评估经得起评估吗？一种基于第一性原理的分类器评估方法

K. Dyrland <kjetil.dyrland@gmail.com> A. S. Lundervold † <alexander.selvikvag.lundervold@hvl.no> P.G.L. Porta Mana <pgl@portamana.org>

（按字母顺序排列）

西挪威应用科学大学计算机科学、电气工程与数学科学系，挪威卑尔根

†& 莫恩医学影像与可视化中心，豪克兰大学医院放射学系，挪威卑尔根

2022年5月27日；更新于2023年2月21日

如果测量工具不符合任何标准，而且其刻度会随着测量对象而扩大或缩小，那么如何才能进行有意义的测量？本文指出，当前对机器学习分类器的评估实践受到这类问题的影响，导致分类器投入实际使用时产生本可避免的负面后果。本文提出，评估应以决策理论为基础，并探讨了这一基础所带来的影响。主要结论是，每个评估指标都必须是混淆矩阵元素的线性组合，其系数——即“效用”——取决于具体的分类问题。对于二分类，这类可能指标的空间实际上是二维的。研究表明，精确率、平衡准确率、马修斯相关系数、Fowlkes-Mallows 指数、F_1-度量和曲线下面积等常用指标从来都不是最优的：它们总会造成一定比例的错误评估，而这种错误原则上可以避免 。这一比例甚至高于使用系数存在中等程度偏差的决策理论指标所导致的比例。

0 序章：一个短篇故事

一家生产某种电子元件的工厂，其经理希望采用机器学习分类器来评估每个生产出的部件的耐久性。耐久性决定了该部件将用于两种可能的设备中的哪一种。该

This document is designed for screen reading and two-up printing on A4 or Letter paper

分类器应以部件的一些复杂特征作为输入，并根据部件类型输出两个类别标签之一："0"表示"长耐久性"，或"1"表示"短耐久性"。

两个候选分类器，分别记作 A 和 B，使用现有训练数据进行训练。在单独的评估集上使用时，它们得出以下混淆矩阵，按以下格式书写

$$\begin{array}{cc} & \begin{array}{cc} \multicolumn{2}{c}{\textit{true class}} \\ 0 & 1 \end{array} \\ \begin{array}{c} \textit{classifier} \\ \textit{output} \\ \begin{array}{c} 0 \\ 1 \end{array} \end{array} & \begin{bmatrix} \text{True } 0 & \text{False } 0 \\ \text{False } 1 & \text{True } 1 \end{bmatrix} \end{array}$$

并按评估数据总数进行归一化：

$$\text{classifier A:} \quad \begin{bmatrix} 0.27 & 0.15 \\ 0.23 & 0.35 \end{bmatrix}, \tag{1}$$

$$\text{classifier B:} \quad \begin{bmatrix} 0.43 & 0.18 \\ 0.07 & 0.32 \end{bmatrix}. \tag{2}$$

这些矩阵表明，工厂生产的部件中，平均有 50% 为短耐久性部件，50 % 为长耐久元件。

上述混淆矩阵得出了两种分类器的一些常见评估指标的以下数值[1]。类别 0 为"正类"，1为"负类"。**蓝色粗体**表示该指标更青睐的分类器，红色表示较不青睐的分类器：

表 1

度量	分类器 A	分类器 B
准确率（亦称平衡准确率）	0.62	**0.75**
精确率	0.64	**0.70**
F_1 度量	0.59	**0.77**
马修斯相关系数	0.24	**0.51**
Fowlkes-Mallows 指数	0.59	**0.78**
真正例率（召回率）	0.54	**0.86**
真负率（特异度）	**0.70**	0.64

这些指标大多更倾向于分类器 B，其中一些指标的相对差异相当大。只有真负率更倾向于分类器 A，但相对差异仅为 9%。

[1] 平衡准确率：Brodersen 等人 2010； F_1 度量：van Rijsbergen 1974；马修斯相关系数：Matthews 1975；Fowlkes-Mallows 指数：Fowlkes 与 Mallows1983。

因此，这些分类器的开发者建议采用分类器 B。

工厂经理并不完全信任这些指标，便问道："我怎么知道它们是否合适？"开发人员保证，这些指标已得到广泛应用。这位（有工程背景的）经理评论说："我不记得'广泛应用'曾是科学正确性的判据——至少伽利略之后就不是了。"于是决定让两种分类器都试用一段时间，看看哪一种实际上能带来最高收入。这两种分类器分别集成到两条彼此独立、但其他方面完全相同的并行生产线上。

在试用期内，这些分类器的表现符合上述混淆矩阵 (1) 和 (2) 的分类统计量。试用期结束时，工厂经理发现，两种分类器在每个经评估部件上带来的平均净收益分别为[2]

$$\begin{array}{cc} \text{classifier A} & \text{classifier B} \\ \hline 3.5\,\$ & -3.5\,\$ \end{array} \tag{3}$$

也就是说，分类器 B 实际上导致了收入损失 。因此，经理决定采用分类器 A，并带着得意的笑容评论说，永远不该轻信开发人员的建议，因为他们并不了解商业运营中的实际细节。

根据一些补充信息，很容易计算出上述平均收益。对一个电子元件进行正确分类或错误分类所带来的最终净收益如下：

$$\begin{array}{c} \\ \\ \textit{classifier output} \end{array} \begin{array}{c} \textit{true class} \\ \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 15\,\$ & -335\,\$ \\ -35\,\$ & 165\,\$ \end{bmatrix} \end{array} \tag{4}$$

这些价值观背后的原因是：短耐久元件（类别 1）功率更大，常用于高端、昂贵的设备；但如果用在需要长耐久元件（类别 0）的设备中，就会造成严重损坏，进而产生维修费用和退款。长耐久元件功率较小，常用于低端、价格较低的设备；如果用在需要短耐久性部件的设备中，也会造成一定损坏，但因此产生的费用较低。

[2] "$"表示通用货币或价值单位，因此没有写在收益前面。

将上述收益分别乘以各自的发生百分比（即混淆矩阵中的元素），再将所得乘积相加，即可得出最终平均收益。例如，使用分类器 A 得出的最终平均收益为

$$15\,\$ \times 0.27 - 335\,\$ \times 0.15 - 35\,\$ \times 0.23 + 165\,\$ \times 0.35 = 3.5\,\$\,.$$

在当前案例中，混淆矩阵（1）和（2）得出了经理所发现的金额（3）。

1 分类器评估中的问题

上述故事说明了当前机器学习分类器评估程序中一些众所周知的问题:

（a）我们被大量可能的评估指标淹没。通常，人们并不清楚哪一个最具说服力。例如，在上述故事中，鉴于类别 1 的正确分类与错误分类之间的收益差异远大于类别 0，人们可以认为真负率是适当的指标。但这种定性推理究竟会在何时失效？设想净收益如下:

$$\begin{array}{c} \\ \textit{classifier output} \end{array} \begin{array}{c} \textit{true class} \\ \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 45\,\$ & -335\,\$ \\ -65\,\$ & 165\,\$ \end{bmatrix} \end{array}\,. \tag{5}$$

在这种情况下，也可以认为，类别 1 的正确分类与错误分类之间的经济差异大于类别 0。因此，真负率理应再次成为合适的度量。然而，进行与第 § 0 节类似的简单计算后可知，分类器 B 实际上带来的平均收入最高：7.3$/部件，而分类器 A为 4.7 $/部件。因此，在这种情况下，真负率并非合适的度量：我们的定性推理失效了。

(b) 多数现有指标所青睐的分类器，在实践中仍可能并非最佳选择。

（c）大多数流行指标都是通过直觉推理、临时的数学运算、特殊假设（例如高斯性[3]

[3] 例如，Fisher1963年第 31 节，第 183 页，关于马修斯相关系数。

或其他统计假设）以及仅对特殊情况进行分析而提出的。遗憾的是，这类推导既不能保证对所有情况都具有泛化能力，也不能保证所提出的度量由所选假设唯一确定，或满足更一般的一致性要求。

相比之下，可以考察这样一种推导：从具体的定性要求出发，通过数学证明，满足这些要求的特定公式具有独特性。例如，将香农熵推导为满足信息量一组基本要求的唯一普适度量[4]。又如，将概率演算推导为满足归纳推理、学习和预测的一般理性要求的唯一规则集[5]。再如，将决策理论推导为一种框架，能够保证在不确定性下作出理性且最优的决策[6]。

(d) 假设某些广受欢迎的指标能够“在大多数情况下”选出最佳算法——尽管很难从统计学上定义这样的多数，而且从未进行过真正的调查来证实这一假设。然而，我们是否指望最终用户只能祈祷自己不属于那不幸的少数？这种不确定性是否不可避免？

在这个问题上，我们不能掉以轻心：在某些机器学习应用中，生死攸关[7]。想象一个与工厂故事类似的情境，不过这次发生在医疗环境中。分类器需要区分两种肿瘤类型，而它们分别需要不同类型的医疗干预。混淆矩阵相同（1）和（2）。正确分类和错误分类会使特定年龄段患者的预期剩余寿命如下：[8]

$$\begin{array}{c} \\ \textit{classifier output} \end{array} \begin{array}{c} \textit{true class} \\ \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} \begin{bmatrix} 350\,\text{months} & 0\,\text{months} \\ 300\,\text{months} & 500\,\text{months} \end{bmatrix} \end{array} . \tag{6}$$

[4] Shannon 1948；Woodward 1964 § 3.2；另见 Good 与 Toulmin 1968。[5] Cox 1946；Fine 1973；Halpern 1999；Snow 1998；2001；Jaynes 2003 第1–2章；另见 Self 与 Cheeseman 1987；Cheeseman 1988；Russell 与 Norvig 2022 第12章。[6] Russell 与 Norvig 2022 § 15.2；von Neumann 与 Morgenstern 1955 第2–3章。[7] 参见 Howard 1980。[8] 参见 Sox 等，2013 § 11.2.9。

这些价值观可能出现在几种情境中。例如，类别 0 和类别 1 的肿瘤可能需要截然不同的治疗方式。如果类别 0 肿瘤被误诊且未得到适当治疗，患者会立即死亡（0个月）；如果诊断正确，治疗通常会成功，患者的预期寿命也会很长（500个月）。类别 0 的肿瘤可以治疗，但患者的预期寿命会较短（350个月）。然而，如果它们被误诊为类别 1，类别 1 治疗造成的损害会使预期寿命进一步缩短（300个月）。

上面的这个矩阵在数值上等价于（4），仅相差一个共同的加性常数335，因此最终的净收益也偏移了这个数值。很容易看出，这些指标与表1中的完全相同，大多数指标都偏向分类器 B。然而，使用分类器 A得出的预期剩余寿命比使用分类器 B长六个多月。

（e）通常，我们无法像虚构故事中的经理那样，暂时部署所有候选分类器，以观察哪一种实际上能带来最佳结果。这样做甚至可能有违伦理：想想上文的医疗情境，某个分类器可能比另一个导致更多人立即死亡。

（f）最后，上述所有问题都不是由类别不平衡（某一类别出现的频率高于另一类别）造成的。例如，在我们的故事中，两个类别的数量完全均衡。类别不平衡会使所有这些问题恶化[9]。

不过，我们的故事也指出了一个可能的解决方案，可以应对所有这些问题。最终证明对经理而言与之相关的“度量”，是使用候选分类器所获得的平均净货币收益。在上文问题 (d) 中讨论的医疗情境变体里，这一度量是平均预期寿命。在这两种情况下，收集正确分类和错误分类的平均收益与损失信息（汇总于矩阵（4）或（6）中），再将这些信息与分类器对应的混淆矩阵中的统计数据结合起来，就可以轻松地预先计算出这一度量。将前一种矩阵记为(U_{ij})，将混淆矩阵记为 (C_{ij})，其中 i 表示分类器的输出

[9] Jeni 等人2013年；Zhu2020年。

（行），j 真实类别（列），则该度量的公式为

$$\sum_{i,j} U_{ij}\, C_{ij} \tag{7}$$

对所有矩阵元素求和。

在本文中，我们论证，公式（7）确实是评估和比较两个或更多分类器性能的唯一可接受度量，每个分类器都有其自身的混淆矩阵（C_{ij}），这些矩阵是在相关测试数据上收集的。系数 U_{ij}称为效用，取决于具体问题。该公式是具有混淆矩阵的分类器的效用收益（C_{ij}）。

我们的论证基于决策理论，其概述见第 §2节。

效用收益（7）是混淆矩阵元素的线性组合，其系数独立于元素本身。在第 § 3 节中，我们探讨该公式以及二分类中这类指标空间的一些性质。我们还将证明，一些常见指标，如精确率、 F_1-度量、马修斯相关系数、平衡准确率和 Fowlkes-Mallows 指数无法写成这种形式的线性组合（或其一一函数）。这一不可能性带来两个后果。首先，这意味着这些指标很可能受到某种认知偏差的影响。其次，不存在任何分类问题，使得这些指标能够正确地对所有分类器对的性能进行排序。使用其中任何一个指标，都无法排除评估先验地出错的可能性。在第 § 5 节中，我们将证明，接收者操作特征曲线下面积同样如此，并从决策理论的角度对其作出一些补充说明。

另一方面，准确率、真正例率、真负例率等指标可以写成以下形式（7）。因此，对于每一项指标，都存在一组分类问题，在这些问题中，它能够正确排列所有分类器对的性能。

如果我们不确定分类问题所适用的效用，会发生什么？如果效用评估有误，又会发生什么？我们在 § 4 中表明，对效用的不确定性仍会导出形式为（7）的度量。我们还表明，即使效用评估有误，且相对误差高达最大效用的 20%，使用这些效用进行评估仍能正确排列更多分类器，

其数量仍高于使用上述任何一种流行指标时正确排列的分类器数量。

我们在最后的第6节总结并讨论研究结果。

2 决策理论概述

2.1 参考文献

这里简要概述决策理论。我们只关注与分类器评估问题相关的概念，并简要陈述该理论的规则。这些规则相当直观，但必须指出，它们的构建旨在确保逻辑和数学上的自洽：请参阅以下参考文献。关于从人工智能和机器学习的角度介绍决策理论，参见 Russell & Norvig 2022 第15章。North 1968；Raiffa 1970；Lindley1988；Tribus1969 第8章；J effrey 1965；Steele & Stefánsson 2020对其基础与历史的讨论。更详尽的论述参见 Raiffa & Schlaifer 2000；Fenton & Neil 2019；Berger 1985；Savage 1972；以及 Sox等 2013；Hunink等 2014（从医学视角出发）。另请参阅 Ramsey 1926 年富有洞见且妙趣横生的开创性论述。

2.2 决策与类别

决策理论区分了

- 我们无法确定的可能情形：在本例中，即可能的类别；
- 我们可以作出的可能决策。

这一区分很重要，因为它可以防止出现各种认知偏差[10] 一方面，在评估可能情形的概率和频率时，另一方面，在评估我们决策的价值时。例如稀缺偏差[11] “这一类别很罕见，因此其正确分类必然会带来高收益”，以及简单的愿望式思维：“这一事件会带来高收益，因此它发生的概率更高”。

类别数量和决策数量往往甚至也不相同。但在使用机器学习分类器时，人们通常考虑

[10] 卡尼曼等人 2008；吉洛维奇等人，2009年；卡尼曼2011年。 [11] 卡梅勒与库恩罗伊特1989年；金与马库斯1999年；Mittone与Savadori2009年。

可用决策集合与可能类别集合之间存在某种自然对应关系，且两者基数相同的情况。例如，在“猫与狗”图像分类中，类别是“猫”和“狗”，而决策可以是“放入猫文件夹”与“放入狗文件夹”。在医疗应用中，类别可以是“患病”和“健康”，决策则是“治疗”与“让其离开”。如前所述，为了简单起见，我们的大多数讨论和示例都侧重于二分类。

2.3 效用与期望效用最大化

对于每个决策，我们都会根据实际成立的可能类别，为其赋予若干效用。例如，效用可以等于金钱、能量、客户数量、预期寿命或生活质量方面的收益或损失，并以适当的单位衡量；也可以等于这些量的某种组合。

这些效用汇总为效用矩阵(U_{ij})，如公式（4）、（5）和（6）所示。部件 U_{ij} 表示类别 i 成立时，与决策相对应的效用，或者简单地说，是类别 i 在类别 j成立条件下的效用。

在单个分类实例中，如果我们知道哪个类别为真，那么最优决策就是在以该真实类别为条件的各项效用中效用最大的那个。另一方面，如果我们不确定哪个类别为真，而类别 j 的概率为p_j ，且满足 $\sum_j p_j = 1$，那么决策理论指出，最优决策就是具有**最大期望效用** $\bar{U}_i$的决策，其定义为决策 i 的效用相对于各类别概率的期望值：

$$\bar{U}_i \coloneqq \sum_j U_{ij}\, p_j \,. \tag{8}$$

用公式表示，这一**期望效用最大化原则**是

$$\text{choose class} \quad i^* = \arg\max_i \{\bar{U}_i\} \equiv \arg\max_i \left\{ \sum_j U_{ij}\, p_j \right\} . \tag{9}$$

决策理论中的一个非常重要的结论是，理性决策的基本要求意味着，理性智能体的决策背后必须存在一组效用，并且这些决策必须遵循最大化期望效用的原则[12]。

[12] Russell & Norvig 2022 § 15.2；von Neumann & Morgenstern1955 第 2–3 章。

效用是如何确定的?显然,效用取决于具体问题,不能由理论给出(否则它就是一个模型,而不是理论)。在涉及金钱或能量等可测量量的收益或损失的决策问题中,效用可能是显而易见的(货币的效用通常不等于货币的数量,二者之间的关系在某种程度上呈对数关系[13])。在医学问题中,效用可能对应于预期寿命和生活质量;例如参见Sox等 2013年,尤其是第8章和§ 11.2.9,以及Hunink等 2014年,尤其是第4章,其中介绍了如何将这类健康因素转化为效用。

单次分类的最终效用在某些情况下可能取决于一连串后续的不确定事件和决策。例如,在 § 0 的故事中,将短耐久性部件误分类为长耐久性部件,会导致最终设备仅在较大比例的情况下发生故障;而在这些情况下,终端客户又会在较大比例的子情形中要求退款。退款金额甚至可能取决于其他情况。表格($U_{01} = -3354$)中的负效用 $ 来自对所有这些可能最终结果中的损失取统计平均值。这就是所谓的决策网络或影响图[14]这一主题。决策理论的一个分支——效用理论——给出了规则,确保一组效用在单项决策或决策网络中彼此一致。关于效用理论的简要介绍,可参阅 Russell & Norvig 2022 § 15.2、North 1968年 第201–205页,以及本节开头列出的参考文献。

在本研究中,为避免讨论变得复杂,我们不考虑这类规则:只要问题的效用经过仔细评估,这些规则就应大致得到满足。

3 从决策理论视角评估分类器

3.1 分类问题的可采纳评估指标

最大化期望效用是理性决策的基本准则[15]。在本文中,我们关注这样一个阶段:分类器已经在一个测试

[13] 例如 North 1968年 第203–204页;Raiffa1970 第4章。 [14] 除前文已列出的通用参考文献外:Russell & Norvig 2022 § 15.5;Howard & Matheson 2005;分步教程见:Raiffa 1970。 [15] 我们在配套研究 Dyrland 等人 2022中讨论并使用该方法。

数据集上对大量数据进行了 N 分类。记 F_{ij} 为分类器选择类别 i 且真实类别为 j的实例数量。那么，(F_{ij})就是分类器在这一特定测试集上的混淆矩阵。对于分类器选择类别 i 且真实类别为 j的所有实例，最终可获得效用 U_{ij}。因此，分类器在测试集上产生的总效用为 $\sum_{ij} U_{ij}\, F_{ij}$。将其除以 N，即可得到每个数据点的平均效用，我们称之为效用收益；它可以写成

$$\boxed{\sum_{ij} U_{ij}\, C_{ij}} \tag{10}$$

其中，$C_{ij} \coloneqq F_{ij}/N$ 是选择 i 且真实类别为 j的相对频率，而(C_{ij})是归一化混淆矩阵。

因此，对于由效用矩阵(10)所表征的分类问题，效用收益公式(U_{ij})是评估和比较分类器在测试集上性能的自然度量。

请注意，效用 U_{ij} 确实取决于频率 F_{ij} 或 C_{ij}。如果它们取决于频率，就意味着我们要等到所有分类实例都完成后，才能评估每个单独实例的价值。在我们能想到的几乎所有分类问题中，这都会造成评估偏差，例如§ 2.2中提到的稀缺性偏差。此外，在单次分类的后果会在下一次分类进行之前显现的情境中，这种做法根本无法实施。

如果我们将效用矩阵的各个元素加上同一个加性常数，或乘以同一个正乘法常数，

$$U_{ij} \mapsto a\, U_{ij} + b \qquad a > 0\,, \tag{11}$$

因此，具有特定混淆矩阵的分类器所产生的最终效用会受到相同常数的修正。因此，任何一组分类器的排名都会保持不变。毕竟，加性常数或正因子只代表效用尺度的零点或计量单位发生了变化[16]。此类变化不应影响决策问题。事实上，这些变化不会造成影响，本身就是决策理论逻辑一致性的又一例证。

[16] 参见 Russell & Norvig2022§ 15.2.2。

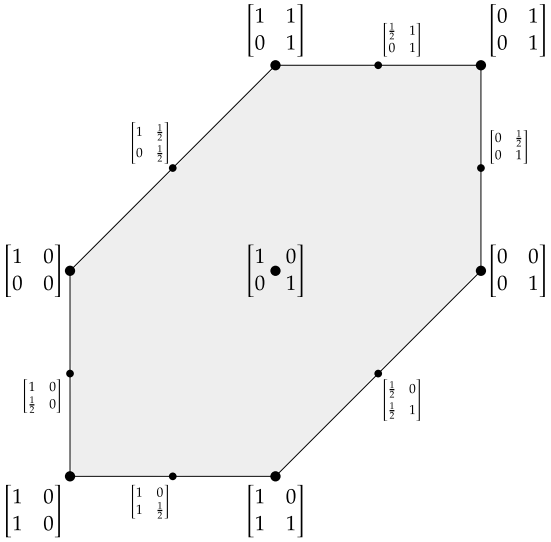


图 1 二分类效用矩阵空间。

3.2 二分类效用矩阵空间

我们来考虑一个二分类问题。它由一个 2×2 效用矩阵刻画。我们假设这些效用并非全都相等；否则，选择哪个类别都无关紧要，分类问题也就毫无难度。我们可以利用选择零点和计量单位的自由，将效用矩阵化为标准形式。我们选择使最大效用为 1、最小效用为 0 的零点和计量单位（请注意，这个值仍可能对应实际的货币损失，例如）。也就是说，我们进行如下变换：

$$U_{ij} \mapsto \frac{U_{ij} - \min(U_{ij})}{\max(U_{ij}) - \min(U_{ij})} \,. \tag{12}$$

按照这一约定，显然，在为二分类问题选择效用矩阵时，我们只有两个自由度。因此，二分类可能采用的评估指标空间（更准确地说，是流形）是二维的。为了评估二分类问题的候选分类器，我们必须从这个空间中选择一个点。

我们可以像图 1 中那样表示这个空间。中心点对应的效用矩阵中，正确分类的最高效用相等，错误分类的最低效用也相等；我们稍后会看到，这对应于使用准确率作为评估指标。从中心向左移动时，正确分类类别 1 的效用相对于类别 0 的效用降低；向右移动则相反。从中心向上移动时，误分类类别 1 的效用增加；向下移动时，误分类类别 0 的效用增加。我们排除了误分类的效用高于正确分类的效用的效用矩阵（尽管某些情况下可能会出现这种情况）；它们会出现在缺失的左上角和右下角。固定经过集合中心的 $(x,\ y)$ 坐标轴后，效用矩阵的坐标为

$$\begin{bmatrix} 1 - x\,\delta(x>0) & y\,\delta(y>0) \\ -y\,\delta(y<0) & 1 + x\,\delta(x<0) \end{bmatrix}, \qquad |x|, |y| \leqslant 1\,. \tag{13}$$

请注意，这种表示方式并非旨在反映任何凸性或度量性质。不过，效用矩阵空间中并未定义度量或距离。如果去掉归一化（12），则可以定义凸组合，但图1的表示方式无法正确反映这一点。

3.3 与常用指标的关系

在 § 3.1 中，我们发现，根据决策理论，最一般的评估指标必须是混淆矩阵元素的线性组合。这种线性组合的系数并不取决于混淆矩阵元素本身；这类依赖通常反映了某种认知偏差。哪些常见的热门指标符合这种数学形式？我们希望在二元分类的情形下回答这个问题，同时尽可能考虑热门指标的典型使用场景。

考虑这样一种情况：我们在同一个测试集上比较多个分类器 *onthesame test set*。数据数量 N 以及测试集中两个类别“0”和“1”出现的相对频率 $f_0,\ f_1$对于所有待评估的分类器而言都是固定且恒定的。

某个分类器会生成一个归一化混淆矩阵（C_{ij}），我们将其写成以下格式：

$$\begin{array}{cc} & \begin{array}{cc} \multicolumn{2}{c}{\textit{true class}} \\ 0 & 1 \end{array} \\ \begin{array}{c} \textit{classifier} \\ \textit{output} \\ \end{array} \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \end{array} .$$

由于约束条件 $C_{00} + C_{10} \equiv f_0$ 和 $C_{01} + C_{11} \equiv f_1$，我们总能从任何公式中消去混淆矩阵的两个元素，或让它们出现，并用涉及其余两个元素和类别频率的表达式替代它们。为避免在解读数学公式的函数形式时产生歧义，我们约定始终只用 C_{00} 和 C_{11} 来表示，并在必要时进行替换：$C_{10} = f_0 - C_{00}$、$C_{01} = f_1 - C_{11}$。

请注意，给定一个效用矩阵，我们总可以用一个共同的正乘法常数来修改其元素 a，也可以用一个共同的加性常数来修改 b，式（11），因为这种修改相当于改变效用尺度的单位和零点。经过这样的修改，评估指标（10）可写成等价形式

$$a \sum_{ij} U_{ij}\, C_{ij} + b \tag{14}$$

因为 $\sum_{ij} C_{ij} \equiv 1$。将求和显式写出，并按照上文所述，用 C_{00}、C_{11} 表示元素 C_{10}、C_{01}，该公式变为

$$a\,(U_{00} - U_{10})\,C_{00} \;+\; a\,(U_{11} - U_{01})\,C_{11} \;+\; a\,f_0\,U_{10} + a\,f_1\,U_{01} + b\;. \tag{15}$$

由于在当前语境下，N，f_0，f_1 都是常数，因此我们可以用任意方式从它们构造任意常数 $a > 0$ 和 b：

$$a = a(N, f_0, f_1) > 0\;, \qquad b = b(N, f_0, f_1)\;. \tag{16}$$

我们也可以利用这种自由，将 $a\,f_0\,U_{10} + a\,f_1\,U_{01}$ 这一项纳入b 上述公式。由此我们得出结论：二分类的评估指标符合决策理论，当且仅当它可以写成如下的一般形式：

$$a(N, f_0, f_1)\; X\; C_{00} + a(N, f_0, f_1)\; Y\; C_{11} + b(N, f_0, f_1) \tag{17}$$

其中，X, Y是实常数，不依赖于 $C_{00}, C_{11}, N, f_0, f_1$；而$a(\cdot) > 0, b(\cdot)$ 只是 N, f_0, f_1 的任意函数。

如果我们只要求比较分数能将多个分类器从优到劣排序，那么上述表达式的单调函数（例如指数函数）也可接受。

从这个角度出发，让我们考察二元分类中的一些常见评估指标。我们用C_{00}, C_{11}来表示它们的公式。

以下指标是公式（17）的具体实例：

√ 准确率： $C_{00} + C_{11}$。我们有 $a = 1$、 $X = Y = 1$、 $b = 0$。事实上，它对应于基于单位效用矩阵（U_{ij}） $= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$的效用收益（或者等价地，使用这样的效用矩阵：对任何类别的正确分类赋予相同的效用，对任何类别的误分类赋予相同且较低的效用）。√ 真正例率（召回率）： C_{00}/f_0。这里 $a = 1/f_0$、 $X = 1$、 $Y = 0$、 $b = 0$。这对应于使用效用矩阵$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$。√ 真负率（特异度）： C_{11}/f_1。这里 $a = 1/f_1$、 $X = 0$、 $Y = 1$、$b = 0$。这对应于使用效用矩阵$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$。

以下这些指标 无法 写成 (17) 的形式，也无法写成该形式的单调函数：

精确率： $C_{00}/(C_{00} - C_{11} + f_1)$。关于 C_{00}, C_{11} 是非线性的。 F_1-度量： $2C_{00}/(C_{00} - C_{11} + 1)$。关于 C_{00}, C_{11} 是非线性的。更一般的 F_β-度量也同样如此。 马修斯相关系数： $\frac{f_1 C_{00} + f_0 C_{11}}{\sqrt{f_0 f_1 f_1 + C_{00} - C_{11}}}(\ )(\ f_0 + C_{11} - C_{00})$。关于 C_{00}, C_{11} 是非线性的。 Fowlkes-Mallows 指数： $C_{00}/\sqrt{f_0}(f_1 + C_{00} - C_{11})$。关于$C_{00}, C_{11}$ 是非线性的。 平衡准确率： $C_{00}/(2f_0) + C_{11}/(2f_1)$。尽管它关于$C_{00}, C_{11}$ 是线性的，并且是两个度量（真正率和真负率）的平均值，而这两个度量都是公式 (17) 的实例，但它本身并非该公式的实例，因为参与平均的两个度量涉及不同的$a(\cdot)$ 函数。

我们发现，许多流行的评估指标并不符合决策理论的原则。这类指标都存在两个问题。

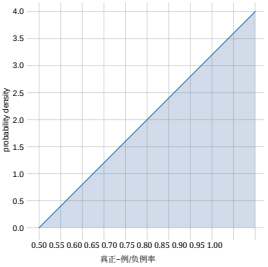
首先，正如第 2 节所讨论的，该度量涉及效用与分类频率之间的相互依存关系，这通常意味着存在某种形式的认知偏差[17]。

其次，该度量得出的混淆矩阵排序，与任何效用矩阵得出的排序都不完全一致——否则，完全一致就意味着该度量可以写成（17）的形式。因此，某些混淆矩阵必然会被错误排序。由于任何理性分类问题都由某个潜在效用矩阵所刻画，这意味着不合规指标总会导致一些错误评估。相比之下，准确率等合规指标，在特定分类问题集合中，对所有成对的混淆矩阵都能给出完全正确的排序。

第二种现象见图 2–3的图示。每幅图中的蓝点代表二分类中从测试数据集得到的一个假想混淆矩阵。点的坐标分别表示：根据分类问题背后的某个效用矩阵计算出的该混淆矩阵的效用收益，以及根据另一种度量计算出的该混淆矩阵得分。所有左列图示所依据的效用矩阵为 $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$，所有右列图示所依据的效用矩阵为 $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$。所考察的其他度量按图示行排列，分别为准确率、真正例率（召回率，其中类别 0 为“正例”）、F_1 度量、马修斯相关系数。

通过首先固定数据集中的类别比例来选取混淆矩阵。对于图 2 中的所有图，类别比例为 50%/50%（平衡数据集）；对于图 3中的所有图，类别比例为 90%/10%（不平衡数据集）。随后，分别选取真正例率和真负率

其取值范围为 $[1/2, 1]$, ，且选取概率随比率线性递增（中位数为 0.85，下四分位数和上四分位数分别为 0.75 和 0.93；见侧图）。因此，这些混淆矩阵代表了倾向于具有较高真负率和真正例率的分类器所产生的分类统计量——从图中各点倾向于聚集在右上角这一事实可以清楚看出这一点。



[17] Hand&Christen2018讨论了与 F_1-度量相关的这类偏差。

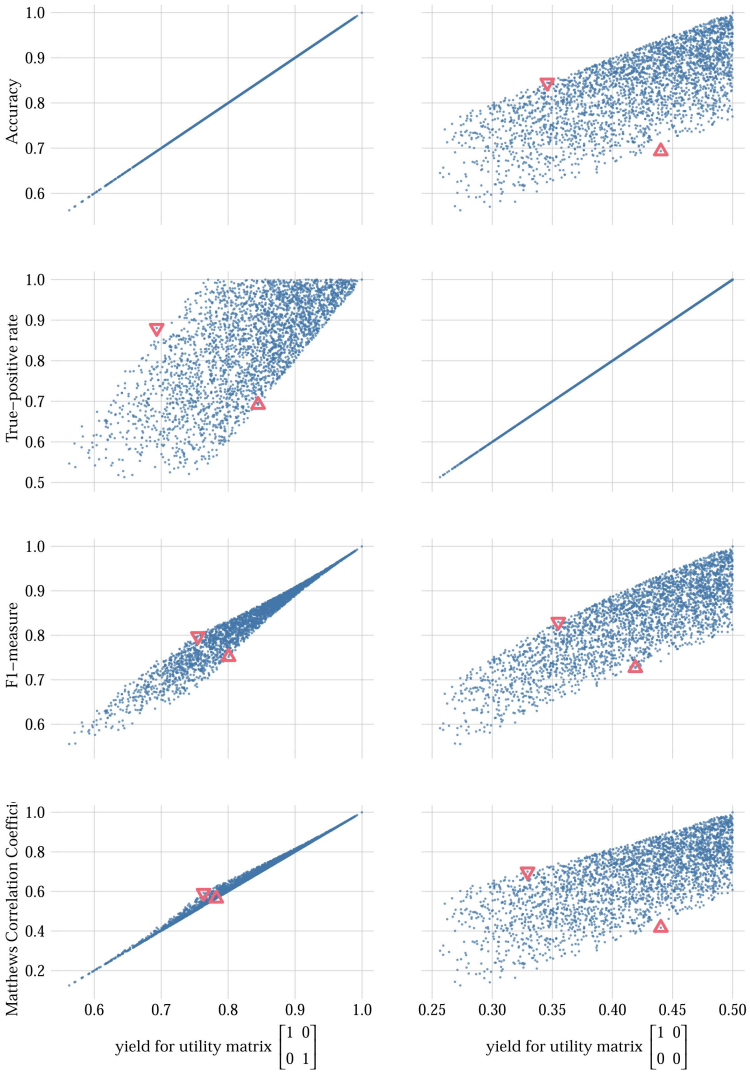


图 2 不同评估指标与两个具有基础效用矩阵的不同二元分类问题的实际效用收益之间的关系 $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$（左列）和 $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$（右列）。所有混淆矩阵（蓝点）均来自类别平衡为 50%/50% 的数据集。图中的成对红色三角形表示两个混淆矩阵：相对于实际效用收益（x 轴），它们被指标（y 轴）错误地排序。显然，指标甚至可能将三个或更多混淆矩阵完全逆序排列。对于左列中的分类问题，准确率能够给出正确的评估；而对于右列中的分类问题，则是真正例率。

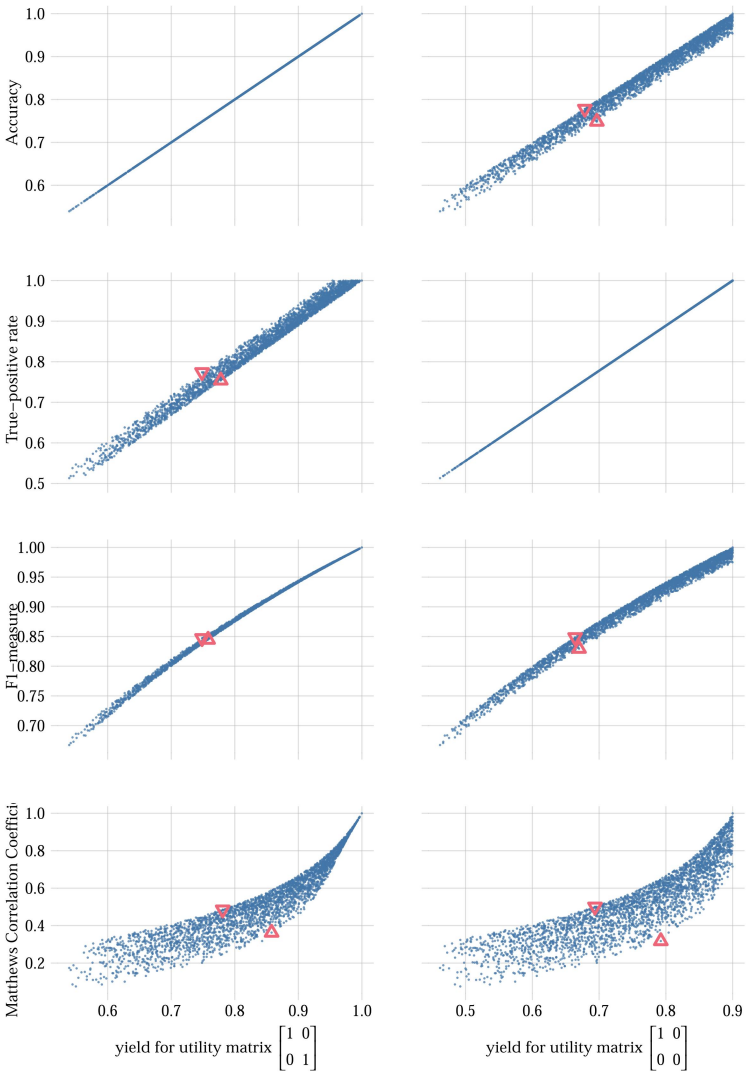


图 3 与图 2 相同，但这里的混淆矩阵来自不平衡数据集，其中类别 0（“正类”）出现的比例为 90%，类别 1（“负类”）出现的比例为 10%。

我们可以看到，当底层效用矩阵等同于 $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$（左列）时，准确率（第一行图）总能对所有混淆矩阵给出正确的相对评估：y 坐标是 x 坐标的单调递增函数——事实上是线性函数。准确率确实是与单位效用矩阵相对应的效用收益。当底层效用矩阵等同于 $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$（右列）时，真正例率（第二行图）总能给出正确的相对评估（前提是测试集相同）。

另一方面，如果将这两个指标中的任意一个用于底层效用矩阵不同的问题，那么指标得分与实际效用产出之间就不存在确定性关系。在这种情况下，总能找到两个或更多混淆矩阵，使该指标给出的评估与实际效用产出完全相反。换言之，根据真实效用表现最差的混淆矩阵及其对应算法，会被该指标排为最佳；反之亦然。图中成对出现的红色三角形，就是被 y 轴指标错误排序的此类混淆矩阵的示例。

准确率和真正例率等指标符合公式（17），因此只有在将它们用于错误的分类问题时，我们才需要依赖评估中的运气。

F_1度量（第三行图）和马修斯相关系数（第四行图）的绘图表明，这两个指标与实际效用产出之间不存在任何函数关系。再次可以找到两个或更多混淆矩阵，使得任一指标给出的评估相对于实际效用产出完全相反。但对于这两个指标，与准确率和真正例率不同，在每一个分类问题中都会始终出现错误评估。

诸如 F_1度量和马修斯相关系数等指标不符合公式（17），因此在评估时始终需要依赖运气。不存在任何能让这些指标始终给出正确评估的分类问题。

不符合决策理论的度量，可能会在某些分类问题和测试集中得到大量正确结果。例如，图2左下方的图显示，当底层效用矩阵为单位矩阵且数据集均衡时，马修斯相关系数几乎是效用收益的单调递增确定性

函数（但当底层效用矩阵为 $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$ 或数据集不均衡时则并非如此；见相应的图）。然而，这种偶尔出现的部分一致性毫无用处。依赖这种部分一致性的前提是掌握效用矩阵；但既然已掌握该矩阵，我们便可以直接使用实际效用产出，它与效用收益完全一致，而且更容易计算。

4 未知或不正确的效用

到目前为止，我们论证了，对于分类器而言，自然的评估指标是其混淆矩阵的效用收益，具体取决于相关分类问题所依据的效用。我们还论证了，许多流行指标（即不符合公式(17)的指标）必然会先验地导致评估不正确的情况。我们的论证基于决策理论的原则。

不过，我们的论证也引出了几个相互关联的问题：

- 当我们不确定某个分类问题所依据的效用时，该怎么办?

- 如果我们使用的效用实际上是错误的，也就是说，并非问题背后真正的效用，会发生什么?
- 不合规指标（例如 F_1-度量或 Matthews 相关系数）平均而言会多频繁地导致错误结果?

事实上，如果对效用的评估中一个小小的误差会导致大量错误评估，而不合规指标平均只会导致少量错误评估，那么决策理论度量的所有严谨性在实践中都毫无用处，不合规指标反而最适合实际应用。

然而，情况并非如此。现在我们来讨论如何处理对效用的不确定性，并给出一项重要结论：即使使用错误的效用，且相对误差几乎达到最大效用的 20%，平均而言，其导致的错误相对评估仍少于使用许多当前流行的度量。

4.1 未知效用；若干分类问题上的平均性能

处理未知效用很简单。假设我们不确定某个分类问题适用的效用矩阵是 $\boldsymbol{U}^{(1)} \equiv (U^{(1)}_{ij})$、$\boldsymbol{U}^{(2)}$、$\boldsymbol{U}^{(3)}$等，其中备选项的数量甚至可以是无限的或连续的。每个备选项$\boldsymbol{U}^{(a)}$ 都有一个概率 q_a；若为连续情况，则有概率密度 $q(a)\,\mathrm{d}a$ 。那么，对于这个分类问题，我们应使用期望效用矩阵

$$\hat{\boldsymbol{U}} \coloneqq q_1\,\boldsymbol{U}^{(1)} + q_2\,\boldsymbol{U}^{(2)} + q_3\,\boldsymbol{U}^{(3)} + \cdots \tag{18}$$

或 $\hat{\boldsymbol{U}} \coloneqq \int q(a)\,\boldsymbol{U}^{(a)}\,\mathrm{d}a$ 在连续情形下。

我们仅简要证明这一符合直觉的结果[18]。如果效用矩阵存在不确定性，那么我们面临一个双重决策问题：既要选择最优效用，也要选择最优类别。例如，若真实的效用矩阵为$\boldsymbol{U}^{(2)} \equiv (U^{(2)}_{ij})$，且真实类别为类别 0，那么选择类别1 会产生效用 $U^{(2)}_{10}$，选择类别0则会产生效用 $U^{(2)}_{00}$，依此类推。因此，我们的双重决策问题可由一个矩形效用矩阵来刻画，该矩阵由效用矩阵$\boldsymbol{U}^{(a)}$按行拼接而成。我们作出一个符合实际的判断：效用矩阵的概率 q_a 与类别的概率p_j 相互独立，因此 $q_a \cdot p_j$ 表示真实效用矩阵为$\boldsymbol{U}^{(a)}$ 且真实类别为 j的概率。最大期望效用原则（§ 2.3 式（9））由此导出对期望效用的最大化

$$\bar{U}_i \coloneqq \sum_{j,a} U^{(a)}_{ij}\, q_a \cdot p_j \equiv \sum_j \underbrace{\left[\sum_a q_a\, U^{(a)}_{ij}\right]}_{\hat{U}} p_j \tag{19}$$

其中，期望效用矩阵（18）表现为一种“有效”效用矩阵，仅用于类别决策问题。

如果对于不同类别的条件效用，我们的不确定性是对称的——例如，我们对类别 0 条件效用的不确定性与对类别 1 条件效用的不确定性相同——那么期望效用矩阵就等同于单位矩阵。在这种情况下，效用收益

[18] 例如参见 Raiffa1970，尤其是第 3 章。

等于准确率。因此，如果我们对底层效用完全不确定，准确率自然就是应采用的评估指标。这一点确实反映在 § 4.2 中讨论的一些结果中。

对于二分类，可能的效用矩阵集合可以如图 1所示表示，正如第3.2节所述。然而，我们对真实底层效用矩阵的不确定性对应于在该集合上的离散或连续概率分布。但请注意，期望效用矩阵（18）并不对应于该分布的质心，因为该图使用了特殊的坐标系。实际的质心需要将效用矩阵集合表示为三维空间中的二维曲面（一个四面体）才能获得。为简洁起见，本文不讨论这种表示方法。

对效用进行平均的过程，即公式（18），同样适用于我们希望评估某个分类器在多个分类问题上的平均表现的情况，这些分类问题的效用矩阵各不相同。同样，我们需要使用的是它们效用矩阵的平均值。

4.2 错误效用评估的后果及与常见指标的比较

我们对某个分类问题的效用矩阵的评估可能有误，尤其是在信息不足、只能依据半定量理由进行评估时。这样一来，我们对分类器的比较评估也可能出错。在这种情况下，比较评估出错的平均概率是多少？它与效用中的误差大小有何关系？这种概率是否高于使用其他指标进行评估时出错的概率？

要准确回答这些问题极其困难，甚至是不可能的，因为要定义“平均而言”，我们需要对各种分类问题进行调查，收集有关其底层效用矩阵、用于解决这些问题的候选分类算法的混淆矩阵，以及效用评估中所犯误差的统计数据。我们试图基于以下假设和判断，对上述问题在二分类情形下作出粗略回答：

（i）图中真实效用矩阵集合上的两种可能分布。1（采用该坐标系）：1. 均匀分布；2. 双变量

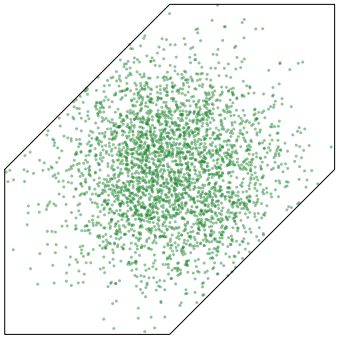


图4 图1中效用矩阵空间里的截断高斯分布1，见条目(i)。

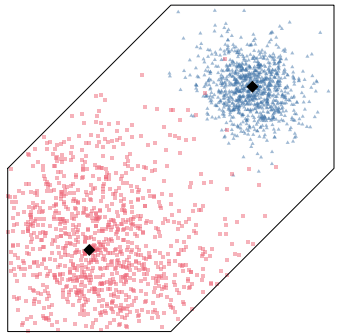


图5 以效用矩阵为中心、标准差为 0。1(蓝色三角形) 和 0。2(红色方块) 的误差范围 $\left[\begin{smallmatrix} 0.5 & 0 \\ 0 & 0.5 \end{smallmatrix}\right]$ 和 $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ (黑色菱形)。

（截断的）以单位矩阵为中心的高斯分布 $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$，其在1/3 的标准差适用于 x 和 y 坐标（见式（13））， 如图4所示。

(ii) 混淆矩阵的一种分布：一个类别的比例在 $[0, 1]$, 中均匀分布，而真阳率和真负率在 $[0。5, 1]$ 中独立分布，且概率线性递增（中位数为 0.85，下四分位数和上四分位数分别为 0.75 和 0.93；见第 16 页的侧图）。这意味着，我们认为不平衡数据问题与平衡数据问题同样常见（根据我们的经验，这是一个现实的假设），而候选分类器通常表现良好。

(iii) 以各效用矩阵元素的真实效用值为中心，在其周围对误差进行截断高斯分布建模。我们考虑的标准差范围为 0 至 0.3。高斯分布必须截断，因为每个真实效用值都介于 0 和 1之间，而且我们要求正确分类的效用高于错误分类的效用。图 5 展示了在标准差分别为0.1 （蓝色三角形）和 0.2 （红色方块）时，效用矩阵空间中的误差范围。

在这些假设下，我们计算一对分类器在具有相同类别比例的两个混淆矩阵时，

当使用错误的效用矩阵或其他度量进行评估时，按照其实际效用产出，评估结果出现逆序的频率。这个计算是一个积分问题，我们通过蒙特卡洛抽样求解。该过程很直观：

1. 根据分布选择一个“真实”的效用矩阵(i)。2. 根据分布选择真实效用矩阵各元素周围的误差(iii)，并将这些误差加到矩阵中。

3. 选择一个类别比例，然后根据分布 (ii)，选择两个具有该类别比例的混淆矩阵（类别比例必须相同，因为这两个矩阵来自同一批数据）。

4. 使用步骤 1 中的真实效用，计算第二个混淆矩阵与第一个混淆矩阵的真实效用收益之差（带符号）。如果该差值为正，则第二个混淆矩阵的效用较高；如果为负，则第一个混淆矩阵的效用较高。

5. a. 考虑若干指标（精确率、马修斯相关系数等）。对于每项指标，计算其对第二个混淆矩阵给出的得分与对第一个混淆矩阵给出的得分之差（带符号）。b. 考虑步骤 2 中的错误效用矩阵。使用该错误效用矩阵，计算第二个混淆矩阵与第一个混淆矩阵的效用收益之差（带符号）。

> 无论哪种情况，差值为正都表示第二个混淆矩阵被排为“最佳”，第一个被排为“最差”；差值为负时则相反。

6. 接下来，逐一检查步骤5中得到的带符号差值，并将它们与步骤4中得到的带符号差值进行比较。如果步骤5 的差值与步骤4的差值符号相反，那么相应的度量或错误的效用矩阵就会对这两个混淆矩阵作出相反且错误的排序。

在真实效用矩阵服从均匀分布、采用多种度量，且效用受到标准差为0.1 的误差影响时，该抽样过程的结果如图6所示。每个点

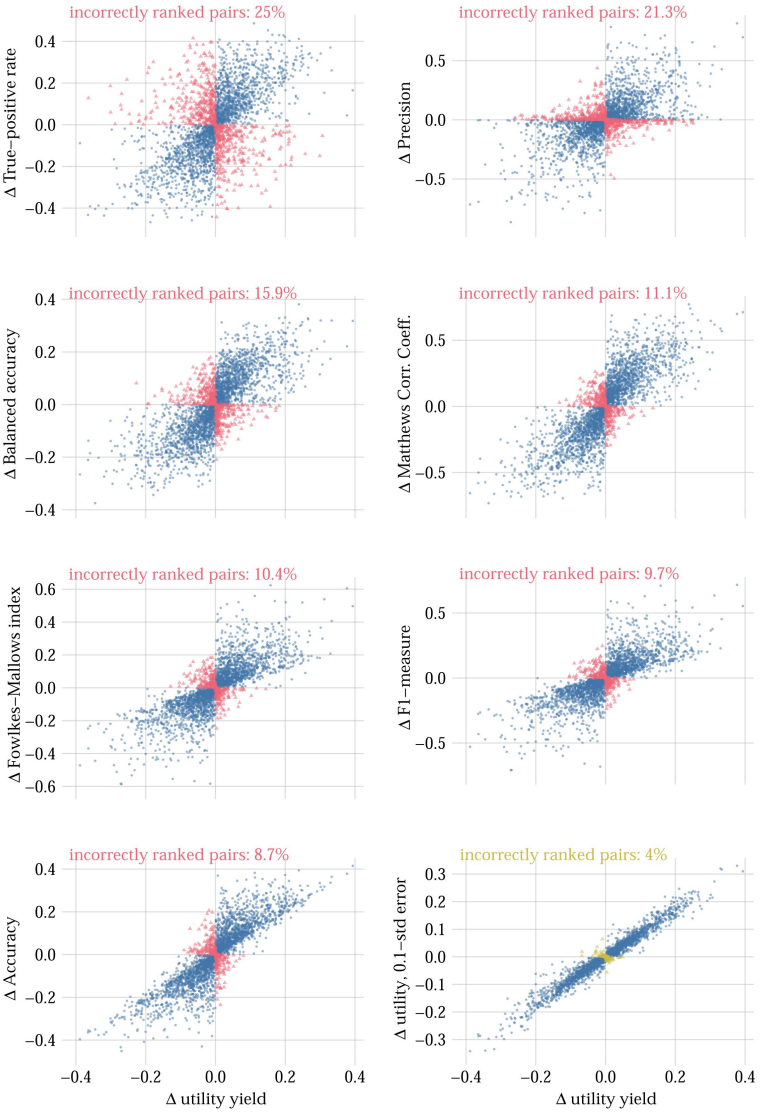


图6 根据“真实”效用矩阵得出的效用收益差异，与根据其他指标（包括评估错误的效用矩阵，误差的标准差为 0.1）得出的得分差异之间的关系。落在第二或第四象限的点代表被错误比较的混淆矩阵对。

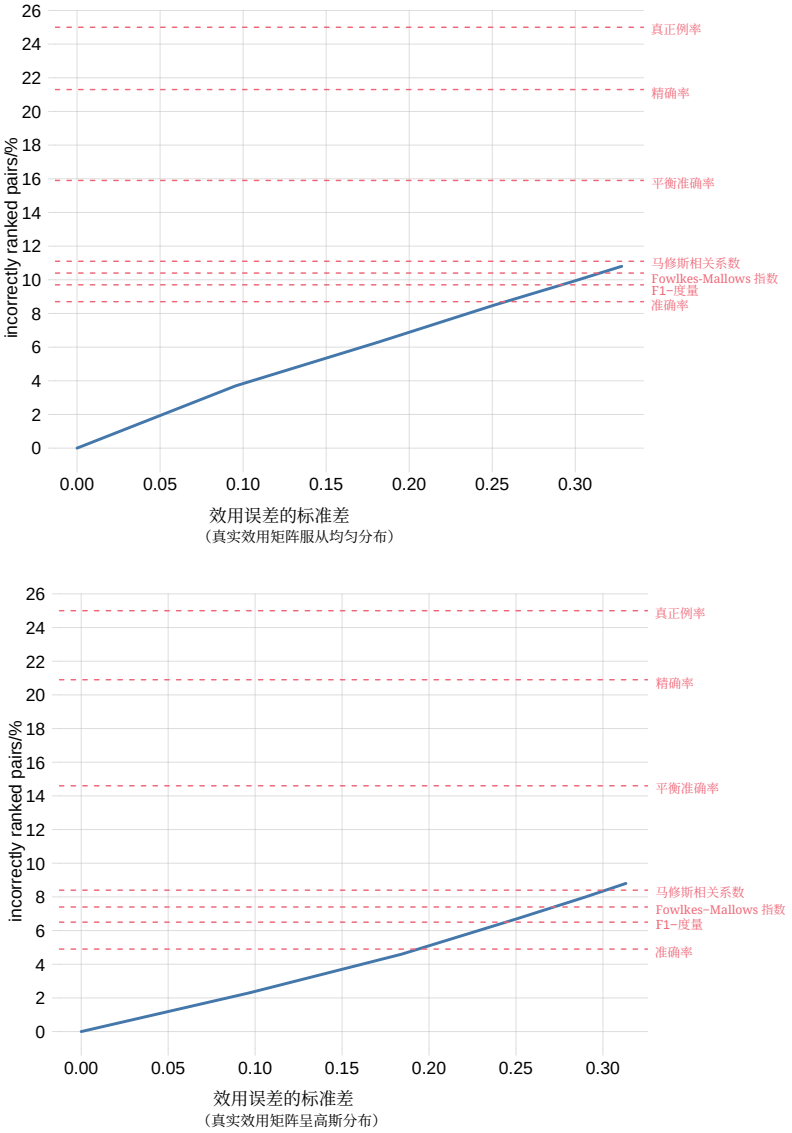


图7 混淆矩阵排序错误的配对比例与效用评估误差标准差之间的关系。上图：真实效用矩阵服从均匀分布的情况。下图：真实效用矩阵服从高斯分布的情况，与图4.

表示一对混淆矩阵（步骤3）；其坐标分别是真实效用收益，以及某个指标给出的得分，或（最后一幅图）根据错误效用矩阵得到的收益。第 II 和 IV 象限中红色或黄色的三角形点（符号不一致）代表排序错误的成对项目。错误排序的百分比根据10^6 个样本计算，结果保留略多于一位小数的有效数字；图中显示的样本数较少。

这些图按错误排序百分比从高到低排列（从左到右、从上到下）。在所考察的指标中，准确率指标表现最佳，成对排序错误率为 8。7%。不过，我们发现，受标准差为 0。1 的高斯误差影响的效用矩阵表现更好，成对排序错误率仅为 4%。

图7展示了错误影响效用时，错误排序比例与误差标准差之间的关系，分别对应真实效用矩阵服从均匀分布（上图）和高斯分布（下图）的情况。两者近似呈线性关系。图中还报告了其他指标的错误排序比例。可以看到，即使效用矩阵受到标准差高达0。15 甚至0。25的误差影响，基于该矩阵的评估仍比基于其他所报告指标的评估更加可靠。考虑到这类标准差对应的误差相当大（如图5所示），这一点尤其值得注意。

效用误差的标准差约为 0.25 时，几乎均匀地覆盖了整个效用矩阵空间（参见图5）。如此大的误差意味着，我们一开始就几乎完全不确定这些效用。因此，当误差达到这一水平时，准确率（相当于使用单位效用矩阵）成为更可靠的度量，这是合理的：正如我们在 § 4.1中所见，在对效用完全不确定的情况下，单位效用矩阵是自然的选择。这个结果再次体现了决策理论的内在一致性。

5 受试者工作特征曲线下面积又如何?

另一个非常常见的二元分类器评估指标是受试者工作特征曲线下面积，简称“曲线

下面积”。这项指标仅适用于特定的分类算法，其含义也不同于迄今为止讨论过的指标。出于这些原因，我们将完整讨论留待未来研究，本文仅在此提出几点说明。

曲线下面积只能针对输出连续变量而非类别的分类器计算。该变量的阈值决定其取值预测的是哪一个类别。在给定的测试集中，阈值选择不同，会产生不同的假阳性率f（即 1 –真负率）和真正例率 t。这些数值对可以绘制成一条曲线 $f \mapsto t\,(f)$，并显示在坐标轴对应的图表上。给定测试集中的类别比例，曲线上的每个点都对应分类器根据所选阈值可能生成的一个混淆矩阵 $C_{ij}\,(f)$。曲线所围成的面积是以一种特殊权重计算的真正例率加权平均值；例如，权重相对于假阳性率是均匀的，但相对于阈值通常并不均匀。Metz 的经典著作（1978）讨论了受试者工作特征的含义及其恰当用法，尤其可参见第 290 页。

从决策理论的角度来看，可以作出两点说明[19]。首先，根据最大期望效用原则，§ 2.3，我们应选择一个阈值及相应的假阳性率 f^*，使效用收益最大化，该收益由公式（10）给出：

$$\text{choose} \quad f^* = \arg\max_{f} \left\{ \sum_{i,j=0}^{1} U_{ij}\, C_{ij}(f) \right\} . \tag{20}$$

任何其他的 f 取值以及阈值都无关紧要。因此，对 f 取值求平均同样无关紧要。其次，假设我们的目标是评估多个可能分类问题上的平均性能。在这种情况下，应进行平均的量是这些分类问题的效用矩阵，如§ 4.1中所述，从而得到唯一的期望效用矩阵。计算出该矩阵后，我们根据第一点说明，重新回到对单个 f 的选择。

由于存在这些问题，曲线下面积也存在与 § 3.3 中讨论的不合规指标相同的问题：对于每一个分类问题，它总会导致评估错误的情况。

[19] Baker 和 Pinsky 2001以及 Lobo 等人 2008也提出了类似的观点。

不过，可以正确使用受试者工作特征曲线 $t(f)$。Metz 1978 年《成本效益分析》第 295 页对此作了解释；Sox等在 2013年§ 5.7.4 中也有说明（奇怪的是，Sox等还提到了曲线下面积这一通常错误的判据）。

用 B 表示测试集中类别 0（正类）的比例。混淆矩阵作为 f 的函数为

$$\begin{bmatrix} C_{00}(f) & C_{01}(f) \\ C_{10}(f) & C_{11}(f) \end{bmatrix} = \begin{bmatrix} B\,t(f) & (1-B)\,f \\ B\,[1-t(f)] & (1-B)\,(1-f) \end{bmatrix} . \tag{21}$$

因此，上式（20）中的求和可以通过重新排列部分项，明确写成

$$\sum_{i,j=0}^{1} U_{ij}\, C_{ij}(f) \equiv (U_{00}-U_{10})\, B\, t(f) - (U_{11}-U_{01})\,(1-B)\, f + U_{10}\, B \; + U_{11}\,(1-B)\,. \tag{22}$$

最大期望效用原则（20）等价于以下条件。该条件由上面的显式求和式推得，为简便起见，省略了第二行中的常数项：

$$\text{choose} \quad f^* = \arg\max_f \bigl\{(U_{00}-U_{10})\, B\, t(f) - (U_{11}-U_{01})\,(1-B)\, f\bigr\}\,. \tag{23}$$

大括号中的函数单调递增，因为 $t(f)$ 是（我们一如既往地假设，某个类别的正确分类效用高于误分类效用，因此 $U_{00}-U_{10} \geqslant 0$ 和 $U_{11}-U_{01} \geqslant 0$）。因此，令其导数为零即可求得其最大值：

$$\text{choose } f^* \text{ such that} \quad t'(f^*) = \frac{(U_{11}-U_{01})\,(1-B)}{(U_{00}-U_{10})\, B}\,. \tag{24}$$

如果我们有多个分类器，每个分类器都有自己的曲线 $t(f)$，那么最佳的就是与截距最高的直线相切的那个

$$t = \frac{(U_{11}-U_{01})\,(1-B)}{(U_{00}-U_{10})\, B} f + \text{const.} \tag{25}$$

具有最高截距的那条线。

从这一准则可以从几何角度看出，如果一个分类器的曲线 $t(f)$ 完全位于 另一个分类器的曲线上方，那么它的效用收益必定更高。但一般而言，如果两个分类器的曲线相交，就无法得出任何结论。重要的是受试者工作特征曲线的切线，而非其所围成的面积。图 8 展示了一个这样的例子。

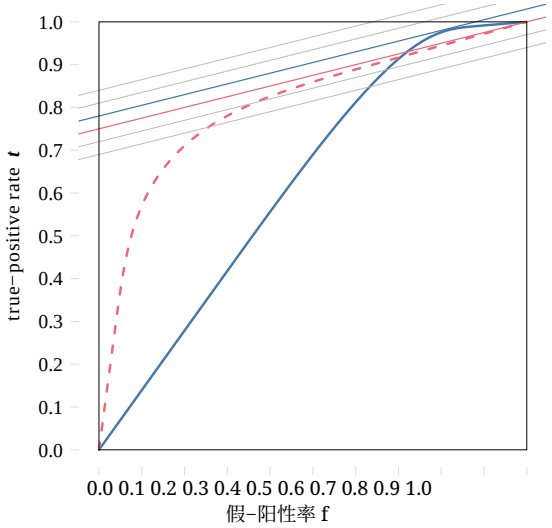


图8 两个分类器的受试者工作特征曲线。红色虚线曲线 所覆盖的面积显然大于 蓝色实线曲线。然而，后者曲线对应的分类器具有更高的效用，因为它在更高的位置与平行线族相切，见式（25）。当效用矩阵等于 $\left[\begin{smallmatrix}4&0\\0&1\end{smallmatrix}\right]$ 且测试集具有$B = 0$时，该示例便会出现。5（均衡），或者当效用矩阵等于 $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ 且测试集具有 $B = 0$时8。

6 总结与讨论

分类算法的评估与排名是其开发和部署过程中的关键阶段。没有这样的评估，我们甚至无法判断某个算法是否优于另一个算法，也无法判断某个算法的一组参数值是否优于另一组。

然而，目前我们拥有的不是评估理论，而只是评估民间经验：各种方法仅凭直觉和对特殊案例的分析提出，判断应采用哪种方法的标准模糊不清，也缺乏严谨的理论基础来保证独特性、普适性以及不存在偏差。我们认为，机器学习在实际应用中[20] 出现的一些令人惊讶的失败，不仅源于选择测试数据集时存在的偏差，

[20] 例如参见 Varoquaux & Cheplygina2022。

以及其他类似偏差，也源于开发阶段使用了错误的评估指标。

在本文中，我们论证了，评估过程的理论基础可以在决策理论中找到。其主要概念和原则——效用及其最大化——都非常直观，正如导言中的故事所展示的那样（我们希望如此）。

以下是决策理论应用于分类器评估的主要结果：

- 评估指标必须取决于具体的分类问题。

- 此类度量由 n^2 个参数完全定义，这些参数称为效用，汇集在一个效用矩阵中； n 是类别数。两个参数是任意设定的，分别表示效用尺度的零点和计量单位。在二分类问题中，这意味着我们有一个二维的可能度量集合。

- 分类器在测试集上的得分由其效用收益直接给出：将效用矩阵与该分类器的混淆矩阵中对应元素相乘后求总和。它是混淆矩阵元素的一个简单线性表达式。

- 当我们不确定分类问题背后的效用时，或希望考察多个分类问题的平均性能时，也会使用通过取平均值获得的效用矩阵。

- 一些常用指标，如精确率、平衡准确率、Matthews 相关系数、Fowlkes-Mallows 指数、 F_1-measure 以及受试者工作特征曲线下面积，并不符合决策理论。因此，在每个分类问题中，即使所有效用和频率都得到了正确评估，它们也总会导致对分类器进行一些错误的比较评估；它们还可能受到认知偏差的影响。

- 即使效用矩阵中的效用评估有误，使用它平均而言仍比使用其他常用指标更少导致错误的比较评估。

我们认为，分类器的决策理论评估也有显著优势：

首先，它将“众多得分中我该依赖哪一个？”这一模糊问题，转化为一个结构更清晰、因而更容易应对的问题：至少以半定量的方式评估，某个类别的正确分类比错误分类有价值得多、更加理想或有用；与另一个类别的正确分类相比又如何，等等。与其他指标相比，这类效用通常有更直接、依赖具体问题的解释。

其次，这会导出一种数学上简单、计算上便利的度量：混淆矩阵元素的线性组合——无需使用非线性函数或对曲线进行积分。

第三，当我们面对新的特殊问题时，基础理论的原则会为我们指引方向。举例来说，设想这样一个分类问题：我们无法笼统地判断真正例是否比真负例更重要，等等，因为这种价值判断可能因受测项目而异。在这种情况下，决策理论要求逐项评估效用，并仍然为每个项目提供一个得分；将这些得分汇总起来，即可得到候选分类器性能的总体评估得分。

本研究的理论、论述和结果不仅适用于分类，也可推广至回归以及更复杂的类分类问题，例如图像分割；这些问题在医学领域有重要应用[21]。值得探究的是，后者领域中常用的指标（如 Dice 系数[22]和豪斯多夫距离[23]）是否符合决策理论原则；如果不符合，又可采用哪些备选项。

在一项配套研究中[24]，我们将本文提出的一般性思路用于提升机器学习分类器的性能。

作者贡献

所有作者对本研究的贡献均等。

[21] Lundervold & Lundervold 2019年。 [22] Dice 1945；Fleiss 1975；Zijdenbos 等1994年。 [23] Alt&Guibas2000年。 [24] Dyrland 等2022。

致谢

KD 和 ASL 感谢 Trond Mohn研究基金会（资助编号：BFS2018TMT07）的支持，PGLPM 感谢挪威研究委员会（资助编号：294594）的支持。

KD 感谢家人无尽的支持；感谢伴侣 Synne 始终如一的爱、支持与鼓励；也感谢 Python、FastAi、PyTorch、scikit-learn、NumPy 和 RDKit 的开发者与维护者提供免费开源软件，并让这些实验得以实现。

PGLPM 感谢 Iván Davidovich 对本作品先前草稿提出的宝贵意见；感谢 Maja、Mari、Miri、Emma 一直以来的鼓励与关爱；感谢 Buster Keaton 和 Saitama 让生活充满敬畏与灵感；也感谢 LaTeX、Emacs、AUCTEX、开放科学框架、R、Python、Inkscape、LibreOffice、Sci-Hub 的开发者与维护者，让自由、公正的科学交流成为可能。

参考文献

（“de *X*”列在 D 项下，“van *X*”列在 V 项下，依此类推，无论各国惯例如何 .
) Alt, H., Guibas, L. J. (2000)：离散几何形状：匹配、插值与近似。载于：Sack, Urrutia (2000)：第3章：121–153。doi:10.1016/B978-044482537-7/50004-8. Baker, S. G., Pinsky, P. F. (2001)：一项比较用于癌症筛查的数字乳腺摄影与模拟乳腺摄影特殊受试者工作特征方法的设计与分析方案。《美国统计学会杂志》 **96**^{454}, 421–428。doi:10.1198/016214501753168136. Berger, J. O. (1985)：《统计决策理论与贝叶斯分析》，第2版（Springer，纽约）。doi:10.1007/978-1-4757-4286-2。初版于1980年。Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. (2010)：平衡准确率及其后验分布。国际模式识别会议论文集20，3121–3124。doi:10.1109/ICPR.2010.764. Camerer, C. F., Kunreuther, H. (1989)：低概率事件的决策过程：政策影响。《政策分析与管理杂志》 **8**^{4}, 565–592。doi:10.2307/3325045. Cheeseman, P. (1988)：探究计算机理解。《计算智能》 **4**^{2}, 58–66。doi:10.1111/j.1467-8640.1988.tb00091.x. Cox, R. T. (1946)：概率、频率与合理期望。《美国物理学杂志》 **14**^{1}, 1–13。doi:10.1119/1.1990764. Dice, L. R. (
1945 **26**^{3},):
297 302 10.2307/1932409
衡量物种间生态关联程度的指标。《生态学》–。doi: 。Dyrland, K., Lundervold, A. S., Porta Mana, P. G. L. (2022)：面向机器学习分类器的概率变换器与决策理论增强。开放科学框架 doi:10.31219/osf.io/vct9y. Fenton, N., Neil, M. (2019)：《基于贝叶斯网络的风险评估与决策分析》，第2版（CRC出版社，博卡拉顿，美国）。doi:10.1201/b21982。初版于2013年。

Fine, T. L. (1973)：《概率理论：基础考察》（学术出版社，纽约）。doi: 10.1016/C2013-0-10655-1. Fisher, R. A. (1963)：《研究人员统计方法》，修订第13版（Hafner，纽约）。初版于1925年。Fleiss, J. L. (1975)：衡量两位评判者对某项特征存在与否的判断一致性。《生物统计学》 **31**[3], 651–659。doi:10.2307/2529549. Fowlkes, E. B., Mallows, C. L. (1983)：比较两种层次聚类的方法。《美国统计学会杂志》 **78**[383], 553–569。doi:10.1080/01621459.1983.10478008. Gilovich, T., Griffin, D., Kahneman, D., 编 (2009)：《启发式与偏差：直觉判断心理学》，第 8 次印刷（剑桥大学出版社，剑桥，美国）。doi: 10.1017/CBO9780511808098. 初版于2002年。Good, I. J., Toulmin, G. H. (1968)：编码定理与证据权重。《IMA应用数学杂志》 **4**[1], 94–105。doi: 10.1093/imamat/4.1.94. Halpern, J. Y. (1999)：重新审视Cox定理。《人工智能研究杂志》 11，429–435。doi: 10.1613/jair.644. 另见Snow (1998)。Hand, D., Christen, P. (2018)：关于使用F度量评估记录链接算法的说明。《统计计算》 **28**[3], 539–547。doi: 10.1007/s11222-017-9746-6. Howard, R. A. (1980)：论生死决策。载于：Schwing, Albers (1980)：89–113。附有讨论。doi:10.1007/978-1-4899-0445-4_5. 重印于 Howard, Matheson (1984)，第481–506页。Howard, R. A., Matheson, J. E., 编 (1984)：《决策分析原理与应用读本》。第II卷：专业文集。（战略决策集团，门洛帕克，美国）。— (2005)：影响图。《决策分析》 **2**[3], 127–143。doi:10.1287/deca.1050.0020. 初刊于 Howard, Matheson (1984)，第719–762页。Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., Glasziou, P. P. (2014)：《卫生与医学决策：整合证据与价值观》，第 2 版 10.1017/CBO9781139506779 2001（剑桥大学出版社，剑桥）。doi: 。初版于。Jaynes, E. T. (2003)：《概率论：科学的逻辑》（剑桥大学出版社，剑桥）。G. Larry Bretthorst 编。初版于1994年。doi: 10.1017/CBO9780511790423, https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html. Jeffrey, R. C. (1965)：《决策逻辑》（麦格劳-希尔，纽约）。Jeni, L. A., Cohn, J. F., De La Torre, F. (2013)：面对不平衡数据：关于使用性能指标的建议。国际情感计算与智能交互会议论文集 **2013**, 245–251。doi: 10.1109/ACII.2013.47。Kahneman, D. (2011)：《思考，快与慢》（法勒、斯特劳斯和吉鲁出版社，纽约）。Kahneman, D., Slovic, P., Tversky, A., 编 (2008)：《不确定性下的判断：启发式与偏差》，第 24 次印刷 10.1017/CBO9780511809477 1982（剑桥大学出版社，剑桥）。doi: 。初版于。Kim, H., Markus, H. R. (1999)：偏离还是独特性，和谐还是从众？一项文化分析。《人格与社会心理学杂志》 **77**[4], 785–800。doi:10.1037/0022-3514.77.4.785. Kyburg Jr., H. E., Smokler, H. E., 编 (1980)：《主观概率研究》，第2版（Robert E. Krieger，亨廷顿，美国）。初版于1964年。Lindley, D. V. (1988)：《决策制定》，第2版（Wiley，伦敦）。初版于1971年。Lobo, J. M., Jiménez-Valverde, A., Real, R. (2008)：曲线下面积：一种会误导人的预测分布模型性能度量。《全球生态学与生物地理学》 **17**[2], 145–151。doi:

10.1111/j.1466-8238.2007.00358.x, https://www2.unil.ch/biomapper/Download/Lobo-GloEcoBioGeo-2007.pdf. Lundervold, A. S., Lundervold, A.（2019）：《聚焦于MRI的医学影像深度学习概述》。Z. Med. Phys. **29**[2], 102–127。doi:10.1016/j.zemedi.2018.11.002。Matthews, B. W.（1975）：《T4噬菌体溶菌酶预测与观测到的二级结构比较》。Biochim. Biophys. Acta **405**[2], 442–451。doi:10.1016/0005-2795(75)90109-9。Metz, C. E.（1978）：《ROC分析基本原理》。Semin. Nucl. Med. VIII[4], 283–298。doi:10.1016/S0001-2998(78)80014-2。Mittone, L., Savadori, L.（2009）：《稀缺偏差》。Appl. Psychol. **58**[3], 453–468。doi:10.1111/j.1464-0597.2009.00401.x。North, D. W.（1968 **4**[3], ）：《决200 210 10.1109/TSSC.1968.300114 https://stat.duke.edu/~scs/Courses/策理论教程导论》。IEEE Trans. Syst. Sci. Cybern. – 。doi: ,STAT102/DecisionTheoryTutorial.pdf。Raiffa, H.（1970）：《决策分析：不确定条件下选择问题导论讲义》，第2次印刷版（Addison-Wesley，雷丁，美国）。初版于1968年。Raiffa, H., Schlaifer, R.（2000）：《应用统计决策理论》，重印版（Wiley，纽约）。初版于1961年。Ramsey, F. P.（1926）：《真理与概率》。载于：Ramsey（1950）：第VII章：156–198。重印于Kyburg, Smokler（1980），第23–52页。写于1926年。—（1950）：《数学基础：及其他逻辑论文集》（Routledge & Kegan Paul，伦敦）。doi:10.4324/9781315887814, https://archive.org/details/in.ernet.dli.2015.46352。R. B. Braithwaite编。初版于1931年。Russell, S. J., Norvig, P.（2022）：《人工智能：一种现代方法》，第4版全球版（Pearson，哈洛，英国）。http://aima.cs.berkeley.edu/global-index.html, https://archive.org/details/artificial-intelligence-a-modern-approach-4th-edition。初版于1995年。Sack, J.-R., Urrutia, J. 编（2000）：《计算几何手册》（Elsevier，阿姆斯特丹）。doi:10.1016/B978-0-444-82537-7.X5000-1。Savage, L. J.（1972）：《统计学基础》，第2次修订增补版（Dover，纽约）。初版于1954年。Schwing, R. C., Albers Jr., W. A. 编（1980）：《社会风险评估：多安全才算足够安全？》（Springer，纽约）。doi:10.1007/978-1-4899-0445-4。Self, M., Cheeseman, P. C.（1987）：《面向人工智能的贝叶斯预测》。载于：第三届人工智能中的不确定性会议论文集（uai'87），J. Lemmer, T. Levitt, L. Kanal编（AUAI Press，阿灵顿，美国）：61 69 10.48550/ – 。arXiv.1304.2717 arXiv重印版doi: 。Shannon, C. E.（1948 **27**[3, 4], ）：《通379 423, 623 656 https://archive.org/details/bstj27-3-379 https://archive.信的数学理论》。Bell Syst. Tech. J. – – 。,org/details/bstj27-4-623, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf。Snow, P.（1998）：《关于Cox定理在有限域上的正确性与合理性》。Comput. Intell. **14**[3], 452–459。doi:10.1111/0824-7935.00070。—（2001 **17**[1], ）：《从178 192 10.1111/0824-7935.00138 Cox的视角看可能性的合理性》。Comput. Intell. – 。doi: 。Sox, H. C., Higgins, M. C., Owens, D. K.（2013）：《医学决策》，第2版（Wiley，纽约）。doi:10.1002/9781118341544。初版于1988年。

Steele, K., Stefánsson, H. O.（2020）：决策理论。载于：斯坦福哲学百科全书，E. N. Zalta编（形而上学研究实验室，斯坦福）。https://plato.stanford.edu/archives/win2020/entries/decision-theory 2015 。初版于 。

Tribus, M.（1969）：理性描述、决策与设计。（Pergamon，纽约）。DOI: 10.1016/C2013-0-01558-7。van Rijsbergen, C. J.（ 1974 **30**[4], 365 373 10.1108/ ）：评估基础。eb026584
J. Doc. – 。DOI: 。Varoquaux, G., Cheplygina, V.（2022）：医学影像机器学习：方法论缺陷与未来建议。npj Digit. Med. **5**[1], 48 10.1038/s41746- 。DOI: 。022-00592-y
von Neumann, J., Morgenstern, O.（1955）：博弈论与经济行为，第3版，第 6 https://archive.org/details/in.ernet.dli.2015.215284 次印刷（普林斯顿大学出版社，普林斯顿）。。初版于 1944 。Woodward, P. M.（1964）：概率与信息论：及其在雷达中的应用，第2版（Pergamon，牛津）。DOI: 10.1016/C2013-0-05390-X。初版于1953年。
Zhu, Q.（2020）：马修斯相关系数（*MCC*）在不平衡数据集上的性能。Pattern Recognit. Lett. **136**，71–80。DOI: 10.1016/j.patrec.2020.03.030。Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., Palmer, A. C.（1994）：磁共振图像中白质病变的形态计量分析：方法与验证。IEEE Trans. Med. Imaging **13**[4], 716–724。DOI: 10.1109/42.363096。