求标签外延和最大语义信息分类

鲁晨光

摘要 如何通过人口在年龄上的先验概率分布 $P(x)$和后验概率分布 $P(x|$成年人$)$求解“成年人”的外延，并用它做新的概率预测？为此，本文提出和证明了第三种贝叶斯定理。如果样本不够大，则用最大语义信息准则优化标签的外延（即隶属函数）。该准则兼容最大似然准则和正则化最小误差平方准则。多标签分类分两步：1)标签学习——求外延，非常简单；2)标签选择——即分类。分类时考虑到类别不平衡和实例概率分布改变，能解释“老年人”分类随人口年龄分布变化。

关键词 贝叶斯定理，真值函数，概念外延，语言通信，语义信息，最大似然准则，自然语言处理，翻译

Solving Labels' Denotations and Maximum Semantic Information Classifications

Abstract For given prior age population distribution $P(x)$ and the posterior distribution $P(x|$adult$)$, how do we obtain the denotation of "adult"? With the denotation, we can make new probability prediction. For this purpose, we propose and proves the Third Kinds of Bayes' Theorem. If samples are not big enough, we can use maximum semantic information criterion to optimize labels' denotations, e.g., membership functions. This criterion is compatible with Maximum likelihood criterion and Regularized Least Squares (RLS) criterion. Multi-label Classification needs two steps: 1) label learning for denotations，which is very simple; 2) Label selection or classification. The classification concerns class-imbalance and instance probability distribution's change and can explain how the classification for "old people" changes with age population distribution.

Keywords Bayes' theorem, Truth function, Conceptual denotation， Linguistic communication, Semantic information, Maximum likelihood criterion，Natural language processing, Translation.

1 引言

假设我们知道普通人群不同年龄 x 的先验概率分布 $P(x)$（连续的），又知道成年人年龄概率分布 $P(x|y_1$ 是真的$)$ (也是连续的，y_1 等于词或标签“成年人”)。 我们是否能得到合适的“成年人”的概念外延？换为不同人群(比如在学校，工厂， 部队)，其先验概率分布是 $P'(x)$, 我们能求出相应的后验分布 $P'(x|y_1$ 是真的$)$吗？ 我们能写出一个通用预测模型，用来求解这样的问题吗？

已有的似然方法【1】不能解决这个问题，因为似然函数不能加进先验知识，从一个人群得到的似然函数， 换一个人群就会失效。贝叶斯主义推理即 Bayesian Inference (BI)【2】声称使用主观概率和逻辑概率，也考虑先验知识。用它可以吗？也不行！因为 1)虽然 BI 用到贝叶斯定理（Bayes' theorem）【3】, 但是实际上没有用到逻辑概率。因为逻辑概率不是归一化的(后面详谈)，而 BI 用到的先验 $P(\theta)$ (θ 是模型参数) 和后验 $P(\theta|x)$都是归一化的；2)BI 的先验是 θ 或 y(词或标签)的先验，它是主观的；而不是 x 的先验——它比较客观。

然而，到目前为止， 我们缺少一个求概念外延的简单方法，也缺少一个的通用的预测模型，以便根据不同的先验分布 $P'(x)$预测后验分布。本文首先提供一个新的贝叶斯公式，用以求解概念外延——可能是模糊的。

进一步，如果我们只知道不太大的带标签样本，不能得到连续的后验概率分布，我们是

否可以像构造连续的似然函数那样，用参数构造一个概念的外延——即真值函数或隶属函数，并且用样本优化？这就是标签学习问题。

再进一步，已知一组词汇的外延或语义，发信人如何在给定 x 或 x 的概率预测情况下选择一个语义信息量最大的词？这也是多标签分类问题。

为解决后面两个问题，我们将使用新的语义信息公式。

虽然优化自然语言通信已经有很多富有成效的方法【4】，但是都没有严格按信息准则。主要是因为 Shannon 信息论不考虑语义。笔者以为合理的标准还是信息标准。按 Shannon 信息论定义【5】， 信息就是节省的平均码长。为了解释语义通信，鲁晨光把它改为：语义信息是因预测而节省的平均码长【6】。比如有人告诉我“这些是成年人”，成年人对我来说后验不确定性就小了，为不同年龄编码的平均码长就短了。

鲁晨光用平均对数标准似然度(log-normalized-likelihood)定义语义信息（量）【6-8】，还提出信道匹配算法——用于最大似然估计【9】和混合模型【10】。本研究基于鲁晨光的语义信息论研究。本文的贡献是：

1） 提出并证明第三种贝叶斯定理，用以简化语义信道(即一组概念外延)求解。

2） 证明最大语义信息准则是一种特殊的正则化的最小误差平方(RLS)准则，它和最大似然准则兼容，同时兼顾类别不平衡问题【11】。

3） 提出标签学习和分类要分两步走: 标签学习(求隶属函数)和标签选择(分类)，即先利用 Shannon 信道得到语义信道，再用最大语义信息（或 log 标准似然）准则分类。因为不需要多个二元分类【12,13】，学习不需要多个数据集，对样本没有特别要求，从而能简化多标签学习。分类也兼顾到标签之间的相关性，不会标注了“老年人”，还会标注“成年人”和“非年轻人”——这是二元关联法要解决的问题【12】。

4） 考察了分类边界如何随先验分布 $P(x)$变化， 比如“老年人”的分类随老年人口比例的变化，已及外延和分类的相互影响。

下面第 2 节介绍数学方法，包括第三种贝叶斯定理证明；第 3 节先分析语义信息准则如何和 RLS 准则兼容；然后从语义通信的角度讨论机器学习——包括多标签学习和分类，以及二分类中划分边界随实例先验概率分布改变。最后是总结。

2 求标签或名词的外延的数学方法

2.1 数学定义——兼容统计概率和逻辑概率的概率系统

已有的概率主要有 4 种【14】: 统计概率，主观概率，逻辑概率和模糊逻辑概率。模糊逻辑概率也就是 Zadeh 提出的模糊事件的概率【15,16】，其条件概率分布就是类别的隶属函数或标签的真值函数。本研究合并逻辑概率和模糊逻辑概率(省去“模糊”)，并且把主观概率看做是统计概率和逻辑概率的杂交， 所以，基本的概率只有两种：统计概率和逻辑概率。

一个假设或标签有两种概率，一是它被选择的概率——统计概率，二是它被判定为真的概率——逻辑概率。两者通常不同。 比如， 考虑 “明天有小雨”和“明天有小到大雨”… 后者逻辑概率较大， 而被选择概率较小。考虑“他是老年人”和“他不是老年人”，后者逻辑概率大于前者，但是它被选择的概率小于前者。一个永真句“他可能是也可能不是老年人”，其逻辑概率更大，是 1， 而选择概率接近 0.

标签被选择的概率是归一化的，即所有标签被选择的概率相加等于 1，**而标签的逻辑概率不是归一化的**。比如，考虑描述人的年龄的标签：“小孩”、“年轻人”、“中年人”、“成年人”、“老年人”… 它们的被选择的概率之和为 1，但是逻辑概率之和远大于 1. 原因在于，一个年龄可能使几个标签为真。 比如, 年龄是 25 岁，标签“年轻人”、“成年人”、

“非老年人”…都是真的。

下面我们定义并存且相关的两种概率，并通过第三种贝叶斯定理把统计概率和隶属函数(反映概念外延)联系起来，使得转移概率函数和隶属函数可以相互转换。

定义 2.1 设论域 U 中有元素 $x_1, x_2, \ldots, x_m$; X 取值于 U 中某个元素 x 的随机变量(按照信息论习惯，随机变量用大写字母)，即 $X\in U=\{x_1, x_2, \ldots, x_m\}$. 再设论域 V 中有元素 $y_1, y_2, \ldots, y_n$; Y 是取值于 V 中某个元素 y 的随机变量，即 $Y\in V=\{y_1, y_2, \ldots, y_n\}$. 对于每个假设 y_j, 存在一个集合 $A_j\in 2^U$, y_j=“$X\in A_j$”.

定义 2.2 我们用等号“=”表示的随机事件的统计概率(简称概率)——比如 $P(X=x_i)$——是统计概率，后面简写为 $P(x_i)$; 如果 X 的值没有给定，我们用 $P(x)$或 $P(X)$表示 $P(X=x)$. 同样地，我们用 $P(y)$或 $P(Y)$表示 $P(Y=y)$. **我们再用属于符号“∈”表示随机事件的逻辑概率**。比如 $P(X\in A_j)$是逻辑概率。

我们把 $P(X\in A_j)$称之为逻辑概率，是因为根据 Tarski 的真理论【17】，$P(X\in A_j)=P$(“$X\in A_j$”是真的)=$P(y_j$ 是真的). 于是，一个假设 y_j 有两种概率：统计概率和逻辑概率. 为了更清楚区分两者，后面我们用 $T(A_j)$或 $T(y_j)$表示 y_j 的逻辑概率，即

$$T(A_j)=T(y_j)= P(y_j \text{ 是真的})= P(X\in A_j) \tag{2.1}$$

以 x 为条件的 y_j 的逻辑概率就是集合 A_j 的特征函数或 y_j 的真值函数，记为 $T(A_j|x)$， 于是

$$T(A_j)=\sum_i P(x_i)T(A_j\mid x_i) \tag{2.2}$$

根据 Davidson 的真值条件语义学【18】，上述真值函数确定了假设 y_j 的语义.

统计概率分布——比如 $P(y)$和 $P(y|x_i)$——是归一化的, 即

$$P(y_1)+P(y_2)+\ldots+P(y_n)=1，\ P(y_1|x_i)+P(y_2|x_i)+\ldots+P(y_n|x_i)=1 \tag{2.3}$$

而逻辑概率不是归一化的，比如在$\{A_1, A_2, \ldots, A_n\}$是 U 的一个覆盖的情况下，

$$T(A_1)+T(A_2)+\ldots+T(A_n)\geqslant 1 \tag{2.4}$$

只有在$\{A_1, A_2, \ldots, A_n\}$是 U 的划分并且 y 总是被正确地使用的情况下，两种概率才相等.

后面用 θ_j 表示相应 y_j 的模糊集合(包括清晰集合)，则相应的逻辑概率是 $T(\theta_j)$. 逻辑概率分布——即后面的 $T(\theta)$——虽然很像流行的贝叶斯推理中的 $P(\theta)$， 但是，$T(\theta)$既不是横向归一化的也不是纵向归一化的；它本身就是归一化系数。真值函数或隶属函数 $T(\theta_j|x)$的最大值是 1。可以说它是纵向归一化的. 即：

$$\max(T(\theta_j|x_1), T(\theta_j|x_2), \ldots, T(\theta_j|x_m))=\max T(\theta_j|X)=1 \tag{2.5}$$

这一重要性质将给求解真值函数带来方便.

2.2 三种贝叶斯定理

第一种贝叶斯定理是关于两个逻辑概率之间关系的定理【3】。

贝叶斯定理 1：设集合 $A, B\in 2^U$. A’是 A 的补集，B’是 B 的补集. $T(A)=P(x\in A)$, $T(B)$等同理. 则：

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)= T(A|B)T(B)+T(A|B')T(B') \tag{2.6}$$

类似地， 也可对称地求出 $T(A|B)$。

第二种贝叶斯定理是 Shannon 使用的关于两个统计概率之间关系的定理【5】。

贝叶斯定理 2：设事件是 $X=x$ 和 $Y=y_j$. $P(x)=P(X=x)$，$P(y_j)=P(Y=y_j)$. 则

$$P(x\mid y_j)=P(x)P(y_j\mid x)/P(y_j),\ P(y_j)=\sum_i P(x_i)P(y_j\mid x_i) \tag{2.7}$$

类似地，也可以对称地求出 $P(y_j|x)$.

之所以说上面两个定理是不同定理，是因为贝叶斯定理 1 中随机变量和论域只有一个，而贝叶斯定理 2 中随机变量和论域都是两个。贝叶斯定理 3 是笔者提出的，是关于一个统计

概率和一个逻辑概率之间关系的定理； 其中随机变量是一个， 概率有三种：统计概率、 逻辑和两者的杂交——预测的概率或似然度.

贝叶斯定理 3：设两个事件是 $X=x$ 和 $P(x\in A)$，$P(x)=P(X=x)$, $T(A_j)=P(x\in A_j)$, 则

$$P(x\mid A_j)=P(x)T(A_j\mid x)/T(A_j),\ T(A_j)=\sum_i P(x_i)T(A_j\mid x_i) \tag{2.8}$$

$$T(A_j\mid x)=T(A_j)P(x\mid A_j)/P(x),\ T(A_j)=1/\max[P(x\mid A_j)/P(x)] \tag{2.9}$$

这个定理包含的两个公式是不对称的，所以两个都要写出来。解释：(2.8)中 $P(x|A_j)$是似然函数；$T(A_j)$是 $P(x|A_j)$的横向归一化系数. 而在(2.9)中，$T(A_j)$是 $T(A_j|x)$的纵向归一化系数，它使 $T(A_j|x)$的最大值等于 1。max[ ]表示其中函数最大值。 (2.9)就是求概念外延公式，它还考虑到模糊概念.

贝叶斯定理 3 证明：设联合概率 $P(X=x，\ x\in A_j)$，则

$P(x|A_j)T(A_j)=P(X=x|\ x\in A_j)P(x\in A_j)=P(X=x,\ x\in A_j)=P(x\in A_j\mid X=x)P(X=x)=T(A_j|x)P(x)$

于是有

$$P(x\mid A_j)=P(x)T(A_j\mid x)/T(A_j),\ T(A_j|\mathrm{x})=T(A_j)P(x\mid A_j)/P(x)$$

因为 $P(x|A_j)$是横向归一化的，所以 $T(A_j)=\sum_i P(x_i)\ T(A_j|x_i)$. 因为 $T(A_j|x)$是纵向归一化的，把(2.5)代入上式，可以得到

$$1=\max[T(A_j)P(x|A_j)/P(x)]=T(A_j)\max[P(x|A_j)/P(x)]$$

所以 $T(A_j)=1/\max[P(x|A_j)/P(x)]$。**证毕.**

第三种贝叶斯定理可以无条件推广到集合模糊的情况(证明不赘)，即

$$P(x\mid\theta_j)=P(x)T(\theta_j\mid x)/\sum_i P(x_i)T(\theta_j\mid x_i) \tag{2.10}$$

$$T(\theta_j\mid x)=[P(x\mid\theta_j)/P(x)]/\max[P(x\mid\theta_j)/P(x)] \tag{2.11}$$

2.3 用第三种贝叶斯定理求概念外延和通用预测模型

例 1 x 表示不同年龄，y_1=“成年人”，我们从一个群体得到连续的先验分布 $P(x)$和后验分布 $P(x|y_1)$。现在换一个群体，y_1 使用规则不变，先验分布变为 $P'(x)$。参看图 1.

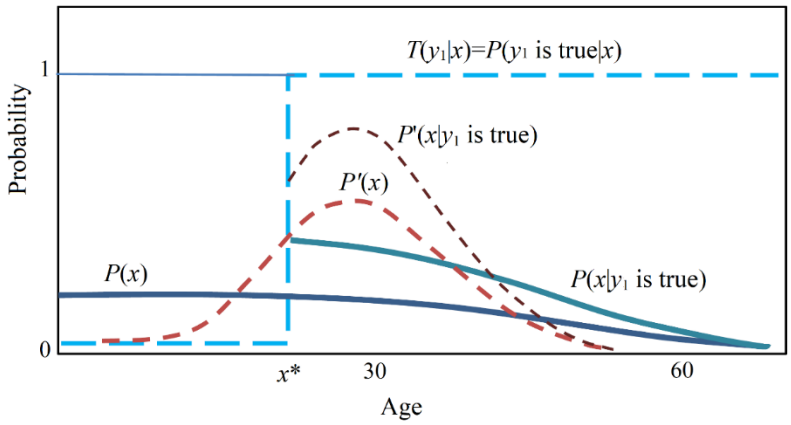

图 1 求 y_1=“成年人”的外延和后验分布 $P'(x|y_1$ is true).

1)假设“成年人”表示年满 x^*岁(未知)，标签都是真的。求“成年人”的外延和适应不同 $P'(x)$的预测模型。

2)“成年人”的外延是在没有法定成年人界限 x^*时自然形成的，如何求其模糊外延和

相应的预测模型？

3)求上面两种情况下的后验分布 $P'(x|y_1$ 是真的)。

如果集合是清晰的，画出 $P(x|y_1)$的分布图，人眼看了之后，人脑很快就知道，可以先求出“成年人”外延(即 y_1 的真值函数)，**外延就是适合不同 $P'(x)$的预测模型**！然后把外延内的 $P'(x)$归一化就得到 $P(x|y_1$ is true).

但是集合模糊时又怎么求模糊外延和预测模型呢？人脑还是没法写出通用的预测模型。利用第三种贝叶斯定理(2.11)，我们一下子就可以得到 1)和 2)的回答—— y_1 的真值函数或“成年人”的外延是：

$$T(y_1 \mid x) = T(\theta_1 \mid x)=[P(x \mid \theta_1) / P(x)] / \max[P(x \mid \theta_1) / P(x)] \tag{2.12}$$

它就是通用的预测模型，在 $P(x)$变为 $P'(x)$时，根据(2.10)，我们有

$$P'(x \mid y_1 \text{ is true})=P(x \mid \theta_1) = P'(x)T(\theta_1 \mid x) / \sum_i P'(x_i)T(\theta_1 \mid x_i) \tag{2.13}$$

现在可以看出， 为什么说真值函数或概念外延作为预测模型是通用的。以前大家都用似然函数作为预测模型，但是信源 $P(x)$变了, 以前的似然函数就失效了。而标签外延独立于信源也 $P(X)$独立于信宿 $P(Y)$， 反映信道特性，具有相对的稳定不变性。

2.4 从 Shannon 信道到语义信道

Shannon 信息论【5】中称 $P(X)$为信源，称 $P(Y)$为信宿，把转移条件概率函数组成的转移概率矩阵 $P(y_j|X), j=1,2,\ldots$叫做信道(后面称之为 Shannon 信道)。它有一个重要性质：在信源 $P(X)$变为 $P'(X)$后，我们可以用它和 $P'(X)$做贝叶斯预测, 得到 X 的后验概率分布 $P'(X|y_j)$; 而且 $P(y_j|X)$ 乘上一个系数 k，预测不变, 即

$$\frac{P'(X)kP(y_j \mid X)}{\sum_i P'(x_i)kP(y_j \mid x_i)} = \frac{P'(X)P(y_j \mid X)}{\sum_i P'(x_i)P(y_j \mid x_i)} = P'(X \mid y_j) \tag{2.14}$$

若干真值函数(隶属函数，外延)$T(\theta_j|X), j=1,2,\ldots$构成一个语义信道【10】。一个语义信道后面总有一个 Shannon 信道。以天气预报为例，转移概率函数 $P(y_j|X)$反映预报语句 y_j 的选择规律，因预报员而异——有人错的少，有人错的多。而 $T(\theta_j|X)$反映听众理解的语义，可能来自语言的定义，也可能来自过去的样本的训练。不同的人理解的语义 $T(\theta_j|X)$是大体相同的。

由式(2.14)可知，当真值函数和转移概率函数成正比时，语义贝叶斯预测和经典贝叶斯预测——用 $P(y_j|X)$和 $P(X)$产生 $P(X|y_j)$——等价。所以我们可以从 Shannon 信道得到与之等价的语义信道：$T(\theta_j|X) \propto P(y_j|X)(j=1, 2, \ldots)$。令 $T(\theta_j|X)$的最大值是 1， 于是有

$$T^*(\theta_j|X)=P(y_j|X)/\max[P(y_j|X)], j=1, 2, \ldots, n \tag{2.15}$$

再根据**贝叶斯定理 2**，我们得到优化的真值函数的数值解

$$T^*(\theta_j|X)= [P(X|y_j)/P(X)]/\max[P(X|y_j)/P(X)] \tag{2.16}$$

汪培庄教授用随机集落影(即集值统计)定义隶属函数【19】，可以证明，贝叶斯定理 3 得到的结果和集值统计结果一致(别处讨论)。

以上结论都假设样本很大，我们能从样本得到连续的联合概率分布 $P(X,Y)$或 Shannon 信道 $P(Y|X)$。如果样本不够大，我们需要用语义信息准则优化真值函数和语义信道。

3 语义通信优化——标签外延和使用规则相互匹配

3.1 语义信息准则及其和 RLS 准则之间的关系

关于语义信息，已有不少研究【20-22】，但是只有鲁晨光把隶属函数和由它产生的似然函数带进语义信息公式【6-8】。下面介绍鲁晨光的语义信息公式。

y_j 提供关于 x_i 的语义信息(量)被定义为对数标准(normalized)似然度：

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

其中用到贝叶斯定理 3 中前一个公式。这个公式就能反映 Popper 的思想【23】：(先验)逻辑概率越小，并能经得起检验(后验逻辑概率越大) ，信息量就越大; 永真句在逻辑上不能被证伪，因而不含有信息。

假设一组实例 $x(1), x(2), \ldots, x(N_j)\in U$ 构成一个同标签样本 $\mathbf{X}_j$。这些实例来自 N_j 个独立同分布随机变量(即 IID 假设). 假设样本中有 N_{ji} 个 x_i, 则 x_i 在其中出现的频率是 $P(x_i|y_j)=N_{ji}/N_j$, θ_j 的标准对数似然度是

$$\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}=N_j\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=N_jI(X;\theta_j) \tag{3.2}$$

其中 $I(X;\theta_j)$是广义 Kullback-Leibler (KL)信息【24，6】，即

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.3}$$

设有带标签样本 $\mathbf{D}=\{(x(t),y(t))\}$, t=1 to N; 从 $\mathbf{D}$ 可以产生 n 个带有 n 个不同标签的样本, y_j 出现的频率是 $P(y_j)$。则对 $I(X;\theta_j)$求平均，就得到语义互信息公式：

$$\begin{aligned}I(X;\theta)&=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_j\sum_i P(x_i,y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\\&=H(X)-H(X\mid\theta)=H(\theta)-H(\theta\mid X)\end{aligned} \tag{3.4}$$

其中 $H(X)$是 Shannon 熵，其他三种熵是交叉熵【6】。因为优化模型参数的时候 $P(X)$不变， 所以最大语义信息准则等价于最大似然准则。容易证明，在语义贝叶斯预测和样本分布一致时，即 $P(x_i|\theta_j)=P(x_i|y_j)$ (对于所有 i,j)时，或真值函数正比于转移概率函数时，即 $T(\theta_j|X)\propto P(y_j|X)$(对于所有 j)时，上述广义 KL 信息达到其上限——KL 信息；语义互信息也达到其上限——Shannon 互信息。

笔者发现，这一信息准则是一个特殊的 Regularized Least Square (RLS)准则【25】。把高斯分布真值函数代入式(3.3), 就得到

$$\begin{aligned}I(X;\theta)&=H(\theta)-\sum_j\sum_i P(x_i,y_j)(x_i-x_j)^2/(2d_j^{\,2})\\H(\theta)&=-\sum_j P(y_j)\log T(\theta_j)\end{aligned} \tag{3.5}$$

其中 $H(\theta|X)$就是误差项，$H(\theta)$就可以理解为正则化项。每个标准偏差 d_j 大了， 则相对偏差就小；但是逻辑概率就大了，$H(\theta)$就小了。我们也可以把 $H(\theta)$理解为潜在信息，把 $H(\theta|X)$理

解为惩罚项。因为 $T(\theta_j)$反映类别的相对比例， 所以语义信息准则也兼顾到分类的类别不平衡问题【11】。看来正则化除了解决过度拟合问题，也解决类别不平衡问题。

3.2 用于机器学习语义通信模型

图 2 所示通信模型是在鲁晨光的语义通信模型【7】基础上加进了机器学习考虑。从语义通信的角度看机器学习，也分标签发送者和标签接收者。接受者通过样本学习得到语义，即真值函数或隶属函数——一个实例可能隶属于多个类别。而发送者划分实例空间给实例分类，每个实例只属于一个类别——可能带有多个标签或一个复合标签。

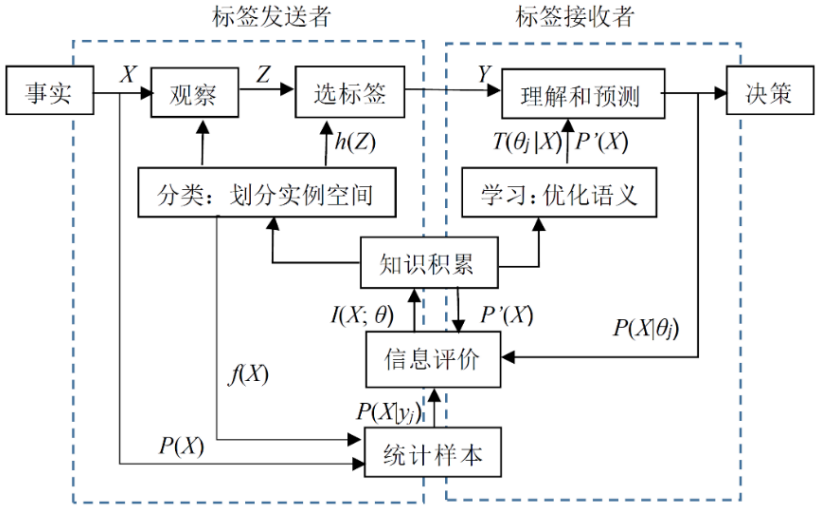

图 2 用于机器学习的语义通信模型。标签学习和分类分为两部分: 1)标签学习是让概念外延或隶属函数匹配 Shannon 信道；2）标签选择或分类是让 Shannon 信道匹配语义信道。

其中 X 是一个实例，Y 是一个原子标签或复合标签。假设发送者根据观察条件 $Z\in C$ 发送标签 $Y=h(Z)$。实际情况也可能 $Z=X$, 划分函数是 $Y=f(X)$。根据上面通信模型，学习(求词的外延)和分类(发送标签)是分开的。

3.3 标签学习（外延匹配用法）——语义信道匹配 Shannon 信道

优化一个语义信道(一组词的外延)等价于优化一组真值函数. 设 $P(x|y_j)$是从样本 $\boldsymbol{X}_j$ 得到的样本分布，先验分布是 $P(x)$。于是有

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \quad (3.6)$$

不难证明因为当 $P(X|\theta_j)=P(X|y_j)$时，$I(X;\theta_j)$最大，等于 KL 信息 $I(X;y_j)$。根据贝叶斯定理 3 和 $P(X|\theta_j)=P(X|y_j)$可以得到

$$T^*(\theta_j|X)=P(y_j|X)/\max[P(y_j|X)]= [P(X|y_j)/P(X)]/\max[P(X|y_j)/P(X)] \quad (3.7)$$

上式适合有大样本时的非参数估计，而式(3.6)也适合只有小样本时的参数估计。当样本足够大时，用 $T^*(\theta_j|X)$做语义贝叶斯预测和用转移概率函数 $P(y_j|X)$做贝叶斯预测，结果相同。所以语义信息方法和贝叶斯定理 3 兼容。在所有标签中，可能有些标签是逻辑互补的。 假设 y_j’ 是 y_j 的否定，则我们可以用两个条件样本分布 $P(X|y_j)$和 $P(X|y_j')$训练一个真值函数：

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j)$$
$$= \arg\max_{T(\theta_j|X)} \sum_i [P(x_i, y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} + P(x_i, y_j') \log \frac{1\text{-}T(\theta_j \mid x_i)}{1\text{-}T(\theta_j)}] \quad (3.8)$$

上式实际上把所有实例分为三类：正例、反例和不清楚实例（流行方法中分为两种）【24】。因为优化的真值函数 $T^*(\theta_j|X)$只和转移概率函数 $P(y_j|X)$和 $P(y_j'|X)$有关，而和 $P(X)$无关， 所以上式实际上把所有不清楚实例排除在外了。我们也可以全用（3.6）而不用（3.8）。用（3.8）所得隶属函数覆盖范围较大。如果有标签“年轻人”，（3.8）就不合适标签“成年人”。

流行的多标签学习方法中，二元关联解法有很大优势【24, 25】。上面方法有类似地方；差别是：1）标签学习时不做二元分类，选择标签时才分类；2）对样本没有特别要求，也不需要准备多个数据集；3）学习和选择时都用语义信息或 log 标准似然度作为准则；4）$P(X)$改变后，所得隶属函数仍然有效，从而能提高泛化性能。4）用语义信息准则选择标签时考虑到标签的相关性，比如不会有了标签“老年人”还要加“成年人”和“非年轻人”。

翻译时，目标词汇集合和原来词汇集合不同，有些词很难找到相同的外延。比如，虽然英语中 cyan(介于绿蓝之间，和红色互补)通常翻译成青色， 但是“青色”在中文中外延更广，可能有“青草”(其实是绿草)，“青天”(浅蓝色)，“青石板”。我们可以用 $P(X|\theta_j)$取代样本分布 $P(X|y_j)$，用下面公式把 y_j 翻译成 y_k*：

$$y_k^* = \arg\max_k I(X;\theta_k|y_j) = \arg\max_k \sum_i P(x_i \mid \theta_j) \log \frac{T(\theta_k \mid x_i)}{T(\theta_k)} \quad (3.9)$$

把“青草”翻译成“cyan grass”肯定错误. 上面公式要用到草的颜色的先验概率分布，所以“青草”应翻译成“green grass”。上面公式能保证在外延不超出原意的情况下，利用先验知识提供正确且更精确的翻译。

3.4 标签选择（用法匹配外延）——Shannon 信道匹配语义信道

在实例可见时，使用最大语义信息准则，选择分类的分类函数是

$$y_j^* = f(X) = \arg\max_{y_j} \log[T(\theta_j \mid X) / T(\theta_j)] \quad (3.10)$$

如果语义信息最大的复合标签不止一种，则选择其中原子标签较少的一种。当所有集合清晰时，上面信息准则就退化为最大 Bar-Hillel-Carnap 信息准则【17】或最小逻辑概率准则：

$$y_j^* = f(X) = \arg\max_{y_j \text{ with } T(\theta_j|X)=1} \log[1/T(\theta_j)] = \arg\min_{y_j \text{ with } T(\theta_j|X)=1} T(\theta_j) \quad (3.11)$$

可以说，标签学习得到标签的(模糊)外延，而标签选择依据标签提供的信息——信息可以理解为内涵。可以说，最大语义信息准则就是最丰富内涵准则。以上方法提供了一个简单的多标签分类方法。 它对样本没有特别要求，只要能从样本得到联合分布 $P(x,y)$就行。又因为每个标签学习是单独的，设计每个标签的真值函数的参数形式时不用考虑其他标签。这种方法自然解决了二元关联方法需要解决的标签相关性问题， 比如标注了“老年人”就不会添加“成年人”和“非年轻人”标签。

另外， 这种方法考虑到 $P(X)$变化对分类的影响。我们以标签“老年人”为例说明类别划分随实例的先验分布 $P(X)$变化(参看图 4)。其中设 X 表示年龄，标签 y_1=“老年人”。假设

$$T(\theta_1 \mid X) = \frac{1}{1+e^{-0.2(X-75)}}, \quad P(X) = 1\text{-}\frac{1}{1+e^{-0.15(X-c)}} \quad (3.12)$$

图 4 显示了有关函数和优化的划分点 x^*。表 1 显示划分点随 X 的先验分布 $P(X)$中参数 c 的变化而变化。

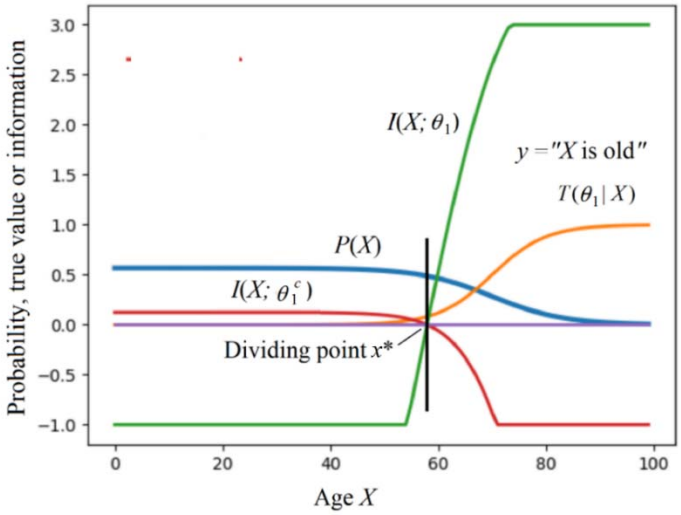

图 4 老年人分类划分点 x*. 它在老年人人口比例增加时右移。

表 1 “老年人”分类器 $f(X)$ 随人口密度分布 $P(X)$中的参数 c 的变化

$P(X)$中参数c	人口密度下降区域	划分点($y_1=f(X\geq x^*)$
50	40-60	49
60	50-70	55
70	60-80	58

注意：这时候“老年人”的外延并没有变化。划分点右移会影响新样本的 $P(x,y)$以及转移概率函数 $P(y_1|X)$，从而改变词的接收者理解的真值函数。这些变化也会反过来影响划分函数 $f(X)$。“老年人”的语义和使用规则就是这样进化的。没有人规定老年人的严格分界年龄在哪里，语言交流过程中会自动形成一个模糊分界，它随人口的年龄分布变化而变化。老年人多的群体，可能 70 岁才算老年人，50 岁算年轻人。以前人的寿命短，50 岁就算老年人了。“大雨”类似，雨水多的地区可能日降雨量 50mm 才算大雨；而雨水少的地区可能日降雨量 20mm 就算大雨了。

如果实例是不可见的(比如我们根据人的声音把人分成男女，或者根据西瓜的外观把西瓜分为好瓜和差瓜，我们只知道观察数据 Z，则可用平均语义信息准则选择标签或划分观察数据空间， 即

$$y_j^*=h(Z)=\arg\max_{y_j}\sum_i P(x_i\mid Z\in C_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.13}$$

在预测不准时，模糊集合外延较大可以减小误差带来的信息损失， 所以用上式选出来的 y_j^*, 其逻辑概率 $T(\theta_j)$未必最小。这也是为什么天气预报经常报“小到中雨”。对于不可见实例分类，划分函数 $h(Z)$(确定每个 C_j)会改变 Shannon 信道和与之相匹配的语义信道，所以又要求重新划分… 因此需要迭代方法。参看鲁晨光的求最大似然估计的信道匹配算法【10】。

4 总结

本文提出和证明了第三种贝叶斯定理， 举例说明了如何用它，通过的样本的先验和后验分布，简单地求出词的外延——即真值函数或隶属函数；通过外延可以根据新的先验分布做新的概率预测。最大语义信息准则被证明是一种特殊的 RLS 准则，它等价于最大似然准则。本文提出多标签分类应该分两步走：多标签学习(语义信道匹配 Shannon 信道)和多标签

选择(Shannon 信道匹配语义信道)。文中提供了一个非常简单的多标签分类方法：1)从样本得到 Shannon 信道， 从 Shannon 信道得到语义信道(即一组外延)；2)用最大语义信息准则选择最好标签，划分实例空间。这种方法继承了流行的二元关联学习方法【12】的优点，但是不做二元分类，所以标签学习非常简单。分类时考虑到类别不平衡和先验概率分布的变化，所以具有较好的泛化性能。另外也能克服标签相关性问题。文中还介绍了用于翻译的广义 Kullback-Leibler 公式，讨论了自然语言中“老年人”分类如何随年龄结构变化——说明分析结论和实际符合。

本文方法也有其局限性，它只适合样本较大，样例较多场合。本文求标签外延方法可谓“逻辑贝叶斯推理”，关于它和贝叶斯主义推理(Bayesian Inference)的关系，有待进一步讨论。

参考文献

1 Fisher, R. A. On the mathematical foundations of theoretical statistics[J]. Philo. Trans. Roy. Soc., A222, 1922: 309-368.

2 Stephen E. Fienberg, When Did Bayesian Inference Become “Bayesian?"[J], Bayesian Analysis，2006，1(1)，1-40.

3 Bayes T, Price R. An essay towards solving a problem in the doctrine of chance[J]. *Philos*ophical Transactions of the Royal Society of London, 1763, 53(0), 370–418

4 冯志伟，自然语言处理的形式模型[M]，中国科学技术大学出版社，2012

5 Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*[J]. 1948, 27(3):379–429 and 623–656.

6 鲁晨光，广义熵和广义互信息的编码意义[J]，通信学报，1994, 15(6): 37-44

7 鲁晨光，广义信息论【M】. 合肥：中国科学技术大学出版社，合肥， 1993

8 Lu C. A generalization of Shannon's information theory[J], Int. J. of General Systems, 1999, 28 (6): 453-4.

9 Lu C. Channels’ matching algorithm for mixture models[A], in IFIP International Federation for Information Processing 2017, Shi et al. (Eds.), Springer International Publishing, Switzerland, 2017, 321–332.

10 Lu C. Semantic channel and Shannon channel mutually match and iterate for tests and estimations with maximum mutual information and maximum likelihood[A]，in Proceedings of 2018 IEEE International Conference on Big Data and Smart Computing，Shanghai, Du et al (eds.), January 2018, 15-18.

11 Jaynes E T. Probability Theory: The logic of Science[M], Edited by Larry Bretthorst, Cambridge University press，New York, 2003

12 Zadeh, L. A.: Fuzzy Sets[J]. Information and Control, 1965, 8(3), 338–353

13 Zadeh, L. A. Probability measures of fuzzy events[J], J. of Mathematical, Analysis and Applications, 1986, 23(2), 421-427.

14 Tarski, *A.:* The semantic conception of truth: and the foundations of semantics[J]. *Philosophy and Phenomenological Research*, 1944, *4(3):* 341–376

15 Davidson, D.: Truth and meaning[J]. Synthese, 1967, 17(1): 304-323

16 汪培庄, 模糊集合随机集落影[M], 北京：北京师范大学出版社，1985.

17 Bar-Hillel Y, Carnap R. An outline of a theory of semantic information[M]. Tech. Rep. No. 247, Research Lab. of Electronics, MIT , 1952)

18 Floridi L. Outline of a theory of strongly semantic information[J]. Minds and Machines. 2004，14:197-221.

19 钟义信，信息科学原理[M], 北京：邮电大学出版社, 2013.

20 Popper K. Conjectures and Refutations[M]. Repr. Routledge, London and New York, 1963/2005

21 Kullback, R.A. Leibler. On information and sufficiency[J], Annals of Mathematical Statistics , 1951, 22 (1), 79–86.

22 Goodfellow, Ian; Bengio, Yoshua， Deep Learning, The MIP Press, Cambridge, 2016.

23 周志华，机器学习，清华大学出版社，北京， 2016

24 Zhang, M. L.; Li,Y. K.; Liu, X. Y.; Geng, X. Binary relevance for multi-label learning: an overview， Frontiers of Computer Science，2018, 12(2), 191–202

25 Zhang, M. L.; Zhou, Z. H.: A review on multi-label learning algorithm. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8), 1819-1837.