“关于贝叶斯决策与机器学习目标选择的讨论”

学 术 前 沿

关于贝叶斯决策与机器学习目标选择的讨论

胡包钢

首先声明这是一篇不成熟的科学问题讨论式文章，其中包括个人学习新知识后的研究思路总结。目的是探讨与交流学术思想，而非只是技术方法的介绍。前不久读到黄合来博士的博文《科研思维与论文写作之5C原则》[1]，深表认同他的说法“科学研究可以笼统的用胡适先生提出的‘大胆假设，小心求证’进行概括，是一个开拓求新与严谨求实的有机结合”。从研究人员角度出发，我冒昧地引申一下：“大胆假设学术新思想，小心求证理论新方法”。当研究中“求新”与“求实”不可偏废时，我理解“求新”可能更具挑战。新思想来源于对问题的洞见和发现，然而如何从学术层面提出或定义出新问题并非是件易事。特别是当我进入机器学习研究领域后，更加体会到迎接挑战的艰难，这其中还确实要包含大胆的勇气：“为什么我们不能提出新的学术问题或思想？由此创造可以进教科书的理论知识，或引导学术发展潮流呢？”

为此，本文以“大胆假设学术新思想”出发对机器学习开展问题探讨。首先提出一个一般性问题“为什么信息处理研究中机器学习被认为是目前最为热门的课题，而不是模式识别或人工智能”？其中包括的原因可以有多种。但我理解的根本原因之一有：统计学在机器学习研究中为重要而主流的理论基础，并成为新方法发展的源泉（如支持向量机的产生）。该原因的探讨使我们可以看到，研究课题是否称为机器学习或模式识别并非至关重要，而应用统计学方法已成为信息技术发展的必然路径。这也是为什么最近10多年来模式识别或人工智能研究开始靠拢或融合统计机器学习的研究主题。而其中的重要发展趋势是广泛应用贝叶斯决策理论方法。下面的问题则是“贝叶斯决策与机器学习目标的关联是什么”？让我们先简短回顾贝叶斯定理或决策原理及存在问题。贝叶斯计算公式与语义解释[2]似乎很简单：

$$\left.\begin{aligned} P(H \mid E) &= \frac{P(H)P(E \mid H)}{P(E)} \\ \text{后验概率} &= \frac{\text{先验概率} \times \text{似然函数}}{\text{归一化因子}} \end{aligned}\right\} \tag{1}$$

其中，H 为假设（如可以理解为模型参数）；E 为证据（如观测数据）。式(1)右端仅包含三项内容，给定先验信息（即先验概率 $P(H)$）和观测数据（即似然函数 $P(E|H)$），贝叶斯定理计算后验概率来推断假设（可理解为确定模型参数）。计算中的归一化因子是为了保证后验概率满足规定的归一化性质。如图1所示，贝叶斯决策的主要思想是：① 应用概率来描述推断；② 推断是基于先验信息和观测数据两部分相融合而实现的。让人感叹的是简洁表述的贝叶斯定理已成为现代统计学中的基础理论，有时又简称为“贝叶斯统计学”[3]。

它不仅已广泛应用在估计、决策、预测等各种问题中，而且渗透到农业、工程、经济、医学等多种学科领域[4]。该理论方法也被认为是获得科学知识或结论的基本手段。行文至此，我更加体会到理论中简洁性(simplicity)的威力。越简洁表达的理论方法，越有生命力，可以在更为广阔的空间生存。

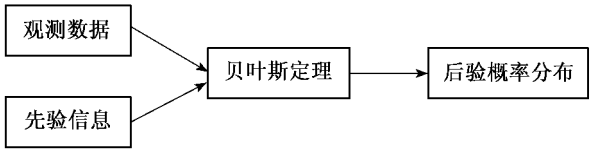

图 1　贝叶斯方法示意图(修改于文献[3])

然而，贝叶斯理论同样面临着诸多挑战。不久前，国际贝叶斯统计学会(ISBA)新当选主席 Jordan 教授根据分发给国际统计学界著名学者的问卷调研结果，公布了贝叶斯理论方法中的五个重大待解问题[5]，它们的排序分别是：① 模型选择与假设检验；② 计算与统计；③ 贝叶斯与频率学派的关系；④ 先验；⑤ 非参数与半参数。在先验概率给定方面，分别产生了主观和客观贝叶斯学派。应用主观方式赋予先验概率可能并不构成什么技术难点，但是保证客观方式赋值先验概率将遇到麻烦。本文后面还将就客观性的问题予以讨论。另一方面，当对模型参数考虑为常数时，一般称为频率学派。常数设定是需求大数渐近假设的。而贝叶斯学派则将模型参数考虑为随机变量，它不是一个常数，而是一个分布。如某物理对象模型被假定为伯努力分布时，该离散概率分布有一个参数，取值范围是 $0\leqslant p\leqslant 1$。在给定小样本训练学习的情况下，为了避免主观选定或有偏学习到 p 值的缺陷，完全的客观贝叶斯学派一般是应用 Beta 分布(为伯努力分布的共轭分布)来描述模型参数 p。Beta 分布中包含两个参数，分别是 α 和 β。由于它们是“关于参数的参数”，因此称为“超参数(hyperparameters)”。超参数的应用似乎更为符合贝叶斯决策思想：参数本身也是一种统计推断，应该由概率分布(而非常数)表示。但新的问题是，超参数 α 和 β 又将派生出推断的求解。对此若再产生一层超参数，那么何处终止呢?

目前关于贝叶斯理论的研讨同样也引起了哲学领域研究人员的关注，并成为研究主题。如何看待统计学及其与其他学科的关系，在我最近读完国际著名统计学家 Rao 教授《统计与真理：怎样运用偶然性》[6]一书之后，才更加深理解了统计学的潜在威力与研究魅力。他认为统计学不仅是一门科学，还是一门艺术：“因为依赖于归纳推理统计学的方法论不是完全能编成条例或是没有争议”。其中他强调“不确定性”的思维应该融入到人们的一切活动，并认为这是人类生活中最为正常的部分。据此思维引申，我理解应该将任何理论方法本身的缺陷视为一种“不确定性”。该不确定性一般可以分解为两种不同成分，一种是通过人们努力可以消除的部分；另一种是自然界的属性而人们无法消除的部分。如果将贝叶斯理论与物理学科中已有的理论相比，它会在整体的“不确定性”方面更为显著。如对于同一问题应用贝叶斯理论原理求解可能会有多种方式解法。该特征也就体现了研究与应用贝叶斯理论的特殊魅力。

下面开始讨论贝叶斯决策与机器学习目标的关联。以二值分类学习问题为例，此时式(1)的 E 变量可以改变为分类器的输入特征变量 X，H 变量变为分类器输出变量 Y。若分类器应用贝叶斯定理实现决策，相当于选择最大后验概率 $P(Y|X)$ 为分类学习目标。这也称为“贝叶斯分类器”，具有最小分类误差的特征。所谓最小分类误差是相对于输入特征变量确定后而言。如果改变输入特征变量取得更优的结果将是另一回事。理论上已经证明贝叶斯分类器在分类误差上是最优的[2]。该问题理论解表明最大后验概率与最小分类误差在分类学习目标上是等价的。由此等价关系，学术界也自然地将贝叶斯决策设定到机器学习目标中。

但是，在我们最近的分类问题研究中，开始提出如下疑问："人们生活中应用'物以稀为贵'分类方式的数学解释机理是什么？是贝叶斯决策机理吗"？在二值分类学习问题中，人们经常对小样本类别更为关注。最为典型的例题是医学诊断。当对小样本类别的癌症病人发生错判时，一般会给出更大的代价。相比而言，对于大样本类别中正常人发生分类错判时，会应用较小的代价。类似的应用还可以包括许多实例，如海量网络数据中的有用信息查询、食品或药品中的次品检测等应用。"物以稀为贵"的常识体现在相当多的分类问题中。具体应用中会发现反例。但如果抛开具体应用来考察，对于以下的分类结果（其中大类样本共有 90 个、小类样本是 10 个、非对角元素为错分个数、两个分类器均有分类误差为 1%）

$$\boldsymbol{C}_1 = \begin{bmatrix} 90 & 0 \\ 1 & 9 \end{bmatrix} \quad \boldsymbol{C}_2 = \begin{bmatrix} 89 & 1 \\ 0 & 10 \end{bmatrix} \tag{2}$$

人们通常会认为 $\boldsymbol{C}_2$ 分类结果优于 $\boldsymbol{C}_1$ 分类结果。这里反映了人们对于误差类别是有区别的。区别的依据反映人们应用"物以稀为贵"常识的内涵：一是小样本更为昂贵；二是样本越小就越昂贵。对于误差类别，我应用贝叶斯分类器方法进行了考察[7]。为了尽量符合客观贝叶斯分析的要求，假设两类别的分布信息已知（避免主观先验而不失相关理论结果的一般性），代价矩阵为缺省（这相当于应用了"0-1"代价函数）。理论推导结果表明，在两类样本不平衡比增大时，该贝叶斯分类器趋向于将全部小样本类别判断为大样本类别。与此分类器相比较，还应用了基于互信息为学习目标的分类器。该分类器无需代价矩阵（还不包括缺省情况）作为输入信息给定。半解析方法与仿真计算结果表明，基于互信息学习目标的分类器可以实现上述"物以稀为贵"的两个内涵特征。此项研究工作[7]首次推导了贝叶斯分类器在区分误差类别与拒识类别情况下的理论公式。有关理论推导结果激发我对贝叶斯理论方法在机器学习中的角色与应用的下列思考和假设。

首先，机器学习目标选择中是否存在一个可以实现"客观性"分析（或解释）的通用理论框架？如果存在，是否它就是贝叶斯理论框架？所谓"客观性"的需求可以参照 Berger 教授于 2006 年发表的文章[8]。前不久在上海华东师范大学刚刚举办的"第八届客观贝叶斯国际研讨会 "①同样包括学者们对不同客观贝叶斯方法的介绍。虽然 Rao 教授曾将统计学视为通向获得真理的必然路径，但是还可以看到另一种见解。美国著名作家马克·吐温曾经评论到"世上有三种谎言：谎言，该死的谎言，以及统计（There are three kinds of lies: lies, damned lies, and statistics.）"。在基于归纳推理的统计学理论应用中，可能会对同一现实问题得到完全相反的统计分析结论，那么应该如何避免魔术师般的统计学应用呢？可以理解，客观性将是实现获得真理的必然条件。然而，目前学界应用客观分析方法时缺乏对"客观性"在技术层面上可以简单考察的定义，而不应只是从哲学或语义层面上的定义（可能会产生歧义理解）。

其次，基于不平衡样本分类问题研究考察结果，可以看到贝叶斯方法并不能胜任为机器学习目标选择中的通用理论框架。因为它无法获得对人们分类常识的"客观性"数理基础解释。为此有必要回顾贝叶斯理论发展的"源"与"流"是怎样形成的，由此理解方法提出和应用中的起因。贝叶斯定理最初是按照统计推断工具提出来的，其中采用了最大后验概率为推理（或学习）目标。后来著名统计学家 Wald 教授将主观代价因子引入变为风险函数为学习目标[9]。为此风险泛函又成为机器实习中的通用学习目标表达形式[10]。为保证"客观性"机器学习结果，我理解贝叶斯理论将是在"如何学（How to Learn）"方面更具主要角色，而非是在"学什么（What to Learn）"方面。

① http://www.sfs.ecnu.edu.cn/Obayes2011/index.html.

最后是我个人的大胆推测："实现机器学习目标选择的'客观性'分析(或解释)的通用理论框架可能是信息(或熵)理论"。这一想法也是来自于目前对不平衡样本分类问题研究考察结果的直观理解。这里需要区别"基于工程应用的机器学习目标"与"基于数理解释机器学习目标"的不同。当前者表现为特定问题对应特定目标时(如分类误差最小),在此更强调后者研究的重要性。这不仅是因为目前机器学习研究中缺乏这方面系统性探讨,更源自于机器学习最终目标是需求建立人们学习机理的数学解释理论基础。虽然熵的概念在工程应用中并非是常规机器学习目标,但是在"学什么"方面,基于熵理论的学习目标函数设定不仅可能建立起与常规机器学习目标的关联,而且可以为我们探讨新的机器学习方法带来独特的路径。

在给出上述假设之后,将面临"小心求证理论新方法"的重要挑战。为此还需要去做大量的具体研究工作。特别是在机器学习研究中发展熵理论与贝叶斯理论有机融合的理论方法。这里也包含一直困惑我的问题:"熵理论与贝叶斯理论分别代表了处理信息不确定性的两种不同理论体系,它们将有怎样的关联? 两者的融合能否为我们对机器学习机理带来更为基础和本质性的数学理论方面的认知?"

相信在求知的路上"行者常至"。

参考文献:

[1] 黄合来. 科研思维与论文写作之"5C"法则 [EB/OL]. [2011-07-02]. http://blog.sciencenet.cn/home.php?mod=space&uid=421679&do=blog&id=335259.

[2] Duda R O, Hart P E, Stork D. Pattern classification[M]. 2nd ed. John Wiley: New York, 2001.

[3] Kotz S, 吴喜之. 现代贝叶斯统计学[M]. 北京:中国统计出版社,2000.

[4] Berger J O. Bayesian analysis: a look at today and thoughts tomorrow[J]. Journal of the Americant Statistical Association, 2000, 95:1269-1276.

[5] Jordan M I. What are the open problems in Bayesian statistics? [J]. The ISBA Bulletin, 2011,18(1): 1-4.

[6] Rao C R. 统计与真理:怎样运用偶然性[M]. 北京:科学出版社, 2004.

[7] Hu Baogang. What are the differences between Bayesian classifiers and mutual-information classifiers[EB/OL]. (2011-04-30) [2011-07-02]. http://arxiv.org/abs/1105.0051.

[8] Berger J O. The case for objective Bayesian analysis[J]. Bayesian Analysis, 2006,1:385-402.

[9] Wald A. Statistical Decision Functions[M]. New York: John Wiley and Sons, 1950.

[10] Vapnik V. 统计学习理论[M]. 许建华,张学工,译. 北京:电子工业出版社, 2004.

作者简介:胡包钢,中国科学院自动化研究所模式识别国家重点实验室研究员,博士生导师;主要研究方向为模式识别和植物生长建模。E-mail:hubg@nlpr.ia.ac.cn

(完稿于 2011 年 7 月 22 日)

补记:建议应用"物以稀为贵"的英文翻译为"Less costs more"。该翻译改编于"Less is more"的英文说法。该说法对应中文可以是"少即多"或"少寓多",反映了建筑学中的"极简主义"设计理念。