

用语义信息 G 测度解释和推广信息率-失真函数和最大熵分布

(英文原文发表于 Entropy, 见 <https://www.mdpi.com/1099-4300/23/8/1050>)

鲁晨光

摘要: 在信息率-失真函数和最大熵方法中, 表示最小互信息分布和最大熵分布的公式就像是贝叶斯公式——包含负指数函数和划分函数. 为什么这些非概率函数存在于类贝叶斯公式之中? 另一方面, 信息率-失真函数有三个缺点: (1)失真函数是主观定义的; (2)实例和标签之间的失真的定义往往非常困难; (3)不能根据标签的语义做数据压缩. 作者曾提出语义信息 G 测度——它同时含有统计概率和逻辑概率. 据此, 我们能把负指数函数解释为真值函数, 把划分函数解释为逻辑概率, 把类贝叶斯公式解释为语义贝叶斯公式, 把最小互信息解释为语义互信息, 把最大熵解释为极度最大熵减去语义互信息. 为了克服上述缺点, 本文建立真值函数和失真函数之间的联系, 通过机器学习从样本分布获得真值函数, 用真值函数构造约束条件从而推广信息率-失真函数. 其中用两个例子帮助读者理解最小互信息迭代并支持理论结果. 使用真值函数和语义信息 G 测度, 我们可以结合机器学习和数据压缩 (包括语义压缩). 为探讨一般数据的语义压缩和恢复, 我们需要做进一步研究.

关键词: 信息率-失真函数; Boltzmann 分布; 语义信息测度; 机器学习; 最大熵; 最小互信息; 贝叶斯公式

1. 引言

贝叶斯(Bayes)公式用于概率预测, 从一个联合(概率)分布 $P(x, y)$ 或一个条件分布 $P(x|y)$ 和先验分布 $P(y)$ (x 和 y 是变量), 我们能得到 y 的后验分布:

$$P(y|x) = \frac{P(x, y)}{\sum_y P(x, y)} = \frac{P(y)P(x|y)}{\sum_y P(y)P(x|y)} = \frac{P(y)P(x|y)}{P(x)}. \quad (1)$$

类似的带有负指数函数和划分函数的表达式经常出现在信息率-失真理论【1-3】、统计力学、最大熵方法【4,5】、机器学习 (见有限波尔茨曼机, Restricted Boltzmann Machine (RBM) 【6,7】和 Softmax 函数中【8,9】). 比如, 在信息率-失真理论中, 最小互信息分布是:

$$P(y|x) = \frac{P(y) \exp[sd(x, y)]}{\sum_y P(y) \exp[sd(x, y)]} \quad (2)$$

其中 s 是负的参数, $d(x, y)$ 是失真函数.

然而, 一个负指数函数, 比如 $\exp[sd(x, y)]$, 并不是一个 (统计) 概率函数, 因为 $\sum_x \exp[sd(x, y)]$ 和 $\sum_y \exp[sd(x, y)]$ 并不是 1。其主要特点是其最大值 $\exp(0) = 1$ 。一个负指数函数更像是一个隶属函数【10】、相似性函数或分布约束函数——它软约束一个概率分布。为什么我们可以把负指数函数放进类贝叶斯公式? 我们能找到一个简单解释——解释为什么负指数函数和划分函数存在于不同领域? 虽然信息率-失真函数和最大熵之间的关系已有人研究【11】, 我们仍然需要一个更一般的概率理论或框架去解释上述现象。

另一方面, 虽然信息率-失真理论已经取得很大成功【2,3,12–14】, 它仍然有下面缺点:

- 失真函数是主观定义的, 缺乏客观标准;
- 机器学习中我们用标签取代实例, 前者数目远小于后者, 要定义标签和实例之间的失真函数是困难的。比如, 我们需要用“小雨”, “中雨”, “大雨”, “小到中雨”等取代日降水量, 或者用“儿童”, “年轻人”, “成年人”, “老年人”等取代人的年龄。在这些情况下, 构造失真函数是不容易的;
- 我们不能把信息率-失真函数用于语义压缩, 即根据标签的语义做数据压缩。

前两个缺点提示我们;我们需要通过机器学习从样本或样本分布获得许可的失真函数。

我们考虑涉及年龄的语义压缩。比如, x 表示年龄(实例), y_i 表示标签: y_1 = “非成年人”, y_2 = “年轻人”, y_3 = “成年人”, and y_4 = “老年人”。所有使 y_i 为真的 x 构成一个模糊集合。这些标签的真值函数也就是模糊集合的隶属函数(见图 1)。根据 Davidson 的真值条件语义学【15】, 真值函数 $T(y_i|x)$ 确定了 y_i 的语义。对于给定的 $P(x)$ 和约束, 我们需要找到最小化互信息的香农信道 $P(y|x)$ 。最小互信息就是我们把 x 编码为 y 的最小平均码长的下限。约束条件是: 标签选择应当符合语言使用规则——它由真值函数确定。在这种情况下, 信息率-失真函数不好使用, 这是因为很难定义兼容标签语义的失真函数。

关于语义压缩, 已有一些有意义研究【16–18】。然而, 这些研究要么没有使用联系到信息率-失真函数或类似函数的信息理论方法, 要么没有使用语义信息测度。联系到感觉的数据压缩或聚类也有人研究【19,20】, 但是他们没有使用分辨率函数(类似于真值函数)和感觉信息测度。考虑语义压缩, 我们需要一个适当的信息测度度量语义信息和感觉信息。我们也需要一个函数——类似于信息率-失真函数——它涉及标签的语义。

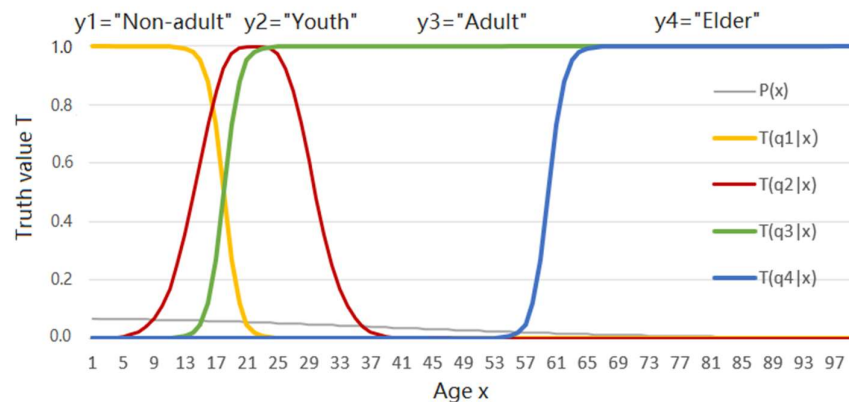


图 1. 关于年龄的 4 个标签的（模糊）真值函数。 $T(y_j|x)$ 表示命题 $y_j(x)$ ($j=1, 2, 3, 4$) 的真值函数，作为约束条件。

为度量语义信息，研究者已经提出许多语义信息测度【21–25】或联系到语义的信息测度【26,27】。然而，要把这些测度用于机器学习和语义压缩还是不容易的。为了类似的目的，作者于上个世纪 90 年代提出语义信息 G 理论（或简单地说，G 理论）【28–30】。字母“G”意味着香农信息论推广。语义信息测度（即 G 测度）被定义为 $\log(\text{真值函数}/\text{逻辑概率})$ 或其平均。其中真值函数和逻辑概率类似于负指数函数和划分函数。但是真值函数不仅能用负指数函数表示，也能用 Logistic 函数和其他数值在 0 和 1 之间的函数表示。语义信息 G 测度可用来度量自然语言的信息，也能用来度量感觉信息和测量仪器读数的信息，比如温度表、秤和 GPS 装置的信息【31】。如果度量感觉或读数的信息，真值函数就变成混淆概率函数【30】。

G 测度度量标签语义只根据标签的外延——它由真值函数确定——而不考虑其内涵。因此，我们可以称此语义信息为形式语义信息。为了简化问题，本文主要考虑标签 y 和实例 x 之间的语义信息。6.4 节将简单讨论两个标签之间的语义信息。

G 理论使用 P-T 概率框架【32】，它同时包含统计概率（P）和逻辑概率（T）。P-T 概率框架和 G 理论已经被应用于好几个领域，比如：联系到视觉分辨率的数据压缩【30】、机器学习【31】、贝叶斯确证【33】和科学哲学【32】。

为了克服信息率-失真函数的缺点，本文建立真值函数和失真函数之间的转换关系，使用真值函数代替失真函数作为约束条件，从而获得信息率-真函数。因为真值函数也能被解释为隶属函数、相似性函数、混淆概率函数、分布约束函数...【32】，所以它能用作学习函数——通常用负指数函数或 Logistic 函数表示。这样，我们就能克服信息率失真函数的三个缺点，因为：

- 真值函数是学习函数，我们能从样本或样本分布得到它【31】，因此它可以不是主观定义的；
 - 使用转换关系，我们能间接表达任何实例和标签之间的失真函数——通过机器学习得到真值函数；
 - 真值函数表示标签的语义，因此可以作为语义压缩的约束条件。
- 结合作者先前的研究【31, 32】，我们可以解释：
- 负指数函数和划分函数分别是真值函数和逻辑概率；

- 导致最小互信息分布或最大熵的分布 (如公式 (2) 表示) 所用的公式就是语义贝叶斯公式【31】;
- 最小互信息 $R(D)$ 可用语义互信息表示;
- 最大熵等于极度最大熵减去语义互信息.

新的解释并不冲突而是补充现存的解释。

我们能模糊真准则代替失真准则使得信息率-失真函数变成信息率-真函数 $R(\Theta)$, 其中 Θ 是一组模糊集合或模糊真值函数。我们也能使用语义信息准则——它兼容似然准则——取代失真准则, 使得信息率-失真函数变成信息率-逼真函数 $R(G)$, 其中 G 是语义互信息下限。两种函数都可以用于语义压缩。信息率-逼真函数已经在【30,31】中介绍, 而信息率-真函数于本文首次提出。

本文目的是:

- 帮助读者从新生的角度理解信息率-失真函数和最大熵分布;
- 结合机器学习 (得到失真函数) 和数据压缩;
- 揭示: 信息率-失真函数可以推广到信息率-真函数和信息率-逼真函数, 用于通信数据的语义压缩。

本文将提供一个例子显示, 对于给定信源和一组真值函数, 香农信道如何匹配语义信信道从而实现最小互信息 $MMI R(\Theta)$ 和 $R(G)$ 。它也提供了另一个例子——显示信息率-失真函数可以用于数据简化 (即降低数据精度的数据压缩)。结果支持理论分析。

新的解释应能更紧密结合经典信息论、语义信息 G 理论、最大熵理论、似然方法、模糊集合论等理论。信息率-失真函数的推广应能适用于语义压缩并且有助于解释使用负指数函数和 Soft-max 的机器学习。新的解释反过来也支持 P-T 概率框架和语义信息 G 理论。

2. 背景

2.1. 香农熵和互信息

定义 1.

- 变量 x 表示一个实例; X 表示一个离散随机变量, 它取值是 $x \in U = \{x_1, x_2, \dots, x_m\}$.
- 变量 y 表示一个假设或标签; Y 表示一个离散随机变量, 它取值 $y \in V = \{y_1, y_2, \dots, y_n\}$.

香农称 $P(X)$ 为信源, $P(Y)$ 为信宿, $P(Y|X)$ 为信道, $P(y_j|x)$ (y_j 一定而 x 可变) 为转移概率函数【34】 (p. 18)。一个香农信道是一个转移概率矩阵或一组转移概率函数:

$$P(y|x): P(y_j|x), j = 1, 2, \dots, n.$$

香农互信息是:

$$\begin{aligned}
 I(X; Y) &= \sum_j \sum_i P(x_i, y_j) \log \frac{P(x_i | y_j)}{P(x_i)} = H(X) - H(X | Y) \\
 &= \sum_j \sum_i P(x_i, y_j) \log \frac{P(y_j | x_i)}{P(y_j)} = H(Y) - H(Y | X),
 \end{aligned} \quad (3)$$

其中 $H(X)$ 和 $H(Y)$ 是 X 和 Y 的香农熵; $H(X|Y)$ 和 X 和 Y 的 $H(Y|X)$ 是香农条件熵.

给定 $Y = y_j$ 时, $I(X; Y)$ 就变成 Kullback–Leibler (KL) 离散度:

$$I(X; y_j) = \sum_i P(x_i | y_j) \log \frac{P(x_i | y_j)}{P(x_i)}. \quad (4)$$

它大于或等于 1。当 $P(x|y_j) = P(x)$, 它等于 0.

2.2. 信息率-失真函数 $R(D)$

香农在【1, 2】中提出信息率-失真函数。因为对于独立同分布信源且失真有界的率-失真函数等于相应的信息率-失真函数, Cover 和 Thomas (见【14】 (p. 307) 在大多数情况下并不区分两种函数。他们统称两者为信息率-失真函数, 并且用 $R(D)$ 表示。本文效仿他们。下面是信息率-失真函数定义。

令失真函数是 $d(x, y)$ 或 $d_{ij} = d(x_i, y_j)$, $i = 1, 2, \dots; j = 1, 2, \dots; \bar{d}$ 是 $d(x, y)$ 的平均; D 是 $\bar{d}$ 的上限。经常使用的失真函数是 $d(x, y) = (x - y)^2$ 。但是这一失真函数只适合 x 和 y 有相同的论域(即 $U = V$)且失真仅和 x 和 y 的距离有关。

给定信源 $P(X)$ 和 D 的信息率-失真函数被定义为:

$$R(D) = \min_{P(y|x): \bar{d} \leq D} I(X; Y) \quad (5)$$

通过变分方法【2, 3】, 我们能获得信息率-失真函数的参数解。约束条件是:

$$D = \sum_j \sum_i P(x_i) P(y_j | x_i) d_{ij} \quad (6)$$

$$\sum_j P(y_j | x_i) = 1, \quad i = 1, 2, \dots, m \quad (7)$$

$$\sum_j P(y_j) = 1. \quad (8)$$

拉格朗日函数因此是:

$$F = I(X; Y) - sD - \mu_i \sum_j P(y_j | x_i) - \alpha \sum_j P(y_j). \quad (9)$$

因为 $P(y|x)$ 和 $P(y)$ 是相互依赖的, 我们需要固定一个优化另一个。为优化 $P(y|x)$, 我们固定 $P(y)$ 并令 $\partial F / \partial P(y_j | x_i) = 0$ 。然后得到优化的转移概率函数或香农信道:

$$P(y_j | x_i) = P(y_j) \lambda_i \exp(sd_{ij}), i = 1, 2, \dots; j = 1, 2, \dots \quad (10)$$

$$\lambda_i = 1 / \sum_k P(y_k) \exp(sd_{ik}), \quad (11)$$

其中 $\exp()$ 是 $\log()$ 的反函数; λ_i 被定义为 $\lambda_i \equiv \exp(\mu_i/P(x_i))$ 。为优化 $P(y)$, 我们固定 $P(y_j | x_i)$ 并令 $\partial F / \partial P(y_j) = 0$ 。于是推导出 $\alpha = 1$ 和 $P(y)$:

$$P(y_j) = \sum_i P(x_i) P(y_j | x_i) \quad (11)$$

因为 $P(y|x)$ 和 $P(y)$ 相互依赖, 我们首先假设 $P(y_1) = P(y_2) = \dots = 1/n$ 并重复等式 (10) 和 (11) 直至 $P(y)$ 不变。我们称这一迭代为最小互信息迭代。

值得注意的是, 我们的目的是找到适当的转移概率函数 $P(y_j | x)$, $j = 1, 2, \dots$, 但是直接找是困难的——因为约束条件等式 (7) 和 (8) 存在。因此, 我们只能通过最小互信息迭代, 先找到适当的后验分布 $P(y | x_i)$, $i = 1, 2, \dots$, 然后间接地得到转移概率函数。

带参数的信息率-失真函数 $R(D)$ 是【3】(p. 32):

$$D(s) = \sum_i \sum_j d_{ij} P(x_i) P(y_j) \exp(sd_{ij}) / Z_i, \quad (12)$$

$$R(s) = sD(s) - \sum_i P(x_i) \log Z_i, \quad (13)$$

$$Z_i = 1 / \lambda_i = \sum_k P(y_k) \exp(sd_{ik}).$$

其中 Z_i 是划分函数。

因为参数 $s = dR/dD$ 是负的, 因此 $\exp(sd_{ij})$ 是负指数函数。于是, 较大的 $|s|$ 将导致分布较窄的 $\exp(sd_{ij})$ 、较大的 R 、和较小的 D 。

香农已经证明, 对于独立同分布信源 $P(X)$ 和平均失真上限 D , $R(D)$ 是平均码长下限。信息率-失真理论是用于数字通信的数据压缩基本理论。

因为 $I(X; Y) = H(X) - H(X|Y)$ 且 $P(X)$ 不变, 最小化 $R = I(X; Y)$ 等价于最大化 $H(X|Y)$ 。因此, 导致最小互信息分布的信道 $P(y|x)$ 也最大化后验熵 $H(X|Y)$ 。

2.3. 最大熵方法

Jaynes【4,5】首先提出最大熵方法, 并且争论说统计力学熵可以看作是信息熵的特殊应用。

对于给定的信源 $P(x)$, 最大化联合熵 $H(Y, X)$ 等价于最大化 $H(Y|X)$:

$$H(Y|X) = - \sum_i \sum_j P(x_i) P(y_j | x_i) \log P(y_j | x_i). \quad (13)$$

进一步, 假定有特征函数 $f_k(x, y)$, $k = 1, 2, \dots$ 约束条件是:

$$\sum_i \sum_j P(x_i, y_j) f_k(x_i, y_j) \geq F_k, \quad k = 1, 2, \dots \quad (14)$$

$$\sum_j P(y_j | x_i) = 1, \quad i = 1, 2, \dots \quad (15)$$

拉格朗日函数是:

$$F = H(Y|X) - \sum_k \alpha_k \sum_i \sum_j P(x_i) P(y_j | x_i) f_k(x_i, y_j) - \mu_i \sum_j P(y_j | x_i) \quad (16)$$

令 $\partial F / \partial P(y | x_i) = 0$, 我们推导出优化的香农信道 $P(y|x)$:

$$P(y|x_i) = \exp[\sum_k \alpha_k f_k(x_i, y)] / Z_i, \quad i=1,2,\dots \quad (17)$$

$$Z_i = \sum_k \exp(\sum_j \alpha_k f_k(x_i, y_j)).$$

这一 $P(y|x)$ 能最大化联合熵 $H(X, Y)$.

3. 作者相关工作

3.1. P-T 概率框架

语义是 G 理论基于 P-T 概率框架【29,30】。该框架包含两种概率：统计概率（用 P 表示）和逻辑概率（用 T 表示）。

定义 2. (定义 P-T 概率框架):

- y_j 是一个标签或假设, $y_j(x_i)$ 是一个命题; θ_j 是论域 U 中的模糊子集, 其元素使 y_j 为真。我们有 $y_j(x) \equiv "x \in \theta_j" \equiv "x \text{ 属于 } \theta_j"$ ("≡" 意思是根据定义逻辑等价)。 θ_j 也代表一个预测模型或一组模型参数。
- 通过 "=" 定义的是统计概率, 比如 $P(y_j) \equiv P(Y = y_j)$ 是一个统计概率; 而通过 " $\in$ " 定义的是逻辑概率, 比如 $P(X \in \theta_j)$ 是一个逻辑概率。为区分 $P(Y = y_j)$ 和 $P(X \in \theta_j)$, 我们定义 $T(y_j) \equiv T(\theta_j) \equiv P(X \in \theta_j)$ 为 y_j 的逻辑概率。
- $T(y_j|x) \equiv T(\theta_j|x) \equiv P(X \in \theta_j | X = x)$ 是 y_j 的真值函数, 也是模糊集合 θ_j 的隶属函数。它在 0 和 1 之间变化, 并且 $T(y|x)$ 的最大值是 1。

一个语义信道由一组真值函数构成:

$$T(y|x): T(\theta_j|x), j = 1, 2, \dots, n.$$

根据上面定义, 我们有逻辑概率:

$$T(y_j) \equiv T(\theta_j) \equiv P(X \in \theta_j) = \sum_i P(x_i) T(\theta_j | x_i) \quad (18)$$

Zadeh 称这一概率是模糊事件的概率【35】。如果 θ_j 是一个清晰集合, $T(\theta_j)$ 就变成累计概率或两个累计概率的差【36】。

一般情况下, $T(y_1) + T(y_2) + \dots + T(y_n) > 1$ 。比如, 图 1 中四个标签的逻辑概率只和大于 1, 因为 $T(y_1) + T(y_3) = 1$ 。关于统计概率和逻辑概率的区别和联系的详细讨论见【32】。

我们可以把 $T(\theta_j|x)$ 和 $P(x)$ 带入贝叶斯公式得到似然函数:

$$P(x|\theta_j) = \frac{T(\theta_j|x)P(x)}{T(\theta_j)}, \quad T(\theta_j) = \sum_i T(\theta_j|x_i)P(x_i) \quad (19)$$

我们称 $P(x|\theta_j)$ 为语义贝叶斯预测。 $P(x|\theta_j)$ 在流行的方法中经常被写作 $P(x|y_j, \theta)$ 。我们称上述公式为语义贝叶斯公式。

因为 $T(y|x)$ 的最大值是 1, 从 $P(x)$ 和 $P(x|\theta_j)$, 我们能推导出:

$$T(\theta_j|x) = \frac{T(\theta_j)P(x|\theta_j)}{P(x)}, T(\theta_j) = 1/\max[P(x|\theta)/P(x)], \quad (20)$$

其中 $\max[P(x|\theta)/P(x)]$ 意味着 $P(x|\theta)/P(x)$ 的最大值——对于不同的 x 和 y 。在作者较早的文章中[31], $T(\theta_j) = 1/\max[P(x|\theta_j)/P(x)]$ 。等式(20)中的变化能保证真值函数是对称的, 即 $T(x|y) = T(y|x)$, 就像失真函数是对称的一样。这一变化也是为了更好比较两个真值函数, 以便根据 x 和 y 之间的相关性进行分类(参看等式(23))。因为等式(19)中 $P(x|\theta_j)$ 在我们用 $cT(\theta_j|x)$ (c 是一个正的常数) 代替 $T(\theta_j|x)$ 时是不变的, 这一变化并不影响 $T(\theta_j|x)$ 的其他使用。

等式(19)和(20)构成第三贝叶斯定理[31], 可以用来把似然函数和真值函数从一个转换为另一个。

3.2. 语义信息 G 测度

作者【28】用标准(normalized)似然度定义 y_j 传递的关于 x_i 的语义信息(量):

$$I(x_i; \theta_j) = \log \frac{P(x_i|\theta_j)}{P(x_i)} = \log \frac{T(\theta_j|x_i)}{T(\theta_j)} \quad (21)$$

$I(x_i; \theta_j)$ 的值或其平均就是语义信息 G 测度, 或简单地说, G 测度。如果 $T(\theta_j|x)$ 总是 1, G 测度就变为 Carnap 和 Bar-Hillel 的语义信息测度【21】。

图 2 图解了上式。图 2 表明, 逻辑概率越小(即图中水平线越低), 信息就越多; 偏差越大, 信息就越少; 一个错误假设或预测传递的信息是负的。这些结论符合 Popper 的思想【37】(p. 294)。因此, 我们也把 $I(x_i; \theta_j)$ 解释为 y_j 和 x_i 之间的逼真度【32】。

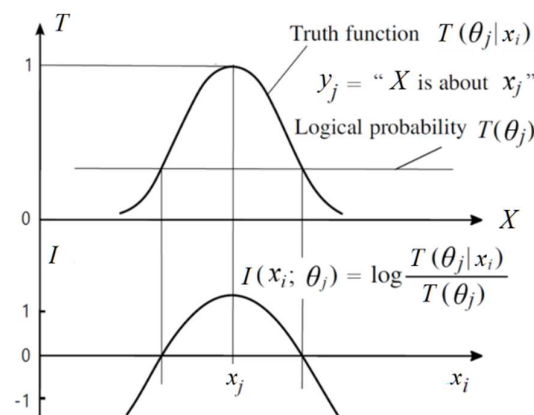


图 2. y_j 传递关于 x_i 的语义信息(量)。

我们也能用上面公式度量感觉信息, 为此, $T(\theta_j|x)$ 就是 x_j 和 x 的混淆概率函数或 x_j 的分辨率函数【31】。

通过平均 $I(x_i; \theta_j)$, 我们得到推广的 KL 信息:

$$I(X; \theta_j) = \sum_i P(x_i | y_j) \log \frac{P(x_i | \theta_j)}{P(x_i)} = \sum_i P(x_i | y_j) \log \frac{T(\theta_j | x_i)}{T(\theta_j)}, \quad (22)$$

其中 $P(x_i | y_j)$, $i = 1, 2, \dots$, 是样本分布, 它可以是不平滑或不连续的。容易证明: $I(X; \theta_j) \leq I(X; y_j)$ 【31】。

当样本很大时, 以至于 $P(x | y_j)$ 是平滑的, 我们可以令 $P(x | \theta_j) = P(x | y_j)$ or $T(\theta_j | x) \propto P(y_j | x)$, 从而获得优化的真值函数:

$$\begin{aligned} T^*(\theta_j | x) &= \frac{P^*(x | \theta_j)}{P(x)} \bigg/ \max \left(\frac{P^*(x | \theta)}{P(x)} \right) \\ &= \frac{P(x | y_j)}{P(x)} \bigg/ \max \left(\frac{P(x | y)}{P(x)} \right) = \frac{P(x, y_j)}{P(x)P(y_j)} \bigg/ \max \left(\frac{P(x, y)}{P(x)P(y)} \right). \end{aligned} \quad (23)$$

根据上式, $T^*(\theta_j | x_i) = T^*(\theta_{xi} | y_j)$ 或 $T^*(y_j | x_i) = T^*(x_i | y_j)$, 其中 θ_{xi} 是 V 的模糊子集。进一步, 我们有:

$$T^*(\theta_j | x) = cP(y_j | x) \quad (24)$$

其中 c 是常数。上式意味着优化的真值函数 $T^*(\theta_j | x)$ 正比于转移概率函数 $P(y_j | x)$ 。上式符合维特根斯坦思想: 意义在于使用 【38】 (p. 80)。

如果 $P(x | y_j)$ 是不平滑的, 我们可以用下式得到带参数的平滑的 $T^*(\theta_j | x)$:

$$T^*(\theta_j | x) = \arg \max_{T(\theta_j | x)} \sum_i P(x_i | y_j) \log \frac{T(\theta_j | x_i)}{T(\theta_j)}. \quad (25)$$

对不同的 y 求 $I(X; \theta_j)$ 的平均, 我们就得到语义互信息:

$$\begin{aligned} I(X; Y_\theta) &= \sum_j P(y_j) \sum_i P(x_i | y_j) \log \frac{P(x_i | \theta_j)}{P(x_i)} \\ &= \sum_i \sum_j P(x_i) P(y_j | x_i) \log \frac{T(\theta_j | x_i)}{T(\theta_j)} = H(Y_\theta) - H(Y_\theta | X), \end{aligned} \quad (26)$$

其中:

$$H(Y_\theta) = - \sum_j P(y_j) \log T(\theta_j), \quad (27)$$

$$H(Y_\theta | X) = - \sum_j \sum_i P(x_i, y_j) \log T(\theta_j | x_i) \quad (28)$$

$H(Y_\theta)$ 是交叉熵。因为 $\sum_j T(\theta_j) \geq 1$, 我们也称它为广义熵或语义熵。我们称 $H(Y_\theta | X)$ 为模糊熵。

当我们固定香农信道 $P(y | x)$ 并令 $P(x | \theta_j) = P(x | y_j)$ 或 $T(\theta_j | x) \propto P(y_j | x)$ (对每个 j) (匹配 I), 语义互信息 $I(X; Y_\theta)$ 达到其最大值即香农互信息 $I(X; Y)$ 。如果我们使用一组真值函数或语义信道 $T(y | x)$ 作为约束条件寻求最小互信息, 我们需要尽可能让 $P(x | y_j) = P(x | \theta_j)$ 或 $P(y_j | x) \propto T(\theta_j | x)$ (对于每个 j) (匹配 II)。4.3 和 4.4 节进一步讨论匹配 II。

如果 $T(\theta_j | x) = \exp[-(x - x_j)^2 / (2\sigma^2)]$, 我们有:

$$\begin{aligned}
 I(X; Y_\theta) &= H(Y_\theta) - H(Y_\theta | X) \\
 &= - \sum_j P(y_j) \log T(\theta_j) - \sum_j \sum_i P(x_i, y_j) (x_i - \mu_j)^2 / (2\sigma_j^2). \quad (29)
 \end{aligned}$$

容易发现，上面互信息就像是正则化最小平方（Regularized Least Square (RLS)）。 $H(Y_\theta)$ 就像是正则化项，另一个是相对误差平方项。因此，我们可以把最大语义互信息准则看作特殊的 RLS 准则。

4. 理论结果

4.1. 最小互信息分布和信息率-失真函数的新解释

我们可以把信息率-失真函数中的 $\exp[sd(x_i, y)]$ 当做真值函数，把划分函数 Z_i 当做逻辑概率。令 θ_{xi} 是 V ($V = \{y_1, y_2, \dots\}$) 的模糊子集且 $T(x_i | y) \equiv T(\theta_{xi} | y)$ 是 $y(x_i)$ 的真值函数。再令 $T(x_i) = T(\theta_{xi}) = \sum_j P(y_j) T(y_j | x_i)$ 是 x_i 的逻辑概率。于是，我们能看到：信息率-失真函数中的最小互信息分布 $P(y | x)$ 是通过语义贝叶斯公式产生的：

$$P(y | x_i) = P(y) \exp[sd(x_i, y)] / Z_i = P(y) T(x_i | y) / T(x_i), \quad i = 1, 2, \dots, m. \quad (30)$$

我们能用语义互信息公式表示 $R(D)$ 函数，因为：

$$\begin{aligned}
 I(Y; X_\theta) &= \sum_j \sum_i P(x_i, y_j) \log [T(x_i | y_j) / T(x_i)] \\
 &= \sum_j \sum_i P(x_i, y_j) \log [\exp(sd(x_i, y_j))] - \sum_i P(x_i) \ln Z_i \quad (31) \\
 &= sD(s) - \sum_i P(x_i) \log Z_i = R(D).
 \end{aligned}$$

4.2. 建立真值函数和失真函数之间的联系

现在我们通过建立真值函数和失真函数之间的联系改进 G 理论。

图 1 中四个真值函数也解释了用 y_j 代表 x_i 时的失真（量）。如果 $y_j(x_i)$ 是 1，失真 $d(x_i, y_j)$ 就是 0。失真随真值减小而增加。因此，我们使用下面定义：

定义 3. 失真函数和真值函数之间的转换关系被定义为：

$$d(x, y) \equiv \log[1/T(y | x)]. \quad (32)$$

据此，我们有 $T(y | x) = \exp[-d(x, y)]$ 和 $H(Y_\theta | X) = \bar{d}$ 。因此，我们能用模糊熵 $H(Y_\theta | X)$ 代替平均失真 $\bar{d}$ 构造约束条件。

因为 $T(y | x) = T(x | y)$ ，我们也有 $d(x, y) = \log[1/T(x | y)]$ 。

4.3. 信息率-真函数 $R(\Theta)$

作者曾提出限误差信息率函数【30】，但是这是一个不成熟的研究。这一函数更像是复杂性-失真函数【39】的推广；而信息率-真函数不同，它是信息率-失真函数的推广。

下面我们使用和推导信息率-失真函数 $R(D)$ 相同的方法得到信息率-真函数 $R(\Theta)$ ，其中 Θ 代表一组真值函数或模糊集合。约束条件 $\bar{d} \leq D$ 变成：

$$H(Y_{\Theta} | X) = -\sum_i \sum_j P(x_i) P(y_j, x_i) \log T(y_j | x_i) \leq D \quad (33)$$

仿照信息率-失真函数的推导(见公式(10)),我们得到 y 的后验分布 $P(y | x_i)$:

$$\begin{aligned} P(y | x_i) &= P(y) T(y | x_i)^{|s|} / \sum_k P(y_k) T(y_k | x_i)^{|s|} \\ &= P(y) T(x_i | y)^{|s|} / T(x_i) = P(y | \theta_{xi}), \end{aligned} \quad i=1, 2, \dots \quad (34)$$

这里我们用正的 $|s|$ 代替 $-s$ 。 $|s|$ 越大, 模糊集合的边界就越清晰, R 因此越大。

然后我们得到优化的 $P(y)$ (参看公式(11))。

我们通过重复式(11)和(34)直至 $P(y)$ 收敛(即通过最小互信息迭代), 获得最小化互信息的香农信道 $P(y | x)$ 。

把公式(34)中的 $P(y | x_i)$ 带入互信息公式, 我们就得到信息率-真函数 $R(\Theta)$:

$$\begin{aligned} R(\Theta) &= \sum_i \sum_j P(x_i) P(y_j | x_i) \log [P(y_j | x_i) / P(y_j)] \\ &= \sum_i \sum_j P(x_i) P(y_j | x_i) \log [T(x_i | y_j)^{|s|} / T(x_i)] \\ &= H(X_{\Theta}) - H(X_{\Theta} | Y) = I(Y; X_{\Theta}), \end{aligned} \quad (35)$$

其中:

$$\begin{aligned} H(X_{\Theta} | Y) &= -\sum_i \sum_j P(x_i, y_j) \log T(x_i | y_j)^{|s|} \\ T(x_i) &= \sum_j T(x_i | y_j)^{|s|} P(y_j), \\ H(X_{\Theta}) &= -\sum_i P(x_i) \log T(x_i). \end{aligned} \quad (36)$$

我们有 $R(\Theta) = H(X_{\Theta}) - H(X_{\Theta} | Y) = I(Y; X_{\Theta})$, 这意味着我们能用语义互信息 $I(Y; X_{\Theta})$ 表示最小香农互信息。当 $|s| > 1$ 时, 约束更紧, 而 $|s| < 1$ 时, 约束更松。当约束条件是原始的真值函数(或者说没有 s 或 $s = -1$)时, 我们有:

$$R(\Theta) = I(X; Y) = I(Y; X_{\Theta}) \geq I(X; Y_{\Theta}). \quad (37)$$

不等式 $I(Y; X_{\Theta}) \geq I(X; Y_{\Theta})$ 存在的理由是: 一般情况下, 迭代只能使 $P(y_j | x)$ 近似地而不是严格地正比于 $T(y_j | x)$ (参看等式(34)和 5.2 节)。

根据香农离散无记忆信源的有失真编码定理【14】(p. 307), 对于给定的 $P(X)$ 和 D , 我们能分组编码实现最小平均码长, 其下限是 $R(D)$ 。因为 $H(X_{\Theta} | Y)$ 能被理解为平均失真, $R(\Theta)$ 也意味着平均码长下限。

对于任何失真函数 $d(x, y)$, 我们怎能表示相应的真值函数为 $\exp[-d(x, y)]$ 。但是对于真值函数, 比如图 1 中的那些真值函数, 我们未必能表示相应的失真函数。因此, 信息率-失真函数可以被考虑为信息率-真函数在真值函数是 $\exp[-d(x, y)]$ 时的特例。

4.4. 信息率-逼真函数 $R(G)$

如果我们把平均失真准则改为语义互信息准则——它和似然准则兼容，那么信息率-失真函数就变成信息率-逼真函数【31】。这时，用 $I_{ij} = I(x_i; \theta_j)$ 代替 $d_{ij} = d(x_i, y_j)$ 。约束条件 $\bar{d} \leq D$ 就变成 $I(X; Y_\theta) \geq G$ ，其中 G 表示语义互信息的下限。仿照信息率-失真函数的推导，我们得到：

$$\begin{aligned} G(s) &= \sum_i \sum_j P(x_i) P(y_j | x_i) I_{ij} = \sum_i \sum_j I_{ij} P(x_i) P(y_j) \exp(s I_{ij}) / Z_i, \\ R(s) &= s G(s) - \sum_i P(x_i) \log Z_i, \end{aligned} \quad (38)$$

其中 s 是正的,且:

$$\begin{aligned} P(y_j | x_i) &= P(y_j) \left[\frac{T(y_j | x_i)}{T(y_j)} \right]^s / Z_i, \quad i = 1, 2, \dots; j = 1, 2, \dots \\ Z_i &= \sum_k P(y_k) \left[\frac{T(y_k | x_i)}{T(y_k)} \right]^s. \end{aligned} \quad (39)$$

我们也需要用最小互信息迭代优化 $P(y)$ 和 $P(y|x)$ 。函数 $P(y|x_i)$ 现在就像是 Softmax 函数，其中分子 $[T(y_j|x_i)/T(y)]^s$ 可能大于 1。在 EM 算法用于混合模型时，其 E-步所用公式和上式类似，也是为了最小化互信息【40】。

当 y 是一个预测时，比如是一个天气预报时，信息比真更重要，这时 $R(G)$ 比 $R(D)$ 和 $R(\Theta)$ 更加合适。关于 $R(G)$ 函数的更多讨论见【30,31】。

4.5. 最大熵分布的新解释和最大熵方法推广

考虑给定 $P(x)$ 时最大化联合熵 $H(X, Y)$ 。

考虑表示最大熵分布的等式 (17)，我们可以假定 $P(y)$ 是常数 $(1/n)$ ，这样 $H(Y)$ 最大。等式(17)变为：

$$\begin{aligned} P(y | x_i) &= P(y) \exp(\sum_k \alpha_k f_k(x_i, y)) / Z_i', \quad i = 1, 2, \dots \\ Z_i' &= \sum_k P(y) \exp(\sum_j \alpha_k f_k(x_i, y_j)). \end{aligned} \quad (40)$$

上面 $\exp(\cdot)$ 是一个负指数函数。我们可以把 $\exp(\cdot)$ 解释为真值函数，把 Z_i' 解释为逻辑该，把上面公式解释为语义贝叶斯公式。

然后，最大联合熵就可以表达为：

$$\begin{aligned} H(X, Y) &= H(X) + H(Y | X) = H(X) - \sum_i P(x_i) \sum_j P(y_j | x_i) \log P(y_j | x_i) \\ &= H(X) - \sum_i \sum_j P(x_i, y_j) \log [P(y_j) \exp(\sum_k \alpha_k f_k(x_i, y_j)) / Z_i'] \\ &= H(X) + H(Y) - I(Y; X_\theta), \end{aligned} \quad (41)$$

其中 $H(Y)$ 等于 $\log n$ ， $I(X; Y_\theta)$ 是语义互信息， $H(X) + H(Y)$ 是 $H(X, Y)$ 的极度最大值。因此，我们可以把最大熵解释为极度最大熵减去语义互信息。

如果我们用一组真值函数或隶属函数（如图 1 所示）作为约束条件推广最大熵方法，上面分布就变为：

$$P(y | x_i) = T(y | x_i)^{|s|} / \sum_k T(y_k | x_i)^{|s|}, \quad i = 1, 2, \dots \quad (42)$$

这一分布和等式 (34) 中的分布在 $P(y) = 1/n$ 时相同。

4.6. 波尔茨曼分布和最大熵定律的新解释

使用 Stirling 公式 $\ln N! = N \ln N - N$ (在 $N \rightarrow \infty$ 时), Jaynes 证明波尔茨曼熵和香农熵有简单联系【4,5】:

$$S = k \ln W = k \ln \frac{N!}{\prod_i N_i!} = -kN \sum_i P(x_i | T) \ln P(x_i | T) = kNH(X | T), \quad (43)$$

其中 k 是波尔茨曼常数, x_i 是第 i 种微观状态, N 是粒子数, $P(x_i | T)$ 表示给定温度 T 时状态为 x_i 的粒子出现的概率或分布密度。

波尔茨曼分布【41】是:

$$P(x_i | T) = \exp(-\frac{e_i}{kT}) / Z, \quad Z = \sum_i \exp(-\frac{e_i}{kT}), \quad (44)$$

其中 Z 是划分函数。

如果我们用能量 e_i 代替 x_i , G_i 是含有能量 e_i 的微观状态数, G 是所有状态个数, 则 $P(x_i) = G_i/G$ 是 x_i 的先验概率。于是等式(43)和(44)就变成:

$$\begin{aligned} S &= k \ln \frac{N!}{\prod_i N_i! G_i^{N_i}} = -kN \sum_i P(x_i | T) \ln \frac{P(x_i | T)}{G_i} \\ &= -kN \sum_i P(x_i | T) \ln \frac{P(x_i | T)}{P(x_i)} + kN \ln G, \end{aligned} \quad (45)$$

$$P(x_i | T) = P(x_i) \exp(-\frac{e_i}{kT}) / Z', \quad Z' = \sum_i P(x_i) \exp(-\frac{e_i}{kT}). \quad (46)$$

现在我们能将 $\exp[-e_i/(kT)]$ 解释为真值函数 $T(\theta_j | x)$, Z' 解释为逻辑概率 $T(\theta_j)$, 等式 (46) 解释为语义贝叶斯公式。

考虑局域平衡系统, 系统不同区域 y_j ($j=1, 2, \dots$) 有不同温度 T_j ($j=1, 2, \dots$)。上面 G 就变成 G_j 。然后我们能推导出(详见附录 A):

$$S / (kN) = \sum_j P(y_j) \ln G_j - I(X; Y_\theta) \quad (47)$$

这意味着热力学熵正比于极度最大熵 $\sum_j P(y_j) \ln G_j$ 减去最小互信息 $I(X; Y_\theta)$ 。这一公式表明, 物理学中最大熵定律可以等价地表述为最小互信息定律; 最小互信息 $I(X; Y)$ 等于语义互信息 $I(X; Y_\theta)$ 。

5. 实验结果

5.1. 一个实验显示求 $R(\Theta)$ 和 $R(G)$ 的最小互信息迭代中香农信道的变化

作者通过下面例 1 的实验检查了两个理论结果:

- 在求 $R(D)$ 、 $R(\Theta)$ 和 $R(G)$ 时, 最小互信息迭代使得香农信道匹配语义信道, 即让 $P(y_j | x)$ 尽可能正比于 $T(y_j | x)$ ($j=1, 2, \dots$) ;

- 最小互信息迭代能减少互信息。

例 1. 真值函数和人口年龄分布如图 1 所示 (附录 B 含有产生这些曲线的公式)。任务是: 使用上面真值函数作为约束条件, 获得产生最小互信息 $R(\Theta)$ 和 $R(G)$ 的分布 $P(y|x)$ 。

首先, 作者使用上面真值函数求 $|s| = 1$ 时的 $R(\Theta)$ 。求 $P(y|x)$ 的迭代过程如图 3 所示 (补充材料说明了产生图 3 的数据是怎样得到的)。

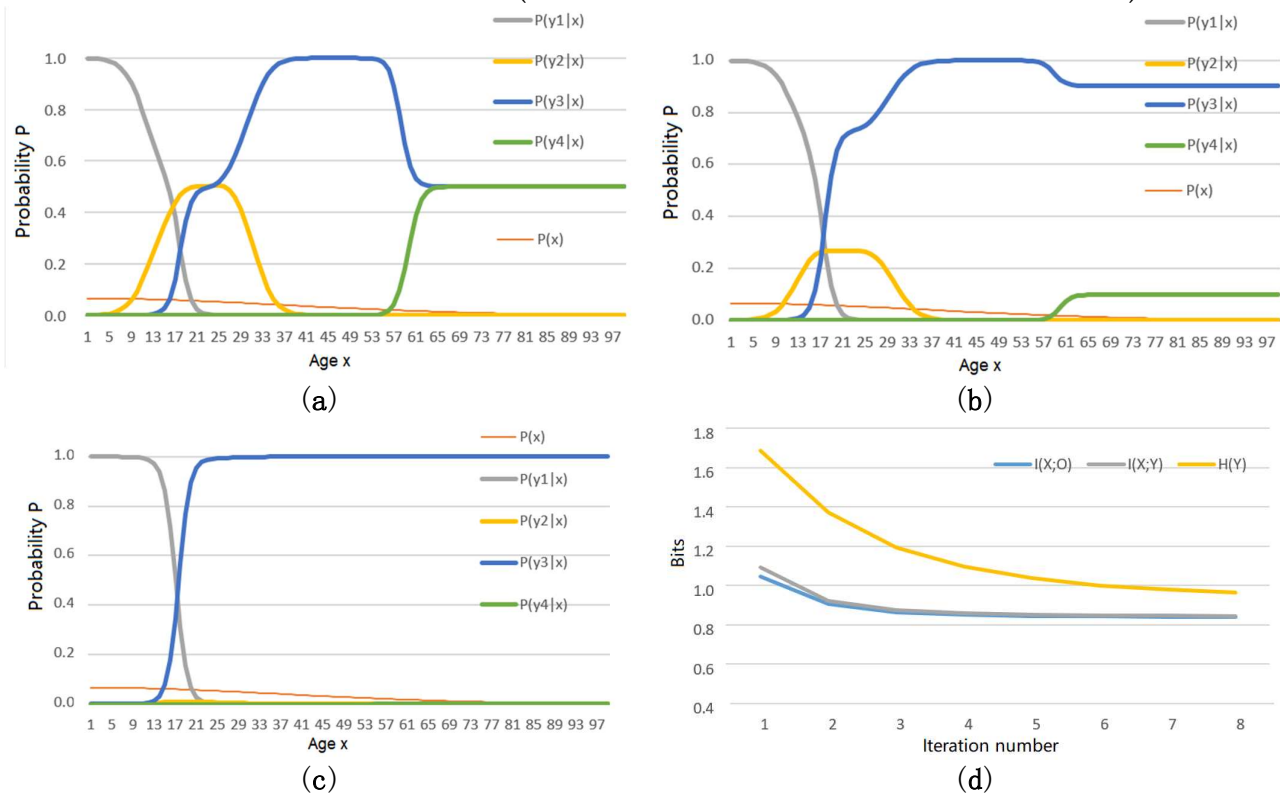


图 3. $P(y|x)$ 和 $I(X; Y)$ 在求 $R(\Theta)$ 的迭代过程中的变化。(a) 第一次迭代后的 $P(y|x)$ (它也是导致最大熵 $H(Y|X)$ 和 $H(X, Y)$ 的分布); (b) 第二次迭代后的 $P(y|x)$; (c) 八次迭代后的 $P(y|x)$; (d) $I(X; Y)$ 、 $I(Y; X)$ 和 $H(Y)$ 在迭代过程中的变化。

收敛的 $P(y)$ 是 $\{P(y_1), P(y_2), P(y_3), P(y_4)\} = \{0.3499, 0.0022, 0.6367, 0\}$ 。最小互信息是 0.845 比特。迭代过程不但让每个转移概率函数 $P(y_j|x)$ 正比于 $T(y_j|x)$, 也让 $P(y_4)$ 接近 0。 $P(y_4)$ 变成 0 是因为 y_4 蕴含 y_3 , 因而被 y_3 取代。因为后者的逻辑概率较大, 用 y_3 代替 y_4 能节省平均码长。

图 3c 表明 $H(Y)$ 也随迭代次数而减小。因为 $H(Y)$ 远小于 $P(X)$, 我们也可以通过用 y 代替 x 压缩数据。

对于给定的 $P(x)$, 要最大化熵 $H(X, Y)$ 或 $H(Y|X)$, 我们并不需要通过迭代求 $P(y|x)$ 。图 3a 中函数 $P(y|x)$ 也导致最大熵 $H(Y|X)$ 和 $H(X, Y)$, 其中 $P(y|x)$ 匹配 $T(y|x)$ 只有一次。

下一步, 作者使用上面真值函数求 $|s| = 1$ 时的 $R(G)$ 。为此, $d(x_i, y_j)$ 被 $I(x_i; \theta_j) = \log[T(y_j|x_i)/T(y_j)]$ (对于每个 i 和 j) 取代。图 4 显示了迭代结果和过程(补充材料含有图 4 中所需数据的产生说明)。

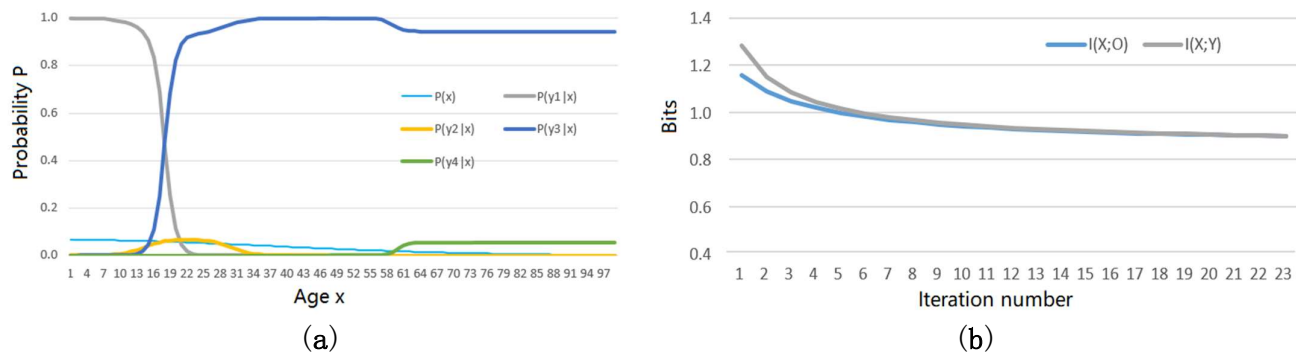


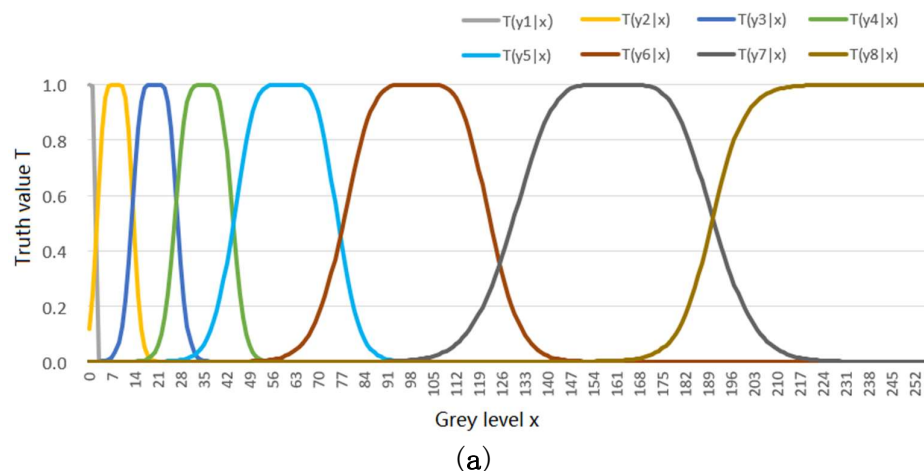
图 4. 求 $R(G)$ 函数的迭代结果。(a) 收敛的香农信道 $P(y|x)$; (b) $I(X; Y_\theta)$ 和 $I(X; Y) = I(Y; X_\theta)$ 在迭代过程中的变化

图 4a 直观地显示了 $P(y|x)$ 如何符合四个标签的语义。收敛的分布 $P(y)$ (为 $R(G)$) 是 $\{P(y_1), P(y_2), P(y_3), P(y_4)\} = \{0.3619, 0.0200, 0.6120, 0.0057\}$ 。最小互信息是 0.883 bits。这一 $P(y_4)$ 不是 0。这时的 $P(y_2)$ 和 $P(y_4)$ 都大于求 $R(\Theta)$ 时的数值。原因是：一个带有较小逻辑概率的标签能传递更多信息，因此应该被更多选择——如果我们使用信息准则而不是失真准则。对于同样的 $s = 1$ 和 Θ , $R(G) = 0.883$ 比特，它大于 $R(\Theta) = 0.845$ bits。

5.2. 一个灰度图像压缩的例子显示求 $R(\Theta)$ 的香农信道 $P(y|x)$

这个例子用来解释：在通过降低图像灰度水平（或精度）的数据压缩时，我们能用真值函数代替失真函数获得 $R(\Theta)$ 和相应的 $P(y|x)$ 。结论也适合数据简化——较高精度是不必要的——的场合。

例 2. 目的是压缩 8 比特图像（有 256 个灰度等级 $x_i, i = 0, 1, \dots, 255$ ）为 3 比特图像（有 8 个灰度等级 $y_j, j = 1, 2, \dots, 8$ ）【42】。考虑人眼视觉分辨率随灰度水平变化（灰度等级越高，分辨率越低），我们使用图 5a 所示 8 个真值函数，它们代表 8 个模糊类别。附录 C 显示了这些曲线是怎样产生的。任务是令 $s = 1$ ，求最小互信息 R 和相应的香农信道。



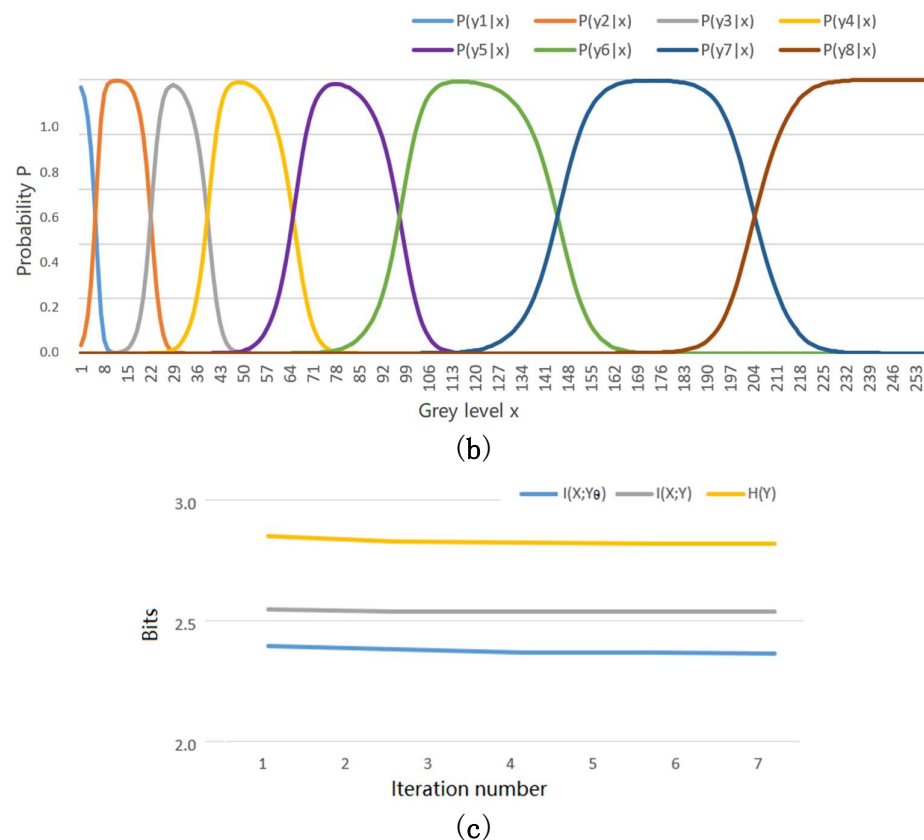


图 5. 例 2 的迭代结果。(a) 8 个真值函数或语义信道 $T(y|x)$; (b) 收敛的香农信道 $P(y|x)$; (c) $I(X; Y_\theta)$, $I(X; Y)$ 和 $H(Y)$ 在迭代过程中的变化。

作者获得收敛的 $P(y|x)$ 如图 5b 所示。图 5c 显示了 $R(\Theta)$ 、 $I(X; Y_\theta)$ 和 $H(Y)$ 在迭代过程中的变化(补充材料说明了图 5 所用数据是如何产生的)。

图 5 显示了香农信道如何匹配语义信道从而实现最小互信息。比较图 5a 和 b, 可以发现, 通过 $T(y|x)$ 控制 $P(y|x)$ 是容易的。如果我们用失真函数 $d(x, y_j)$ 而不是真值函数 $T(y_j|x)$ ($j = 1, 2, \dots, 8$), 定义 $d(x, y_j)$ 是不容易的。要通过 $d(x, y)$ 预测收敛的 $P(y|x)$ 也是不容易的。

在这一例子中, $I(X; Y) = I(Y; X_\theta)$ 要接近 $I(X; Y_\theta)$ 是困难的, 因为对于 $j = 2, 3, \dots, 6$, $P(y_j|x)$ 不能严格地正比于 $T(y_j|x)$ 。

作者用过不同的先验分布和语义信道, 用上面方法计算产生最小互信息的香农信道, 得到类似结果。 $s = 1$ 时, 每个收敛的转移概率函数 $P(y_j|x)$ 覆盖的 U 中区域和 $T(y_j|x)$ 覆盖的区域类似, 因此, $P(y|x)$ 满足通过 $T(y|x)$ 定义的约束。

6. 讨论

6.1. 用真值函数连接机器学习和数据压缩

已有研究者把信息率-失真函数应用于机器学习获得有意义结果【43–45】。然而, 我们还需要通过学习得到失真函数。

在信息率-失真理论中, 失真函数是主观定义的, 缺少客观标准。当实例 x 和标签 y 有不同论域 U 和 V 且 $|U| \gg |V|$ 时, 定义 x 和 y 之间的失真

是困难的。现在我们能通过机器学习从样本分布得到标签的真值函数(见等式(23)–(25))并用真值函数间接表示失真函数(见等式(32))。在信源 $P(x)$ 改变后,语义信道仍然可用【31】。我们能从 $T(y|x)$ 和新的 $P(x)$ 得到产生最小互信息 $R(\Theta)$ 的新的香农信道 $P(y|x)$ 。我们也能使用 $T(y|x)^{|s|}$ 和 $|s| > 1$ 减少平均失真,或用 $T(y|x)^{|s|}$ 和 $|s| < 1$ 放松失真限制。这样,我们就能克服信息率-失真函数的三个缺点(如第1节所述)。另外,一组真值函数描述一组标签的外延或语义,因此是直观的。和失真函数比,真值函数有更强的描述力,因为我们总能用一个真值函数 $T(y_j|x) = \exp[-d(x, y_j)]$ 代替一个失真函数 $d(x, y_j)$, 但是可能不能用一个失真函数代替一个真值函数,比如代替图1中 $y_3 = \text{“Adult”}$ 的真值函数。

两个经常使用的学习函数是真值函数(或相似性函数或隶属函数)和似然函数。我们也能用 $\log$ 似然函数 $\log P(x|\theta)$ 代替失真函数 $d(x, y)$, 或等价地用 $I(X; Y_\theta) \geq G$ 代替 $\bar{d} \leq D$, 获得信息率-逼真函数 $R(G)$ 。函数 $R(G)$ 使用似然准则或语义信息准则, 它能减少较小逻辑概率事件的漏报。比如, 图1中 y_2 和 y_4 的选择概率 $P(y_2)$ 和 $P(y_4)$ 在求 $R(G)$ 时(见图4a)比求 $R(\Theta)$ 时(见图3c)更大。函数 $R(G)$ 比 $R(\Theta)$ 更适合于预测场合, 其中信息比真更重要。

如同 $R(D)$, $R(\Theta)$ 和 $R(G)$ 也意味着平均码长下限。

6.2. 从语义信息 G 理论的角度看信息率-失真函数和最大熵分布

在信息率-失真函数的最小互信息分布和最大熵分布之间有不少类似性。根据4.1和4.5节的分析,我们可以把负指数函数当做真值函数,把划分函数当做逻辑概率。这些分布或香农信道,比如公式(2)中的 $P(y|x)$, 可看作是语义贝叶斯分布——通过语义贝叶斯公式(见等式(19))产生。

信息率-失真函数和最大熵方法的主要差别是:

- 对于信息率-失真函数 $R(D)$, 我们需要最小化互信息 $I(X; Y) = H(X) - H(X|Y)$ ——等价于最大化 X 的后验熵 $H(X|Y)$ 。为了求 $R(D)$, 我们需要最小互信息迭代, 找到合适的 $P(y)$ 。然而, 在最大熵方法中, 我们在给定 $P(X)$ 时最大化 $H(X, Y)$ 或 $H(Y|X)$, 不要求 $P(y)$ 的迭代;
- 信息率-失真函数可以用语义信息 G 测度表示(见等式(31)); 而最大熵等于或正比于极度最大熵减去语义互信息(见等式(41)和(47))。

6.3. 实验怎样支持对最小互信息迭代的解释

4.2–4.4节的理论分析表明: 求 $R(D)$ 、 $R(\Theta)$ 和 $R(G)$ 的最小互信息迭代促使香农信道匹配语义信道, 即尽可能让 $P(y_j|x) \propto T(y_j|x)$ ($j=1, 2, \dots$)。5.1节的例1显示求 $R(\Theta)$ 和 $R(G)$ 的迭代过程不仅找出适当的概率分布 $P(y)$ (使具有较大逻辑概率的 y_j 更频繁地被选择), 还更改每一个转移概率函数 $P(y_j|x)$, 使得 $P(y_j|x)$ 近似地正比于 $T(y_j|x)$ 。因此, 图3中结果支持上面理论分析。

图4表明, 求 $R(G)$ 的香农信道和求 $R(\Theta)$ 的香农信道有些不同, 比如, 具有较小逻辑概率的两个标签 y_2 和 y_4 在求 $R(G)$ 时比求 $R(\Theta)$ 时有较大的被选择概率 $P(y_2)$ 和 $P(y_4)$ 。这一结果也符合理论分析——它表明语义信息准则相对失真或真准则能减少具有较小逻辑概率事件的漏报。

图 5 显示了一个简化数据的例子。对于这个例子，失真函数是很难定义的，但是使用真值函数表示模糊类别是容易的。结果表明，我们可以容易地用语义信道 $T(y|x)$ 控制香农信道 $P(y|x)$ 。

6.4. 考虑其他种类的语义信息

Wyner 和 Debowski 独立地定义了一个信息测度（见【27】中的等式（1））。它对于度量两个语句或标签之间的语义信息是有意义的。结合这一定义和 G 理论，我们可以从这一测度发展出一个有些不同的两个标签 y_j 和 y_k 之间的语义信息测度：

$$I(\theta_j; \theta_k) = \sum_i P(x_i | y_j, y_k) \log \frac{T(\theta_j \cap \theta_k | x_i)}{T(\theta_j)T(\theta_k)}. \quad (48)$$

和【27】中的等式（1）不同，上面定义同时包含统计概率 $P(x)$ 和 $P(x|y_j, y_k)$ 和逻辑概率，并且它并不要求 U 中子集 $\theta_1, \theta_2, \dots, \theta_n$ 形成 U 的一个划分。我们用一个例子说明使用 $P(x)$ 的理由。如果 $P(x)$ 是人口年龄分布，我们用两个标签 $y_1 =$ “50 多岁的人”和 $y_2 =$ “老年人”代替年龄为 x 的人。人的平均寿命（用 t 表示）在不同地区和不同年代可能在 50 年和 80 年之间变化。上面公式能保证：算出的 y_1 和 y_2 之间的语义信息在 $t = 50$ 时比 $t = 80$ 时更多。

求 $I(\theta_j; \theta_k)$ 的平均就得到语义互信息：

$$I(Y_{\theta_1}; Y_{\theta_2}) = \sum_j \sum_k \sum_i P(x_i, y_j, y_k) \log \frac{T(\theta_j \cap \theta_k | x_i)}{T(\theta_j)T(\theta_k)}. \quad (49)$$

这个公式能保证错误翻译或替代可能传递负的语义信息。两个标签之间的失真函数就变成：

$$d(y_j, y_k) = \sum_i P(x_i | y_j, y_k) \log [1 / T(\theta_j \cap \theta_k | x_i)]. \quad (50)$$

我们能把这一失真函数用于标签或语句的数据压缩。

考虑设计、决策、控制计划等的语义信息，G 测度和 $R(G)$ 函数也是有用的。在这些情况下，真值函数 $T(\theta_j|x)$ 就变成分布约束函数， $P(x|y)$ 就变成实现的分布，信息就意味着控制复杂性或控制量，G 意味着有效控制量而 R 意味着控制成本。要增大 $G = I(X; Y_\theta)$ ，我们需要在固定 $T(\theta_j|x)$ 时优化控制，从而改进 $P(x|y)$ 。R(G) 函数告诉我们，我们可以用等式（15）改进 $P(y|x)$ 并通过加大 s 增加 G。在【32】中，为了上述目的，作者不适当地用了一个有些不同的信息公式（见【32】中等式（24））。那个公式看来只适合于控制结果是连续分布的场合。作者将在其他文章中进一步讨论这一话题。

G 测度也有其局限性，包括：

- G 测度只涉及标签的外延而不涉及其内涵。比如，“老人”意味着衰老和接近死亡，而 G 测度只和真值函数 $T(\text{“老人”}|x)$ （确定“老人”的外延）有关；
- 要度量根据上下文做模糊推理的语义信息，上述方法是不够的，虽然作者已经讨论了怎样计算复合命题的真值函数和利用模糊大前做模糊推理【32】。

因此，要度量更多种类的语义信息，我们需要更多研究..

6.5. 关于语义信息研究的重要性

香农并不考虑语义信息，他写道【34】 (p. 3):

“通信的语义方面和工程问题不相关。意义方面在于：实际的消息 (message) 是从一组消息中选一个。而系统的设计一定要考虑每一种可能的选择，并不是实际上选一个，因为选哪个事先是不知道的。”

然而，作者并不认为香农反对语义信息研究，因为在他和 Weaver 合著的书【34】 (pp. 95 - 97) 中，Weaver 就倡导语义信息研究。如果我们推广香农理论到机器学习——其中频繁涉及消息的选择，我们就一定要考虑语义信息。

4.2 节解释，失真就关系到语义，因此关系到语义信息。等式(31)表明：信息率-失真函数 $R(D)$ 可以被表达为语义互信息 $I(Y; X_\theta)$ 。

随着机器学习的发展，语义信息理论正在变得更加重要。从语义信息 G 理论的视角看，用于混合模型的 EM 算法——其 E 步使用了和等式 (39) 类似的公式——可以用语义信道和香农信道相互匹配（使两种信息近似相等）来解释【40】。有限波尔茨曼机 (Boltzmann Machine (RBM))——带有负指数函数和划分函数的的 Softmax 函数——已经成为流行的机器学习方法。已有研究者讨论了 RBM 和 EM 算法的相似性【46】。看来我们也能用两种信道的相互匹配解释 RBM。关于应用 G 理论到使用神经网络的机器学习，我们需要更多研究。

7. 结论

在引言中，我们提出三个问题：

- 为什么带有负指数函数和划分函数的类贝叶斯公式 (见等式 (2)) 广泛存在于信息率-失真理论、最大熵方法和机器学习中？
- 我们能结合机器学习（得到失真函数）和数据压缩（使用信息率-失真函数）？
- 我们能将信息率-失真函数或类似函数应用到语义压缩？

使用基于 P-T 概率框架的语义信息 G 测度，我们解释了：负指数函数和划分函数是真值函数和逻辑概率，类贝叶斯公式是语义贝叶斯公式。我们也解释了：语义互信息公式能表示最小香农互信息（见公式(35)）；统计力学或最大熵方法中的最大熵正比于或等于极度最大熵减去语义互信息（见公式(41)和(47)）。

我们已经显示：通过机器学习可以从样本分布得到真值函数(见公式(23)和(25))。进一步，我们可以建立真值函数和失真函数之间的联系(见 4.2 节)，用 $\log(1/\text{truth_function})$ 代替失真函数。因此，我们能结合机器学习（求失真函数）和数据压缩。

因为真值函数能表示标签的语义【15】，我们可以解释推广信息率-失真函数 $R(D)$ 到信息率-真函数 $R(\Theta)$ ——通过用模糊熵 $H(Y_\theta|X)$ 代替平均失真 $\bar{d}$ (见 4.3 节)。我们也能推广 $R(D)$ 到信息率-逼真函数 $R(G)$ ——通过用语义互信息下限 G 代替平均失真上限 D (见 4.4 节)。我们已经使用两个例子显示了，在语义约束下，该迭代算法能促使香农信道 $P(y|x)$ 匹配语义信道

$T(y|x)$, 实现最小互信息 R 。例 1 揭示了：该迭代算法能利用标签之间的逻辑蕴含减少香农互信息。例 2 表明：为了简化数据，用语义信道 $T(y|x)$ 控制香农信道 $P(y|x)$ 是容易的。

然而，我们也需要压缩和恢复文本数据（根据上下文）、图像数据和语音数据（根据模式识别或物种识别结果）。上面推广是有用的，但还是不够的；因此，我们有必要做进一步研究。

补充材料：产生图 3–5 的 Excel 文件可从下面地址得到：
<http://survivor99.com/lcg/cm/mmi-iteration.zip>

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: 本文初始版本经两位审稿人评审；重新提交的版本经三位审稿人评审。他们的评审意见带来了许多有意义的更改。特别是前两位审稿人的批评提醒我强调机器学习和数据压缩相结合。我感谢五位审稿人的修改意见、批评和支持；我也感谢客座编辑和助理编辑——为她们鼓励我重新提交手稿。

Conflicts of Interest: The author declares no conflict of interest.

附录 A. 等式(47)的推导

从等式(45), 我们推导出 X 和 Y 之间的互信息和熵 S 的关系:

$$I(X;Y) = \sum_j P(y_j) \sum_i P(x_i | y_j) \log \frac{P(x_i | y_j)}{P(x_i)} = \sum_j P(y_j) \log G_j - S / (kN). \quad (A1)$$

进一步, 根据等式(46), 我们得到最小互信息:

$$\begin{aligned} I(X;Y) &= \sum_j \sum_i P(x_i, y_j) \log \frac{\exp[-e_i / (kT_j)]}{Z'} = \sum_j \sum_i P(x_i, y_j) \log \frac{T(\theta_j | x_i)}{T(\theta_j)} \\ &= H(Y_\theta) - H(Y_\theta | X) = I(X;Y_\theta). \end{aligned} \quad (A2)$$

从上面等式, 我们得到:

$$S / (kN) = \sum_j P(y_j) \log G_j - I(X;Y_\theta). \quad (A3)$$

附录 B. 产生第 1 节的图 1 和 5.1 节的例 1 中数据的公式

$$\begin{aligned} T(y_3 | x) &= T(\text{"adult"} | x) = \frac{1}{1 + \exp[-1.5(x-18)]}, \\ T(y_1 | x) &= T(\text{"non-adult"} | x) = 1 - T(y_3 | x), \\ T(y_2 | x) &= T(\text{"youth"} | x) = 1 - [1 - \exp(-\frac{(x-22)^2}{2 \times 5^2})]^2, \\ T(y_4 | x) &= T(\text{"elder"} | x) = \frac{1}{1 + \exp[-(x-60)]}, \end{aligned} \quad (A4)$$

$$P(x) = \exp(-\frac{x^2}{2 \times 37^2}) / \sum_{x \geq 0} \exp(-\frac{x^2}{2 \times 37^2}), x \geq 0. \quad (A5)$$

附录 C. 图 5a 和 5.2 节用到的公式和参数

$$T(y_0 | x) = 1 - \frac{1}{1 + \exp[-(x - 2)]},$$

$$T(y_j | x) = 1 - [1 - \exp(-\frac{(x - \mu_j)^2}{2\sigma_j^2})]^3, j = 2, \dots, 6, \quad (A6)$$

$$T(y_7 | x) = \frac{1}{1 + \exp[-0.2(x - 200)]},$$

$$P(x) = \exp(-\frac{x^2}{2 \times 37^2}) / \sum_{x \geq 0} \exp(-\frac{x^2}{2 \times 37^2}), x \geq 0. \quad (A7)$$

表 A1 6 个真值函数 $T(y_j | x)$ ($j = 2, 3, \dots, 7$) 用到的 6 对参数.

y_j	$T(y_2 x)$	$T(y_3 x)$	$T(y_4 x)$	$T(y_5 x)$	$T(y_6 x)$	$T(y_7 x)$
μ_j	14	30	52	80	120	170
σ_j^2	16	24	50	80	160	240

上面公式中只有表示 $T(y_j | x)$ ($j = 2, 3, \dots, 6$) 的公式没见别人使用。这一公式表示一组带圆角的梯形——通常用分段函数表示。用上面公式就可以不用分段函数。

References

- Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–429+623–656.
- Shannon, C.E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.* **1959**, *4*, 142–163.
- Berger, T. *Rate Distortion Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1971.
- Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620.
- Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev. II* **1957**, *108*, 171.
- Smolensky, P. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*; Rumelhart, D.E., McClelland, J.L., Eds.; MIT Press: Cambridge, UK, 1986; pp. 194–281.
- Hinton, G.E. A Practical Guide to Training Restricted Boltzmann Machines. In *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science*; Montavon, G., Orr, G.B., Müller, K.R., Eds.; Springer: Berlin, Germany, 2012; Volume 7700, pp. 599–619. https://doi.org/10.1007/978-3-642-35289-8_32.
- Salakhutdinov, R.; Hinton, G.E. Replicated softmax: An undirected topic model. *Neural Inf. Process. Syst.* **2009**, *22*, 1607–1614. <https://doi.org/10.5555/2984093.2984273>.
- Goodfellow, I.; Bengio, Y.; Courville, A. Softmax Units for Multinoulli Output Distributions. In *Deep Learning*; MIT Press: Cambridge, UK, 2016; pp. 180–184.
- Zadeh, L.A. Fuzzy Sets. *Inf. Control.* **1965**, *8*, 338–353.
- Harremoës, P. Maximum Entropy on Compact Groups. *Entropy* **2009**, *11*, 222–237. <https://doi.org/10.3390/e11020222>
- Berger, T.; Gibson, J.D. Lossy Source Coding. *IEEE Trans. Inf. Theory* **1998**, *44*, 2693–2723.
- Berger, T.; Gibson, J.D. Lossy Source Coding. *IEEE Trans. Inf. Theory* **1998**, *44*, 2693–2723. [CrossRef]
- Gibson, J. Special Issue on Rate Distortion Theory and Information Theory. *Entropy* **2018**, *20*, 825. <https://doi.org/10.3390/e20110825>.
- Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2006.
- Davidson, D. Truth and meaning. *Synthese* **1967**, *17*, 304–323.

16. Willems, F.M.J.; Kalker, T. Semantic compaction, transmission, and compression codes. In *Proceedings of the International Symposium on Information Theory ISIT 2005*, Adelaide, Australia, 4–9 September 2005; pp. 214–218, doi:10.1109/ISIT.2005.1523325.
17. Babu, S.; Garofalakis, M.; Rastogi, R. SPARTAN: A Model-Based Semantic Compression System for Massive Data Tables. *ACM SIGMOD Rec.* **2001**, *30*, 283–294. <https://doi.org/10.1145/376284.375693>.
18. Ceglarek, D.; Haniewicz, K.; Rutkowski, W. Semantic Compression for Specialised Information Retrieval Systems. *Adv. Intell. Inf. Database Syst.* **2010**, *283*, 111–121.
19. Blau, Y.; Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA, 18 June 2018; pp. 6228–6237.
20. Bardera, A.; Bramon, R.; Ruiz, M.; Boada, I. Rate-Distortion Theory for Clustering in the Perceptual Space. *Entropy* **2017**, *19*, 438. <https://doi.org/10.3390/e19090438>.
21. Carnap, R.; Bar-Hillel, Y. An outline of a theory of semantic information. In *Technical Report No. 247; The Research Laboratory of Electronics*, MIT: 1952. <http://dspace.mit.edu/bitstream/handle/1721.1/4821/RLE-TR-247-03150899.pdf;sequence=1> (assessed on 1 July, 2021)
22. Klir, G. Generalized information theory. *Fuzzy Sets Syst.* **1991**, *40*, 127–142.
23. Floridi, L. Outline of a theory of strongly semantic information. *Minds Mach.* **2004**, *14*, 197–221.
24. Zhong, Y.X. A theory of semantic information. *China Commun.* **2017**, *14*, 1–17.
25. D'Alfonso, S. On Quantifying Semantic Information. *Information* **2011**, *2*, 61–101.
26. Bhandari, D.; Pal, N.R. Some new information measures of fuzzy sets. *Inf. Sci.* **1993**, *67*, 209–228.
27. Dębowski, Ł. Approximating Information Measures for Fields. *Entropy* **2020**, *22*, 79, <https://doi.org/10.3390/e22010079>.
28. Lu, C. *A Generalized Information Theory*; China Science and Technology University Press: Hefei, China, 1993; ISBN 7-312-00501-2. (In Chinese)
29. Lu, C. Meanings of generalized entropy and generalized mutual information for coding. *J. China Inst. Commun.* **1994**, *15*, 37–44. (In Chinese)
30. Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* **1999**, *28*, 453–490.
31. Lu, C. Semantic information G theory and logical Bayesian inference for machine learning. *Information* **2019**, *10*, 261, <https://doi.org/10.3390/info10080261>.
32. Lu, C. The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning. *Philosophies* **2020**, *5*, 25.
33. Lu, C. Channels' Confirmation and Predictions' Confirmation: From the Medical Test to the Raven Paradox. *Entropy* **2020**, *22*, 384, <https://doi.org/10.3390/e22040384>.
34. Shannon C.E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1963.
35. Zadeh, L., A.: Probability measures of fuzzy events. *J. Math. Anal. Appl.* **1986**, *23*, 421–427.
36. Cumulative_Distribution_Function. Available online: https://en.wikipedia.org/wiki/Cumulative_distribution_function (accessed on 10 April 2021).
37. Popper, K. *Conjectures and Refutations*; Routledge: London, UK; New York, NY, USA, 2002.
38. Wittgenstein, L. *Philosophical Investigations*; Basil Blackwell Ltd: Oxford, UK, 1958.
39. Sow, D.M.; Eleftheriadis, A. Complexity distortion theory. *IEEE Trans. Inf. Theory* **2003**, *49*, 604–608, doi:10.1109/TIT.2002.808135.
40. Lu, C. Understanding and Accelerating EM Algorithm's Convergence by Fair Competition Principle and Rate-Verisimilitude Function. Available online: <https://arxiv.org/abs/2104.12592> (accessed on 27 April 2021).
41. Boltzmann Distribution. Available online: https://en.wikipedia.org/wiki/Boltzmann_distribution (accessed on 10 April 2021).
42. Binary Images. Available online: <https://www.cis.rit.edu/people/faculty/pelz/courses/SIMG203/res.pdf> (accessed on 25 June 2021).
43. Kutyniok, G. A Rate-Distortion Framework for Explaining Deep Learning. Available online: https://maths-of-data.github.io/Talk_Edinburgh_2020.pdf (accessed on 30 June 2021).
44. Nokleby, M.; Beirami, A.; Calderbank, R. A rate-distortion framework for supervised learning. In *Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, USA, 17–20 September 2015; pp. 1–6, doi:10.1109/MLSP.2015.7324319.

-
45. John, S.; Gadde, A.; Adsumilli, B. Rate Distortion Optimization Over Large Scale Video Corpus With Machine Learning. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1286–1290, doi:10.1109/ICIP40778.2020.9191120.
 46. Song, J.; Yuan, C. Learning Boltzmann Machine with EM-like Method. Available online: <https://arxiv.org/abs/1609.01840> (accessed on 20 April 2021).