语义信息 G 理论和逻辑贝叶斯推断用于机器学习[①②]

鲁晨光

摘要：机器学习的一个重要问题是：在标签数目 n>2 时，构造和优化一组学习函数并且希望它在先验分布 $P(x)$(x 是一个实例)改变时仍然有用。 要解决这个问题是非常困难的。为解决这一问题，语义信息 G 理论、逻辑贝叶斯推断(Logical Bayesian Inference，缩写为 LBI)和一组信道匹配(Channel Matching，缩写为 CM)算法一起构成一个系统解决方案。在语义信息 G 理论(或 G 理论)中，一个语义信道由一组真值函数或隶属函数构成。和流行的方法中使用的似然函数、贝叶斯后验和 Logistic 函数相比，隶属函数能更方便地用作学习函数，并且能避免上述困难。使用 LBI，每个标签的学习是独立的。多标签学习时，我们能从一个足够大的带标签样本直接获得一组优化的真值函数，不需要为不同的标签准备不同的样本。基于 G 理论和 LBI，我们得到一组用于机器学习信道匹配算法。在最大互信息分类的例子中，两维空间有三个呈高斯分布的类别，对于大多数初始划分，2-3 次迭代就能使三个标签和三个类别之间的互信息达到其最大值的 99%。在混合模型的例子中，期望-最大(Expectation-Maxmization，缩写为 EM)算法被改进为信道匹配 EM (CM-EM)算法，它对于某些容易导致局部收敛的混合模型有较好表现。G 理论中的信息率逼真函数能帮助我们证明 EM 算法和 CM-EM 算法收敛。对于高维特征空间的最大互信息分类，我们还需要结合 CM 迭代算法和合神经网络做进一步研究。为了统一统计和逻辑，LBI 也需要进一步研究。

关键词：语义信息论；贝叶斯推断；机器学习；多标签分类；最大互信息分类；混合模型； 确证测度；真值函数。

1 引言

机器学习需要学习函数和分类器。1922，费希尔(Fisher) [1] 提出似然推断(Likelihood Inference，缩写为 LI) ——它使用最大似然(Maximum Likelihood，缩写为 ML) 准则优化学习函数和分类器。然而，在后验分布 $P(x)$ (x 是一个实例)改变时

[①] 本文译自英文文章(提供 CC Common 许可证, 略有修改)：Semantic Information G Theory and Logical Bayesian Inference for Machine Learning，*Information* 2019，*10*(8)，261; doi: 10.3390/info10080261

② 随后一章使用同一数学框架和语义通信模型，涉及科学哲学问题和概率的基本问题，两章分别独立成篇。但是一并阅读或交叉阅读有助于全面理解。

，优化的似然函数将是无效的。因为 LI 不能利用先验知识，上世纪五十年代，贝叶斯主义者提出贝叶斯推断[①](Bayesian Inference，缩写为 BI) [2，3]——它使用贝叶斯后验作为学习函数。

然而，我们通常只有实例而不是标签或模型参数的先验知识。在这些情况下，BI 仍然不够好。一对 Logistic (或 Sigmoid) 函数经常被用作二分类的学习函数。使用 Logistic 函数，我们可以用最大后验概率准则（等价于最大正确率准则）分类。我们也可以用 Logistic 函数和贝叶斯定理，利用新的 $P(x)$做新的概率预测，使用 ML 分类器。然而，在标签数量 $n>2$ 时，我们找不到像 Logistic 函数那样的用于多标签学习的学习函数。我们称上述问题为 “可变 $P(x)$的多标签学习问题”。

机器学习是为了获取和传递信息。信息准则在很多情况下应该是较好的准则。1974，Akaike [4] 证明 ML 准则等价于 最小 Kullback-Leibler (KL) [5]距离准则(KL 距离有时被译为“KL 离散度”，有时被解释为 KL 信息)。随后，信息准则——特别是兼容 ML 准则的信息准则——引起了普遍关注[6]。然而，随着似然度增加，KL 距离是减少的，所以 KL 距离用作信息准则并不理想。我们能使用香农(Shannon)互信息或其他信息测度作为信息准则？

香农[7] 于 1948 年提出经典信息论。Weaver [8] 于 1949 年提出通信的三个水平——分别联系到：技术问题，如香农所解决的；语义问题，涉及意义和真；效果问题，涉及信息价值。卡尔纳普(Carnap)和 Bar-Hillel [9]于 1952 年提出一个语义信息论概要。现在我们能发现不同的信息理论[10-13]。我们也能看到模糊信息理论[14-16] 和广义信息理论[17，18]，它们和语义信息理论相关。最近，也有人用带参数的香农互信息测度优化神经网络[19，20]。

然而，就作者所指，在作者之前，没有人把带参数的学习函数带入香农互信息公式，进而用样本分布优化学习函数和分类。因此，作者试图发展一个语义信息论——它能结合香农信息论和费希尔的似然方法。语义信息 G 理论或 G 理论就是为了这一目的。这一理论主要由作者在最近 30 年发展的[21-26]。它使用扎德(Zadeh) [27]提出的集合的隶属函数作为学习函数，并且把一个模糊集合的隶属函数当做一个假设的真值函数。根据塔斯基(Tarski)的真理论 [28] 和 Davidson 的真值条件语义学[29]，真值函数能表示假设的语义。

使用“ G 理论”是因为其中语义互信息是香农互信息的自然推广(“G” 意味着推广)，使得香农互信息是语义互信息的上限。G 也表示语义互信息——如同香农的信息率失真函数中 D 表示平均失真 [30]。作者用 G 代替 D，把信息率失真函数 $R(D)$ 改造为信息率逼真函数 $R(G)$ [24-25]。这样做不仅是为了数据压缩，也是为了机器学习。

[①] Inference 也被译为“推理”。为了和 Reasoning 相区别，我们只把 Bayesian Reasoning 译为“贝叶斯推理”。Reasoning 强调过程，Inference 强调结论；可以认为 Bayesian reasoning 产生或包括 Bayesian inference。

G 理论有两个来源：香农信息论和波普尔(Popper)的假设检验理论 [31] (p. 96，269) [32] (p. 294)，后者强调：逻辑概率较小的假设如能经得起经验的检验，它就能传递更多信息，因而更加可取。

卡尔纳普 and Bar-Hillel [9] 提出的语义信息公式包含了波普尔思想。它是

$$I(p)=\log(1/m_p), \tag{1.1}$$

其中 p 是命题，m_p 是它的逻辑概率。然而，这一公式并不涉及假设是否经得起经验的检验。为此，本文作者提出改进公式——见 2.1.1 节公式(2.15)。进一步，把似然函数和真值函数带入 香农互信息公式，我们就能获得语义互信息公式。

交叉熵逐渐流行于机器学习 [33]。 G 理论不仅使用交叉熵，还使用交互交叉熵 [22，25]。 G 理论中的语义互信息就是交互交叉熵。

为解决可变 $P(x)$的多标签学习问题，作者提出新的参数推断方法——逻辑贝叶斯推断(Logical Bayesian Inference，缩写为 LBI)。

贝叶斯主义包括主观贝叶斯主义和逻辑贝叶斯主义。 主观贝叶斯主义者发展出 BI，他们用主观概率作为统计推断工具。逻辑贝叶斯主义者，如 Keynes(凯恩斯)和卡尔纳普[34]，用逻辑概率(包括真值函数)作为归纳逻辑的推理工具。我们使用“逻辑贝叶斯推断”(LBI)是因为(模糊)真值函数而不是贝叶斯后验被用作推断工具。LBI 同时使用统计概率和逻辑概率，使得推断存在于统计和逻辑之间。BI 适合预测模型的先验分布 $P(\theta)$已知时的场合，而 LBI 适合实例先验分布 $P(x)$已知时的场合。

除了香农信息论和波普尔的假设检验理论，G 理论还继承、吸收或兼容

- 费希尔的似然方法(用于假设检验)[1]，
- 扎德的模糊集合论[27，35](涉及假设的语义和逻辑概率)，
- 卡尔纳普和 Bar-Hillel 的语义信息公式(使用逻辑概率)[9]，
- Floridi 语义信息概念[11，36]，
- 塔斯基的真理论(涉及真和逻辑概率定义)[28]，
- Davidson 的真值条件语义学 [29]，
- Kullback 和 Leibler 距离 [5]，
- Theil 的广义 KL 公式[36]，
- Akaike 关于 ML 准则等价于最小 KL 距离准则的证明 [37]，
- Donsker-Varadhan 表示(含有 Gibs 密度分布的广义 KL 公式) [38]，

  维特根斯坦(Wittgenstein) 的思想：语义在于使用 [39] (p.80)，
- Bayes 定理 [40]——其推广形式可以连接似然函数和隶属函数[41]。

基于 G 理论和 LBI，作者为机器学习发展出一组算法：信道匹配算法(Channel Matching 算法，简称 CM 算法) [41-44]。 在 CM 算法中，语义信道和香农信道相

互匹配从而获得含有最大信息的学习函数或分类，或含有最大信息效率 (*G/R*)的混合模型[①]。

这些算法主要用于：

- 利用实例的先验知识做概率预测;
- 多标签学习，属于有监督学习；
- 不可见实例的(或特征空间的)最大互信息(Maximum Mutual Informatioon，缩写为 MMI)分类 ，属于半监督学习；
- 混合模型，属于无监督学习。

上述每个任务都是非常困难的，并且没有被很好解决。

本文目的是完整地介绍 G 理论、 LBI 和信道匹配算法，包括背景知识和应用，以便读者全面理解它们，特别是理解如何用它们解决可变 *P(x)*的多标签学习问题。

本文部分内容已出版于会议文集 [41-44]。 这些内容在本文得到改进。据作者所知，之前没有人用语义信息测度和样本分布优化隶属函数或真值函数，也没有人既区分统计概率和逻辑概率又把它们用在同一个公式中，也没有其他人提出语义信道的数学表示。

本文也比较了信道匹配算法和某些流行的算法以显示信道匹配算法的效率。

2 方法 I：背景

2.1 从香农信息论到语义信息 G 理论

2.1.1 从香农互信息到语义互信息

定义 1:

x 是一个实例或数据点; X 是一个离散随机变量，它取值 $x \in U=\{x_1, x_2, \ldots, x_m\}$。

y 是一个假设或标签; Y 是一个离散随机变量，它取值 $y \in V=\{y_1, y_2, \ldots, y_n\}$。

$P(y_j|x)$ (y_j 是固定，x 是可变的) 是一个转移概率函数(Transition Probability Function) (香农如此称它 [7])。

香农称 $P(X)$为信源，$P(Y)$为信宿，$P(Y|X)$ 为信道。 一个香农信道是一个转移概率矩阵或一组转移概率函数:

[①] 可以通过附录找到产生图 9-15 的 Python 3.6 源文件.

$$P(Y \mid X) \Leftrightarrow \begin{bmatrix} P(y_1 \mid x_1) & P(y_1 \mid x_2) & ... & P(y_1 \mid x_m) \\ P(y_2 \mid x_1) & P(y_2 \mid x_2) & ... & P(y_2 \mid x_m) \\ ... & ... & ... & ... \\ P(y_n \mid x_1) & P(y_n \mid x_2) & ... & P(y_n \mid x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} P(y_j \mid x) \\ P(y_j \mid x) \\ ... \\ P(y_n \mid x) \end{bmatrix}, \quad (2.1)$$

其中⇔表示等价。注意：条件概率函数 $P(y|x_i)$是归一化的，而转移概率函数 $P(y_j|x)$ 并不是归一化的。在条件概率函数 $P(y|x_i)$中，y 是变量，x_i 是常量。我们将讨论转移概率函数可以用于传统的贝叶斯预测(见 2.2.1 节)。

X 和 Y 的香农熵是:

$$H(X) = -\sum_j P(x_i) \log P(x_i), \quad (2.2)$$

$$H(Y) = -\sum_j P(y_j) \log P(y_j). \quad (2.3)$$

X 和 Y 的 香农后验熵是：

$$H(X|Y) = -\sum_j \sum_i P(x_i, y_j) \log P(x_i \mid y_j), \quad (2.4)$$

$$H(Y|X) = -\sum_j \sum_i P(x_i, y_j) \log P(y_j \mid x_i). \quad (2.5)$$

香农互信息是

$$\begin{aligned} I(X;Y) &= -\sum_j \sum_i P(x_i, y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)} = H(X) - H(X \mid Y) \\ &= -\sum_j \sum_i P(x_i, y_j) \log \frac{P(y_j \mid x_i)}{P(y_j)} = H(Y) - H(Y \mid X). \end{aligned} \quad (2.6)$$

如果 $Y=y_j$，互信息 $I(X; Y)$ 就变成 KL 距离:

$$I(X; y_j) = \sum_i P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)} = \sum_i P(x_i \mid y_j) \log \frac{P(y_j|x_i)}{P(y_j)}. \quad (2.7)$$

有些研究者用下面公式度量 x_i 和 y_j 之间的信息(量)：

$$I(x_i; y_j) = \log \frac{P(x_i \mid y_j)}{P(x_i)} = \log \frac{P(y_j \mid x_i)}{P(y_j)}. \quad (2.8)$$

因为 $I(x_i; y_j)$可能是负的，香农并不使用它。香农 解释：信息是减少的不确定性或节省的平均编码长度。作者认为：上面公式是有意义的，因为负信息意味一个坏的预测会增加不确定性或平均编码长度。

因为香农信息论不能度量语义信息，卡尔纳普和 Bar-Hillel 提出一个语义信息公式：$I(p)=\log[1/m_p]$。因为 $I(p)$ 和预测是否正确无关，这一公式并不实用。

钟义信 [12]使用 DeLuca 和 Termini 的模糊上[14]定义语义信息测度：

$$I(y_j) = \log 2 + [t_j \log t_j + (1 - t_j) \log(1 - t_j)], \tag{2.9}$$

其中 t_j 是 y_j 的逻辑真。然而，根据这一公式，不管 t_j=1 或 t_j=0，信息都达到其最大值：1 比特。这一结果并不是所期望的。因此，这一公式是不合理的。其他使用 DeLuca 和 Termini 模糊熵公式的信息测度也遇到这一问题 [14]。

Floridi 的语义信息公式 [11，36]稍许复杂，它能保证永真句和矛盾句的信息是最小值 0，但是，根据常识，错误预测或谎言比永真句更坏。至于语义信息量怎样和偏差和对错相关，从他的公式我们得不到清晰回答。

作者于 1990 年[21]提出改进的语义信息测度，随后发展成 G 理论。

根据塔斯基关于真的理论[28]，$P(X \in \theta_j)$等价于 P(“$X \in \theta$” 是真的)=$P(y_j$ 是真的)。根据 Davidson 的真值条件语义学[29]，y_j 的真值函数确定了 y_j 的语义。为此，我们做如下定义。

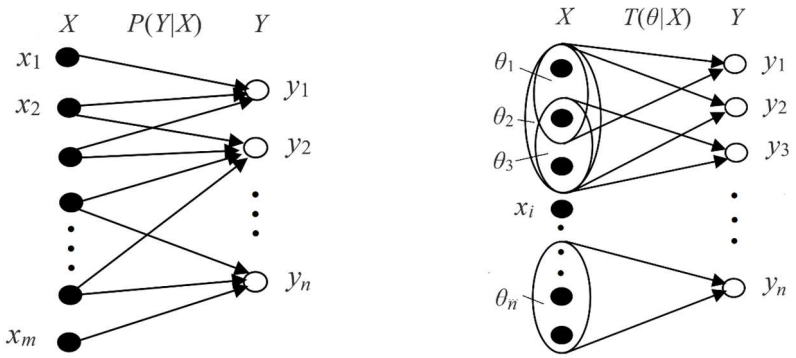
定义 2:

- θ_j 是 U 的模糊子集，是用来解释谓词 $y_j(X)$=“$X \in \theta_j$”的语义(形式语义，涉及外延而不涉及内涵)。我们也可以把 θ_j 当做一个预测模型或一组模型参数。如果 θ_j 是非模糊集合，我们也用 A_j 表示。
- 用等号“=”定义的概率，比如 $P(y_j)=P(Y=y_j)$，是统计概率; 而用属于符号 “$\in$” 定义的概率，比如 $P(X \in \theta_j)$，是逻辑概率。为了区分 $P(Y=y_j)$和 $P(X \in \theta_j)$, 我们定义 $T(\theta_j)=P(X \in \theta_j)$是 y_j 的逻辑概率。
- $T(\theta_j|x)=P(x \in \theta_j)=P(X \in \theta_j|X=x)$是 y_j 的条件逻辑概率，也被称之为 y_j 的(模糊)真值函数或模糊集合 θ_j 的隶属函数。

一组转移概率函数 $P(y_j|x)(j=1，2，\ldots，n)$构成一个香农信道，而一组真值函数或隶属函数 $T(\theta_j|x)(j=1，2，\ldots n)$构成一个语义信道:

$$T(\theta \mid X) \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid x_1) & T(\theta_1 \mid x_2) & \ldots & T(\theta_1 \mid x_m) \\ T(\theta_2 \mid x_1) & T(\theta_2 \mid x_2) & \ldots & T(\theta_2 \mid x_m) \\ \ldots & \ldots & \ldots & \ldots \\ T(\theta_n \mid x_1) & T(\theta_n \mid x_2) & \ldots & T(\theta_n \mid x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1 \mid x) \\ T(\theta_2 \mid x) \\ \ldots \\ T(\theta_n \mid x) \end{bmatrix}. \tag{2.10}$$

图 1 图解了香农信道 $P(Y|X)$和语义信道 $T(\theta|X)$。



(a) 香农信道。 (b)语义信道。

图 1. 香农信道(a)和语义信道(b)。x 和 θ_j 之间的隶属关系确定 y_j 的语义。两个模糊集合可以相互搭接或包含。

香农信道表明 X 和 Y 的相关性，而语义信道表明一组标签的模糊外延。对于气象台和听众之间的天气预报，香农信道表示气象台选择标签或预报的规则，而语义信道表示听众如何理解的这些预报的语义。

真值函数的期望是逻辑概率：

$$T(\theta_j)=\sum_i P(x_i)T(\theta_j\mid x_i)\,, \tag{2.11}$$

它就是扎德 [35]曾提出的模糊事件的概率。该逻辑概率和卡尔纳普和 Bar-Hillel 用的逻辑概率 m_p[9]有些不同。后者只取决于假设的外延。比如，y_j 是一个假设：“X 被 HIV 感染” (HIV 即 Human Immunodeficiency Virus——人类免疫缺陷病毒) 或一个标签“被 HIV 感染”。其逻辑概率 $T(\theta_j)$对于普通人来说很小，因为被感染的人很少。然而，m_p 和 $P(x)$无关; 它可以是 1/2。

注意：统计概率是归一化的，而逻辑概率不是归一化的。当 θ_0，θ_1，…，θ_n 构成 U 的覆盖时，$P(y_0)+P(y_1)+\ldots+P(y_n)=1$ 而 $T(\theta_0)+T(\theta_1)+\ldots+T(\theta_n)\geq 1$。比如，$U$ 是不同年龄集合，其中有非模糊子集 A_1={成年人}={x|$x\geq$18}、A_0={非成年人}={x|x<18}=A_1^c (^{c} 是集合的补运算)和 A_2={年轻人}={x|15≤x≤35}。三个集合构成 U 的一个覆盖。于是有 $T(A_0)+T(A_1)=1$。如果 $T(A_2)=0.3$，则三个逻辑概率是 1.3>1。然而，三个统计概率的和 $P(y_0)+P(y_1)+P(y_2)$一定小于或等于 1。如果 y_2 的用法总是正确的，$P(y_2)$将在 0 和 0.3 之间变化。如果 A_0、A_1 和 A_2 变成模糊集合，结论是一样的。考虑永真句“明天有雨或无雨”，其逻辑概率是 1，而其统计概率则接近 0——因为很少有人这样说。

我们可以把 $T(\theta_j|x)$和 $P(x)$带进贝叶斯公式得到似然函数 [21]:

$$P(x\mid\theta_j)=\frac{T(\theta_j\mid x)P(x)}{T(\theta_j)},\ T(\theta_j)=\sum_i T(\theta_j\mid x_i)P(x_i),\qquad (2.12)$$

上式可谓语义贝叶斯预测，$P(x|\theta_j)$ 可谓似然函数。根据 Dubois 和 Prade 的文章[45]，Thomas [46]和其他人曾提出过类似公式。

令 $T(\theta_j|x)$的最大值是 1，我们能从 $P(x)$和 $P(x|\theta_j)$得到

$$T(\theta_j\mid x)=\frac{T(\theta_j)P(x\mid\theta_j)}{P(x)},\ T(\theta_j)=1/\max[P(x\mid\theta_j)/P(x)],\qquad (2.13)$$

其中 max[.]表示[.]中函数的最大值(对于不同 x)。作者[41]提出第三种贝叶斯定理，它由式（2.12）和（2.13）组成。使用这一定理，给定 $P(x)$，我们可以使似然函数和真值函数(或隶属函数)相互转换。上式和汪培庄教授的随机集落影理论兼容[41，47]。

图 2 显示了 $P(x|\theta_j)$和 $T(\theta_j|x)$的关系——给定 $P(x)$时，其中 x 是年龄，y_j=“年轻人”是标签; 因为 θ_j 是非模糊集合，我们用 A_j 表示。

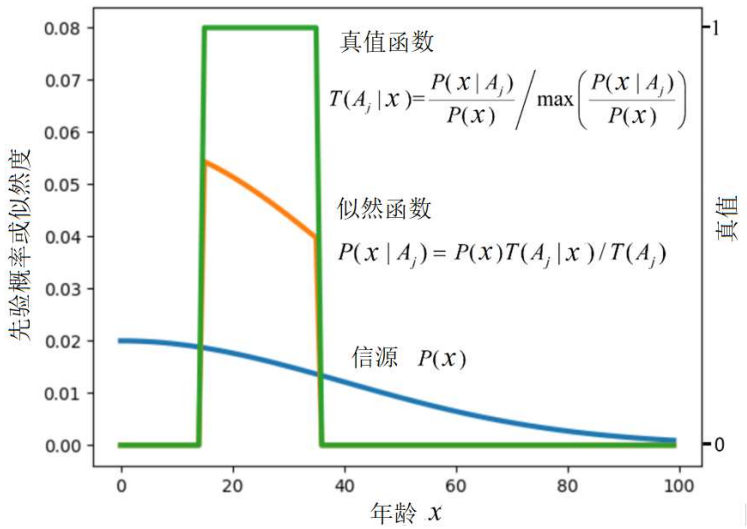


图 2 给定信源时真值函数和似然函数的转换。

我们以定位系统(Positioning System，缩写为 PS)为例说明语义贝叶斯预测。

例 2.1 一个 PS 设备用在列车上，于是 $P(x)$均匀分布在一条线上(见图 3)，PS 指针有偏差。请找出 PS 设备最可能的位置。

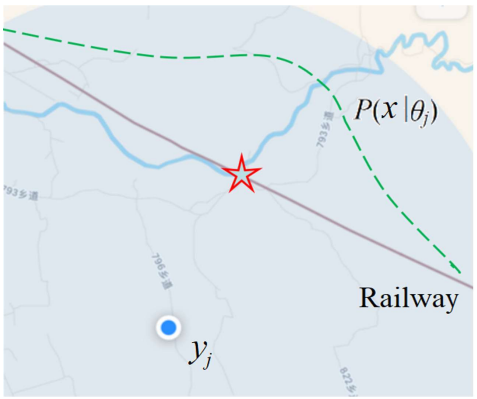


图 3. 图解 PS 定位。圆点是所指带有偏差的位置，星号是最可能位置。

PS 指针的语义可以表述为

$$T(\theta_j|x) = \exp[-(x-x_j)^2/(2\sigma^2)], \tag{2.14}$$

其中 x_j 是 y_j 所指位置，σ 是差平方根(Root Mean Square，缩写为 RMS)。为简单起见，这里假设 x 是一维的。

根据式(2.12)，我们能预测星号所在位置是最为可能位置。大多数人不用数学公式也能做同样的预测。看来人脑能自动用类似的方法预测——根据 y_j 的模糊外延。

语义通信中，我们经常见到这样的假设或预测："温度大约是 10 摄氏度"、"时间大约是 7 点"、"股市指数下个月会上涨约 10%"。它们每个都可以用 y_j="X is about x_j"表示。我们也能用式(2.14)表示它们的真值函数。

作者用对数标准似然度(log-normalized-likelihood)定义 y_j 传递的关于 x_i 的语义信息量：

$$I(x_i;\theta_j) = \log\frac{P(x_i|\theta_j)}{P(x_i)} = \log\frac{T(\theta_j|x_i)}{T(\theta_j)}. \tag{2.15}$$

把式(2.14)带进上式就得到

$$I(x_i;\theta_j) = \log[1/T(\theta_j)] - (x_i - x_j)^2/(2\sigma^2), \tag{2.16}$$

据此，我们可以解释：该语义信息量等于卡尔纳普和 Bar-Hillel 的语义信息量减去相对偏差平方。图 4 图解了上面语义信息公式。

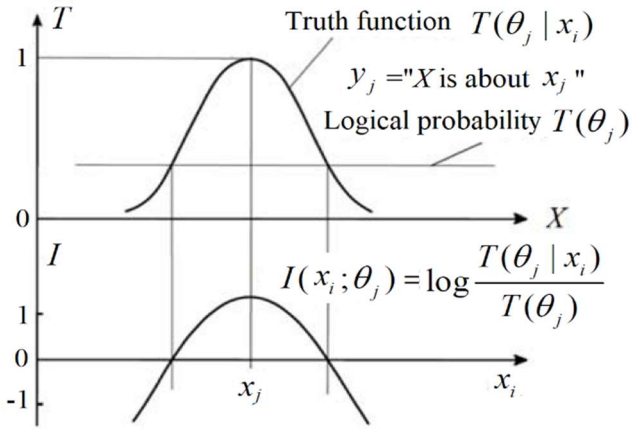


图 4. y_j 传递的关于 x_i 的语义信息（量）

图 4 表明：逻辑概率越小，信息量越大； 偏差越大，信息量越小；错误假设传递的信息是负的。这些结论符合波普尔思想 [32] (p. 294)。

平均 $I(x_i; \theta_j)$就得到广义 KL 信息：

$$I(X;\theta_j)=\sum_i P(x_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}, \quad (2.17)$$

其中 $P(x_i|y_j)$ (i=1，2，…) 是样本分布，它可以是不平滑或不连续的。

Theil 较早提出带有三个概率的广义 KL 信息公式[37]。然而，式(2.17)中 $T(\theta_j)$是常数。如果 $T(\theta_j|x)$是以 e 为基数的指数函数，式 (2.17)就变成 Donsker-Varadhan 表示[19，38]。

Akaike[4]证明了最小 KL 距离准则等价于最大似然准则。仿照 Akaike 的证明，我们也可以证明最大语义信息准则——即最大广义 KL 信息准则——等价于最大似然准则。

定义 3. **D** 是一个带标签样本：{($x(t)$，$y(t)$)| t=1 to N; $x(t) \in U$; $y(t) \in V$}，它包括 n 个带有不同标签的子样本，含有标签 y_j 的字样本 $\mathbf{X}_j$ 包括数据点 $x(1)$，$x(2)$，…，$x(N_j)\in U$。如果 $\mathbf{X}_j$ 足够大，我们能从 $\mathbf{X}_j$ 获得样本分布 $P(x|y_j)$。如果 y_j 是未知的，我们用 $\mathbf{X}$ 代替 $\mathbf{X}_j$ 且用 $P(x|.)$代替 $P(x|y_j)$。

假定 $\mathbf{X}_j$ 中有 N_j 个数据点，其中 N_{ji} 个数据点是 x_i，N_j 个数据点来自 N_j 个自独立同分布随机变量。于是有对数似然度

$$\begin{aligned}\log P(\mathbf{X}_j \mid \theta_j) &= \log P(x(1),x(2),...,x(N)|\theta_j) = \log\prod_i P(x_i \mid \theta_j)^{N_{ji}} \\ &= N_j\sum_i P(x_i \mid y_j)\log P(x_i|\theta_j) = -N_j H(X \mid \theta_j).\end{aligned} \quad (2.18)$$

因为

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}$$
$$=-H(X\mid\theta_j)+\sum_i P(x_i\mid y_j)\log P(x_i)\quad,\qquad(2.19)$$

随着 θ_j 变化，$I(X;\theta_j)$和 $\log P(\mathbf{X}_j|\theta_j)$同时达到最大，因此两个准则是等价的。容易证明，当 $P(x|\theta_j)=P(x|y_j)$，两者同时达到最大。

当 $\mathbf{X}_j$ 极大时，令 $P(x|\theta_j)=P(x|y_j)$，我们可以得到优化的真值函数:

$$T^*(\theta_j|x)=[P^*(x|\theta_j)/P(x)]/\max[P^*(x|\theta_j)/P(x)]$$
$$=[P(x|y_j)/P(x)]/\max[P(x|y_j)/P(x)].\qquad(2.20)$$

利用贝叶斯定理可得

$$T^*(\theta_j|x)=P^*(\theta_j|x)/\max[P^*(\theta_j|x)]=P(y_j|x)/\max[P(y_j|x)].\qquad(2.21)$$

这一公式清楚表明语义信道匹配香农信道。这正好兼容维特根斯坦的思想：意义在于用法[39]（p. 80）。

对不同的 y_j 求 $I(X;\theta_j)$的平均，就得到语义互信息公式

$$I(X;\theta)=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}$$
$$=\sum_i\sum_j P(x_i)P(y_j\mid x_i)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}.\qquad(2.22)$$

如果对于不同的 y_j, $P(x|\theta_j)=P(x|y_j)$或 $T(\theta_j|x)\propto P(y_j|x)$——意味着语义信道匹配香农信道，语义互信息就等于香农互信息。

把式(2.13)带入上式就得到

$$I(X;\theta)=H(\theta)-H(\theta\mid X)$$
$$=-\sum_j P(y_j)\log T(\theta_j)-\sum_j\sum_i P(x_i,y_j)(x_i-\mu_j)^2/(2\sigma_j^2).\qquad(2.23)$$

容易发现：最大语义互信息准则是一个特殊的正则化的最小误差平方(Regularized Least Square，缩写为 RLS)准则[33]。$H(\theta|X)$ 就是平均误差平方，$H(\theta)$就是负的正则化项。不同的是：$H(\theta)$并不惩罚每个参数，它只惩罚相对偏差。重要的是，最大语义互信息准则也兼容最大似然准则。

2.1.2 从信息率失真函数 $R(D)$到信息率逼真函数 $R(G)$

$R(G)$函数将被用来证明最大互信息分类和混合模型的收敛。

香农提出信息率失真函数 $R(D)$[30]。$R(G)$[25]函数是 $R(D)$的推广。在 $R(D)$，R 是信息率，即互信息 $I(X;Y)$，D 是平均失真上限。$R(D)$意味着给定 D 时 R 的最小值。

带参数 s 的信息率失真函数[48] (P. 32)包括两个公式:

$$D(s)=\sum_i\sum_j d_{ij}P(x_i)P(y_j)\exp(sd_{ij})/\lambda_i,$$
$$R(s)=sD(s)-\sum_i P(x_i)\ln\lambda_i, \tag{2.24}$$

其中 $\lambda_i=\sum_j P(y_j)\exp(sd_{ij})$ 是划分函数(保证 $P(y|x_i)$是归一化的)。

我们用 $I_{ij}=I(x_i;y_j)=\log[T(\theta_j|x_i)/T(\theta_j)]=\log[P(x_i|\theta_j)/P(x_i)]$代替 d_{ij}，令 G 是 $I(X;\theta)$的下限，则可定义信息率逼真函数：

$$R(G)=\min_{P(Y|X):I(X;\theta)\geq G} I(X;Y). \tag{2.25}$$

波普尔 [32] 提出用逼真度(verisimilitude 或 truthlikeness)代替正确性，用以评价假设。因为逼真度包含正确性和精确度，语义信息量 $I(x_i;\theta_j)$ 是一个很好的测度——y_j 反映 x_i 的逼真度。我们称 $R(G)$为信息率逼真函数。

仿照 $R(D)$函数的推导[48] (p. 31)，我们可以获得 $R(G)$函数的参数形式：

$$G(s)=\sum_i\sum_j I_{ij}P(x_i)P(y_j)2^{sI_{ij}})/\lambda_i=\sum_i\sum_j I_{ij}P(x_i)P(y_j)m_{ij}^s/\lambda_i,$$
$$R(s)=sG(s)-\sum_i P(x_i)\log\lambda_i, \tag{2.26}$$
$$m_{ij}=T(\theta_j\mid x_i)/T(\theta_j),\ \lambda_i=\sum_j P(y_j)m_{ij}^s,$$

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; $\lambda_i=\sum_j P(y_j)m_{ij}^s$ 是划分函数。所有 $R(G)$函数都是碗状的，二阶导数大于 0，其中有一点和直线 $R=G$ 相切，如图 5 所示。

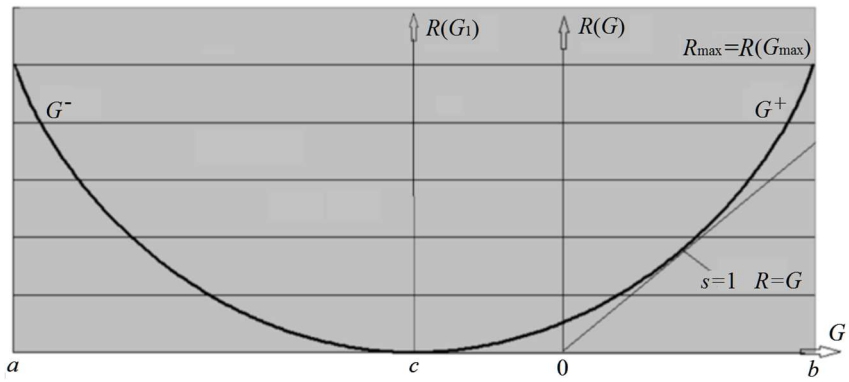


图 5. 一个二元通信的信息率逼真函数 $R(G)$。任一 $R(G)$函数都含有匹配点 $R(G)=G$。

图 5 中，s= dR/dD。当 s=1 时，R = G，这意味着语义信道匹配香农信道。G/R 表明语义通信的效率。我们将在 3.4 节中看到：求解混合模型就是找到模型参数 θ, 它最大化 G/R ，使得 G/R 接近 1 或 $G\approx R$。

当 $s\to\infty$，R 和 G 同时达到最大值 $R_{\max}$ 和 $G_{\max}$。随着 s 增大，转移概率函数 $P(y_j|x)(j=1，2，\ldots，n)$将变得陡峭，香农信道将减少噪声。 $R(G)$函数的这一性质可以用来证明信道匹配最大互信息分类算法——用于不可见实例(或特征空间)的最大互信息分类——的收敛。

R(G)函数不同于 $R(D)$函数。对于给定 R，我们有最大值 G^+ 和最小值 G^-。G^- 是负的，这意味着要想给敌人带来一定的信息损失，我们也需要一定的客观信息。当 R=0，G 是负的，这意味着如果我们听信谁随机预测，我们已有的信息就会减少。

作者当初提出 $R(G)$函数主要是为了根据视觉分辨率做图像压缩[25]。现在我们也可以把它用于最大互信息分类和混合模型的收敛证明。

2.2 从传统的贝叶斯预测到逻辑贝叶斯推断

2.2.1 传统的贝叶斯预测、似然推断和贝叶斯推断

为更好理解逻辑贝叶斯推断(Logical Bayesian Inference，缩写为 LBI)，我们回顾传统的贝叶斯预测(Traditional Bayes' Prediction，缩写为 TBP)、似然推断(LI)和贝叶斯推断(BI)。

我们称使用转移概率函数 $P(y_j|x)$做为工具的预测是 TBP。对于给定的 $P(x)$和 $P(y_j|x)$，我们可以做概率预测：

$$P(x|y_j)=P(x)\,P(y_j|x)/P(y_j)。\tag{2.27}$$

我们用 $kP(y_j|x)$取代 $P(y_j|x)$ (k 是常数) 时，预测结果 $P(x|y_j)$不变，因为

$$\frac{P(x)kP(y_j\,|\,x)}{\sum_i P(x_i)kP(y_j\,|\,x_i)}=\frac{P(x)P(y_j\,|\,x)}{\sum_i P(x_i)P(y_j\,|\,x_i)}=P(x\,|\,y_j).\tag{2.28}$$

使用这一公式，我们可以容易解释：我们能用和转移概率函数成正比的真值函数做同样的概率预测。

对于给定的 $P(y_j)$、$P(x|y_j)$和 $P(x)$，我们能获得预测模型，即转移概率函数：

$$P(y_j|x)=P(y_j)\,P(x|y_j)/P(x).\tag{2.29}$$

我们用医学检验(或信号检测)做例子解释转移概率函数和香农信道如何用作预测模型。

定义 4：令 z 是不可见实例的观察特征(见图 5); Z 是取值 $z\in C=\{z_1，z_2，\ldots\}$的随机变量。对于特征分类(或不可见实例分类)，$x$ 表示真类别或真标签。

假设我们根据实例的特征 $z\in C$ 为 x 分类，也就是提供分类器 $y=f(z)$得到标签 y (见图 5)。

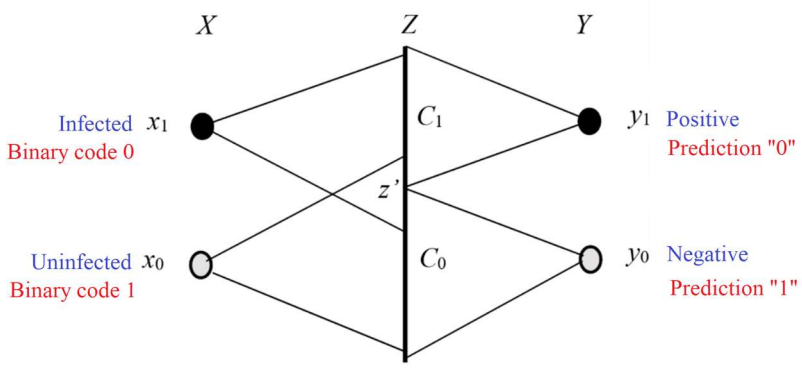


图 5. 图解医学检验和信号检测。我们根据特征 $z \in C_j$ 选择 y。$\{C_0, C_1\}$是 C 的划分。Positive：阳性；Negative：阴性。Infected:被感染；Uninfected：未感染。

我们用 HIV 检验为例说明转移概率函数可以用作概率预测，且适合不同的 $P(x)$。对于感染者 x_1，阳性 y_1 的条件概率 $P(y_1|x_1)$被称之为敏感性(sensitivity)，它也就是真阳性比率。对于未感染者 x_0，阴性 y_0 的条件概率 $P(y_0|x_0)$被称之为特异性，它也就是真阴性比率[49]。敏感性和特异性确定了一个香农信道，如表 1 所示。

表 1. 医学检验的敏感性和特异性确定了一个香农信道 $P(Y|X)$

	感染者 x_1	**未感染者 x_0**
阳性 y_1	$P(y_1\|x_1)$=敏感性=0.917	$P(y_1\|x_0)$=1-特异性=0.001
阴性 y_0	$P(y_0\|x_1)$=1-敏感性=0.083	$P(y_0\|x_0)$=特异性=0.999

＊ 数据来自 OREQuick HIV 检验 [50]。

例 2.2 通过表 1 中转移概率函数 $P(y_1|x)$计算 $P(x_1|y_1)$在 $P(x_1)$=0.0001、0.002 (对于普通人群)和 0.1 (对于高危人群)时的值。

解：使用式(2.28)，当 $P(x_1)$=0.0001、0.002 和 0.1 时，我们得到 $P(x_1|y_1)$=0.084、0.65 和 0.99。

使用似然方法或似然推断，求解上例反而是不容易的。不过，当 x 是很多数值中的一个时，转移概率函数通常是不平滑的。这时我们就需要似然推断。如果样本不很大，以致于转移概率函数不平滑或不连续，我们不能使用转移概率函数获得连续的概率预测 $P(x|y_j)$。这就是为什么我们需要似然推断——它使用参数构造平滑的似然函数。使用最大似然估计(Maximum Likelihood Estimation，缩写为 MLE)，我们可以用一个样本 $\mathbf{X}$ 训练一个似然函数，获得最优参数 θ_j：

$$\theta_j^* = \arg\max_{\theta_j} P(\mathbf{X}|\theta_j) = \arg\max_{\theta_j}[\sum_i P(x_i|.)\log P(x_i|\theta_j)], \qquad (2.30)$$

其中 $P(x_i|.)$ 意味着 y_j 是未知的。似然推断的主要缺点是：它不能利用先验知识 $P(x)$ 或 $P(y)$；在 $P(x)$改变时，以前优化的似然函数是无效的。

为了利用先验知识，主观贝叶斯主义者发展了贝叶斯推断(BI)[2，3]。他们把参数的先验分布 $P(\theta)$带进贝叶斯公式得到贝叶斯后验：

$$P(\theta \mid \mathbf{X}) = \frac{P(\theta)P(\mathbf{X} \mid \theta)}{P_\theta(\mathbf{X})} \text{ ，} P_\theta(\mathbf{X}) = \sum_j P(\theta_j)P(\mathbf{X} \mid \theta_j) \text{ ，} \tag{2.31}$$

其中 $P_\theta(\mathbf{X})$是归一化常数且和 θ 相关；$P(\theta|\mathbf{X})$是参数的后验分布，叫做贝叶斯后验。使用 $P(\theta|\mathbf{X})$，我们可以推导出最大后验估计(Maximum A Posterior estimation，缩写为 MAP):

$$\begin{aligned} \theta_j^* &= \arg\max_{\theta_j} P(\theta_j \mid \mathbf{X}_j) = \arg\max_{\theta} P(\theta_j)P(\mathbf{X}_j \mid \theta_j) \\ &= \arg\max_{\theta_j} [\sum_i P(x_i \mid .) \log P(x_i \mid \theta_j) + \log P(\theta_j)], \end{aligned} \tag{2.32}$$

其中 $P_\theta(\mathbf{X})$被忽略。

BI 有下面优点

- 它特别适合 Y 是表示频率发生器(比如骰子)的随机变量；
- 随着样本尺寸增大，后验分布 $P(\theta|\mathbf{X})$将逐渐收缩到最大似然估计得到的 θ_j^*；
- BI 能更好利用先验知识。

然而，BI 也有下面缺点：

- 根据 BI [3]所做的概率预测和传统的概率预测不兼容；
- $P(\theta)$ 是主观选择的；
- BI 不能利用常有的关于 x 的先验知识 $P(x)$。

如果我们试图使用 BI 求解例 2.1 和 2.2，我们将发现贝叶斯后验 $P(\theta|\mathbf{X})$不如转移概率函数 $P(y_j|x)$好用。因此，要利用 x 的先验知识，我们需要参数化的转移概率函数 $P(\theta_j|x)$。

2.2.2 从费希尔的反概率函数 $P(\theta_j/x)$到逻辑贝叶斯推断

德. 摩根（De Morgan）在谈到拉普拉斯（Laplace）的概率方法[2]时首先称转移概率函数 $P(y_j|x)$为反概率(inverse probability)。后来 费希尔称似然函数为 $P(x|\theta_j)$为概率，称参数化的转移概率函数 $P(\theta_j|x)$为反概率[2]。我们用 θ_j 而不是 θ ，用 x 而不是 x_i，是要强调 θ_j 是常数而 x 是变量，因此，$P(\theta_j|x)$应该是函数。后面我们称 $P(\theta_j|x)$是反概率函数。根据贝叶斯定理，应有

$$P(\theta_j|x) = P(\theta_j)\, P(x|\theta_j)/P(x), \tag{2.33}$$

$$P(x|\theta_j) = P(x_i)\, P(\theta_j|x)/P(\theta_j). \tag{2.34}$$

反概率函数 $P(\theta_j|x)$能很好利用先验知识 $P(x)$ 。当 $P(x)$变成 $P'(x)$，我们仍然能从 $P'(x)$和 $P(\theta_j|x)$获得 $P'(x|\theta_j)$。

当 n=2 时，我们可以容易地用参数构造 $P(\theta_j|x)$ (j=1，2)。比如，我们可以用一对 Logistic (或 Sigmoid)函数作为反概率函数。不幸的是，当 $n > 2$ 时，构造 $P(\theta_j|x)$(j=1，2，…，n)是及其困难的，因为存在归一化限制：对于每个 x，要求$\sum_j P(\theta_j|x)$=1。这就是为什么多标签分类通常要被转换为多个二元分类 [51，52]。

$P(\theta_j|x)$和 $P(y_j|x)$作为预测模型还有一个严重问题。在许多场合，我们只知道 $P(x)$和 $P(x|y_j)$，而不知道 $P(\theta_j)$或 $P(y_j)$，以致于不能获得 $P(y_j|x)$ 或 $P(\theta_j|x)$。但是在这些情况下我们能获得真值函数 $T(\theta_j|x)$。对于 LBI，不存在上述归一化限制，因而构造一组真值函数 $T(\theta_j|x)$(j=1，2，…)并且用 $P(x)$和 $P(x|y_j)$(j=1，2，…，n)优化它们是容易的，$P(y_j)$或 $P(\theta_j)$是不需要的。这是我们需要 LBI 的一个重要原因。

当样本 $\mathbf{X}_j$ 巨大时，我们可以直接从式(2.20)获得优化的真值函数 $T^*(\theta_j|x)$ 。而对于一个有限尺寸的样本，我们可以用广义 KL 公式获得

$$T^*(\theta_j \mid x) = \arg\max_{T(\theta_j|x)} I(X;\theta_j) = \arg\max_{T(\theta_j|x)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}$$
$$= \arg\max_{T(\theta_j|x)} \sum_i P(x_i \mid y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)}. \quad (2.35)$$

这是 LBI 的主要公式。LBI 提供最大语义信息估计(Maximum Semantic Information Estimation，缩写为 MSIE) :

$$\theta_j^* = \arg\max_{\theta_j} I(X;\theta_j) = \arg\max_{\theta_j} [\sum_i P(x_i \mid .) \log[P(x_i \mid \theta_j) / P(x_i)]$$
$$= \arg\max_{\theta_j} [\sum_i P(x_i \mid .) \log T(\theta_j \mid x_i) - \log T(\theta_j)], \quad (2.36)$$

它和 MLE 兼容。如果样本极大，MSIE、MLE 和 MAP 是等价的。

作者建议在一些场合使用 LBI——用真值函数作为推断和预测工具——是因为真值函数有下面优点:

- 我们可以用优化的真值函数 $T^*(\theta_j|x)$做概率预测——允许不同的 $P(x)$，就像用转移概率函数 $P(y_j|x)$和反概率函数 $P(\theta_j|x)$；
- 我们能用小样本优化带参数的真值函数，如同优化似然函数；
- 真值函数可以表示假设的语义或标签的外延；
- 真值函数也就是模糊集合的隶属函数，适合用于分类；
- 要获得优化的真值函数 $T^*(\theta_j|x)$，我们只需要 $P(x)$和 $P(x|y_j)$，而不需要 $P(y_j)$或 $P(\theta_j)$；
- 令 $T(\theta_j|x) \propto P(y_j|x)$，我们能在统计和逻辑之间建起桥梁；

下面一组信道匹配算法将进一步揭示这些优点。

3 方法 II：一组信道匹配(CM)算法

3.1 CM1：解决可变$P(x)$的多标签学习问题

3.1.1. 优化真值函数或隶属函数

我们用 CM1 表示基本的信道匹配——语义信道匹配香农信道——算法，其中真值函数用作学习函数。

假定：x 是一年龄; y_j 是一标签“年轻人”；θ_j 是一模糊集合$\{x|x$ 是年轻人$\}$。从人口统计，我们能得到年龄的先验(密度)分布 $P(x)$和后验分布 $P(x|y_j)$。

如果样本巨大以至 $P(x)$和 $P(x|y_j)$是平滑的，我们可以直接使用式(2.20)获得优化的真值函数 $T^*(\theta_j|x)$ ——不带参数。如果 $P(x)$和 $P(x|y_j)$不是平滑的，我们可以使用式(2.35)获得带参数 $T^*(\theta_j|x)$。因为不需要 $P(y_j)$，用 CM1 做每个标签的学习是独立的。

如果给定的样本分布是转移概率函数 $P(y_j|x)$，我们可以假定 $P(x)$是常数。这样式(2.35)就变成

$$T^*(\theta_j \mid x) = \arg\max_{T(\theta_j|x)} \sum_i \frac{P(y_j|x_i)}{\sum_k P(y_j|x_k)} \log \frac{T(\theta_j \mid x_i)}{\sum_k T(\theta_j|x_k)}. \tag{3.1}$$

如果 $P(y_j|x)$是平滑的，我们能用式(2.21)直接得到无参数 $T^*(\theta_j|x)$。对于多标签学习，我们能直接从香农信道 $P(Y|X)$(或一个样本或一个样本分布 $P(x，y)$)获得一组优化的真值函数。然而，使用流行的二元关联(Binary Relevance)方法[51，52]，我们不得不为若干 Logistic 函数准备若干样本。

在 $P(x)$改变后，$T^*(\theta_j|x)$仍然可以用于语义贝叶斯预测。

3.1.2 大前提的确证

我们用“确证度”(Degree of confirmation)表示证据或样本支持的对一个大前提的“信任度”。

贝叶斯主义者用“信任度”解释假设的主观概率，该信任度在 0 和 1 之间变化。然而归纳研究者用“信任度”表示我们对 if-then 叙述或大前提的信任度，该信任度在-1 和 1 之间变化。本文使用“信任度”表示后者——对 if-then 叙述的评价。我们知道两个随机变量的相关系数也在-1 和 1 之间变化，差别是，if-then 叙述是不对称的，在 X 和 Y 之间有不止一个信任度。

我们用医学检验作为例子，解释怎样用真值函数代替转移概率函数作为语义信道，并做概率预测。

从表 1 所示香农信道我们能推导出语义信道，如表 2 所示。设两个非模糊大前提是 $y_1 \to x_1$ (“如果 $Y=y_1$ 则 $X=x_1$”)和 $y_0 \to x_0$(“如果 $Y=y_0$ 则 $X=x_0$”)，四个真值是

$T(y_1|x_1)=T(y_0|x_0)=1$ 和 $T(y_1|x_0)=T(y_0|x_1)=0$。则相应的两个模糊假设 y_1 和 y_0 的真值函数是

$$T(\theta_1|x)= b_1' +b_1T(y_1|x), \tag{3.2}$$

$$T(\theta_0|x)= b_0' +b_0T(y_1|x), \tag{3.3}$$

其中 $b_1=b(y_1\rightarrow x_1)$是我们对大前提 $y_1\rightarrow x_1$ 的信任度；$b_1'=1-|b_1|$是对 $y_1\rightarrow x_1$ 的不信度，也是谓词 y_1 中永真句的比例。类似地，我们有 $b_0=b(y_0\rightarrow x_0)$和 $b_0'=1-|b_0|$。

表 2. 医学检验的两个不信度构成一个语义信道

Y	感染者 x_1	未感染着 x_0
阳性 y_1	$T(\theta_1\|x_1)=1$	$T(\theta_1/x_0)=b_1'$
阴性 y_0	$T(\theta_0\|x_1)=b_0'$	$T(\theta_0\|x_0)=1$

根据式(2.20)和(2.21)，两个优化的不信度是

$$b_1'^*= P(y_1|x_0)/ P(y_1|x_1)=[P(x_0|y_1)/P(x_0)]/[P(x_1|y_1)/P(x_1)], \tag{3.4}$$

$$b_0'^*= P(y_0|x_1)/P(y_0|x_0)= [P(x_1|y_0)/P(x_1)]/[P(x_0|y_0)/P(x_0)]. \tag{3.5}$$

给定 y_1 时，我们可以用 $b_1'^*$ 和不同的 $P(x)$做语义贝叶斯预测:

$$P(x_1|\theta_1)=P(x_1)/[P(x_1)+b_1'^*P(x_0)], \tag{3.6}$$

$$P(x_0|\theta_1)= b_1'^*P(x_0) /[P(x_1)+b_1'^*P(x_0)]. \tag{3.7}$$

上述预测结果和传统的贝叶斯预测结果相同，即使我们只知道 $P(x|y_1)$和 $P(x)$ 而不知道 $P(y_1)$，我们仍然可以做上述概率预测。容易证明，使用式(3.6)求解例 2.2，结果和用传统的贝叶斯预测的结果相同。

如果我们使用 LI 或 BI 求解例 3.1，我们将发现，要使模型适合不同的 $P(x)$是不可能的。和表 1 中的香农信道比较，表 2 中的语义信道更容易理解和记忆。要记忆 $P(y_j|x)$，我们需要记忆两个数字，而要记忆 $T^*(\theta_1|x)$，我们只需要记忆一个数字 $b_1'^*$。

在文献[41]，作者为正的和负的确证度提供了两个公式。现在我们可以把他们合并为一个公式[①]:

$$b_1^* = b^*(y_1 \rightarrow x_1) = \frac{P(y_1|x_1)-P(y_1|x_0)}{\max(P(y_1|x_1),P(y_1|x_0))} = \frac{LR^+-1}{\max(LR^+,1)}, \tag{3.8}$$

其中 $LR^+=P(y_1|x_1)/\ P(y_1|x_0)$ 是阳性似然比。随着 LR 从 0 变到正无穷大，b^*从 -1 变到 1，如图 7 所示。

[①] 发表于 *Information* 的原文中用置信水平 CL 是不对的，这里改成似然比.

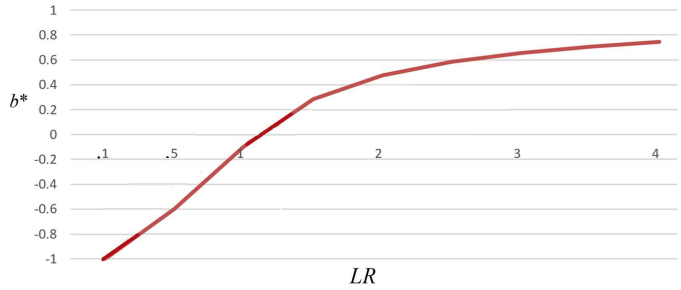


图 7. 确证度和似然比之间的关系。当 *LR* 从 0 变到无穷大时，*b**从-1 变到 1。

3.1.3 矫正定位系统(PS)设备参数

假如我们不知道一个 PS 设备的真实参数，我们可以假定

$$T(\theta_j|x)=\exp[-(x-x_j-\triangle x))^2/(2\sigma^2)], \tag{3.9}$$

其中 x 是一个两维矢量。然后我们可以用一个样本发现系统偏差$\triangle x$ 和 σ。我们可以驾驶带有 PS 设备的小车在一个广场上，记录相对偏差 $x'=x-x_j$，许多偏差构成一个样本，由此得到样本分布 $P(x'|y_j)$。因为在一个广场上，$P(x)$应该是常数。然后我们可以使用广义 KL 公式得到优化的参数$\triangle x^*$ 和 σ^*。然后我们可以根据$\triangle x^*$ 矫正 y_j ——用 $y_k=y_j+\triangle x^*$取代 y_j。

我们也可以假定一个 PS 设备经常出错，用

$$T(\theta_j|x)=b\ \exp[-(x-x_j-\triangle x)^2/(2\sigma^2)]+1-b \tag{3.10}$$

作为学习函数，获得确证度 b^*。

如果我们试图使用反概率函数 $P(\theta_j|x)$或贝叶斯后验 $P(\theta|\mathbf{X})$解决相同任务(见图 3)，我们将发现这是非常困难的，因为我们只能从 PS 地图得到先验知识 $P(x)$，而得不到 $P(y)$或 $P(\theta)$。

3.2 CM2：语义信道和香农信道相互匹配用于多标签分类

我们用 CM2 表示两步：

- **匹配 I**：令语义信道匹配香农信道，或使用 CM1 实现多标签学习;
- **匹配 II**：令香农信道匹配语义信道。

两步都使用最大语义信息准则或最大似然准则。

关于多标签学习，我们可以用式(2.35)或(3.1)训练每一个标签的真值函数。我们也可以学习流行的方法[52]，同时用样本分布 $P(x，y_j)$ 和 $P(x，y_j')$ (y_j'是 y_j 的否定)训练真值函数 $T(\theta_j|x)$[①]：

[①] 式(3.11)在英文原文基础上有所改变。

$$T^*(\theta_j \mid x) = \arg\max_{T(\theta_j|X)}[P(y_j)I(X;\theta_j) + P(y_j')I(X;\theta_j^c)]$$
$$= \arg\max_{T(\theta_j|x)} \sum_i [P(x_i, y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} + P(x_i, y_j')\log\frac{1-T(\theta_j \mid x_i)}{1-T(\theta_j)}], \quad (3.11)$$

其中 θ_j^c 是 θ_j 的补集。这一真值函数 $T^*(\theta_j|x)$可以是 Logistic 函数，和式(2.35)和(3.1)中的 $T^*(\theta_j|x)$比，它覆盖较大区域。

对于可见实例分类，x 是给定的。对于匹配 II，最大语义信息分类器是

$$y_j^* = h(x) = \arg\max_{y_j} \log I(x;\theta_j) = \arg\max_{y_j} \log[T(\theta_j \mid x) / T(\theta_j)]. \quad (3.12)$$

使用 $T(\theta_j)$可以克服类别不平衡问题[50]并减少小概率事件漏报。如果 $T(\theta_j|X)\in\{0, 1\}$，语义信息测度就变成卡尔纳普和 Bar-Hillel 的语义信息测度，上述分类器变成最小逻辑概率分类器:

$$y_j^* = h(x) = \arg\max_{y_j \text{ with } T(A_j|x)=1} \log[1/T(A_j)] = \arg\min_{y_j \text{ with } T(A_j|x)=1} \log T(\theta_j). \quad (3.13)$$

上面分类器鼓励我们选择具有较小逻辑概率的复合标签。

对于不可见实例或不确定 x，我们只知道 $P(x|z)$。这时最大语义信息分类器变成

$$y_j^* = f(z) = \arg\max_{y_j} \sum_i P(x_i \mid z)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \quad (3.14)$$

为简化多标签学习，我们可以选择较少的标签，用它们和兼容布尔代数的模糊逻辑[22]产生复合标签，用于多标签分类[44]。

在流行的使用最大后验概率准则的多标签分类中，对于给定的 x，上述分类器比较两个反概率函数 $P(\theta_j|x)$ 和 $P(\theta_k|x)$——比如两个 Logistic 函数或相似度函数，然后选择一个有较大反概率函数的标签。这个方法和最大信息准则或最大似然准则不兼容，适合用正确率重要而信息不重要的场合。

3.3 CM3：用于不可见实例最大互信息分类的信道匹配算法

我们用 CM3 表示信道匹配迭代算法——它重复两个步骤匹配 I 和匹配 II。CM2 不是迭代算法，而 CM3 是。这一算法可用于最大互信息分类。流行的最大互信息分类算法是梯度下降算法。

我们用图 6 所示医学检验解释不可见实例最大互信息分类问题。

我们需要通过优化 z'实现最大互信息分类。问题是：没有分类器或 z'，我们不能表达互信息 $I(X; Y)$；而没有互信息表达式，我们不能优化 z'。这一问题也是香农信道不确定时最大似然估计遇到的问题。要解决这一问题，研究者通常使用参数构造划分边界，然后使用梯度下降算法或牛顿方法搜索产生最大互信息的参数。CM 迭

代算法不同，它使用数值边界和信息增益函数(作为奖励函数)。它重复更新信息增益函数和划分边界，从而实现最大互信息分类。

令 $y_j=f(z|z\in C_j)$(C_j 是 C 的子集)。于是 $S=\{C_1, C_2, \ldots\}$是 C 的一个划分。我们的目的是：给定来自样本的样本分布 $P(x, z)$，找到最优划分的 S，它是

$$S^* = \arg\max_S I(X;\theta|S) = \arg\max_S \sum_j \sum_i P(C_j)P(x_i \mid C_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \quad (3.15)$$

匹配 I (Matching I)：让语义信道匹配香农信道——设置奖励函数。

首先，通过给定的 S 得到转移概率函数或香农信道:

$$P(y_j \mid x) = \sum_{z_k \in C_j} P(z_k \mid x), j = 1,2,...,n. \quad (3.16)$$

根据这一信道，我们得到语义信道和语义信息 $I(x_i; \theta_j)$。然后，对于不同的 z 求奖励函数:

$$I(X_i;\theta_j|z) = \sum_i P(x_i \mid z) I(x_i;\theta_j) \quad , \quad j=0, 1, \ldots, n, \quad (3.17)$$

当 z 是两维时，$I(X; \theta_j |z)$是两维平面上面的曲面。我们也可以直接令 $I(x_i; \theta_j)=I(x_i; y_j)=\log[P(y_j|x)/P(y_j)]$。但是，使用语义信道的概念，我们能更好理解这一算法并证明其收敛。

匹配 II (Matching II)：让香农信道匹配语义信道，即使用分类器：

$$y_j^* = f(z) = \arg\max_{y_j} I(X_i;\theta_j|z)，j=0, 1, \ldots, n. \quad (3.18)$$

重复匹配 I 和匹配 II 直至 S 不变。收敛的 S(记为 S^*)就是要求的。

匹配 II 在优化香农信道时减少噪声。我们可以这样理解两种匹配：匹配 I 得到奖励函数，而匹配 II 做贝叶斯判决。

给定 $P(x)$时，一个语义信道确定一个 $R(G)$函数。一个改进的 $R(G)$函数有一个更高的匹配点 $R(G)=G$。信道匹配算法就是找到匹配点 $R(G)=G=R_{max}=G_{max}$，见图 8。

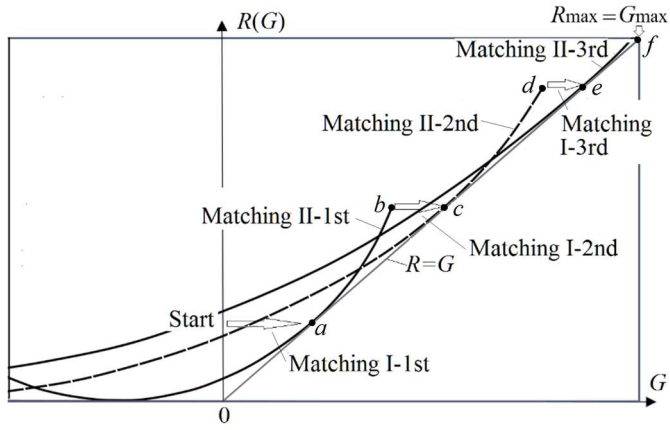


图 8. 图解不可见实例最大互信息分类的收敛。在迭代过程中，坐标 (G, R)逐步从开始点移到 a、b、c、d、e…、f。

我们可以证明上述迭代收敛。在迭代 CM 算法过程中，坐标 (G, R)变化如下：

$$\text{Start} \xrightarrow{\text{Matching I-1st}} a \xrightarrow{\text{Matching II-1st}} b \xrightarrow{\text{Matching I-2nd}} c \xrightarrow{\text{Matching II-2nd}} d \xrightarrow{\text{Matching I-3rd}} e \xrightarrow{\text{Matching II-3rd}} \ldots f$$

上述过程直至匹配 II 不能再改进 R 和 G。

(G, R) 能收敛到$(G_{\max}, R_{\max})$是因为每个匹配 I 增大 G 且每个匹配 II 同时增大 G 和 R，而 G 和 R 的最大值是有限的。

给定奖励函数 $I(X; y_j|z)(j=1, 2, \ldots)$时，匹配 II 总能找到最好划分，这是因为它检查每个 z 看奖惩函数哪个最大。我们可以这样理解 CM 迭代算法：一个 $R(G)$函数就像一个楼梯，坐标(G, R)就像登高者。在匹配 I 中，(G, R)建造一个楼梯并爬上去。在匹配 II 中，它爬到楼梯的最高点。如此重复直至到达$(G_{\max}, R_{\max})$。

3.4 CM4：用于混合模型的信道匹配EM算法

我们用 CM4 表示信道匹配 EM(CM-EM)算法——一个改进的用于混合模型的 EM 算法。在 CM3，匹配 II 是为了最大 R，而在 CM4 中，匹配 II 是为了最小 R-G 或最大信息效率 G/R。

CM4 基于完全不同混合模型收敛理论。流行的用于混合模型的 EM 算法收敛理论解释：我们能通过最大化完全数据的对数似然度 Q 最大化观察到的不完全的数据的对数似然度 $L_X(\theta)$，而 CM-EM 算法解释：我们能通过最大化信息效率 G/R 或最小化 R-G 最大化 $L_X(\theta)$。

EM 算法[53]用于混合模型时经常导致缓慢和不正确收敛 [54，55]。现在我们可以通过让语义信道和香农信道相互匹配改进收敛。和上节匹配 II 不同的是，这里匹配 II 是为了最小化香农互信息 R。

如果概率分布 $P_\theta(x)$来自 n 似然函数的混合，即

$$P_\theta(x)=\sum_{j=1}^{n}P(y_j)P(x\mid\theta_j)\,, \tag{3.19}$$

那么我们称 $P_\theta(x)$是一个混合模型。如果每个预测模型 $P(x|\theta_j)$是一个高斯函数，则 $P_\theta(x)$是高斯混合模型。下面我们设 n=2 讨论混合模型算法。

假设 $P(x)$来自 $P(x|\theta_1*)$和 $P(x|\theta_2*)$的混合，混合比例是 $P*(y_1)$和 $P*(y_2)$=1-$P*(y_1)$，即

$$P(x)=P*(y_1)P\,(x|\theta_1*)+P*(y_2)P(x|\theta_2*). \tag{3.20}$$

我们只知道 $P(x)$和 n=2，用猜测的参数和混合比例产生

$$P_\theta(x)=P(y_1)P(x|\theta_1)+P(y_2)P(x|\theta_2). \tag{3.21}$$

那么，观察数据的对数似然度是

$$L_X(\theta)=N\sum_i P(x_i)\log P_\theta(x_i)=-NH_\theta(X); \tag{3.22}$$

相对熵或 KL 距离是:

$$H(P\,||\,P_\theta)=\sum_i P(x_i)\log\frac{P(x_i)}{P_\theta(x_i)}=H_\theta(X)-H(X). \tag{3.23}$$

如果两个分部 $P(x)$和 P_θ(x)相互接近，以至于相对熵接近 0，比如说小于 0.001 比特，那么，我们可以说我们的猜测是对的。因此，我们的任务是：通过改变参数 θ 和混合比例 $P(y)$ 最大化对数似然度 $L_X(\theta)$=log$P(\mathbf{X}|\theta)$或最小化相对熵 $H(P||P_\theta)$。

解释 EM 算法的主要公式是:

$$\begin{aligned}Q&=N\sum_i\sum_j P(x_i)P(y_i\mid x_i)\log P(x_i,y_j\mid\theta)\\&=L_X(\theta)+N\sum_i\sum_j P(x_i)P(y_j\mid x_i)\log P(y_j\mid x_i),\end{aligned} \tag{3.24}$$

其中 Q= -$NH(X，Y|\theta)$被称之为完全数据对数似然度。$P(y_j|x)$来自式(3.26)。于是

$$L_X(\theta)=Q-H, \tag{3.25}$$

其中 H= -$NH(Y|X，\theta)$是香农条件熵。

EM 算法包含两步:

E-step：写出条件概率函数(即香农信道)：

$$\begin{aligned}&P(y_j\mid x)=P(y_j)P(x\mid\theta_j)/P_\theta(x),\\&P_\theta(x)=\sum_j P(y_j)P(x\mid\theta_j).\end{aligned} \tag{3。26}$$

M-step：通过改进 $P(y)$和 θ 最大化 Q。如果 Q 不能被进一步改进，结束迭代；否则转到 E-step。

Neal 和 Hinton[56]提出一种改进 EM 算法：最大-最大(Maximization-Maximization，缩写为 MM)算法，其中用 $F=Q+H(Y)$取代 Q。两步都最大化 F。

大多数 EM 算法研究者相信：Q 和 $L_X(\theta)$是正相关的; E-step 不减小 Q。然而，这并不是事实。作者发现：在 E-step 中，Q 可能减小；在某些情况下，Q 和 F 也应该减小[42]。

使用信道匹配算法改进 EM 算法，我们可以得到 CM-EM 算法,它能更好收敛。

CM-EM 算法包括三步:

E1-step：构造香农信道。这一步和 EM 算法中 E-step 相同。

E2-step：重复下面等式直至 $P^{+1}(y)$等于或近似于 $P(y)$。

$$\begin{aligned} &P^{+1}(y_j)=\sum_i P(x_i)P(y_j\mid x_i)=\sum_i P(x_i)P(y_j\mid x_i),\ j=1,2,...;\\ &P(y_j)=P^{+1}(y_j);\\ &P(y_j\mid x_i)=P(x_i\mid\theta_j)P(y_j)/\sum_k P(y_k)P(x_i\mid\theta_k),\ i=1,2,...;j=1,2,... \end{aligned} \tag{3.27}$$

如果 $H(P||P_\theta)$ 小于一个较小值，则结束迭代。

MG-step：通过优化 log 函数中参数 θ_j^{+1} 最大化 G：

$$G=I(X;\theta)=\sum_i\sum_j P(x_i)\frac{P(x_i\mid\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(x_i\mid\theta_j^{+1})}{P(x_i)}。 \tag{3.28}$$

转到 E1 -step。

因为 G 在 $P(x|\theta_j^{+1})/P(x)=P(x|\theta_j)/P_\theta(x)$时达到最大值，所以新的似然函数是

$$P(x|\theta_j^{+1})=P(x)P(x|\theta_j)/P_\theta(x). \tag{3.29}$$

如果没有 E2-step，上面 $P(x|\theta_j^{+1})$并不是归一化的[57]。对于高斯混合，我们可以容易获得参数:

$$\begin{aligned} &\mu_j^{+1}=\sum_i P(x_i)P(x_i\mid\theta_j)/P_\theta(x_i),\ j=1,\ 2,...,n;\\ &\sigma_j^{+1}=\{\sum_i P(x_i)[P(x_i\mid\theta_j^{+1})-\mu_j^{+1}]^2\}^{0.5},\ j=1,\ 2...,n. \end{aligned} \tag{3.30}$$

如果似然函数并不是高斯分布，我么可以在参数空间搜索最优参数，比如使用梯度下降算法。

我们可以使用 $R(G)$函数的如下性质证明 CM-EM 算法收敛：

- $R(G)$ 函数是凹的，并且 $R(G)$-G 有一个唯一最小点 0——当 $R(G)=G$ 时[25];
- $R(G)$-G 近似于 $H(P||P_\theta)$。

E1 -step 之后，香农互信息变成

$$R=\sum_i\sum_j P(x_i)\frac{P(x_i\mid\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(y_j\mid x_i)}{P^{+1}(y_j)}. \tag{3.31}$$

我们定义

$$R''=\sum_i\sum_j P(x_i)\frac{P(x_i\mid\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(x_i\mid\theta_j)}{P_\theta(x_i)}. \tag{3.32}$$

容易证明：R''-G=$H(P||P_\theta)$。于是

$$H(P\|P_\theta)=R''-G=R-G+H(Y\|Y^{+1}), \tag{3.33}$$

其中

$$H(Y^{+1}||Y)=\sum_j P^{+1}(y_j)\log[P^{+1}(y_j)/P(y_j)]. \tag{3.34}$$

证明 $P_\theta(X)$收敛到 $P(X)$等价于证明 $H(P||P_\theta)$收敛到 0。因为 E2-step 促使 R=R'' 和 $H(Y^{+1}||Y)$=0，我们只要证明每一步最小化 R-G 就行了。很显然，MG-step 最小化 R-G，因为它增大 G 不改 R。剩下的问题就是证明 E1-step 和 E2-step 最小化 R-G。我们学习香农 [30] 和其他人[48]在分析信息率失真函数时使用的变分方法和迭代方法，通过分别优化 $P(y|x)$和 $P(y)$，最小化 R-G=$I(X;\ Y)$-$I(X;\ \theta)$。因为 $P(y|x)$和 $P(y)$是相互依赖的，我们只能固定一个优化另一个。E1-step 和 E2-step 正好可以达到这一目的。详细收敛证明见 [57]。

4 结果

4.1 CM2算法用于多标签学习和分类的结果

我们使用一个先验分布 $P(x)$和一个后验分布 $P(x|y_j)$优化一个真值函数 $T^*(\theta_j|x)$，如图 9 所示。

为得到 $P(x)$和 $P(x|y_j)$，首先我们用高斯随机数发生器产生两个样本 S_1 和 S_2。S_1 中大于 0 的部分具有分布 $P(x)$

$$P(x)\approx\begin{cases}k\exp[-(x/40)^2], & 0\le x\le 86;\\ 0, & \text{otherwise.}\end{cases}$$

其中 k 是归一化常数。S_2 的分布是 $P_2(x)$。$P(x|y_j)$从 $P_2(x)$和下面真值函数得到：

$$T(\theta_2\mid x)=\begin{cases}\exp[-(x-18)^2/(2*3^2)], & x<18;\\ 1, & 18\le 1\le 25;\\ \exp[-(x-25)^2/(2*4^2)], & x>25.\end{cases}$$

即 $P(x|y_j)$=$P_2(x|\theta_2)$=$P_2(x)T(\theta_2|x)/T(\theta_2)$。

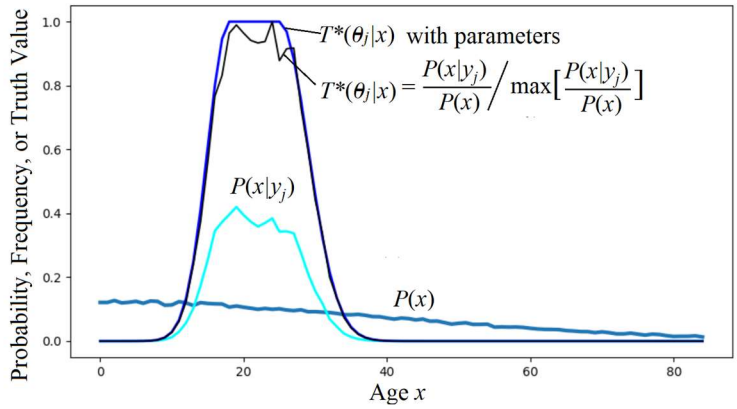


图 9. 使用先验分布 $P(x)$和后验分布 $P(x|y_j)$获得优化的真值函数 $T^*(\theta_j|x)$。图 9-15 来自 Python 3.6 程序——能通过文后附录找到。

然后我们能得到从 $P(x)$和 $P(x|y_j)$得到 $T^*(\theta_j|x)$。如果我们直接使用式(2.20)，$T^*(\theta_j|x)$将是不平滑的。我们可以用参数构造真值函数如下：

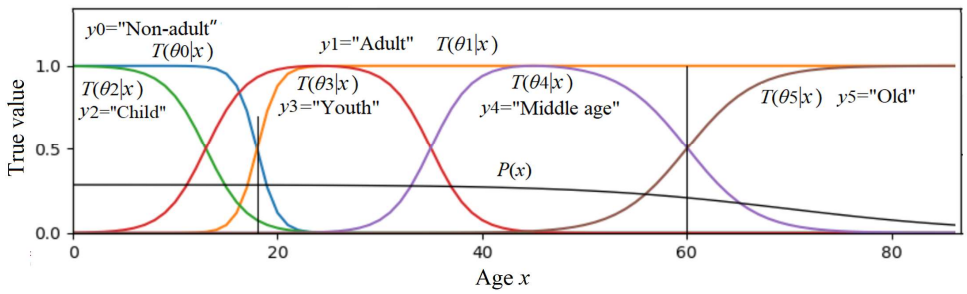
$$T(\theta_j \mid x)=\begin{cases}\exp[-(x-\mu_{j1})^2/(2\sigma_{j1}{}^2)], & x<\mu_{j1};\\ 1, & \mu_{j1}\le 1\le\mu_{j2};\\ \exp[-(x-\mu_{j2})^2/(2\sigma_{j2}{}^2)], & x>\mu_{j2}.\end{cases}$$

然后使用广义 KL 信息公式优化 $T(\theta_j|x)$，得到 $T^*(\theta_j|x)$——它是平滑的。如果 $S_2=S_1$，则 $T^*(\theta_j|x)=P(x|y_j)/P(x)/\max[P(x|y_j)/P(x)]=T(\theta_2|x)$。

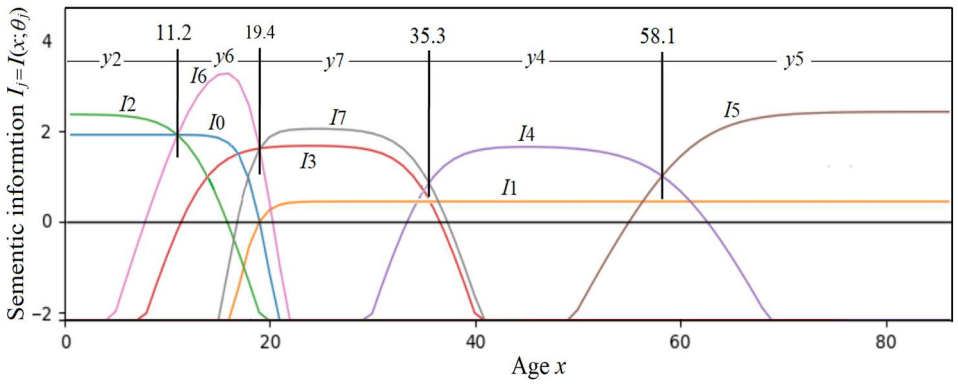
图 10 显示了最大语义信息分类——给定先验分布 $P(x)$和五个标签的真值函数时。五个标签是(y_1，y_2，y_3，y_4，y_5)=(“成人”，“小孩”，“年轻人”，“中年人”，“老年人”)= (“Adult”，“Child”，“Youth”，“Middle-age”，“Old”)。图 10 (a)显示了五个标签的真值函数。其中只有 $T(\theta_3|x)$和 $T(\theta_4|x)$是两个 Logistic 函数的乘积；其他都是一个 Logistic 函数。我们可以把它们当做从流行方法得到的学习函数 $P(\theta_j|x)$，然后使用贝叶斯分类器或最大后验概率准则(等价于最大正确率准则)分类。

在图 10 (a)中，$P(x)$被假定为：$P(x)=k[1-1/\exp[-0.1*(x-70)]$ (对于 $x>0$)，其中 k 是归一化常数。从五个标签我们还可以得到复合标签 $y_0=y_1$’、$y_6=y_3 \wedge y_1$’和 $y_7=y_3 \wedge y_1$。

图 10 (b)显示了最大语义信息分类的效果，见式(3.12)。



(a) 关于年龄的五个标签的真值函数和人口密度先验分布 $P(x)$。



(b) 给 x 贴标签——根据哪个标签的信息量 $I_j=I(x;\theta_j)$ (j=0，1，…7) 最大。

图 10. 根据人群年龄做最大语义信息分类。

图 10 表明最大后验(逻辑)概率准则和最大语义信息准则会导致不同分类。使用最大后验概率准则，对于大多数年龄，我们会选 择 y_0=“非成年人”或 y_1=“成年人”。但是，使用最大语义信息准则，随着年龄增大，我们依次选择 y_2、 y_6、 y_7、 y_4 和 y_5. 最大语义信息准则鼓励我们使用带有较小逻辑概率的复合标签。比如，当 x 在 11.2 和 16.6 之间时，我们应当选择 $y_6=y_3 \wedge y_1$’=“年轻人”和“非成年人”。然而，对于大多数 x，CM2 并不像二元关联法(Binary Relevance [52])那样使用太多的冗余标签。比如，使用最大语义信息准则，我们不需要添加 “非年轻人”到已有标签“老年人”的 x=60 上。

4.2 CM3用于不可见实例最大互信息分类的结果

作者已经用很多例子检验 CM3。

例 4.2.1 z 从 0 到 100 变化；步长是 1。两个高斯分布 $P(z|x_0)$和 $P(z|x_1)$的参数是：μ_0=30，μ_1=70，σ_0=15 和 σ_1=10。先验概率 $P(x_0)$=0.8；$P(x_1$=0.2)。初始划分点 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。

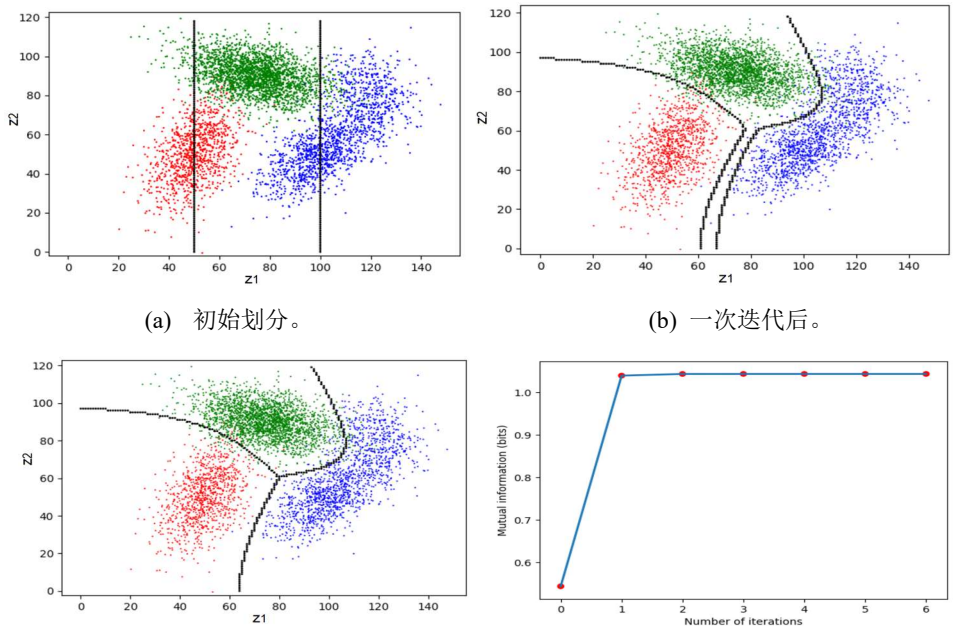
下面是一个两维的例子。

例 4.2.2 (见图 11) 有三个类别，左边两个是高斯分布 $P(z|x_0)$ 和 $P(z|x_1)$；右边一个是两个高斯分布 $P(z|x_{21})$和 $P(z|x_{22})$ 的混合。显示的样本尺寸是 1000。表 3 显示了四个高斯分布的参数。

表 3. 四个高斯分布的参数

	μ_{z1}	μ_{z2}	σ_{z1}	σ_{z2}	ρ	$P(x_i)$
$P(z\|x_0)$	50	50	75	200	50	0.2
$P(z\|x_1)$	75	90	200	75	-50	0.5
$P(z\|x_{21})$	100	50	125	125	75	0.2
$P(z\|x_{22})$	120	80	75	125	0	0.1

迭代过程如图 11 所示。图 11(a)中垂直线是初始划分。



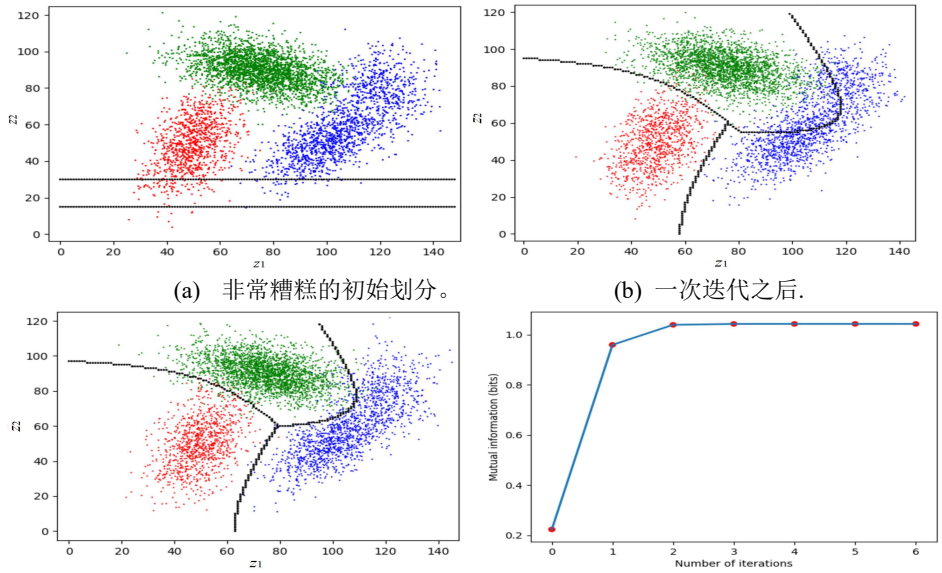
(a) 初始划分。 (b) 一次迭代后。

(b) 两次迭代后。 (c) 互信息随迭代变化。

图 11. 不可见实例最大互信息分类。分类器是 $y=f(z)$。互信息是 $I(X;Y)$。x 是一个真类别，y 是选择的标签。

两次迭代后，互信息是 1.0434 比特。收敛的互信息是 1.0435 比特。两次迭代使互信息达到收敛值的 99.99%。

为检验 CM3 的可靠性，作者使用一个非常糟糕的初始划分。但是收敛也很快，见图 12。



(a) 非常糟糕的初始划分。 (b) 一次迭代之后.

(c) 第二次迭代之后。 (d) 互信息随迭代次数变化。

图 12. 最大互信息分类——用一个很坏的初始划分。

作者也用不同参数和不同初始划分检验 CM3。所有迭代过程都很快，收敛都正确。在大多数情况下，2-3 次迭代就是能使互信息超过最大值的 99%。

4.3 CM4用于混合模型的结果

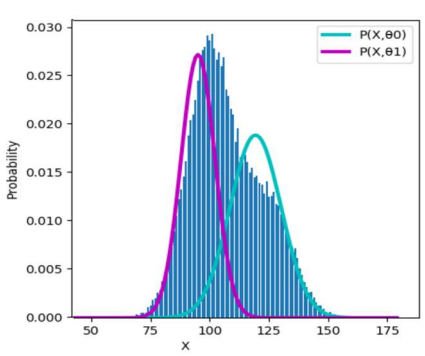
CM4 即信道匹配 EM(CM-EM)算法。下面例子用以说明 CM-EM 算法在某些场合优于 EM 算法[①]。

Ueda 和 Nakano 用一个例子[54] 试图说明某些初始参数导致局部最大 Q 时，局部或不正确收敛是不可避免的。不正确收敛也被 Marin 等人 [55]验证。下面是这个例子。

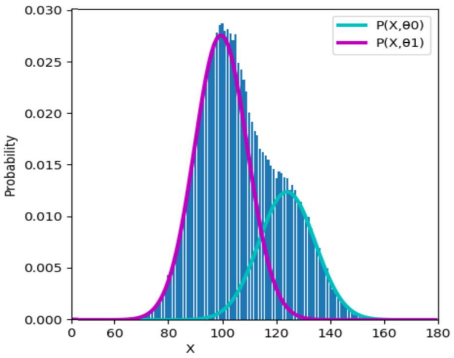
例 4.3.1. 一个混合模型有两个高斯部件(components)。真模型参数是(μ_1*，μ_2*，σ_1*，σ_2*，P*(y_1))=(100，125，10，10，0.7)。不正确收敛中心在 μ_1-μ_2 平面上大约是 (μ_1，μ_2)=(115，95)，这里 Q 局部极大。

笔者实验表明，局部最大是因为样本较小，当样本增大到 50000 万时，不管什么初始(μ_1，μ_2)，EM 和 CM-EM 算法都会收敛。 但是后者需要的迭代次数较少。图 13 显示了一个运行过程，其中初始参数不易收敛，样本尺寸是 50000。

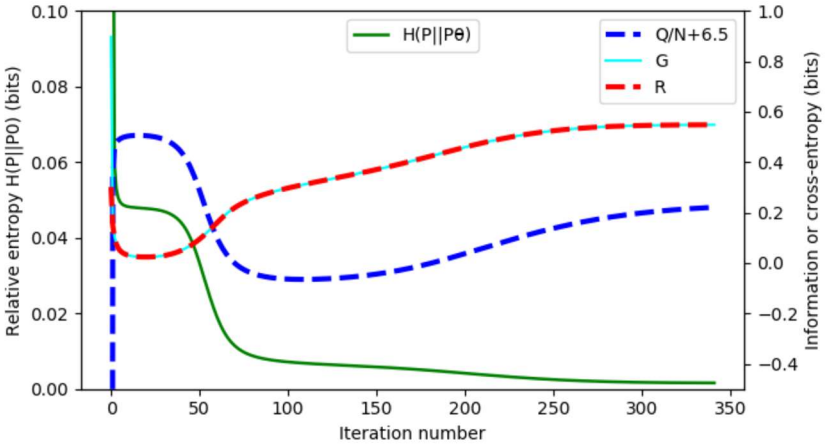
[①] 本节内容较之英文原文有较大改进。所有图片都换了；初始化地图是新加的。



(a) 初始的(μ_1，μ_2)导致 Q 局部收敛。



(b)迭代收敛时的混合模型。



(c) Q 和 $H(P||P_\theta)$随迭代数变化。初始参数是 (μ_1，μ_2，σ_1，σ_2，$P(y_1)$)= (80，95，5，5，0.5))。

图 13. EM 算法和 CM-EM 算法用于例 4.3.1。EM 算法需要约 348 次迭代，而 CM-EM 算法需要约 240 次迭代 (N=50000)。

作者提出用婚配竞争解释不同初始位置(μ_1，μ_2)的收敛难度[62]，并且通过图 14 中计算结果验证了这种解释。其中绿色圆圈是全局收敛区点，红色圆圈是不正确的局部收敛点。 因为下标交换混合效果同样，所以图中收敛点和收敛难度是对称的。图中每一对数字中左边是 EM 算法需要的迭代次数，右边是 CM-EM 算法需要的迭代次数(运行三次取中间的迭代数)。

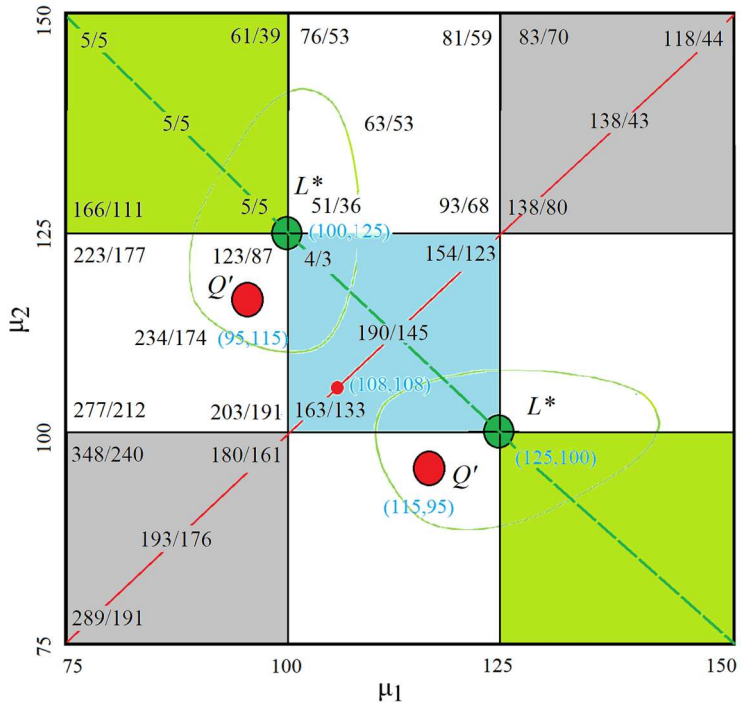


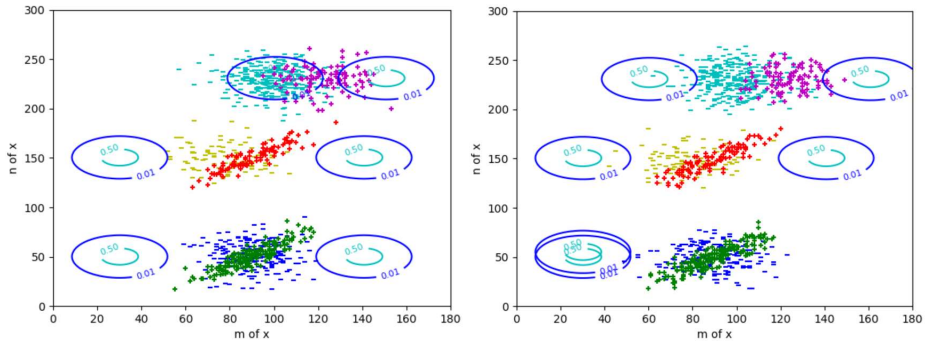
图 14. 迭代数随初始位置(μ_1，μ_2)变化。

用 EM 算法时平均迭代数是 136.7，而用 CM-EM 算法时(每次迭代时 E2-step 重复三次)，平均迭代数是 90.4。后者节省了 26%次迭代和大约 18%的时间。

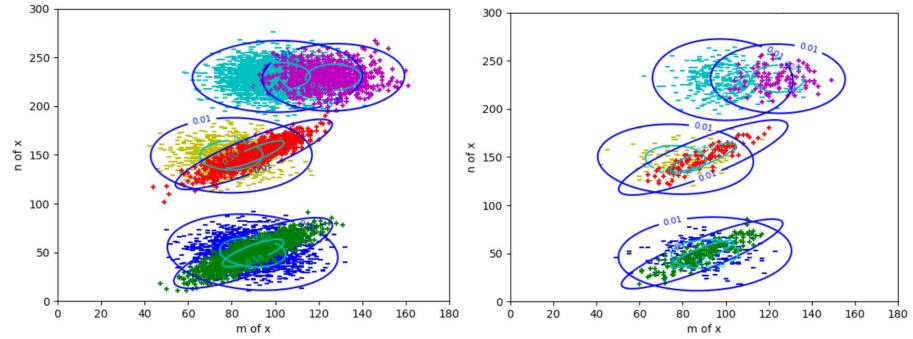
根据婚配竞争解释，图中绿色虚线是公平竞争线，红线是绝对平等线。绿色区域容易收敛，灰色区域收敛较难。路途有局部最大 Q 时，收敛也较难。要想减少迭代次数，初始点(μ_1，μ_2)要靠近公平竞争线并远离绝对公平线，特别是避免在不公平竞争区——从这些区域去收敛点 L^*（绿圈)要经过 Q'(红圈)附近。

例 4.3.2[①] 分别用 EM 算法和 CM-EM 算法求一个二维混合模型，其中有三对部件需要覆盖三对子样本，见图 14。图 14(a)中上下两对部件不能正确收敛。样本尺寸从 1000 增大到 10000 后(见图 14(c))，模型能正确收敛。按公平婚配原则调整一些部件初始中心，即使样本尺寸等于 1000，模型也能正确收敛。收敛条件是$|\mu_i-\mu_i^*|<1$ 和$|\sigma_i-\sigma_i^*|<1$，$j=1$，2。

[①] 本例对应英文原文中例 4.3.3 和图 15，具体参数有所改动。原文中例 4.3.2 被删了。



（a） 上下两对部件不能正确收敛的初始化。 (b)按公平竞争原则调整初始参数。



(c) 样本增大后模型能正确收敛。 (d) 调整初始位置后模型也能正确收敛。

图 15. EM 算法 CM-EM 算法用于两维混合模型。

当初始参数如图 14(a)而样本较大如图 14(c)时，要达到同样的收敛，CM-EM 算法需要 50-60 次迭代(每次迭代时 E2-step 重复三次)，而 EM 算法需要 70-100 次迭代。

5. 讨论

5.1 讨论确证测度b^*

归纳问题在当代已经演变为确证问题[58]。研究者已经提出很多确证测度 [59]。大多数确证测度强调较大的 $P(y_1|x_1)$ (即更多正例)是重要的，而 $b_1^*=b^*(y_1\rightarrow x_1)$ 强调较小的 $P(y_1|x_0)$ (更少反例)是重要的。例如，当敏感性 $P(y_1|x_1)=0.1$ 且特异性 $P(y_0|x_0)=1$ 时，我们能得出 $b_1^*=1$，这是合理的。然而，使用流行的确证公式[59]，这时 $y_1\rightarrow x_1$ 的确证度很小。

当敏感性是 1 的时候，如果特异性很小，比如是 0.1，b_1*=b*($y_1\rightarrow x_1$)=0.1。而用流行的确证测度算出的确证度远远大于 0.1。这是不合理的，因为这时反例的比例是 0.9/1.9≈0.47≈0.5，这意味 $y_1\rightarrow x_1$ 几乎是不可信的。

从上面两个例子我们能看出，测度 b* 强调没有反例(对于非模糊大前提)或较少的反例(对于模糊大前提)比较多的正例更重要。所以这一测度兼容波普尔的证伪思想[31，32]。测度 b*也兼容似然比[①] ，因而能得到医学实践的支持。

Eells 和 Fitelson [60]建议用假设对称性标准评价各种确证测度。如果这一对称性成立，应有 b *($y_1\rightarrow x_1$)= -b *($y_1\rightarrow$not x_1)=-b*($y_1\rightarrow x_0$)。我们能证明这一对称性对于 b*成立，因为:

$$\begin{aligned} b^*(y_1 \rightarrow x_1) &= \frac{P(y_1 \mid x_1) - P(y_1 \mid x_0)}{\max(P(y_1 \mid x_1), P(y_1 \mid x_0))} \\ &= -\frac{P(y_1 \mid x_0) - P(y_1 \mid x_1)}{\max(P(y_1 \mid x_0), P(y_1 \mid x_1))} = -b^*(y_1 \rightarrow x_0). \end{aligned} \tag{5.1}$$

然而，上述确证测度和流行的所有确证测度都不能用来澄清乌鸦悖论。作者在另一篇文章[61]中提出要区分信道确证测度(如 b*)和预测确证测度(取名为 c*)。后者可以用来澄清乌鸦悖论。

5.2 讨论CM2用于多标签分类

和流行的多标签学习方法(比如二元关联方法 [52])相比，CM3 优化 n 对互补标签时并不需要 n 个样本。它能直接从香农信道 $P(Y|X)$或样本分布 $P(x，y)$得到由一组真值函数构成的语义信道。

好最大后验概率准则相比，最大语义信息准则能减少小概率事件的漏报。在信息比正确率重要时，最大语义信息准则更好。

注意：在图 10(b)中，“老年人”(即“Old”)的下界不是 60 而是 58.1。这是因为“老年人”的逻辑概率小于“中年人”(“Middle age”)的逻辑概率。如果大家寿命延长，“老年人”的下界将右移。可以想象，新的划分边界将产生新的条件样本分布 $P(x|y_5)$和新的真值函数 $T(\theta_5|x)$; 新的真值函数进一步使边界右移… “老年人”的真值函数和语义就是这样随人类寿命的延长而改变的。

5.3 讨论CM3用于不可见实例最大互信息分类

在机器学习和经典信息论中[19，20]，求解最大互信息分类都是很困难的任务。香农和很多研究者使用最小平均失真准则 [7，6] 而不是最大互信息准则优化检测

[①] 英文原文中认为兼容置信水平 CL，现在纠正为似然比 LR。

和估计。如果使用最大互信息准则，为预测的残余误差编码将需要较短的平均码长。他们为什么不用最大互信息准则？原因是求解最大互信息划分是非常困难的。现在使用 CM3，我们至少可以容易解决低维特征空间的最大互信息分类。

流行的最大互信息分类和估计方法是使用参数构造转移概率函数或似然函数，然后用梯度下降法或牛顿法优化这些参数。优化的参数确定划分边界。然而，CM3 或 CM 迭代算法分别用参数构造 n 个似然函数——覆盖 n 个类别，然后优化划分边界或不同 z 的标签。表 6 比较了 CM3 和梯度下降法。

表 6. 比较信道匹配算法和梯度下降法用于低维特征空间的最大互信息分类。

关于	梯度下降法	信道匹配算法
不同类别的模型	一起优化	分开优化
边界表达	带参数函数	数值
用于复杂边界	不容易	容易
考虑梯度和搜索	需要	不需要
收敛	不容易	容易
计算	复杂	简单
需要的迭代数	很多次	2-3 次
需要的样本	不必很大	要足够大

信道匹配迭代算法也有两个缺点：1)它需要每个子样本足够大以便为 n 个类别构造 n 个似然函数；2)对于高维特征空间，为每个 z 贴标签是不可行的。因此，对于高维特征空间分类，我们需要结合 CM3 和神经网络。

一个神经网络就是一个分类器 $y=f(z)$。对于给定的神经网络，匹配 I 就是让语义信道匹配香农信道，得到奖励函数 $I(X;\theta_j|z)$ (j=0，1，…)。对于给定的奖励函数，匹配 II 就是让香农信道匹配语义信道，获得新的神经网络参数。重复两种匹配就能使 $I(X;\theta)$ 收敛到最大互信息。匹配 I 和匹配 II 类似于生成竞争网(GAN)中的两个任务。结合 CM3 和流行的深度学习方法[33]，我们应能改进高维特征空间的最大互信息分类。

5.4 讨论CM4用于混合模型

4.3 节中的结果表明：完全数据的对数似然度 Q 和不完全数据的对数似然度 $L_X(\theta)$ 并不像很多人相信的那样总是正相关的。在很多情况下，Q 可以且应该减小，因为 Q 可能大于正确收敛的 $Q^*=Q(\theta^*)$。在例 3.4.1 中，假定真模型参数 $\sigma_1^*=\sigma_2^*=\sigma^*$ 和 $P^*(y_1)=P^*(y_2)=0.5$，我们能证明 $P(y_1|x)$和 $P(y_2|x)$是一对 Logistic 函数，它们在 σ 减小时变得陡峭。因此，H 会因 σ 减小而增大。我们能证明偏导数$\partial H/\partial\sigma^*>0$ [57]。因此，在 $\theta=\theta^*$，

$$\frac{\partial Q}{\partial \sigma}=\frac{\partial L_X(\theta)}{\partial \sigma}-\frac{\partial H}{\partial \sigma}=0-\frac{\partial H}{\partial \sigma}<0.$$

所以，我们总能找到一个正数$\triangle$ 并且用 $\sigma^*-\triangle$ 代替 σ^*，使得 $Q(\sigma^*-\triangle)>Q(\sigma^*)$。

新的收敛理论解释：CM4 能收敛是因为迭代能最大化 G/R 或最小化 $R-G$。

我们使用不同例子检验了用于混合模型的 CM-EM 算法。实验表明，使用图 14 所示初始化地图，在存在 Q 的局部极大的情况下，CM-EM 算法能更明显加速收敛且减少不正确收敛。对于大多数二元高斯混合，使用初始化地图，CM-EM 算法不超过 10 次迭代就能全局收敛。和其他改进的 EM 算法[56，63]相比，结合初始化地图的 CM-EM 算法能更好应对各种可能混合。

CM-EM 不仅能用于高斯混合，也能用于其他混合。对于其他混合，MG-step 要困难些。但是收敛证明是相同的。

CM4 和 CM3 一起可以用于无监督学习。使用 CM4，我们能从样本分布 $P(x)$获得一组模型参数。使用 CM3，我们能做最大互信息分类。

但是对于更多部件的混合，CM-EM 算法不能避免参数收敛到参数空间的边界。为实现混合模型的全局收敛，我们需要结合现存的算法，比如分裂合并(Split and Merge) EM 算法[64] 和竞争(Competitive)EM 算法[65]。

6. 结论

语义信息 G 理论结合了香农、波普尔、费希尔、扎德和卡尔纳普 等人的思想。语义信息测度或 G 测度像卡尔纳普 and Bar-Hillel 的语义信息测度一样随逻辑概率减小而增加；但是 G 测度也随相对偏差增加而减小。因此，G 测度能用于假设检验。

逻辑贝叶斯推断(LBI)使用真值函数而不是贝叶斯后验作为推断工具。使用真值函数 $T(\theta_j|x)$，我们能在先验分布 $P(x)$不同的情况下做概率预测，就像使用转移概率函数 $P(y_j|x)$或反概率函数 $P(\theta_j|x)$。然而，从样本分布获得优化的真值函数要比获得其他两种函数容易得多，因为我们不需要 $P(y_j)$或 $P(\theta_j)$。更重要的是，真值函数能代表假设或标签的语义，因而能更好连接统计和逻辑。一个意外的收获是真值函数的优化为归纳理论带来一个看来合理的确证测度。

一组信道匹配算法 CM1、CM2、CM3 和 CM4 可以用来改进机器学习，特别是用来解决可变 $P(x)$的多标签学习问题。CM1 可以用于改进标签学习和确证；CM2 可以用于改进多标记分类; CM3 可以用于改进低维特征空间的最大互信息分类；CM4 可以用于改进混合模型，特别是解释混合模型收敛。G 理论和逻辑贝叶斯推断应该已经经受了机器学习的检验。

为了把语义信息 G 理论和逻辑贝叶斯推断用于机器学习，未来我们还需要结合信道匹配算法和神经网络。逻辑贝叶斯推断应能进一步发展，从而促进逻辑和统计的统一。

附录：补充材料

补充材料包括若干Python 3.6源程序，可从下面地址下载
http://survivor99.com/lcg/cm/forZWGtheory.zip

表1. 补充材料清单

程序名	任务
Bayes Theorem III 2.py	为图 7——显示贝叶斯定理 III 的第二个公式。
Ages-MI-classification.py	为图 10——显示根据年龄的人群分类。
MMI-v.py	为图 11、12——显示 CM 算法用于最大互信息分类。
LocationTrap63.py	为图 13、14——显示 CM-EM 算法用于一维混合模型。
Ex432.py	为图 15——显示信道匹配 EM 算法用于二维混合模型。

参考文献

1. Fisher R A. On the mathematical foundations of theoretical statistics. Philo. Trans. Roy. Soc., 1922(A222): 309-368.
2. Fienberg S E. When Did Bayesian Inference Become “Bayesian"? Bayesian Analysis, 2006, 1(1):1-40.
3. Bayesian Inference, In: Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Bayesian_inference (Uploaded on 2019-7-17)
4. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6)：716–723.
5. Kullback, S.; Leibler, R. On information and Sufficiency. Annals of Mathematical Statistics 1951, 22(1):79–86.
6. Cover, T. M.; Thomas, J. A. Elements of Information Theory, John Wiley & Sons: New York, USA, 2006.
7. Shannon C E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27, 379–429 and 623–656.
8. Weaver W. Recent Contributions to the Mathematical Theory of Communication// Shannon C.E. and Weaver W. Eds, The Mathematical Theory of Communication. Urbana: The University of Illinois Press, 1963, 93-117.
9. Carnap R., Bar-Hillel Y. An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT, 1952.
10. Bonnevie E. Dretske's semantic information theory and metatheories in library and information science. Journal of Documentation. 2001, 57(4): 519-34.
11. Floridi L. Outline of a theory of strongly semantic information. Minds and Machines, 2004(14): 197-221.
12. Zhong Y X. A theory of semantic information. China Communications, 2017(14):1-17. [CrossRef]

13. D'Alfonso S. On Quantifying Semantic Information. Information, 2011(2): 61-101. [CrossRef]
14. De Luca A.; Termini S. A definition of a non-probabilistic entropy in setting of fuzzy sets, Information and Control, 1972, 20(4): 301-312.
15. Bhandari D.; Pal N R. Some new information measures of fuzzy sets. Information Science 1993, 67(3): 209-228.
16. Kumar T.; Bajaj R K.; Gupta, B. On some parametric generalized measures of fuzzy information, directed divergence and information Improvement. International Journal of Computer Applications, 2011, 30(9): 5-10.
17. Klir G. Generalized information theory. Fuzzy Sets and Systems, 1991, 40(1): 127-142.
18. Wang Y. Generalized Information Theory: A Review and Outlook. Information Technology Journal, 2011, 10(3):461-469.
19. Belghazi I.; Rajeswar S.; Baratin A., Hjelm R D.; Courville A. Mine: Mutual information neural estimation, In Proceedings of International Conference on Machine Learning, Long Beach, USA, June 2018. https://arxiv.org/abs/1801.04062, (accessed on 1 Jan. 2019).
20. Hjelm R D.; Fedorov A. Lavoie-Marchildon S.; Grewal K.; Bachman P.; Trischler A.; Bengio Y. Learning deep representations by mutual information estimation and maximization. https://arxiv.org/abs/1808.06670 (accessed on 22 Feb 2019).
21. Lu C. Shannon equations reform and applications, BUSEFAL, 1990, 44, 45-52.
22. 鲁晨光. B-模糊集合准布尔代数和广义熵公式, 《模糊系统和数学》, 1991, 5(1)：76-80.
23. 鲁晨光. 《广义信息论》, 中国科技大学出版社, 1993。
24. 鲁晨光. 广义熵和广义互信息的编码意义, 《通信学报》, 1994, 5(6): 37-44.
25. Lu C. A generalization of Shannon's information theory. Int. J. of General Systems, 1999, 28(6): 453-490.
26. 鲁晨光. GPS 信息和限误差信息率 ——及其和信息率失真及复杂性失真之间的关系, 《成都信息工程学院学报》, 2012(6): 27-32.
27. Zadeh L A. Fuzzy Sets. Information and Control, 1965(8): 338–53.
28. Tarski A. The semantic conception of truth: and the foundations of semantics. Philosophy and Phenomenological Research, 1994, 4(3): 341–376.
29. Davidson D. Truth and meaning. Synthese, 1967, 17(3): 304-323.
30. Shannon C E. Coding theorems for a discrete source with a fidelity criterion, IRE Nat. Conv. Rec., 1959(4): 142–163.
31. Popper K. The Logic of Scientific Discovery, Hutchinson: London, 1959(德文版 1935).
32. Popper K. Conjectures and Refutations, Routledge: London and New York, 2002.
33. Goodfellow I.; Bengio Y. Deep Learning, The MIP Press: Cambridge, 2016.
34. Carnap R. Logical Foundations of Probability, University of Chicago Press: Chicago, 1950.
35. Zadeh L A. Probability measures of fuzzy events. J. of Mathematical, Analysis and Applications, 1986, 23(2)：421-427.

36. Floridi L. Semantic conceptions of information. In Stanford Encyclopedia of Philosophy, http://seop.illc.uva.nl/entries/information-semantic/ (substantive revision Jan 7, 2015).
37. Theil H. Economics and information theory. Amsterdam and Rand McNally, Chicago: North-Holland Pub. Co., 1967.
38. Donsker M.; Varadhan S. Asymptotic evaluation of certain Markov process expectations for large time IV. Communications on Pure and Applied Mathematics, 1983, 36(2): 183-212.
39. Wittgenstein L. Philosophical Investigations. Oxford: Basil Blackwell Ltd, 1958.
40. Bayes T.; Price R. An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society of London, 1763, 53, 370–418.
41. Lu C. From Bayesian inference to logical Bayesian inference: A new mathematical frame for semantic communication and machine learning//Shi, Z.Z. Ed. Intelligence Science II. Proceedings of ICIS2018, Springer International Publishing: Switzerland, 2018, 11-23.
42. Lu, C. Channels' matching algorithm for mixture models// Shi Z.Z.; Goertel B.; Feng J L. Eds. Intelligence Science I. Proceedings of ICIS 2017, Switzerland: Springer International Publishing, 2017. 321-332.
43. Lu C. Semantic channel and Shannon channel mutually match and iterate for tests and estimations with maximum mutual information and maximum likelihood// Proceedings of 2018 IEEE International Conference on Big Data and Smart Computing, Washington: IEEE Computer Society Press Room. 2018. 15-18.
44. Lu C. Semantic channel and Shannon channel mutually match for Multilabel classification// Shi, Z.Z. Ed. Intelligence Science II, Switzerland: Springer International Publishing, 2018, 37-48.
45. Dubois D.; Prade H. Fuzzy sets and probability: misunderstandings, bridges and gaps. Proceedings 1993 Second IEEE International Conference on Fuzzy Systems, San Francisco, USA, 1993, 1059-1068.
46. Thomas S F. Possibilistic uncertainty and statistical inference. Proceedings of ORSA/TIMS Meeting, Houston, Texas, 1981.
47. Wang P Z. From the fuzzy statistics to the falling fandom subsets// Wang P P. Ed. Advances in Fuzzy Sets, Possibility Theory and Applications, New York: Plenum Press, 1983, 81–96.
48. Berger T. Rate Distortion Theory. New Jersey: Prentice-Hall, Enklewood Cliffs, 1971.
49. Thornbury J R.; Fryback D G.; Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology, 1975, 114(3): 561–565.
50. http://www.oraquick.com/Home (上载于 2016-12-31)
51. Zhang M L; Zhou Z H. A review on Multilabel learning algorithm. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.
52. Zhang M L.; Li Y K.; Liu X Y.; Geng X. Binary relevance for Multilabel learning: an overview. Frontiers of Computer Science, 2018, 12(8): 191-202.

53. 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, 1997, 39(1): 1–38.
54. Ueda N.; Nakano R. Deterministic annealing EM algorithm, Neural Networks, 1998, 11(2), 271-282.
55. Marin J M.; Mengersen K.; Robert C P. Bayesian modelling and inference on mixtures of distributions// Dey, D.; Rao, C. R. Eds., Handbook of Statistics: Bayesian thinking, modeling and computation, 25, Amsterdam: Elsevier, 2011, 459-507.
56. Neal R.; Hinton G. A view of the EM algorithm that justifies incremental, sparse, and other variants// Michael I. Jordan Ed., Learning in Graphical Models, Cambridge: MIT Press, 1999, 355–368.
57. Lu C. From the EM Algorithm to the CM-EM Algorithm for Global Convergence of Mixture Models, https://arxiv.org/abs/1810.11227 (上载于 2018-10-26).
58. James H. Inductive Logic// Zalta E N. Ed., The Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/archives/spr2018/entries/logic-inductive/ (Uploaded on 2018-3-19).
59. Tentori K.; Crupi V.; Bonini N.; Osherson D. Comparison of confirmation measures. Cognition, 2007, 103(1): 107–119.
60. Ellery E.; Fitelson B. Measuring confirmation and evidence. Journal of Philosophy 2000, 97(12): 663–672.
61. Lu C. Channels' Confirmation and Predictions' Confirmation: From the Medical Test to the Raven Paradox, Entropy, 2020, 22(4), 384. https://doi.org/10.3390/e22040384
62. Lu C. Fair Marriage Principle and Initialization Map for the EM Algorithm, https://arxiv.org/abs/2007.12845 (Uploaded on 2020-7-25)
63. Huang W H.; Chen Y G. The multiset EM algorithm. Statistics & Probability Letters, 2017, 126, 41–48.
64. Ueda N.; Nakano R.; Ghahramani Z.; Hinton G E. SMEM algorithm for mixture models. Neural Comput. 2000, 12(9): 2109-2128.
65. Zhang B.; Zhang C., Yi X. Competitive EM algorithm for finite mixture models. Pattern Recognition, 2004, 37(1): 131-144.

Semantic Information G Theory and Logical Bayesian Inference for Machine Learning

Chenguang Lu

Abstract: An important problem with machine learning is that when label number $n>2$, it is very difficult to construct and optimize a group of learning functions, and we wish that optimized learning functions are still useful when prior distribution $P(x)$ (where x is an instance) is changed. To resolve

this problem, the semantic information G theory, Logical Bayesian Inference (LBI), and a group of Channel Matching (CM) algorithms together form a systematic solution. A semantic channel in the G theory consists of a group of truth functions or membership functions. In comparison with likelihood functions, Bayesian posteriors, and Logistic functions used by popular methods, membership functions can be more conveniently used as learning functions without the above problem. In Logical Bayesian Inference (LBI), every label's learning is independent. For Multilabel learning, we can directly obtain a group of optimized membership functions from a big enough sample with labels, without preparing different samples for different labels. A group of Channel Matching (CM) algorithms are developed for machine learning. For the Maximum Mutual Information (MMI) classification of three classes with Gaussian distributions on a two-dimensional feature space, 2-3 iterations can make mutual information between three classes and three labels surpass 99% of the MMI for most initial partitions. For mixture models, the Expectation-Maxmization (EM) algorithm is improved and becomes the CM-EM algorithm, which can outperform the EM algorithm in cases where local convergence is easy to happen. The information rate verisimilitude function in the G theory can help us to prove the convergence of the EM and CM-EM algorithms. The CM iteration algorithm needs to combine neural networks for MMI classifications on high-dimensional feature spaces. LBI needs further studies for the unification of statistics and logic.

Keywords: semantic information theory; Bayesian inference; machine learning; multilabel classifications; maximum mutual information classifications; mixture models; confirmation measure; truth function.