四步创新——推广香农信息论到语义信息论

鲁晨光写于 2026-9

本人推广香农信息论得到的语义信息论简称为语义信息 G 理论。其系统介绍见我于 Entropy 发表的综述文章：https://www.survivor99.com/lcg/information/g-theory/index.html 但是文章太长。为便于大家理解 G 理论，我写了这篇不到三页的短文，介绍其要点。

第一步：建立失真和真的关系

根据 Davidson 的真值条件语义学，命题函数的语义由真值函数确定。香农虽然没有讨论语义信息，但是他使用了失真函数，而失真可以转换为真。所以香农信息论实际上涉及语义，只是没有把失真函数或真值函数带进信息量公式(只带进信息率失真函数)。我的第一步是建立真值函数和失真函数之间的联系：

$$T(y|x)=\exp[-d(y|x)];\quad d(y|x)=-\log T(y|x) \tag{1}$$

其中 y 是标签，x 是实例；比如 y_j=“年轻人”，x_i=20 岁。如果失真函数是张口向上的抛物线，真值函数就是高斯函数(没有归一化系数)。

真值函数也是隶属函数。用 θ_j 表示模糊集合，其中 x 使 y_j 为真，用 $T(\theta_j|x)$表示隶属函数，则按定义 $T(y_j|x)= T(\theta_j|x)$。

如果 x 和 y 的论域相同，真值函数就变成相似函数 $s(x,y_j)$。比如：GPS 指针位置和实际位置之间相似；相似函数也是混淆概率函数。

相似函数是对称的，即 $s(x,y_j)=s(y_j,x)$。但是真值函数不是对称的，比如：20 岁使“成年人”为真，反过来，一个成年人一般不使“大约 20 岁”为真。

第二步：建立似然函数和真值函数之间的关系

通过语义贝叶斯公式：

$$P(x|\theta_j)=\frac{T(\theta_j|x)P(x)}{T(\theta_j)},\ T(\theta_j)=\sum_i T(\theta_j|x_i)P(x_i) \tag{2}$$

其中 $P(x|\theta_j)=P(x|y_j,\theta)$. 这个公式的应用比如：对于 GPS 定位，定位指针的数学含义可用一个高斯函数表述

$$T(y_j|x)=\exp[-(x-x_j)^2/(2\sigma^2)] \tag{3}$$

σ 就是 RMS(标准偏差)。道路就反映实际位置的先验概率分布 $P(x)$，有了 $P(x)$和 $T(\theta_j|x)$就可以用上式算出概率预测 $P(x|\theta_j)$。

令 $T(\theta_j|x)$的最大值是 1，可以推导出

$$T(\theta_j|x)=\frac{P(x|\theta_j)}{P(x)}\Big/\max_x\left[\frac{P(x|\theta)}{P(x)}\right] \tag{4}$$

其中 $T(\theta_j)$是 y_j 的逻辑概率，它就是平均真值。再据最大似然或最大语义信息准则，我们就得到从样本分布求隶属函数的公式(随机点落影公式)：

$$T(\theta_j|x)=\frac{P(x|y_j)}{P(x)}\Big/\max_x\left[\frac{P(x|y_j)}{P(x)}\right]=P(y_j|x)/\max_x P(y_j|x) \tag{5}$$

第三步：得到语义互信息的两种推广方法(异途同归)

方法 1：用似然函数 $P(x|\theta_j)$代替香农互信息公式中 log 右边的 $P(x|y_j)$(左边 $P(x|y_j)$表示样本分布，不变)，就得到语义互信息公式：

$$
\begin{aligned}
I(X;Y_\theta) &= \sum_j \sum_i P(x_i)P(x_i \mid y_j)\log\frac{P(x_i \mid \theta_j)}{P(x_i)} \\
&= \sum_j \sum_i P(x_i)P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \\
&= H(X) - H(X \mid Y_\theta) \\
&= H(Y_\theta) - H(Y_\theta \mid X) = H(Y_\theta) - \bar{d},
\end{aligned}
\tag{5}
$$

$\bar{d}$ 是平均失真，可见语义互信息就是负的正则化失真。

方法 2：在 Carnap 和 Bar-Hillel 的语义信息公式(log[1/逻辑概率])基础上，用真值代替 1(表示预测需要事实检验)，就得到单个语句的语义信息量公式：

$$
I(x_i;\theta_j) = \log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} = \log\frac{P(x_i|\theta_j)}{P(x_i)} \tag{6}
$$

它能反映 Popper 的预测检验思想(参看图 1)。

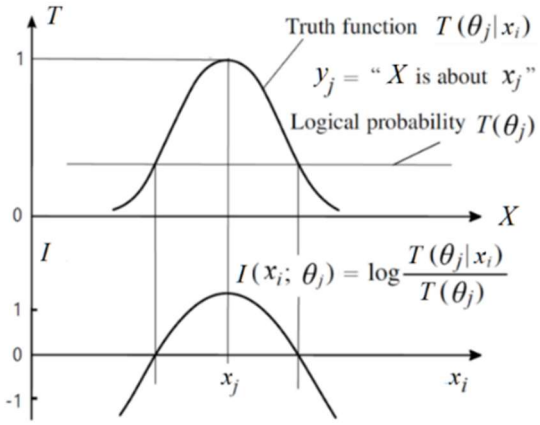

图 1. 语义信息量图解。y_j 传递关于 x_i 的语义信息随逻辑概率减小而增大，随偏差增大而减小。偏差大的一定程度，信息就是负的。

对 $I(x_i;\theta_j)$求平均，就得到语义 Kullback-Leibler(KL)公式(也是广义 KL 公式)：

$$
I(X;\theta_j) = \sum_i P(x_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)} = \sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{7}
$$

其中 $P(x|y_j)$是样本分布。该公式可用于优化似然函数。再对 $I(X;\theta_j)$求平均，就得到语义互信息公式。容易证明：当 $P(x|\theta_j)=P(x|y_j)$或 $T(\theta_j|x) \propto P(y_j|x)$时($j$=1,2,…)(即语义信道匹配香农信道时)，语义信息量达到最大，等于香农互信息。

第四步：推广信息率失真函数 $R(D)$得到信息率逼真函数 $R(G)$

因为语义信息量反映从不太真到真的进步，所以语义信息量反映逼真。用给定语义互信息 G 代替给定平均失真 D，就得到信息率逼真函数 $R(G)$。

$R(G)$函数的参数解(图解如图 2)：

$$G(s)=\sum_i\sum_j P(x_i)P(y_j\mid x_i)I_{ij}=\sum_i\sum_j I_{ij}P(x_i)P(y_j)\left[m_\theta(x_i,y_j)\right]^s/Z_i,$$
$$R(s)=sG(s)-\sum_i P(x_i)\log Z_i,\quad Z_i=\sum_k P(y_k)\left[m_\theta(x_i,y_j)\right]^s. \tag{8}$$

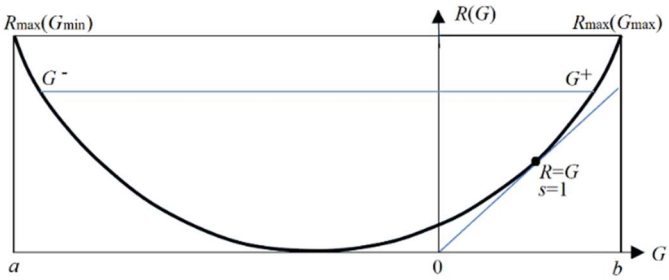

图 2. 二元通信的信息率逼真度函数 $R(G)$。任何 $R(G)$函数都是碗状的，其中有一个 $R(G)=G$ 的点 (s=1)。给定 R 有两个反函数 $G^+(R)$ 和 $G^-(R)$。

轮流用 $P(y|x)$和 $P(y)$作为变分，求 R-sG 的最小值就得到优化的香农信道。迭代公式：

$$P(y_j\mid x_i)=P(y_j)m_{ij}^s\Big/\lambda_i,\ \lambda_i=\sum_j P(y_j)m_{ij}^s,\ i=1,2,...;\ j=1,2,...$$
$$P^{+1}(y_j)=\sum_i P(x_i)P(y_j\mid x_i) \tag{9}$$

类似于 EM 算法。

G/R 就反映信息效率。由图 2 可见，存在一个香农信道，在它和语义信道相匹配时(即 $P(y_j|x) \propto T(y_j|x)$ 或 $P(x|y_j) = P(x|\theta_j)$时)，R=G，信息效率达 G/R 最大，等于 1. 要继续增大 G，G/R 会降低。在数据压缩和约束控制或强化学习中，我们需要在最大语义互信息 G 和最大信息效率 G/R 之间权衡，s 就是用以权衡的超参数，1/s 表示边际语义信息率 dG/dR。

顺便说一下，有人推广信息率失真函数到语义信息率失真 $Is(D)$. 我不能理解，为什么要最小化语义互信息？在我看来，香农互信息是成本，语义互信息是效用——和似然度正相关。有人最小化似然度了吗？

总结：

用上面 4 步推广得到的语义信息 G 理论不同于其他语义信息理论，其特点是：

1. 兼容 Popper 思想——语义信息多少需要用事实检验；
2. 和流行的机器学习方法兼容，主要因为：1）最大语义信息准则兼容最大似然准则和正则化误差准则；2）两种信道匹配算法正好对应模型学习和预测及强化学习。
3. 只在香农信息论框架下作了微小改动——用语义信息量代替平均失真量作为通信质量要求；
4. 使用的 P-T 概率框架在统计和逻辑之间架起桥梁。

注意：上面 4 步的顺序是为了便于理解，作者实际研究顺序是(2)->(3)->(4)-(1). 作者研究过程见：

https://www.survivor99.com/lcg/books/git/

个人网站：https://www.survivor99.com