

EM 算法用到 Jensen 不等式， 使得最大化 $\log P(X^N|\ \Theta)$ 变成最大化 $L=Q-H$。而 CM 算法没有用到 Jensen's 不等式. 原因是，CM 算法使用了样本分布 $P(X)$和子模型 θ_j (j=1, 2, …, n)，以及语义信息测度，得到公式 $H(Q||P)=R_Q-G$。在 Right-step 最大化 G 等价于最小化 $H(Q||P)$。

无论从理论是还是实际例子看，CM 算法都更加简单，收敛解释更加明了。但是其收敛证明还需要改进，使得它在数学上更严格。

6.4.3 收敛速度和模型应用

对于 n=2 的高斯混合模型，我们用不同真实参数测试 CM 算法达到收敛的迭代数。假设 $H(Q||H)\leq 0.001$ 是迭代收敛，则收敛需要的迭代数是 5 的情况最多；少数情况下也收敛缓慢(在 R-G 比 $|G^*-G|$小很多时)，迭代超过 30 次；所需迭代次数的中位数估计在 5 和 10 之间。

根据[18]，标准 EM 算法需要迭代 30 多次，改进的增量算法需要迭代大约 17 次。根据文献[19]，标准 EM 算法需要迭代大约 18 次；改进的 Multi-Set EM 算法需要迭代大约 12 次。在[7]中，作者比较了一种改进的 EM 算法和 Newton 算法，用某种距离标准衡量收敛；改进的 EM 算法需要迭代平均 17 次，Newton 算法需要平均 7 次；但是两者需要的计算时间相近。

看来 CM 算法较之标准 EM 算法收敛更快。但是要知道确切的迭代次数和耗时差别，我们还需要用相同的真实参数和起始参数做运算比较。对少数难以收敛的情况，如果 CM 算法适当结合各种改进 EM 算法的方法[18, 21-24]，比如优化初始参数的方法，这些情况下的收敛应能大大加快。

CM 算法得到的模型可以用于信源 $P(X)$可变场合。 根据式(3.12)， 用 $P(X)$和优化的 $P(X|\theta_j)\approx P^*(X|y_j)$($j$=1, 2, …, n), 可以得到真值函数 $T(A_j|X)$——它类似于 logistic 函数. 在 $P(X)$改变时，我们可以通过语义贝叶斯推理(见式(2.4))得到似然函数。比如， X 表示身高，Y 表示性别. 首先我们用来自普通人群的样本得到优化的似然函数 $P(X|\theta_1)$，然后我们可以得到真值函数模型 $T(A_j|X)$，并且可以把它用到身高分布 $P(X)$不同的人群， 比如中学生(平均身高应矮点)。给定性别， 比如男性 y_1, 我们就可以通过语义贝叶斯推理得到新的似然函数 $P(X|\theta_1)=P(X)T(\theta_1|\mathrm{X})/T(\theta_1)$. 然而， 在流行的似然方法中， 优化的模型不能用到信源 $P(X)$可变的场合。

CM 算法可以用于决策(或分类)函数并继续使用最大似然准则。Left-step c 可以用于这一目的。 而 CM 算法并不提供类似方法。

因为 CM 算法使用样本分布而不是样本序列，它更适合较大样本场合。而 EM 算法使用样本序列，适于较小样本场合。

CM 算法基于一个新的语义信息理论，还有许多其它特点， 比如可以更方便地用于检验和估计；可以用来解释自然语言进化, 如前面讨论。

7．总结

本文重新叙述了鲁晨光的语义信息测度，从而说明它就是平均 log 标准(normalized)似然度. 本文揭示，通过语义信道和 Shannon 信道相互匹配和迭代（信道匹配算法或 CM 算法），我们可以求解检验，估计和混合模型的最大互信息和最大似然度。迭代的收敛可以通过鲁氏 $R(G)$函数得到直观解释和证明。文中提供了检验、估计和混合模型的几个例子。这些例子和理论分析显示，和标准 EM 算法相比，CM 算法有更高效率，更清晰收敛理由，和更广的潜在应用。

本文得到结论：信息论和似然方法更紧密结合对于解决检验、估计和混合模型难题是必要的。结果也表明，通过鲁氏语义信息方法，结合两者是可行的。

致谢：作者感谢汪培庄教授的长期支持和鼓励。因为作者 1993 年发表的专著《广义信息论》就是在北师大当汪培庄教授访问学者期间完成的。没有汪培庄教授的最近鼓励，作者也不会这样努力。作者也要感谢拒绝发表其关于语义信息的文章的期刊——可能是因为作者总想兜售自己的概念和方法。这些拒绝促使作者通过求解最大互信息和最大似然度难题证明其语义信息方法的理论意义和实用价值。

References

1 Shannon, C. E., 1948, A mathematical theory of communication, Bell System Technical Journal, 1948, 27: 379–429 and 623–656.

2 Fisher, R. A., On the mathematical foundations of theoretical statistics, Philo. Trans. Roy. Soc., 1992, A222: 309-368.

3 Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 1974, 19(6):716–723.

4 Cover. T. M., Thomas, J. A., *Elements of Information Theory*, *2nd Edition*, New York: John Wiley & Sons, 2006.

5 Barron, A., Roos, T., and Watanabe, K., Bayesian Properties of Normalized Maximum Likelihood and its Fast Computation, IEEE IT Symposium on Information theory, 2014, 1667-1671.

6 Kullback, S. and Leibler, R.，On information and sufficiency， Annals of Mathematical Statistics，1951，22：79–86.

7 Kok, M., Dahlin, J., B. Schon, T. B., Wills, A., Newton-based maximum likelihood estimation in nonlinear state space models, IFAC-PapersOnLine, 2015，48: 398–403.

8 Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B. 1977, 39: 1–38.

9 Xie, Q., and Barron, A. R., Asymptotic minimax regret for data compression, gambling and prediction. IEEE Trans. on Information Theory, 2000, 46: 431–445.
10 鲁晨光, 广义信息论，合肥, 中国科学技术大学出版社, 1993.
11 鲁晨光, 广义熵和广义互信息的编码意义, 通信学报，1994, 15(6): 37-44.
12 Lu, C. G. (鲁晨光), A generalization of Shannon's information theory, Int. J. of General Systems, 1999, 28: 453-490.
13 Shannon, C. E., Coding theorems for a discrete source with a fidelity criterion, IRE Nat. Conv. Rec., 1959, Part 4:142–163.
14 Zadeh, L. A., Fuzzy Sets, Information and Control, 1965, 8: 338–53.
15 周炯槃，信息理论基础, 北京，中国邮电出版社，1983.
16 Thornbury, J. R., Fryback, D. G., and Edwards, W. 1975. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information, Radiology, 1975, 4: 561–565.
17 Wu, C. F. J., On the Convergence Properties of the EM Algorithm, Annals of Statistics. 1983, 11: 95–10.
18 Neal, R., and Hinton, G., A view of the EM algorithm that justifies incremental, sparse, and other variants, Learning in Graphical Models， (ed. Michael I. Jordan), Cambridge, MA: MIT Press, 1999, 355–368.
19 Huang, W. H., Chen, Y. G., The multiset EM algorithm, Statistics & Probability Letters, 2017, 126: 41–48.
20 Springer, T., Urban, K., Comparison of the EM algorithm and alternatives，Numerical Algorithms, 2014, 67(2): 335–364.
21 Zhao, Y. L., Bi, W. J., WANG, T., Iterative parameter estimate with batched binary-valued observations, SCIENCE CHINA Information Sciences 2016，59: 052201:1-052201:18.
22 Yang, H. L., Pend J. H., Xia, B. R., Zhang D. X., An improved EM algorithm for remote sensing classification, Chinese Science Bulletin 2013，58(9): 1060-1071.

The Semantic Information Method for Maximum Mutual Information and Maximum Likelihood of Tests, Estimations, and Mixture Models

Chenguang Lu
lcguang@foxmail.com

ABSTRACT：It is very difficult to solve Maximum Mutual Information (MMI) or Maximum Likelihood (ML) for all possible Shannon Channels or uncertain rules of choosing hypotheses, so that we have to use iterative methods. According to the Semantic Mutual Information (SMI) and $R(G)$ function proposed by Chenguang Lu (1993) (where $R(G)$ is an extension of information rate distortion function $R(D)$, and G is the lower limit of the SMI), we can obtain a new iterative algorithm of solving the MMI and ML for tests, estimations, and mixture models. The SMI is defined by the average log normalized likelihood. The likelihood function is produced from the truth function and the prior by the semantic Bayesian inference. A group of truth functions constitute a semantic channel. Letting the semantic channel and

Shannon channel mutually match and iterate, we can obtain the Shannon channel that maximizes the Shannon mutual information and the average log likelihood. This iterative algorithm is called Channels' Matching algorithm or the CM algorithm. The convergence can be intuitively explained and proved by the $R(G)$ function. Several iterative examples for tests, estimations, and mixture models show that the computation of the CM algorithm is simple, and can be demonstrated in excel files. For most random examples, the numbers of iterations for convergence are close to 5. For mixture models, the CM algorithm is similar to the EM algorithm; however, the CM algorithm has better convergence and more potential applications in comparison with the standard EM algorithm.

Keywords: Shannon channel; semantic channel; semantic information; likelihood; tests; estimations; mixture; EM algorithm