中国科学: 数学 2024 年 第 54 卷 第 12 期: 2121 ∼ 2140
SCIENTIA SINICA Mathematica
《中国科学》杂志社
SCIENCE CHINA PRESS
综 述

共型预测方法的发展和应用

献给中山大学建校百年暨中山大学数学学科建设 100 周年

金璋, 王学钦*

中国科学技术大学管理学院国际金融研究院, 合肥 230026
E-mail: jz97@mail.ustc.edu.cn, wangxq20@ustc.edu.cn

收稿日期: 2024-03-27; 接受日期: 2024-09-03; 网络出版日期: 2024-09-14; * 通信作者
国家重点研发计划 (批准号: 2022YFA1003803) 和国家自然科学基金 (批准号: 12231017 和 72171216) 资助项目

摘要 共型预测通过将预测点扩展成预测集的方式来严谨量化预测的不确定性, 它以灵活的结构和严格的有限样本理论保证而著称, 可以简单而方便地嵌入几乎任何预测模型中, 并随着机器学习的飞速发展而在近年来获得了越来越多的关注. 本文综述共型预测相关的发展历程, 在回顾共型预测的基础算法和理论同时, 也对共型预测无处不在的应用场景进行了介绍.

关键词 共型预测 机器学习 预测集 不确定性量化

MSC (2020) 主题分类 62J99

1 引言

共型预测 (conformal prediction) 或共型推断 (conformal inference) 是一种形式非常灵活、能应用于各种场景和模型下的统计预测工具. 共型预测最早在 *Algorithmic Learning in a Random World* 一书中提出[88]. 这里的 conformal 和复分析中的共形变换 (conformal mapping) 中的 conformal 毫无关系, 用于强调待预测的样本和已有的训练集中的样本之间的相似性, 把那些因为相似而合理的预测结果囊括到预测集中, 而把不相似的预测结果排除在外.

共型预测在一开始是被作为直推推理方法 (transduction) 的一种而提出的[38]. 所谓直推推理, 是区别于归纳推理 (induction) 和演绎推理 (deduction) 的一个统计学习中的概念. 归纳推理是 “从特殊到一般”, 演绎推理是 “从一般到特殊”, 这在统计学习中体现在先从训练集中用归纳推理总结出一般规律, 再将学习到的一般规律通过演绎推理应用到新样本或新问题中. 作为对比, 直推推理是 “从特殊到特殊”, 即直接从训练集中得到新问题的答案. 在应用直推推理中, 不再需要将解决更普遍的问题作为中间步骤. 显然, 直推推理对样本分布的假设和模型的要求要宽松许多, 但其缺点是当新

英文引用格式: Jin Z, Wang X Q. Development and application of conformal prediction (in Chinese). Sci Sin Math, 2024, 54: 2121–2140, doi: 10.1360/SSM-2024-0086

问题变更时, 不能像归纳 - 演绎推理那样有现成的解法可供使用, 而是需要将算法从头再运行一遍. 除了共型预测之外, 其他一些思想类似且同样以模型无关与样本分布无关作为优点的直推推理方法包括 Venn 预测[90] 与 Venn-Abers 预测[89] 等, 但这些方法所针对的问题的形式较为固定, 适用范围较为狭窄, 例如 Venn-Abers 预测方法主要只用于二元分类预测问题的校准, 因此具有一定的局限性, 远没有共型预测方法的应用广泛.

在如今的大模型时代, 有众多表现优异却暂时缺乏严格理论支撑和保障的预测算法和工具. 很多预测会输出一个点估计的结果, 一个自然而然的问题就是如何去量化预测的不确定性, 也就是去质询所得预测的可信度有多高. 统计学中常规的做法包括给数据添加分布假设, 或者使用一些性质被充分研究的模型来获得预测区间, 如线性回归模型. 但如果不对数据施加分布假设, 或者采用一些难以直接构造预测区间、但点估计却更为精准的先进方法, 那么共型预测就可以派上用场. 简而言之, 共型预测是一种可以将任何模型的预测结果转化为具有严格概率保障的预测集的方法. “任何模型”在以往指的是最近邻方法、岭回归、支持向量机等[77]. 而到了现在, 这些模型包括了目前先进和复杂的机器学习和深度学习方法, 共型预测也随之在自然语言处理、计算机视觉、深度强化学习等领域有了越来越多的应用. 可以预见, 在模型变得更加复杂的未来, 对模型的显式剖析会变得更加困难, 而同时共型预测也仍然会因其简单的形式和严格的概率保障而经久不衰.

本文的余下部分的内容安排如下: 第 2 节介绍共型预测的基础算法; 第 3 节介绍共型预测的评估标准与算法框架的改进和拓展; 第 4 节介绍共型预测的应用场景; 第 5 节给出结论和展望.

2 基础算法

考虑训练集含有独立同分布的样本 $Z_1, Z_2, \ldots, Z_n$, 其中每个 $Z_i = (X_i, Y_i)$ 由协变量 $X_i \in \mathbb{R}^p$ 和研究者所关心的响应变量 Y_i 构成. 这里的响应变量可以是离散的或者连续的, 分别对应着分类和回归的预测问题. 假设样本的分布为 P, 而新样本

$$Z_{n+1} = (X_{n+1}, Y_{n+1})$$

也服从分布 P, 记为

$$Z_{n+1} \sim P.$$

这里, 研究者只能观测到新样本的协变量 X_{n+1}, 但新样本的响应变量 Y_{n+1} 是需要推断的未知变量. 共型预测的目标就是对于一个名义的未覆盖率 α, 寻找一个预测集 $C^\alpha(X_{n+1})$, 使得

$$\mathbb{P}(Y_{n+1} \in C^\alpha(X_{n+1})) \geqslant 1 - \alpha. \tag{2.1}$$

当响应变量是连续变量的时候, 这里的预测集 $C^\alpha(X_{n+1})$ 通常是预测区间的形式.

共型预测作为一种输出预测集的校准方法, 需要先有一个点预测的模型作为基础. 在实际应用中, 常用的分割共型预测 (split conformal prediction) 方法的做法是先将原来的训练集, 即有标签的数据 (响应变量已知的数据) 随机分成两部分, 分别称之为新的训练集 (training set) 和校准集 (calibration set). 取新的训练集中的数据先训练或拟合出一个点预测模型, 再将这个点预测模型用在新的校准集的数据上, 执行共型预测的算法, 从而获得预测集. 不妨假设在执行共型预测算法之前

已经获得了一个点预测模型 $\hat{\mu}$, 这个模型输入 X_i 可以输出 Y_i 的一个点估计, 即 $\hat{\mu}(X_i)=\hat{Y}_i$. 之后, 需要定义一个**不一致得分 (nonconformity score) 函数** $R(x,y)\in\mathbb{R}$ 用来衡量任意一组协变量和响应变量的组合在该点预测模型下的异常程度. 直观上, $R(x,y)$ 越大, 说明 (x,y) 出现在由这个点预测模型所描述的系统中的概率越低. 不一致得分函数的定义非常灵活, 以回归问题为例, $y\in\mathbb{R}$, 此时, 不一致得分函数可以取响应变量的实际值与预测值的差的绝对值, 即

$$R(x,y)=|y-\hat{\mu}(x)|.$$

在定义好不一致得分函数后, 只需将校准集中的所有数据依次计算不一致得分, 得到

$$R_1=R(X_1,Y_1),\quad \ldots,\quad R_n=R(X_n,Y_n).$$

下一步, 求出 $R_1,\ldots,R_n$ 这 n 个量的 $\frac{\lceil(n+1)(1-\alpha)\rceil}{n}$ 分位数, 记为 $\hat{q}$, 这里 $\lceil a\rceil$ 为上取整函数, 表示不小于 a 的整数中最小的一个. 从直觉上来看, 如果有一组新的协变量和响应变量 (X_{n+1},y) 使得 $R(X_{n+1},y)>\hat{q}$, 那么说明这样的组合比校准集中大约 $(1-\alpha)$ 的数据都要更加 "奇怪". 从共型预测的字面意思理解, y 从 "形状" 上来说与 X_{n+1} 的输出 "不一致", 因此这样的 y 会被排除在预测集之外. 而在预测集中留下的, 则是那些使得 $R(X_{n+1},y)<\hat{q}$ 的更为 "正常" 的 y. 从假设检验的角度来说, 共型预测考虑的原假设是新样本和校准集来自同一分布, 而在原假设下, 不一致得分的秩服从离散均匀分布. 不一致得分函数非常重要, 因为它里面包含了非常多的信息, 包括之前提到的点预测模型. 总地来说, 共型预测的基础算法可以总结为如下的算法 1:

算法 1 共型预测算法

输入: 不一致得分函数 $R(x,y)$, 校准集 $(X_i,Y_i), i=1,\ldots,n$, 新样本的协变量 X_{n+1}, 未覆盖率 α

输出: 新样本响应变量的预测集 $C^{\alpha}(X_{n+1})$

1: **for** $i=1,\ldots,n$ **do**

2: $\quad R_i=R(X_i,Y_i)$

3: **end for**

4: $\hat{q}=\{R_1,\ldots,R_n\}$ 中的 $\frac{\lceil(n+1)(1-\alpha)\rceil}{n}$ 分位数

5: $C^{\alpha}(X_{n+1})=\{y: R(X_{n+1},y)<\hat{q}\}$

通过共型预测得到的预测集满足 (2.1) 式中的覆盖率保证, 即

定理 2.1 假设 $(X_1,Y_1),\ldots,(X_{n+1},Y_{n+1})$ 是独立同分布的, 那么, 对于算法 1 得到的预测集 $C^{\alpha}(X_{n+1})$, 有以下式子成立

$$\mathbb{P}(Y_{n+1}\in C^{\alpha}(X_{n+1}))\geqslant 1-\alpha.$$

这里的覆盖率是有限样本下的覆盖率, 是非渐近的结果, 即并不要求 $n\to\infty$ 即可满足. 另外, 值得注意的一点是, 在共型预测的算法中, 不一致得分函数具体的值并不重要, 重要的是它的秩. 因此, 共型预测具有 (不一致得分函数) 的单调递增变换不变性: 也就是说, 在不一致得分函数外再嵌套一个单调递增的函数, 得到的共型预测结果还是会完全相同的.

共型预测框架的灵活性极强, 这体现在它对数据分布几乎没有任何要求, 仅仅要求校准集和新数据来自同一未知分布, 就能巧妙地利用不一致得分函数对样本们的可交换性来得到有限样本下的严格覆盖率. 理论上, 任意奇怪的样本分布、任意错误的点预测模型、任意无信息的不一致得分函数最后都能在共型预测的框架下获得具有严格覆盖率保障的预测集. 当然, 这里的严格覆盖率是有代价

的. 如果因为缺乏对样本分布的了解或者使用了过度错误的点预测模型而在共型预测算法中使用了欠佳的不一致得分函数, 那么得到的预测集可能会非常巨大. 例如, 在分类问题中, 得到的预测集可能包含了所有的类别. 这样的预测集即使有着有限样本严格覆盖率, 也是毫无意义的. 选择好的不一致得分函数可以获得更小、统计意义上更有效率的预测集. 文献 [6, 36, 77] 中介绍了许多在不同场景下的不一致得分函数的取法. 文献 [51] 评估了不同不一致得分函数对共型预测结果的影响, 文献 [4] 考虑了对不一致得分函数用分位数回归森林方法重加权以改进共型预测效果的问题. 最近, 文献 [43] 提出了基于嵌套集序列构造的嵌套共型预测算法, 提供了另一个统一各种不一致得分函数的框架.

3 共型预测算法的推广

3.1 条件覆盖率和自适应性

(2.1) 式给出了共型预测集的严格覆盖率下界. 但考虑到当预测集取全集或接近于全集的足够大的集合时, 覆盖率的下界总可以充分大, 此时的预测集很可能因为过于保守而失去实用价值. 为了使覆盖率随着样本数量的增加而精准地逼近 $1-\alpha$, 覆盖率的上界也需要被给出. 如果假设不一致得分函数 $R(x,y)$ 是连续的, 或者说 $R(x,y)$ 几乎处处不同, 共型预测算法得到的预测集[54, 88] 满足

$$1-\alpha \leqslant \mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1})) \leqslant 1-\alpha+\frac{1}{n+1}. \tag{3.1}$$

需要说明的是, 这里所提到的覆盖率实际上是边缘覆盖率 (marginal coverage rate). 对于一个预测集来说, 边缘覆盖率的精确并不一定说明这个预测集性质优良. 例如, 考虑这样的一个预测集, 无论输入的 X_{n+1} 是什么, 都以 $(1-\alpha)$ 的概率输出全集, 以 α 的概率输出空集. 显然, 这样的一个预测集的边缘覆盖率必然是精确的 $(1-\alpha)$, 但同时它也是全然无用的. 再考虑另一个例子, 假设现在有一个图像识别分类问题, 训练集和新的样本点都是猫或狗的图片, 响应变量的取值只有可能是 “猫” 或者 “狗”. 在这种情况下, 预测集的可能取值一共有四种, 空集、{猫}、{狗} 或者 {猫、狗}. 在绝大多数共型预测方法中, 对于常规的未覆盖率 α 的指定 (例如 $\alpha=0.1$), 空集几乎不会出现, 而全集 {猫、狗} 则很可能大量出现, 尤其是对于那些有一定辨别难度的图片. 注意到, 由于所考虑的问题是两分类问题, 所有样本点非猫即狗, 因此预测集 {猫、狗} 是必然能覆盖正确的响应变量的. 由于这里的 {猫、狗} 输出被计算到正确的边缘覆盖率中, 真正具有信息量的 {猫} 和 {狗} 的输出所对应的覆盖率很可能会远小于标定的边缘覆盖率 $(1-\alpha)$[49].

为了克服边缘覆盖率所带来的种种局限性, 研究者们提出了**条件覆盖率**的概念[85]. 一个预测集 $C^{\alpha}(X_{n+1})$ 满足条件覆盖是指

$$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1}) \mid X_{n+1}) \geqslant 1-\alpha, \tag{3.2}$$

即, 对于任何一个可能的输入 X_{n+1}, 都要求输出的预测集能有 $(1-\alpha)$ 以上的覆盖率. 直觉上看, 条件覆盖的性质要求预测集犯错误而未覆盖的概率在协变量空间上是均匀的. 虽然看起来边缘覆盖和条件覆盖的描述相似, 但实现难度可以说天差地别. 边缘覆盖是所有共型预测集必然能满足的基础要求, 但严格的条件覆盖却不可能在样本分布无关的假设下实现. 例如, 在 $Y \in \mathbb{R}$ 的回归问题下, 如果不对样本分布施加任何假设, 那么在有限样本下, 满足严格条件覆盖率的预测集几乎必然有着无限测

度[37,55,85]. 这是因为当样本分布相当奇怪时, 任意有限的训练集都无法对某个特定的未包含在训练集中的新协变量 x_{n+1} 对应的响应变量提供有效的信息, 当对此特定的协变量的预测集的测度有限时, 总能找到该预测集外的点, 将其与 x_{n+1} 匹配, 用插值的方法构造样本分布. 因此, 在完全不假设数据分布的条件下, 严格满足条件覆盖率的预测集几乎必然是全集, 而这是毫无信息的.

在严格的条件覆盖率无法被满足时, 研究者们倾向于用各种方法来近似条件覆盖率. 其中一种方法是寻求边缘覆盖率和条件覆盖率的折衷, 以保证当协变量属于某些子集的时候, 预测集的覆盖率可以达到预设值[37,42,52], 即

$$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1}) \mid X_{n+1} \in \mathcal{R}(x)) \geqslant 1-\alpha, \tag{3.3}$$

这里的 $\mathcal{R}(x)$ 代表协变量空间上的某些子集类. 例如, 当 $X \in \mathbb{R}^p$ 时, 文献 [85] 考虑事先把协变量空间划分为 K 个互不相交的子集, 即

$$\mathbb{R}^p = \mathscr{X}_1 \cup \cdots \cup \mathscr{X}_K,$$

然后再提出在每个子集上都能保证覆盖率的共型预测算法, 即要求

$$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1}) \mid X_{n+1} \in \mathscr{X}_k) \geqslant 1-\alpha \tag{3.4}$$

对于 $k=1,\ldots,K$ 成立. 另一种取 $\mathcal{R}(x)$ 的方法是取其为 $\mathbb{R}^p$ 上所有概率测度大于 δ 的 ℓ_2 球 $\mathbb{B}(x,r)$ 构成的子集类, 即要求

$$P_X(\mathbb{B}(x,r)) \geqslant \delta.$$

此时, 上述的覆盖率被称为 $(1-\alpha,\delta)$ **受限条件覆盖率**[37]. 在原来的条件覆盖要求下, 任何一个新样本所对应的预测集都必须保证 $(1-\alpha)$ 的覆盖率, 而在放松过后的受限条件覆盖率中, 则只要求任何一个新样本的局部附近 δ 比例的所有样本点共同满足 $(1-\alpha)$ 的边缘覆盖率. 可以发现, 受限条件覆盖率的要求位于边缘覆盖率和条件覆盖率之间, 且有着明确的实际意义, 例如当样本为病人时, 考虑同时满足不同性别、年龄段、人种的覆盖率就可以改写为受限条件覆盖率的要求. 此外, 还有一种考虑逼近条件覆盖率的方法是给样本分布施加一些光滑性条件, 然后在样本量趋于无穷的情况下通过 $Y \mid X$ 条件分布的局部光滑估计量来渐近地实现条件覆盖[20,27,47,55]. 除了受限条件覆盖之外, 还有一些其他的性质介于边缘覆盖和条件覆盖之间, 例如文献 [25] 中提到的二类有效与强二类有效等.

评估一个预测集的好坏或者有用程度的最重要指标往往并不是预测集的平均大小. 因为如果只以预测集的大小来作为优化目标, 可能会让预测集出现这样的现象: 为了保证边缘覆盖率的成立, 算法会致力于对 “简单” 的样本预测并输出较小的预测集, 却直接放弃掉 “困难” 的任务, 并输出空集. 一个真正有用的预测集算法应该具有这样的特征: 它对简单的样本点输出较小的预测集, 对困难的样本点输出较大的预测集, 输出的预测集的大小可以忠实地反映共型预测背后的点预测算法在这个

表 1 三种覆盖率的定义与特点比较

概念	形式	特点
边缘覆盖率	$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1})) \geqslant 1-\alpha$	容易满足, 但实际意义有限
条件覆盖率	$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1}) \mid X_{n+1}=x) \geqslant 1-\alpha$	性质极好, 但难以构造
受限条件覆盖率	$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1}) \mid X_{n+1} \in \mathbb{B}(x,r)) \geqslant 1-\alpha$	实现难度介于前两者之间, 且具有实际意义

样本点做出的预测的不确定性大小. 换言之, 当 $\alpha = 0.1$ 时, 一个良好的预测集应该在简单和困难的任务中都达到 90% 左右的覆盖率, 而不是在简单的任务中有 99% 的覆盖率却同时只在困难的任务有 10% 的覆盖率, 用这种方式来满足总体的 90% 边缘覆盖率. 这样的优良性质被称为预测集的**自适应性 (adaptivity)**. 可以看出, 自适应性和条件覆盖率有着密切的联系, 它要求预测集不止满足边缘覆盖率, 也同时要在不同任务的条件覆盖率上尽可能地均匀.

有一些度量指标可以用来评估一个预测集的条件覆盖率或自适应性. 例如, 在分类问题中, 可以根据预测集的大小将样本划分为 G 类, 然后对这 G 类分别计算覆盖率, 并取其中的最小值作为衡量该预测集自适应性的指标. 这样的指标被称为预测集大小分层覆盖率度量[5]. 具体而言, 假设共有 m 个新样本 $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_m, \tilde{Y}_m)$ 用来评估某个预测集的条件覆盖率, 所得到的预测集 $C^\alpha(\tilde{X}_1), \ldots, C^\alpha(\tilde{X}_m)$ 按照大小分成了 G 类, 且记 $\mathcal{I}_g \subset \{1, \ldots, m\}$ 为被分到第 g 类的新样本的下标所构成的集合. 那么, 此时的预测集大小分层覆盖率度量即为

$$\min_{g \in \{1,\ldots,G\}} \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \mathbb{I}\{\tilde{Y}_i \in C^\alpha(\tilde{X}_i)\}.$$

如果预测集在条件覆盖率上表现得比较好, 研究者们会期待在经过预测集大小的分层之后, 得到的每个覆盖率都在 $1-\alpha$ 附近, 其中的最小值也不会比 $1-\alpha$ 低太多. 反之, 如果得到的指标远低于 $1-\alpha$, 那么就说明这样的预测集在自适应性上表现欠佳, 不能很好地满足条件覆盖率的要求. 除了预测集大小之外, 类似的分层方法也可以采用不同的依据, 例如对协变量的某些特征进行分层, 这样的指标被称为协变量特征分层覆盖率度量[6]. 预测集大小分层覆盖率度量和协变量特征分层覆盖率度量都有着明确的现实意义, 前者侧重于度量预测集在 “简单” 和 “困难” 的任务中能否一视同仁地实现条件覆盖, 而后者侧重于衡量预测集对于各种类别的输入 (例如不同的人种、性别、年龄区间的患者) 能否实现条件覆盖. 此外, 文献 [24,34] 也提出了一些其他用来度量条件覆盖率的指标.

在共型预测算法的框架下, 对于各式各样的实际问题, 可以通过调整不一致得分函数的定义来改进预测集的性质, 得到更具有自适应性的预测集. 接下来将通过介绍一个分类问题和一个回归问题的例子来结束这一节. 对于 K 分类问题, 每个输入 X 都对应 $\{1, \ldots, K\}$ 中的一类. 考虑采用的预测器 $\hat{f}$ 使用了 softmax 输出, 即输出一个归一化的 K 维向量, 而点预测的结果就是该向量最大值对应的类别. 具体而言,

$$\hat{f}(x) = (\hat{f}_1(x), \ldots, \hat{f}_K(x)) \in [0,1]^K, \quad \sum_{i=1}^{K} \hat{f}_i(x) = 1.$$

在这样的设定下, 对于任何协变量 x 和类别 y, 一个最自然的不一致得分函数的定义方式是将不一致得分函数定义为预测集把协变量 x 的输入识别为 y 标签以外的概率, 即

$$R(x, y) = 1 - \hat{f}_y(x).$$

这样的不一致得分函数合理且符合直觉, 得到的预测集也总体较小, 但却不具有自适应性[6]. 这是因为当遇到一些比较 “困难” 而难以分辨的任务时, softmax 输出得到的向量可能会在各维度上比较均匀, 缺少一个值特别大的类别, 这也就导致了各类别所对应的不一致得分函数都比较大. 如果所有类别的不一致得分函数都超过了共型预测框架中的 $\hat{q}$, 那么预测集就将输出空集, 也即放弃了这个困难的任务. 然而, 自适应性要求预测集对困难的任务输出更大的集合, 来尽量满足预测集的边缘覆盖率. 显然, 上述的不一致得分函数的定义是与自适应性背道而驰的.

一种通过修改不一致得分函数的定义而让上述问题的预测集的自适应性大大改善的方法被称为自适应预测集 (adaptive prediction sets, APS) 方法[5,6,74]. 具体而言, 在得到一个 softmax 输出 $\hat{f}(x)$ 后, 该方法会先将该向量中的 K 个值从大到小进行排序, 得到

$$(\hat{f}_{\pi_1(x)}(x),\dots,\hat{f}_{\pi_K(x)}(x)),$$

其中

$$\hat{f}_{\pi_1(x)}(x)\geqslant\cdots\geqslant\hat{f}_{\pi_K(x)}(x).$$

然后, 对于任意类别 $y=\pi_j(x)$, 定义不一致得分函数

$$R(x,y)=\sum_{i=1}^{j}\hat{f}_{\pi_i(x)}(x),$$

即将预测出的各分类概率由高到低排列, 从最高的概率开始一直累加到该类别的概率为止. 假如 (x,y) 在该模型下是不合理的数据, 那么类别 y 在输出结果 $\hat{f}(x)$ 中的概率应该很低, 排位靠后, 当累加到 y 时, 对应的不一致得分函数 $R(x,y)$ 已经非常接近 1 了, 这也与不一致得分函数的定义相符合. 在共型预测算法的分位数 $\hat{q}$ 计算完毕后, 对于新的样本 x, 为了避免输出的预测集是空集, 略微修改预测集为

$$C^{\alpha}(x)=\{\pi_1(x),\dots,\pi_k(x)\},\quad k=\sup\left\{k':\sum_{j=1}^{k'}\hat{f}_{\pi_j(x)}(x)<\hat{q}\right\}+1,$$

也就是说, 该算法选择的预测集就是从 softmax 输出结果值最大的类别开始, 一直包括到累加到值刚好超过 $\hat{q}$ 的类别为止. 值得注意的是, 对于 “简单” 的任务, 正确答案很可能是 softmax 输出结果值最大的类别, 而且这个值会比较大, 可能会直接超过 $\hat{q}$, 此时预测集就只包含了一个类别. 而对于 “困难” 的任务, softmax 输出结果可能比较均匀, 需要好多个类别的累加才能超过 $\hat{q}$, 对应的预测集就会更大. 这样的特性很好地契合了自适应性的要求.

除了修改不一致得分函数的定义, 也可以通过直接改变共型预测框架背后的点预测模型方法来改进预测集的自适应性. 本节将举一个 $y\in\mathbb{R}$ 的回归问题的例子来说明这一点. 在回归问题中, 如果取点预测模型 $\hat{\mu}$ 为任意回归模型, 例如线性回归、岭回归或样条回归、局部多项式回归等非参数回归模型, 则可以将不一致得分函数取成 $R(x,y)=|y-\hat{\mu}(x)|$. 此时的不一致得分函数虽然符合直觉, 但在共型预测算法框架计算得到 $\hat{q}$ 后, 输出的预测集 (预测区间) 是

$$C^{\alpha}(x)=[\hat{\mu}(x)-\hat{q},\hat{\mu}(x)+\hat{q}].$$

尽管此时的预测集可以满足 $(1-\alpha)$ 的边缘覆盖率, 但对于不同的输入 x, 预测区间的宽度都相等, 不具有自适应性. 研究者们希望预测区间能具有自适应性, 在响应变量变化平缓的地方有着较窄的宽度, 在响应变量变化陡峭的地方有着较宽的宽度, 使得预测区间在不同的分段处都能尽量满足条件覆盖率. 在这样的动机下, 共型分位数回归预测方法应运而生, 并且被广泛使用[3,73,76]. 共型分位数回归预测在选用点预测模型的时候, 直接采用分位数回归模型, 分别估计 $\alpha/2$ 和 $(1-\alpha/2)$ 分位数, 在共型预测框架的应用之前就先得到了一个未经校准的 $(1-\alpha)$ 预测区间

$$[\hat{\mu}_{\alpha/2}(x),\hat{\mu}_{1-\alpha/2}(x)].$$

之后, 在应用共型预测框架时, 该方法定义不一致得分函数为

$$R(x,y)=\max(\hat{\mu}_{\alpha/2}(x)-y, y-\hat{\mu}_{1-\alpha/2}(x)),$$

在评估校准集样本点的不一致得分函数时, 其衡量的就是这个未经校准的 $(1-\alpha)$ 预测区间需要如何调整才能恰好覆盖该样本点的响应变量. 如果样本点的 y 值被该区间覆盖, 那么此时的分位数估计区间过于保守, 应该收缩, $R(x,y)$ 取负值. 如果样本点的 y 值未被该区间覆盖, 那么此时的分位数估计区间太窄, 应该扩张, $R(x,y)$ 取正值. 在这样定义完不一致得分函数之后, 利用共型预测框架得到分位数 $\hat{q}$, 最后可以得到经过共型预测校准的预测集 (预测区间)

$$C^{\alpha}(x)=[\hat{\mu}_{\alpha/2}(x)-\hat{q}, \hat{\mu}_{1-\alpha/2}(x)+\hat{q}].$$

值得注意的是, 尽管在共型预测的这一步中, 对于不同的输入 x, 所进行的预测区间的校准宽度都是一样的, 但在点预测模型的那一步中, 对于不同的输入 x, 得到的未经校准的预测区间的宽度有所不同. 因此, 最终获得的的预测集的宽度在协变量变化时并不均匀, 这也说明了共型分位数回归预测有着更好的自适应性. 最近, 还有一些其他的工作通过修改共型预测框架而增强预测集的自适应性, 例如文献 [42] 提出了局部共型预测算法, 对离测试集样本在协变量意义上更近的校准集样本赋予更大的权重, 从而在边际覆盖的基础上得到了额外的局部覆盖保证. 另外, 虽然条件覆盖在共型预测的文献中通常指代如 (3.2) 中对测试数据的协变量取条件, 最近也有一些其他的工作考虑对分类问题中测试数据的类别取条件[30], 或者对训练数据集的分划取条件[19], 从而获得其他形式的条件覆盖率.

3.2 训练和校准

如第 2 节中所述, 共型预测实质上是一种校准方法, 将任意的黑箱预测模型的点预测输出包装为具有严格覆盖率的预测集的输出. 如果研究者在拿到一个数据集的时候已经有了一个训练好了的模型, 那么自然是最理想的情形. 这意味着可以将所有的数据全部用于共型预测中的校准步骤, 计算它们的不一致得分函数, 得出分位数, 进而得出预测集. 但这样的模型很少凭空出现. 绝大多数时候, 研究者必须 “自产自销”, 利用有限的数据, 同时进行训练和校准两项工作. 为了避免用同一个数据来进行预测和校正, 从而导致可能的严重过拟合问题, 最简单的方法就是将训练集随机分成数量大致相同的两部分, 用一部分来作为训练模型 $\hat{\mu}$, 另一部分来进行共型预测算法中的校准. 这样的方法在第 2 节中被提到过, 称为分割共型预测 (split conformal prediction), 是计算效率最高的共型预测方法. 分割共型预测的别名是归纳共型预测 (inductive conformal prediction), 因为当分割共型预测算法完成后, 可以直接适用于任意的新样本点上, 这正与本文第 1 节中所提到的归纳 - 演绎推理的逻辑框架相契合. 而与归纳共型预测相对的就是直推共型预测 (transductive conformal prediction), 因为对任意的新样本点需要重新运行算法而得名. 和分割共型预测相对应, 直推共型预测又名完全共型预测 (full conformal prediction). 虽然其算法结构比起分割共型预测要更为复杂, 但它出现时间却比分割共型预测时间更早. 研究者所关心的响应变量 Y_{n+1} 落在其取值空间的某一处, 那如果将 Y_{n+1} 遍历取值空间中的所有取值, 总能得到那个正确的值, 而此时得到的 (X_{n+1}, Y_{n+1}) 与训练集 $Z_1, Z_2, \ldots, Z_n$ 在分布上可交换. 因此, 完全共型预测的核心思想就是将 Y_{n+1} 遍历取值空间中的所有取值, 对每个特定的取值 $Y_{n+1}=y$ 都将其加入原来的训练集, 用总共 $n+1$ 个样本来训练模型, 然后计算与 y 相关的不一致得分函数与相应的分位数 $\hat{q}^{y}$. 总地来说, 完全共型预测的步骤可以总结如下:

算法 2 完全共型预测算法

输入: 不一致得分函数 R, 训练集 $(X_i, Y_i), i = 1, \ldots, n$, 新样本的协变量 X_{n+1}, 未覆盖率 α

输出: 新样本响应变量的预测集 $C^{\alpha}(X_{n+1})$

for y in $\mathcal{Y}$ **do**
  $Y_{i+1} = y$
  以扩充训练集 $(X_i, Y_i), i = 1, \ldots, n+1$ 训练预测模型 $\hat{\mu}^y$
  通过模型 $\hat{\mu}^y$ 定义不一致得分函数 R^y
  for $i = 1, \ldots, n$ **do**
    $R_i^y = R^y(X_i, Y_i)$
  end for
  $\hat{q}^y = \{R_1^y, \ldots, R_n^y\}$ 中的 $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ 分位数
end for
$C^{\alpha}(X_{n+1}) = \{y : R^y(X_{n+1}, y) < \hat{q}^y\}$

对于分类问题, 响应变量的取值空间 $\mathcal{Y}$ 是有限集, 其个数记为 $|\mathcal{Y}|$, 此时完全共型预测算法中的遍历可以做到, 但需要拟合 $|\mathcal{Y}|$ 个模型, 而作为对比, 分割共型预测算法只需要拟合一个模型. 而在回归问题中, $\mathcal{Y} = \mathbb{R}$, 此时如果模型不是像线性回归、岭回归那样有着显式的参数形式, 通常就只能通过离散采样的方式来完成对遍历取值空间的近似, 即将取值空间分划成有限个子集, 然后对响应变量在每一个子集的情况分别拟合模型. 总之, 完全共型预测的计算量要远大于分割共型预测. 当新样本发生变化的时候, 完全共型预测又要从头开始繁重的训练工作, 而分割共型预测得到的预测集仍然可以直接适用. 最近, 文献 [28, 53, 65, 66] 开发了一些减少完全共型预测的计算量, 加速完全共型预测的方法, 文献 [22] 考虑了共型预测的并行实现, 但完全共型预测仍是高度计算密集的算法.

完全共型预测算法虽然在计算效率上远不如分割共型预测算法, 但其在统计效率上相比分割共型预测算法则具有优势. 在预测模型 $\hat{\mu}$ 是对称模型 (即训练集样本的输入顺序与训练好的预测模型无关) 的前提下, 完全共型预测算法得到的预测集也可以达到 (2.1) 式中的边缘覆盖率, 即:

定理 3.1 假设 $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ 是独立同分布的, 且预测模型 $\hat{\mu}$ 是对称模型. 那么, 对于算法 2 得到的预测集 $C^{\alpha}(X_{n+1})$, 有以下式子成立:

$$\mathbb{P}(Y_{n+1} \in C^{\alpha}(X_{n+1})) \geqslant 1 - \alpha.$$

由于利用上了全部的训练集数据来训练模型, 且避免了引入样本分割步骤中额外的随机性, 完全共型预测算法得到的预测集一般有着更小的集合大小和更低的集合变异性 [6]. 考虑到分割共型预测算法得到的预测集可能受样本的一次性随机分割的影响很大, 文献 [79] 提出了多重分割共型预测算法, 通过多次分割共型预测算法来进一步校准预测集. 文献 [14] 所提出的去随机化分割共型预测算法也是基于同样的考量, 通过聚合对同一数据集多次分割得到的结果来提高分割共型预测算法的稳定性. 由于对同一数据集多次分割会不可避免地引入相关性, 为了更好地将多次分割的结果整合成一个预测集, 文献 [95] 提出了 Cauchy 聚合共型预测算法, 引入了用来整合具有任意相关结构的 p 值的 Cauchy 聚合方法, 使得到的共型预测区间更加有效. 完全共型预测和分割共型预测间的选择代表了统计效率和计算效率间的权衡. 然而, 分割共型预测在计算效率上必然远胜于完全共型预测, 而完全共型预测在统计效率上相比于分割共型预测却远没有那么明显, 有时候甚至还会被反超 [59, 60, 68], 这也解释了为什么现在分割共型预测的使用要远比完全共型预测广泛.

除了分割共型预测和完全共型预测以外, 还有许多介于两者中间的共型预测算法, 常见的包括交叉共型预测 (cross conformal prediction)[86] 与折刀共型预测 (jackknife conformal prediction)[54] 等. 交叉共型预测结合了交叉验证的思想, 在数据分割中将样本分成 K 折, 每次取其中一折作为校准集, 剩下 $(K-1)$ 折作为训练集, 如此重复 K 次之后即可以得到所有样本点的不一致得分函数. 折刀共型预测结合了留一法 (leave-one-out) 的思想, 在计算训练集中每个样本的不一致得分函数时都用训练集中另外 $(n-1)$ 个样本来拟合新的模型, 这样看似能比起分割共型预测利用上更多的样本点来训练模型, 但却不能像分割共型预测那样得到有效的样本外覆盖率, 只能保证弱得多的样本内覆盖率. 文献 [11] 在折刀共型预测的基础上开发了折刀 + 共型预测、折刀 - minmax 共型预测、不对称折刀共型预测等新的算法框架, 在详尽比较了上述这些基于不同样本分割方式而衍生出的不同共型预测算法的同时, 还成功建立了折刀共型预测的严格样本外边缘覆盖率, 填补了理论上的空白. 与分割共型预测能在无分布假设下保证 $(1-\alpha)$ 的边际覆盖率不同, 折刀 + 共型预测与交叉共型预测的无分布假设下的理论边际覆盖率只能达到 $(1-2\alpha)$, 但这两种方法在实验中的经验覆盖率仍能近似达到 $(1-\alpha)$[11].

3.3 分布的假设

共型预测能生成具有严格有限样本覆盖率保证的预测集, 其核心原因就在于校准集和测试集的独立同分布假设, 而这导致在原假设下的校准集和测试集的不一致得分函数服从均匀分布. 事实上, 为了满足这个性质, 定理 2.1 与 3.1 中的校准集和测试集的样本分布假设还能从独立同分布进一步放松, 即只需要满足可交换性 (exchangeability). 如果对于 $1,\ldots,n$ 的任意一个置换 $\sigma(1),\ldots,\sigma(n)$, 都有 $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ 和 $\{(X_{\sigma(1)},Y_{\sigma(1)}),\ldots,(X_{\sigma(n)},Y_{\sigma(n)})\}$ 的联合分布相同, 那么就称数据满足可交换性. 显然, 独立同分布包含了可交换性, 而可交换性却只要求变量同分布, 但不一定需要独立, 文献 [77] 给出了一些可交换但却不独立的变量的例子. 例如, 对三个独立同分布变量 X_1,X_2,X_3 进行两两平均, 得到三个新的变量

$$W_1=(X_1+X_2)/2,\quad W_2=(X_1+X_3)/2,\quad W_3=(X_2+X_3)/2,$$

那么此时的新变量 (W_1,W_2,W_3) 就是可交换却不独立的.

然而无论是独立同分布还是可交换性假设, 在实际数据处理中都经常会被违背. 考虑文献 [6] 提出的这样一个问题: 通过磁共振成像来预测疾病, 但协变量 X 除了磁共振成像之外还包括了患者的年龄. 在获得不一致得分函数的校准集中, 采用的是婴儿和成年人各占一半的数据, 但在进行实际预测的时候, 测试集中的数据却包含了 95% 的成人和仅仅 5% 的婴儿. 显然, 这里的校准集和测试集中的样本分布不可交换, 因此如果直接采用原始的共型预测算法, 得到的预测集将不具有有限样本覆盖率, 由于在校准集中的婴儿数据过多, 婴儿易发的疾病会被高估. 在处理这类问题时, 如果添加合理的假设, 即, 给定协变量时响应变量的条件分布在校准集和测试集之间保持不变, 那么共型预测算法可以通过简单的修改而重新获得有限样本覆盖率. 在这个实际问题中, 这样的假设是指当年龄和磁共振成像确定时, 患病可能的概率分布不变, 这在实际设定下是非常合理的. 这样的问题被称为协变量偏移[78], 考虑这样的设定

$$(X_1,Y_1),\ldots,(X_n,Y_n)\sim P_{Y|X}\times P_X,\quad (X_{n+1},Y_{n+1})\sim P_{Y|X}\times \tilde{P}_X,$$

协变量的边缘分布在校准集和测试集中不同, 但给定协变量后的响应变量条件分布在校准集和测试集中相同, 这被称为协变量偏移问题. 为了处理此问题, 文献 [83] 提出了加权共型预测算法, 对校准集中计算得到的不一致得分函数进行不等权重的加权, 对更有可能出现在测试集中的协变量 X_i 所对应的不一致得分函数赋予更大的权重, 并期待加权后的校准集数据能和测试集中的数据达成等效的可交换的效果. 具体而言, 定义

$$w(X_i) = \mathrm{d}\widetilde{P}_X(X_i)/\mathrm{d}P_X(X_i).$$

如果将具有 n 个样本的校准集和只有一个样本的测试集中的数据全部打乱, 并对所有数据的 X_i 计算 $w(X_i)$ 值, 那么在不知道哪个数据是来自测试集的场景下, 根据贝叶斯公式, 第 i 个数据为测试集中的数据的概率为 $w(X_i)/(\sum_{j=1}^{n+1} w(X_j))$. 因此, 在测试集的新样本的协变量是 $X_{n+1} = x$ 的情况下, 可以对校准集中的样本点 (X_i, Y_i) 添加一个与 x 有关的权重 $p_i^w(x)$, 其中

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(x)}, \quad i = 1, \ldots, n,$$

类似地, 测试集的新样本也进行加权,

$$p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^{n} w(X_j) + w(x)},$$

在进行加权之后, 所有校准集上的点在形式上就可以与测试集中的点 “可交换”, 因此在原始共型预测算法中计算不一致得分函数的分位数时加权, 得到的预测集就重新获得了有限样本的覆盖率保证. 在实际应用中, 这样具有密度比形式的权重也需要足够的训练集样本来估计, 文献 [83] 介绍了逻辑回归和随机森林两种估计权重的办法, 并且展示了在模拟实验下加权共型预测算法可以在协变量偏移设定的问题中重新获得具有有效覆盖率的预测集. 与协变量偏移类似, 另一种放松可交换性的假设被称为标签偏移[75], 指的是以下的设定

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{X|Y} \times P_Y, \quad (X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y,$$

即给定响应变量后, 协变量的条件分布在校准集和测试集中相同. 文献 [70] 考虑了标签偏移下的共型预测算法, 其核心思想仍是用加权的办法重构校准集和测试集的可交换性. 由于真实的校准集标签未知, 标签偏移假设下的加权共型预测算法还需要对响应变量空间进行遍历搜索. 和协变量偏移与标签偏移不同, 文献 [23] 从另一个角度考虑了校准集和测试集中分布不同的问题. 文献 [23] 提出了稳健共型预测, 该方法得到的预测集对任何在校准集分布周围的 f- 散度球内的测试集分布都有着几乎精确的标定覆盖率保证, 动机是利用稳健统计的思想, 建立一个对测试集分布任意方向扰动稳健的算法.

上述的例子中都假设校准集和测试集来自两个不同的静态的分布. 在实际问题中, 研究者们还会关心样本分布随着时间[12,26,39,40,100] 或者空间[64] 而缓慢连续变化的动态情形. 文献 [26] 提出的方法在数据是相互依赖的时间序列时也能得到近似有效的有限样本覆盖率, 但依赖于一些其他样本假设. 文献 [39,100] 考虑在线设定下的共型预测, 允许数据生成的分布以未知的方式随着时间变化, 根据当前的覆盖率不断调整下一个时间节点的预测参数来最终达到长时间间隔内的标定覆盖率. 文献 [40] 在文献 [39] 提出的自适应共型预测算法的基础上引入了一个额外的步骤, 即随着时间调整梯

度下降的步长参数, 从而解决了过度加权历史数据的问题, 得到了更稳健的在线共型预测集. 文献 [12, 64] 都采用加权的方式来刻画样本分布在时间或空间上的变化, 对于时间或空间距离测试集数据更近的样本点赋予更大的权重. 文献 [12] 不仅考虑了样本点不可交换的情况, 也考虑了共型预测背后的预测算法不对称的情形. 除了从不可交换的数据集中开发共型预测的方法, 研究者们也会反过来通过共型预测的工具来检验数据集中的不可交换性[87], 由此来识别变点[84] 和离群值[16, 57] 等. 其中, 文献 [16] 将离群值检验问题视为一分类问题, 只用到了训练集中的正常值数据, 并利用这些数据的可交换性来构造共型 p 值, 来检验新的测试集数据是正常值还是离群值. 而文献 [57] 则将该问题视为二分类问题, 不仅用到了训练集中的正常值数据, 还用到了训练集中的离群值数据. 然而, 由于离群值数据并不是可交换的, 因此用与文献 [16] 相同方法构造出的不一致得分函数的秩统计量不能直接作为一个有效的共型 p 值使用, 而只能用于对原始的共型 p 值的加权调整. 文献 [57] 加权调整后的共型 p 值被称为综合共型 p 值.

在放松独立同分布或可交换性假设的同时, 另一个值得研究的问题是考虑对共型预测中任意分布的假设的修改. 在共型预测的算法框架中, 假设样本分布任意无疑非常具有吸引力, 但这样过于宽泛的假设有时也是一种诅咒, 它会将很多不应被考虑的怪异而不光滑的分布囊括到算法的考量之中. 这些怪异分布的存在使得严格的条件覆盖率无法满足, 也使得当样本量趋近无穷的时候, 共型预测的预测集长度仍然不能收敛到 0[13]. 是否存在某种有趣的性质比 “任意分布” 要稍弱, 但仍具有现实意义? 是否存在这样的一种方法, 能一定程度上平衡共型预测这种完全分布无关的方法与参数模型, 使得这种方法对于一切的样本分布下都能保证一种稍弱一点的性质, 并且对于一些光滑的样本分布能得到更强的性质? 这些都是有待继续研究的开放性的问题.

4 应用

4.1 图像处理

图像识别问题是共型预测应用最多的场景之一, 因为图像识别的算法, 如卷积神经网络等深度学习方法都是难以直接获得不确定性度量的黑箱子算法, 而共型预测提供的有限样本覆盖率刚好可以弥补这个空缺. 先进的深度学习方法提升了图像识别问题的预测准确性, 但它们也在某个时刻以一种无声而无规律的方式犯错[8]. 很多时候, 这样的犯错常常会带来很高昂的代价, 例如在医疗领域中, 这时候就需要采用共型预测来量化不确定性并约束犯错的概率. 在第 3.1 小节中讨论过的自适应预测集 (APS) 方法[74] 以及其推广正则化自适应预测集 (RAPS) 方法[5] 正是针对图像分类问题而开发的. 本节将讨论另一类在图像处理问题中引入共型预测算法的方式, 即, 将控制覆盖率推广为控制风险的做法[15].

无论是分类问题还是回归问题, 响应变量 Y 都是一维的: 在分类问题里, Y 只在有限个标签上取值, 在回归问题里 $Y \in \mathbb{R}$. 在这些问题中, “覆盖率” 作为选择预测集的标准已经足够. 然而, 实际应用中的许多问题里的 Y 都是高维或复杂的. 以文献 [15] 中的肿瘤图像识别为例, 此时的输入 X 为一张 $d \times d$ 大小的医学影像, 通过神经网络, 输出每个像素点的一个 0 到 1 之间的值, 越接近 1 则说明神经网络认为这个像素点越可能是肿瘤. 研究者只需要决定一个阈值 λ, 选取机制则是把所有超过这个阈值的像素点找出. 显然, 当 $\lambda = 0$ 的时候, 整个图都会被选定为肿瘤, 在这种最平凡的情况下, 覆

盖率必然能达到 100%. 但如果直接简单地采用 “覆盖率” 作为不一致得分, 采用共型预测框架计算分位数 $\hat{p}$ 之后选取 λ 之后, 得出的预测集也不会比最平凡的 “选取整图” 要好多少. 事实上, 这样得到的预测集大量是假阳的 (false positive), 也就是本来没有肿瘤的区域被选了进去, 而这只为了保证这个 $(1-\alpha)$ 以上的覆盖率. 直接采用覆盖率作为不一致得分函数的做法会得到极为保守的预测集, 效果很差, 而原因是在这种高维输出问题中这样简单定义不一致得分函数太脆弱和极端. 例如, 可能真实的肿瘤图像包含 100 个像素点, 而某个阈值 λ 所对应的选择只少选了其中一两个像素点, 这样的选择其实已经很好地完成了任务, 找出了绝大部分的肿瘤像素点, 但却因为少选的这一两个像素点, 而让整张图都被判定为 “没有覆盖” 而前功尽弃. 因此一个自然的想法就是能否降低对每个图的判定标准, 例如, 只要找到 95% 以上的肿瘤像素点, 就算是识别成功了. 这样的想法就将覆盖率的控制推广到了风险控制的框架下. 风险是损失函数 L 的期望值

$$R(C^{\lambda}) = \mathbb{E}[L(C^{\lambda}(X_{n+1}), Y_{n+1})],$$

选择特定的损失函数可以得到不同的风险度量, 例如, 如果选择是否覆盖的指示函数, 风险就自然退化成了覆盖率:

$$R_{\text{miscoverage}}(C^{\lambda}) = \mathbb{E}[\mathbb{I}(Y_{n+1} \notin C^{\lambda}(X_{n+1}))] = \mathbb{P}(Y_{n+1} \notin C^{\lambda}(X_{n+1})).$$

在引入了风险控制的框架之后, 研究者想获得的预测集的性质就不再是达到边缘覆盖率, 而是更为一般的达到风险控制率:

$$P(R(C^{\lambda}) \geqslant \gamma) \leqslant \delta,$$

其中这里的 γ 与 δ 都是事先指定好的, 这被称为共型风险控制算法, 是共型预测算法的一种推广 [15]. 当 λ 从最大值 1 逐渐调整到最小值 0 的过程中, 识别出来的肿瘤像素点将从空集逐渐变成全图. 一个合理的假设是损失函数 L 随着 λ 的变化也会单调的变化, 当 λ 调到最小值 0 的时候, 不可能有任何遗漏的肿瘤像素点, 因此这时候的风险也必然是 0, 这也对应着共型预测算法中预测集取全集而使覆盖率必然达到 100% 的平凡情况. 因此, 当 λ 从大到小调整的时候, 风险第一次触碰到 γ 时所对应的 λ 自然就是想要找的 λ. 由于真实的风险函数 $R(C^{\lambda})$ 无法获得, 因此只能通过共型预测的思想以概率 $1-\delta$ 成立的一个风险的上置信界. 除了肿瘤识别问题之外, 还有一些其他的共型预测算法无法解决的图像处理问题, 都可以在共型风险控制算法的帮助下迎刃而解. 例如, 多标签分类问题, 即每张图中有多个物品需要识别, 黑箱子算法会扫描全图之后对所有类别的物品出现的可能性计算概率, 并输出所有超过预设的阈值 λ 的标签. 在这样的问题下, 仅用二值的 “覆盖与否” 准则也无法衡量一个预测集的好坏或共型程度, 必须引入风险函数进行细致的量化. 最近, 文献 [8] 还将共型风险控制算法引入到图像 - 图像回归问题中, 建立了严格的统计保证, 在快速磁共振成像和超分辨率透射电子显微镜成像等生物成像问题中有着广泛的应用.

4.2 因果推断

在因果推断领域中, 对治疗效果的识别和推断一直是最重要的问题之一. 许多治疗的效果具有严重的异质性, 对于不同人群有着不同的效果, 因此基于平均治疗效果 (ATE) 给出对所有人一刀切的治疗建议没有意义. 相比于平均治疗效果, 个体治疗效果 (ITE) 的推断可以帮助个性化和精准医疗的

方案定制, 因此在应用中尤为关键. 当前的许多研究主要关注条件平均治疗效果 (CATE) 的估计, 即在给定协变量值下的个体治疗效果的条件期望. 条件平均治疗效果虽然比起平均治疗效果提供了更丰富的信息, 但却依然忽略了个体的变异性, 也就是给定协变量值下的个体治疗效果的条件方差. 由于每位患者不能同时接受和不接受治疗, 因此两个可能发生的潜在结果中至多只能观测到一个, 对个体治疗效果的推断是反事实推理. 文献 [56] 将针对协变量偏移的加权共型预测与反事实推理框架进行了结合, 提出了共型反事实推理方法. 在潜在结果框架的因果模型下, 强可忽略性假设是指

$$(Y(1), Y(0)) \perp T \mid X,$$

其中 $Y(1)$ 和 $Y(0)$ 是两个可能的潜在结果, X 是协变量, T 是治疗. 这样的假设保证了反事实推理中的分布差异正是协变量偏移的设定. 具体来说, 对于试验组而言, (X, Y^{obs}) 的分布由 $P_{X|T=1} \times P_{Y(1)|X}$ 给出, 而对于对照组来说, 在他们进行试验的反事实世界中, (X, Y^{mis}) 的分布由 $P_{X|T=0} \times P_{Y(1)|X}$ 给出. 强可忽略性假设保证了在这两组分布中 $P_{Y(1)|X}$ 保持不变, 此时的密度比正是协变量 X 在两组间的倾向性得分之比, 而密度比可以直接用来计算加权共型预测中的权重. 共型反事实推理方法可以得到个体治疗效果的具有理论保证的覆盖率的置信区间, 并且还具有如双重稳健性的在因果推断方法中常见的优良特性.

共型反事实推断算法建立在强可忽略性假设下, 排除掉了所有能同时影响试验设计和潜在结果的混杂因素. 然而, 在实际数据中, 混杂总是存在. 文献 [50] 和 [98] 分别以不同视角用共型预测的方法对个体治疗效果进行了敏感性分析. 文献 [98] 通过边际敏感性模型[82] 量化未观测的混杂, 并采用共型预测的框架来估计给定混杂强度下的个体治疗效果的估计区间, 另外, 文献 [98] 将敏感性分析转化为分布偏移下的共型预测问题, 这是对共型反事实推断算法中协变量偏移下的共型预测问题的一种推广. 文献 [50] 定义了 Γ- 值来量化使个体治疗效果的估计区间包含 0 的最小混杂强度, 将个体治疗效果的敏感性分析问题表示为一系列假设检验问题, 并用稳健加权共型预测算法构建了有效的反事实预测集. 文献 [2] 将共型预测方法和因果推断问题中常用的基于机器学习的元学习器 (meta-learner) 相结合, 将元学习器提供的条件平均治疗效果的点估计转化为个体治疗效果的区间估计. 与文献 [56] 不同, 文献 [2] 可以直接对个体治疗效果进行推断, 而不需要估计潜在反事实结果作为中间步骤. 最近, 文献 [92] 还开发了新的共型因果推断方法, 在整群随机试验的框架下得到了稳健的有限样本的治疗效果估计.

4.3 其他应用

在半监督统计学习任务中, 通常存在小部分有标签样本和数量大得多的无标签样本, 其中有标签样本可以同时观测到协变量 X 和响应变量 Y, 而无标签样本中只能观察到协变量 X. 在实际问题中, 对众多无标签样本全部进行响应变量的测量通常是不可行的. 如果感兴趣的无标签样本是 $\{Y \in \mathcal{A}\}$ 的那些无标签样本, 那么在进行资源密集型的具体测量 Y 之前, 通常需要借助有标签样本的帮助. 在庞大的无标签样本池中, 通过协变量 X 的信息, 先进行筛选或子采样, 来选取一部分更加可能满足 $\{Y \in \mathcal{A}\}$ 的样本. 这样的实际问题在疾病筛查、人才招聘、药物设计合成等领域非常常见[1, 33, 49]. 共型预测算法因其无模型、无分布假设的特性, 可以无缝应用到这类样本选择问题之中. 文献 [49] 将共型 p 值和 Benjamini-Hochberg 方法[18] 结合, 在控制错误发现率的保障下选取样本. 共型 p 值由文献 [16] 提出, 是指通过共型预测框架和任意预测模型与不一致得分函数而得到的 p 值, 在测试集

样本分布和校准集相同的原假设下, 该 p 值服从 $[0,1]$ 间的均匀分布. 文献 [94] 则在控制错误发现率的基础上增加了有限预算限制和最大化样本多样性的目标, 即限制了样本选择的数量, 并且希望选择出来的满足 $\{Y\in\mathcal{A}\}$ 的样本在协变量空间上的相似性尽可能的低, 以保证样本足够有代表性. 这种对挑选出来的子样本进行共型预测的问题被称为后选择共型预测问题. 由于样本选择程序可能与校准集数据或测试集数据相关, 数据点之间的可交换性将不再满足, 此时得到的共型预测区间也不再享有边缘覆盖率的保证. 文献 [10] 提出了选择性条件共型预测算法, 同时对校准集和测试集数据应用选择程序, 再利用后选择的校准集数据的条件经验分布为后选择的测试集数据构建共型预测区间, 并得到了模型无关的覆盖率保证. 文献 [9] 则将后选择共型预测问题与在线共型预测框架相结合, 提出了自适应选择后校准算法, 为当前选择的个体自适应地挑选历史数据以构建校准集并以此得到共型预测区间. 文献 [9] 证明了自适应选择后校准算法得到的共型预测区间具有有限样本和分布无关的选择条件覆盖率保证, 并可以嵌入到样本分布随时间变化的在线共型预测算法中, 实现长期的错误覆盖率控制.

除了 p 值以外, 统计学家们还提出了 q 值[81] 和 e 值[72,91,93] 等概念, 以更好地进行复杂和相互依赖的多重假设检验问题中的统计推断, 尤其是控制错误发现率. 因此, 与共型 p 值相对应, 共型 e 值[14] 和共型 q 值[104] 的概念也被用于奇异值检测和结构化的多重假设检验问题中. 前者利用了 e 值可以通过加权平均而合并成为新的 e 值这一优良的数学特性, 整合了对同一数据集多次分割共型预测的结果, 使得最后输出的结果更加稳定和去随机化. 后者则可以放松用共型 p 值进行统计推断中所需的对称决策规则和联合可交换性假设, 而用非对称规则和原假设下不一致得分函数的成对可交换性代替之.

生存分析, 也被称为时间 - 事件分析, 在临床医学、工业寿命测试、经济学、社会学等各种学科中都有着广泛的应用. 文献 [21] 提出了共型生存分析算法, 将针对协变量偏移的加权共型预测算法和生存分析相结合, 在 I 型右删失的设置下构建了生存时间的预测下界. 在条件独立删失假设下, 该方法有着双重稳健性, 只要删失机制和条件生存函数的其中一个能被较好地估计, 共型生存分析算法得到的预测下界就能近似保证边缘覆盖率.

在大数据时代, 隐私保护越来越成为机器学习方法应用在实际问题中的一个重要考虑因素. ϵ- 差分隐私[32] 的概念用数学定义提供了隐私保护机器学习方法的被广泛使用的标准, 但对隐私保护机器学习直接应用共型预测, 所得到的预测集可能会丢失隐私保护的良好性质. 文献 [7] 将分割共型预测算法与隐私保护机器学习相结合, 通过隐私分位数的计算, 同时实现了可靠的不确定性量化与对校准数据集样本隐私的保护两个目标.

另一类在自然语言处理、计算机视觉、计算化学等领域的实际应用中经常遇到的设定被称为小样本学习 (few-shot learning). 在分割共型预测算法中, 需要将有标签的样本分划为训练集和校准集, 但当有标签的样本数量过少时, 训练出准确的模型已经非常困难, 再进行共型预测的校正所输出的预测集更会大到难以想象, 完全无法使用. 此时, 元学习 (meta-learning) 中的 “学习如何学习” 的思想可以帮助解决这类困难. 具体而言, 假如有一些相似的辅助任务 (例如, 目标任务是对不同品种的狗的分类, 辅助任务可以是对不同品种的猫的分类) 有着更多的样本, 那可以学习这些辅助任务的统计学习方式, 来弥补目标任务的样本不足的缺陷. 文献 [35] 将利用辅助任务元学习的范式与共型预测算法结合, 假设不同的辅助任务是可交换的, 首先利用辅助任务来学习小样本学习任务中的不一致得分

函数, 再将第一步学习到的不一致得分函数的模型应用到目标任务中, 由此利用共型预测算法得到的预测集将会比不利用辅助任务的信息所获得的预测集要小得多, 且同样能近似保证边缘覆盖率. 除了小样本学习外, 共型预测最近还被用于自然语言处理中的零样本 (zero-shot) 文本分类问题中以提升预训练语言模型的效率[29].

共型预测算法还可以和许多机器学习中的热门方向相结合, 例如联邦学习[62,63]、强化学习[44]、图神经网络[45,101]、量子机器学习[69] 等. 此外, 针对不同类型的数据, 也有不同的共型预测方法被开发出来, 例如时间序列数据[48,80,96,100]、纵向数据[17]、相依数据[26,67]、含有缺失值的数据[99] 等. 对共型预测的研究也包括了各学科中许多真实问题的应用与解决. 例如, 文献 [71] 利用共型预测算法的思想构建了一种公平性调整的分类算法, 旨在保护某些群体的人不受到人工智能算法的歧视和不公正对待. 文献 [41] 将共型预测和拉曼光谱技术结合以分类白酒, 文献 [102] 将共型预测和电子鼻系统结合以预测肺癌, 文献 [61] 讨论了共型预测在乳腺钼靶检查中的应用, 文献 [97] 将共型预测用在多模态时空数据中以建立实时预测森林火灾的模型, 文献 [46] 将共型预测与传统机器学习模型结合以生成保证边际覆盖率的可靠的岩爆分类, 文献 [103] 将共型预测框架与深度前馈神经网络结合, 生成了具有明确不确定性量化的化合物毒性预测模型, 文献 [31,58] 利用共型预测算法探究自动驾驶中动态环境下的安全路线设计问题. 在广泛而通用的理论推导和算法设计之外, 这些具体问题的落地也是共型预测研究的重要组成部分.

5 结论与展望

本文简要介绍了共型预测的算法实现、理论推广与应用实例. 共型预测可以在最少的分布假设下, 对任意的预测模型给出具有有限样本有效覆盖率的预测集. 现代数据科学的发展趋势是数据分布越来越不规则、预测模型越来越复杂、不确定性量化越来越重要, 而共型预测方法的三大优势正好可以应对这样的趋势. 此外, 共型预测方法的思想简单易懂, 算法操作容易, 使其成为其他领域的研究者和从业者也能够轻松上手的预测工具. 这一特点为共型预测方法在跨学科应用中发挥关键作用提供了便利. 总体来说, 目前关于共型预测的研究主要分为三个类别, 一是针对共型预测自身理论性质的研究, 力求进一步放松共型预测所需的假设并得到更优良的预测集性质; 二是融合共型预测算法与各种统计模型或课题的研究, 将共型预测算法视为强有力的不确定性度量工具, 设计适应各类统计问题的算法; 三是将共型预测算法用于各学科中的具体问题的研究, 为机器学习和黑箱子算法在这些问题中的输出结果添加置信度量, 增强可解释性. 可以预见, 共型预测方法的广泛应用将进一步推动统计学习的发展, 深刻改变各行各业的决策方式. 随着数据科学和其他学科的交叉融合, 共型预测方法将成为解决复杂现实问题的重要工具.

致谢 衷心感谢审稿专家对本文提出的宝贵意见.

参考文献

1 Ahlberg E, Hammar O, Bendtsen C, et al. Current application of conformal prediction in drug discovery: Two useful applications. Ann Math Artif Intell, 2017, 81: 145–154

2 Alaa A M, Ahmad Z, van der Laan M. Conformal meta-learners for predictive inference of individual treatment effects. In: Advances in Neural Information Processing Systems, vol. 36. Long Beach: Curran Associates, 2024, 47682–47703

3 Alaa A M, Hussain Z, Sontag D. Conformalized unconditional quantile regression. In: International Conference on Artificial Intelligence and Statistics. Cambridge: PMLR, 2023, 10690–10702
4 Amoukou S I, Brunel N J. Adaptive conformal prediction by reweighting nonconformity score. arXiv:2303.12695, 2023
5 Angelopoulos A N, Bates S, Malik J, et al. Uncertainty sets for image classifiers using conformal prediction. arXiv:2009.14193, 2020
6 Angelopoulos A N, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511, 2021
7 Angelopoulos A N, Bates S, Zrnic T, et al. Private prediction sets. Harv Data Sci Rev, 2022, 4: 1–16
8 Angelopoulos A N, Kohli A P, Bates S, et al. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning. Cambridge: PMLR, 2022, 717–730
9 Bao Y J, Huo Y Y, Ren H J, et al. CAS: A general algorithm for online selective conformal prediction with FCR control. arXiv:2403.07728, 2024
10 Bao Y J, Huo Y Y, Ren H J, et al. Selective conformal inference with false coverage-statement rate control. Biometrika, 2024, 111: 727–742
11 Barber R F. Is distribution-free inference possible for binary regression? Electron J Stat, 2020, 14: 3487–3524
12 Barber R F, Candès E J, Ramdas A, et al. Predictive inference with the jackknife+. Ann Statist, 2021, 49: 486–507
13 Barber R F, Candès E J, Ramdas A, et al. The limits of distribution-free conditional predictive inference. Inf Inference, 2021, 10: 455–482
14 Barber R F, Candès E J, Ramdas A, et al. Conformal prediction beyond exchangeability. Ann Statist, 2023, 51: 816–845
15 Bashari M, Epstein A, Romano Y, et al. Derandomized novelty detection with FDR control via conformal e-values. In: Advances in Neural Information Processing Systems, vol. 36. Long Beach: Curran Associates, 2024, 65585–65596
16 Bates S, Angelopoulos A N, Lei L H, et al. Distribution-free, risk-controlling prediction sets. J ACM, 2021, 68: 1–34
17 Bates S, Candès E, Lei L H, et al. Testing for outliers with conformal p-values. Ann Statist, 2023, 51: 149–178
18 Batra D, Mercuri S, Khraishi R. Conformal predictions for longitudinal data. arXiv:2310.02863, 2023
19 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol, 1995, 57: 289–300
20 Bian M, Barber R F. Training-conditional coverage for distribution-free predictive inference. Electron J Stat, 2023, 17: 2044–2066
21 Cai T T, Low M, Ma Z M. Adaptive confidence bands for nonparametric regression functions. J Amer Statist Assoc, 2014, 109: 1054–1070
22 Candès E, Lei L H, Ren Z M. Conformalized survival analysis. J R Stat Soc Series B Stat Methodol, 2023, 85: 24–45
23 Capuccini M, Carlsson L, Norinder U, et al. Conformal prediction in spark: Large-scale machine learning with confidence. In: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing. Piscataway: IEEE, 2015, 61–67
24 Cauchois M, Gupta S, Ali A, et al. Robust validation: Confident predictions even when distributions shift. J Amer Statist Assoc, 2024, 1–66
25 Cauchois M, Gupta S, Duchi J C. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. J Mach Learn Res, 2021, 22: 3681–3722
26 Cella L, Martin R. Validity, consonant plausibility measures, and conformal prediction. Internat J Approx Reason, 2022, 141: 110–130
27 Chernozhukov V, Wüthrich K, Zhu Y C. Exact and robust conformal inference methods for predictive machine learning with dependent data. In: Conference on Learning Theory. Cambridge: PMLR, 2018, 732–749
28 Chernozhukov V, Wüthrich K, Zhu Y C. Distributional conformal prediction. Proc Natl Acad Sci USA, 2021, 118: e2107794118
29 Cherubin G, Chatzikokolakis K, Jaggi M. Exact optimization of conformal predictors via incremental and decremental learning. In: International Conference on Machine Learning. Cambridge: PMLR, 2021, 1836–1845
30 Choubey P K, Bai Y, Wu C S, et al. Conformal predictor for improving zero-shot text classification effciency. arXiv:2210.12619, 2022
31 Ding T, Angelopoulos A N, Bates S, et al. Class-conditional conformal prediction with many classes. In: Advances in Neural Information Processing Systems, vol. 36. Long Beach: Curran Associates, 2024, 64555–64576
32 Dixit A, Lindemann L, Wei S X, et al. Adaptive conformal prediction for motion planning among dynamic agents. In: Learning for Dynamics and Control Conference. Cambridge: PMLR, 2023, 300–314

33 Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography: Third Theory of Cryptography Conference. New York: Springer, 2006, 265–284
34 Eklund M, Norinder U, Boyer S, et al. The application of conformal prediction to the drug discovery process. Ann Math Artif Intell, 2015, 74: 117–132
35 Feldman S, Bates S, Romano Y. Improving conditional coverage via orthogonal quantile regression. In: Advances in Neural Information Processing Systems, vol. 34. Long Beach: Curran Associates, 2021, 2060–2071
36 Fisch A, Schuster T, Jaakkola T, et al. Few-shot conformal prediction with auxiliary tasks. In: International Conference on Machine Learning. Cambridge: PMLR, 2021, 3329–3339
37 Fontana M, Zeni G, Vantini S. Conformal prediction: A unified review of theory and new challenges. Bernoulli, 2023, 29: 1–23
38 Gammerman A, Vovk, V, Vapnik V. Learning by transduction. arXiv:1301.7375, 2013
39 Gibbs I, Candès E J. Adaptive conformal inference under distribution shift. In: Advances in Neural Information Processing Systems, vol. 34. Long Beach: Curran Associates, 2021, 1660–1672
40 Gibbs I, Candès E J. Conformal inference for online prediction with arbitrary distribution shifts. J Mach Learn Res, 2024, 25: 1–36
41 Gu J, Liu H B, Ma C Q, et al. Conformal prediction based on raman spectra for the classification of chinese liquors. Appl Spectrosc, 2019, 73: 759–766
42 Guan L Y. Localized conformal prediction: A generalized inference framework for conformal prediction. Biometrika, 2023, 110: 33–50
43 Gupta C, Kuchibhotla A K, Ramdas A. Nested conformal prediction and quantile out-of-bag ensemble methods. Pattern Recogn, 2022, 127: 108496
44 Gupta N, Kahou S E. Cammarl: Conformal action modeling in multi agent reinforcement learning. arXiv:2306.11128, 2023
45 Huang K X, Jin Y, Candès E J, et al. Uncertainty quantification over graph with conformalized graph neural networks. In: Advances in Neural Information Processing Systems, vol. 36. Long Beach: Curran Associates, 2024, 26699–26721
46 Ibrahim B, Ahenkorah I. Classifying rockburst with confidence: A novel conformal prediction approach. Int J Min Sci Technol, 2024, 34: 51–64
47 Izbicki R, Shimizu G, Stern R B. CD-split and HPD-split: Effcient conformal regions in high dimensions. J Mach Learn Res, 2022, 23: 3772–3803
48 Jensen V, Bianchi F M, Anfinsen S N. Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Trans Neural Netw Learn Syst, 2024, 35: 9014–9025
49 Jin Y, Candès E J. Selection by prediction with conformal p-values. J Mach Learn Res, 2023, 24: 1–41
50 Jin Y, Ren Z M, Candès E J. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. Proc Natl Acad Sci USA, 2023, 120: e2214889120
51 Johansson U, Linusson H, Löfström T, et al. Model-agnostic nonconformity functions for conformal classification. In: 2017 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2017, 2072–2079
52 Jung C, Noarov G, Ramalingam R, et al. Batch multivalid conformal prediction. arXiv:2209.15145, 2022
53 Lei J. Fast exact conformalization of the lasso using piecewise linear homotopy. Biometrika, 2019, 106: 749–764
54 Lei J, G'Sell M, Rinaldo A, et al. Distribution-free predictive inference for regression. J Amer Statist Assoc, 2018, 113: 1094–1111
55 Lei J, Wasserman L. Distribution-free prediction bands for non-parametric regression. J R Stat Soc Series B Stat Methodol, 2014, 76: 71–96
56 Lei L H, Candès E J. Conformal inference of counterfactuals and individual treatment effects. J R Stat Soc Series B Stat Methodol, 2021, 83: 911–938
57 Liang Z Y, Sesia M, Sun W G. Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers. arXiv:2208.11111, 2022
58 Lindemann L, Cleaveland M, Shim G, et al. Safe planning in dynamic environments using conformal prediction. IEEE Robot Autom Lett, 2023
59 Linusson H, Johansson U, Boström H, et al. Effciency comparison of unstable transductive and inductive conformal classifiers. In: Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD. New York: Springer, 2014, 261–270
60 Linusson H, Norinder U, Boström H, et al. On the calibration of aggregated conformal predictors. In: Conformal and Probabilistic Prediction and Applications. Cambridge: PMLR, 2017, 154–173
61 Lu C, Chang K, Singh P, et al. Three applications of conformal prediction for rating breast density in mammography. arXiv:2206.12008, 2022
62 Lu C, Kalpathy C J. Distribution-free federated learning with conformal predictions. arXiv:2110.07661, 2021
63 Lu C, Yu Y D, Karimireddy S P, et al. Federated conformal predictors for distributed uncertainty quantification.

In: International Conference on Machine Learning. Cambridge: PMLR, 2023, 22942–22964
64 Mao H Y, Martin R, Reich B J. Valid model-free spatial prediction. J Amer Statist Assoc, 2024, 119: 904-914
65 Ndiaye E, Takeuchi I. Computing full conformal prediction set with approximate homotopy. In: Advances in Neural Information Processing Systems, vol. 32. Long Beach: Curran Associates, 2019
66 Ndiaye E, Takeuchi I. Root-finding approaches for computing conformal prediction set. Mach Learn, 2023, 112: 151–176
67 Oliveira R I, Orenstein P, Ramos T, et al. Split conformal prediction for dependent data. arXiv:2203.15885, 2022
68 Papadopoulos H. Inductive conformal prediction: Theory and application to neural networks. In: Tools in Artificial Intelligence. Rijeka: InTech, 2008, 315–330
69 Park S, Simeone O. Quantum conformal prediction for reliable uncertainty quantification in quantum machine learning. arXiv:2304.03398, 2023
70 Podkopaev A, Ramdas A. Distribution-free uncertainty quantification for classification under label shift. In: Uncertainty in Artificial Intelligence. Cambridge: PMLR, 2021, 844–853
71 Rava B, Sun W G, James G M, et al. A burden shared is a burden halved: A fairness-adjusted approach to classification. arXiv:2110.05720, 2021
72 Ren Z M, Barber R F. Derandomised knockoffs: Leveraging e-values for false discovery rate control. J R Stat Soc Series B Stat Methodol, 2024, 86: 122–154
73 Romano Y, Patterson E, Candès E J. Conformalized quantile regression. In: Advances in Neural Information Processing Systems, vol. 32. Long Beach: Curran Associates, 2019
74 Romano Y, Sesia M, Candès E J. Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems, vol. 33. Long Beach: Curran Associates, 2020, 3581–3591
75 Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new *a priori* probabilities: A simple procedure. Neural Comput, 2002, 14: 21–41
76 Sesia M, and Candès E J. A comparison of some conformal quantile regression methods. Stat, 2020, 9: e261
77 Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res, 2008, 9: 371–421
78 Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference, 2000, 90: 227–244
79 Solari A, Djordjilović V. Multi split conformal prediction. Stat Probab Lett, 2022, 184: 109395
80 Stankeviciute K, Alaa A, van der Schaar M. Conformal time-series forecasting. In: Advances in Neural Information Processing Systems, vol. 34. Long Beach: Curran Associates, 2021, 6216–6228
81 Storey J D. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Statist, 2003, 31: 2013–2035
82 Tan Z Q. A distributional approach for causal inference using propensity scores. J Amer Statist Assoc, 2006, 101: 1619–1637
83 Tibshirani R J, Barber R F, Candès E J, et al. Conformal prediction under covariate shift. In: Advances in Neural Information Processing Systems, vol. 32. Long Beach: Curran Associates, 2019
84 Volkhonskiy D, Burnaev E, Nouretdinov I, et al. Inductive conformal martingales for change-point detection. In: Conformal and Probabilistic Prediction and Applications. Cambridge: PMLR, 2017, 132–153
85 Vovk V. Conditional validity of inductive conformal predictors. In: Asian Conference on Machine Learning. Cambridge: PMLR, 2012, 475–490
86 Vovk V. Cross-conformal predictors. Ann Math Artif Intell, 2015, 74: 9–28
87 Vovk V. Testing randomness online. Statist Sci, 2021, 36: 595–611
88 Vovk V, Gammerman A, Shafer G. Algorithmic Learning in A Random World, vol. 29. Berlin: Springer, 2005
89 Vovk V, Petej I. Venn-abers predictors. arXiv:1211.0025, 2012
90 Vovk V, Shafer G, Nouretdinov I. Self-calibrating probability forecasting. In: Advances in Neural Information Processing Systems, vol. 16. Long Beach: Curran Associates, 2003, 1133–1140
91 Vovk V, Wang R D. E-values: Calibration, combination and applications. Ann Statist, 2021, 49: 1736–1754
92 Wang B K, Li F, Yu M. Conformal causal inference for cluster randomized trials: model-robust inference without asymptotic approximations. arXiv:2401.01977, 2024
93 Wang R D, Ramdas A. False discovery rate control with e-values. J R Stat Soc Series B Stat Methodol, 2022, 84: 822–852
94 Wu X Y, Huo Y Y, Ren H J, et al. Optimal subsampling via predictive inference. J Amer Statist Assoc, 2023: 1–13
95 Wu X Y, Huo Y Y, Zou C L. Multi-split conformal prediction via cauchy aggregation. Stat, 2023, 12: e522
96 Xu C, Xie Y. Conformal prediction interval for dynamic time-series. In: International Conference on Machine Learning. Cambridge: PMLR, 2021, 11559–11569
97 Xu C, Xie Y, Vazquez D A Z, et al. Spatio-temporal wildfire prediction using multi-modal data. IEEE J Sel Areas Inf Theory, 2023, 4: 302–313
98 Yin M Z, Shi C, Wang Y X, et al. Conformal sensitivity analysis for individual treatment effects. J Amer Statist

Assoc, 2024, 119: 122–135
99 Zaffran M, Dieuleveut A, Josse J, et al. Conformal prediction with missing values. In: ICML 2023—40th International Conference on Machine Learning. New York: Omnipress, 2023, 40578
100 Zaffran M, Féron O, Goude Y, et al. Adaptive conformal predictions for time series. In: International Conference on Machine Learning. Cambridge: PMLR, 2022, 25834–25866
101 Zargarbashi S H, Bojchevski A. Conformal inductive graph neural networks. arXiv:2407.09173, 2024
102 Zhan X H, Wang Z, Yang M, et al. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. Measurement, 2020, 158: 107588
103 Zhang J, Norinder U, Svensson F. Deep learning-based conformal prediction of toxicity. J Chem Inf Model, 2021, 61: 2648–2657
104 Zhao Z N, Sun W G. False discovery rate control for structured multiple testing: Asymmetric rules and conformal Q-values. J Amer Statist Assoc, 2024, in press

Development and application of conformal prediction

Zhang Jin & Xueqin Wang

Abstract Conformal prediction has gained increasing attention in recent years with the rapid development of machine learning. Known for its flexible structure and strict finite-sample theoretical guarantees, conformal prediction can be quickly and conveniently embedded into almost any prediction model. It performs rigorous uncertainty quantification by expanding prediction points into prediction sets. In this paper, we summarize the development history related to conformal prediction and review the basic algorithms and generalizations of conformal prediction along with the ubiquitous application scenarios of conformal prediction.

Keywords **conformal prediction, machine learning, prediction set, uncertainty quantification**
MSC(2020) **62J99**
doi: 10.1360/SSM-2024-0086