arXiv:2305.02780v1 [stat.ML] 4 May 2023

可解释区域描述符：基于超盒的局部解释

Susanne Dandl[1,2], Giuseppe Casalicchio[1,2], Bernd Bischl[1,2], 和 Ludwig Bothmann[1,2]

[1] 统计系，慕尼黑路德维希-马克西米利安大学，Ludwigstr. 33, 80539慕尼黑，德国[2] 慕尼黑机器学习中心（MCML），德国慕尼黑Ludwig.Bothmann@stat.uni-muenchen.de

摘要。本文提出了可解释区域描述符（IRD），用于局部、模型无关的解释。IRD 是描述如何改变某个观测的特征值而不影响其预测结果的超盒。它们通过提供一组“即使……也……”的论据（半事实解释）来阐明预测结果，并指出哪些特征会影响预测，以及是否存在逐点偏差或不合理之处。一个具体用例表明，这对机器学习模型开发者和受决策影响的人都很有价值。我们将搜索 IRD 的过程形式化为一个优化问题，并提出了一个用于计算 IRD 的统一框架，涵盖了设计要求、初始化技术和一种后处理方法。我们展示了如何调整现有的超盒方法，使其适用于这一统一框架。一项基准研究基于多项质量指标比较了这些方法，并确定了两种改进 IRD 的策略。

关键词： 可解释机器学习 · 模型无关的局部可解释性 · 半事实解释 · 超盒。

1 引言

监督式机器学习（ML）模型因其良好的预测性能而得到广泛应用，但由于其复杂性，通常难以解释。来自可解释机器学习（IML）领域的事后解释方法有助于得出关于这些模型内部过程的结论。解释方法可分为两类：解释单个预测的局部方法，以及解释模型一般预期行为的全局方法。Doshi-Velez 和 Kim [4] 将模型可解释性定义为“向人类解释或以人类能够理解的方式呈现的能力”。满足这一可解释性概念的一种拓扑形式是超盒。在本研究中，我们考察超盒作为局部解释，用于描述如何改变某个观测的特征值而不影响其预测。我们将这些盒称为可解释区域描述符（IRDs）。IRDs通过实值特征的区间以及类别特征所有可能类别的子集来描述特征空间（见表1）。

表1：基于信贷数据集 [5,9] 的示例，包含9个特征。第二列显示了具有中等信用风险预测的客户 的取值。*IRD* （由 MaxBox 和后处理（第4节）生成）展示了如何同时更改所有特征，使信用仍处于中等风险。*1*维 *IRD* 展示了如何更改单个特征而不改变预测（其他特征保持不变）。对于上半部分的特征，IRD 覆盖了训练数据中观测值的完整范围 。

特征	客户 IRD		一维 IRD	范围
sex	女性	{女性, 男性}	{女性, 男性}	{女性, 男性}
储蓄账户	少	{少, 中等	{少, 中等,	{少, 中等,
		丰富}	丰富}	丰富}
目的	car	{汽车、收音机/电视,	{汽车，收音机/电视,	{汽车，收音机/电视,
		家具，其他}	家具，其他}	家具，其他}
age	22	[19, 22]	[19, 75]	[19, 75]
job	熟练	{熟练，高技能	{非熟练，熟练，	{非熟练，熟练，
		熟练}	高技能}	高技能}
住房	rent	租赁	{自有, 免费, 租赁}	{自有, 免费, 租赁}
支票账户	中等	{少, 中等}	{少, 中等}	{少, 中等,
				rich}
credit.amount	4000	[4000, 5389]	[2127, 8424]	[276, 18424]
持续时间	30	[26, 33]	[6, 44]	[6, 72]

1.1 使用可解释区域描述符的动机示例

以银行贷款为例：一位客户向银行申请 €4000 信用贷款，用于购买一辆新车。她 22 岁，技能熟练，住在租赁住所，储蓄很少，支票账户余额中等。机器学习模型预测这笔贷款的风险属于低、中等还是高。由于预测结果为中等风险，银行拒绝了这项申请。表1中的 IRD 回答了这样一个问题：“特征或多个特征在多大程度上可以改变，同时预测结果仍属于中等风险类别”。通过 IRD，可以从多个角度深入了解预测结果。

首先，IRD 可提供一组半事实解释（SFEs）——也称为举重以明论证——以“即使……”的陈述形式为某项决策提供依据[19]。要让这些陈述令人信服，需要领域知识，例如，储蓄账户和支票账户余额越高，以及从事高技能工作，都会降低个人向银行借款的风险。基于这些知识，可以从表1中的 IRD 推导出多种半事实解释：(1) 说明某人属于中等风险类别而非低风险类别（例如：“即使你的储蓄账户余额中等，而且成为高技能人才，你的信贷风险仍属于中等”）；(2) 说明某人属于中等风险类别而非高风险类别（“即使你的支票账户余额很少，你的信贷风险仍属于中等”）。后一种陈述还揭示了一条“安全界限”：如果未来某些特征朝不利方向（高风险类别）变化，情况也仍在此界限之内。

其次，IRD 中某个特征的区间宽度或基数相对于其整个特征空间的大小，可以表明该特征是否会在局部影响预测（前提是定理1和定理2成立）。例如，在银行贷款示例中，与信贷金额或其支票账户的余额状态相比，储蓄或贷款用途似乎不会在局部影响预测，因为区域描述符在这两个维度上涵盖了整个观测到的特征范围。这些见解还揭示了可以通过哪些选项来改变给定的预测。[3]

第三，IRD 是模型审计的工具。如果从一个盒（例如，半事实解释）中获得的洞见与领域知识一致，用户就会更加信任模型；如果两者不一致，则有助于揭示模型意料之外的逐点偏差或不合理性。例如，一个不涵盖男性客户的 IRD可能 表明，模型会根据性别对个体进行不同分类。[4] 一个涵盖信贷金额为 €300 且支票账户余额较高的客户的 IRD，可能是模型不准确的指标，因为这类客户对银行而言应仅构成低风险。除信用风险外，我们还在附录A中展示了 IRD 的其他实际应用。

1.2 贡献

我们的贡献如下：1）我们将 IRD 引入为一类新的局部解释，用于描述特征空间中不影响某个观测预测结果的区域；2）我们将寻找 IRD 形式化为一个优化问题，并提出 IRD 方法应具备的理想属性；3）我们提出了一个用于计算 IRD 的统一框架，其中包括初始化技术和后处理方法；4）我们展示了如何调整数据挖掘或可解释机器学习领域中现有的超盒方法，使其适用于我们的统一框架；5）我们提出了一组质量度量，并在一项基准研究中据此比较所推导出的方法；6）我们提供了开放访问代码库，其中包含用于实现这些方法的 R 包，以及用于复现该基准研究的代码。[5]

2 方法论

ˆ令 $f : \mathcal{X} \to \mathbb{R}$ 为机器学习模型的预测函数，其中 $\mathcal{X}$ 表示一个 p 维特征空间。对于分类模型，我们考虑一个预定义的ˆ关注类别，f 针对该类别返回预测得分或概率。

2.1 为 IRD 形式化一般任务

我们的目标是找到最大的超盒 B，它覆盖感兴趣点 $\mathbf{x}' \in \mathcal{X}$，其中 B 中的所有数据点的预测都与 $\hat{} f(\mathbf{x}')$ 足够接近。该

[3] 然而，具体策略只能揭示反事实解释 [27]。[4] 请注意，如果所有性别都属于该盒，这并不意味着模型是公平的。[5] 链接将在录用后分享。审阅时，我们附上了一个 zip 文件。

超盒 B 应具有 p 个维度 $B = B_1 \times ... \times B_p$

$$\text{with } B_j = \begin{cases} \{c | c \in \mathcal{X}_j\} & \text{categorical } X_j \\ [l_j, u_j] \subseteq \mathcal{X}_j & \text{numeric } X_j \end{cases}, \tag{1}$$

由数值特征的区间以及类别特征的可能类别子集组成。$\mathcal{X}_j$ 反映第 j 个特征 X_j 的取值空间。根据 Lemhadri 等 [18], 的定义，若预测落入接近区域，则认为该预测足够接近；接近区域是用户定义的预测区间 $Y' = [\hat{f}(\mathbf{x}') - \epsilon_L, \hat{f}(\mathbf{x}') + \epsilon_H]$with $\epsilon_L, \epsilon_H \in \mathbb{R}_{\geq 0}$。[6] 在银行贷款示例中，接近区域应涵盖所有会导致中等风险类别的模型预测，例如违约概率预测为 30-60 %，即 $Y' = [0.3, 0.6]$。为实现上述目标，我们需要三项度量 [22,24]:

1. $coverage(B) = \mathbb{P}(\mathbf{x} \in B | \mathbf{x} \in \mathcal{X})$，用于衡量一个超盒覆盖整个特征空间的程度。由于实际中并非所有$\mathbf{x} \in \mathcal{X}$ 都可观测，我们使用基于给定数据的经验近似（$\mathbf{x}_i$）$_{1 \leq i \leq n}$，其中$\mathbf{x}_i \in \mathcal{X}$

$$\widehat{coverage}(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\mathbf{x}_i \in B). \tag{2}$$

2. $precision(B) = \mathbb{P}(\hat{f}(\mathbf{x}) \in Y' | \mathbf{x} \in B)$，盒内预测位于 Y'中的点所占的比例 B。同样，我们使用经验近似

$$\widehat{precision}(B) = \frac{\sum_{i=1}^{n} \mathbb{I}(\mathbf{x}_i \in B \wedge f(\mathbf{x}_i) \in Y')}{\sum_{i=1}^{n} \mathbb{I}(\mathbf{x}_i \in B)}. \tag{3}$$

3. 用于指示 B 是否覆盖$\mathbf{x}'$的指标

$$locality(B) = \mathbb{I}(\mathbf{x}' \in B). \tag{4}$$

以下内容将搜索 IRD 的过程具体化 [22]:[7]

$$\begin{aligned} &\underset{B \subseteq \mathcal{X}}{arg\,max}(\widehat{coverage}(B)) \\ &\text{s.t. } \widehat{precision}(B) = 1 \text{ and } locality(B) = 1. \end{aligned} \tag{5}$$

定义 1. 当且仅当在完全精确率下无法再添加任何盒时，该盒才是极大的；也就是说，对于所有数值型 X_j，均有 $(\nexists\, x_j \in X_j \wedge x_j < l_j : precision(B \cup [x_j, l_j]) = 1) \wedge (\nexists\, x_j \in X_j \wedge x_j > u_j : precision(B \cup [u_j, x_j]) = 1)$；对于所有类别型 X_j，均有 $(\nexists\, x_j \in X_j \setminus B_j : precision(B \cup x_j) = 1)$。

[6] 对于分类模型，必须满足 $Y' \subset [0, 1]$。[7] 为此，我们扩展了 Ribeiro 等 [22] 提出的优化任务，使其适用于 IRD：将精确率设为 1，并加入局部性约束。

覆盖率最大的盒 B 满足这一极大性性质。我们旨在获得一个极大的 B，因为这样 B 便能检测出对预测 $\hat{f}(\mathbf{x}')$ 局部不相关的特征。我们在附录 B 中证明如下结论。

定理 1. 如果 B 是极大的，那么对于一个未参与模型 f的特征 $X_j\hat{}$，$B_j = [min(X_j)\ max(X_j)]$ 成立。

同样，我们希望得到满足 $precision(B) = 1$ 的同质盒 B。这样，B 就能检测出对于 $\hat{f}(\mathbf{x}')$ 局部相关的特征。我们在附录 C 中证明如下结论。

定理 2. 如果 $precision(B) = 1$，那么对于一个与$\hat{f}(\mathbf{x}')$局部相关的特征，$B_j \subset X_j$ 成立。

2.2 可解释区域描述符的期望性质

在第 3 节中，我们将讨论生成 B 的相关方法。这些方法作为 IRD 方法的适用性，取决于它们是否考虑了公式 (5) 的所有目标，以及是否满足以下关于可解释区域描述符的期望性质。

可解释性 为了使 B 具有可解释性，我们只考虑返回单个 p维超盒的方法。B 的超矩形结构便于自然地进行解释，而超椭球或由半空间构成的多胞体则不具备这一特点 [18]。根据公式（5），B 需要覆盖$\mathbf{x}'$,，满足以下条件即可：$\forall j \in \{1, ..., p\}$：$x'_j \in B_j$。

$\hat{}$模型无关性 f 的定义并未对机器学习模型或特征空间施加任何限制。因此，方法应具备模型无关性，以便能够解释具有各种特征类型（二元、名义型、有序型或连续型）的回归模型或分类模型。

稀疏性约束 Eckstein 等人 [6] 证明，如果未固定用于定义盒的特征，那么最大盒问题的优化任务是$\mathcal{NP}$-难问题。这同样适用于可解释区域描述符的搜索，后者只额外要求$\mathbf{x}' \in B$。由于超盒的搜索空间会随特征数量增加而扩大，因此无法遍历所有潜在解。此外，可解释区域描述符的维度与数据集相同，这也妨碍了其可解释性，而可解释性恰恰是可解释区域描述符的首要目标。为减少特征数量，方法应能够遵循用户定义的稀疏性约束，使得对于某些特征 X_j，$B_j = x'_j$。其他解决方案将在第 7 节讨论。

3 相关工作

式（5）的优化任务可以从数学上理解为：寻找邻域中预测值 $\in Y'$ 的原像 $\mathbf{x}'$。因此，

表 2：在特征空间中搜索超盒的方法概览。

	目标			**期望特性**		
	覆盖率	精确率	局部性	可解释	与模型无关	稀疏
水平集方法						
PBnB [28,29]	√	√	×	×	√	×
数据挖掘						
MaxBox [6]	√	√	×	√	×	×
PRIM [10]	×	×	×	√	×	×
事后可解释机器学习						
Anchors [22]	√	√	√	√	×	×
MAIRE [24]	√	√	√	√	×	×
LORE [12,13]	×	×	√	√	√	×
可解释分类器						
列生成 [2]	√	√	×	×	√	×

IRD 可以看作函数值 $\in Y'$ 的水平集的一个子集。水平集近似通常由点 [7]，组成，只有少数方法通过超盒 [28,29] 对其进行近似。这些方法会生成多个盒，而不是单个盒，并且不要求包含给定的 $\mathbf{x}'$。因此，从我们的意义上说，它们并不具备可解释性，因而也不适合生成 IRD。

在数据挖掘领域，[6] 提出了一种针对具有二元结果的数据集的最大盒（MaxBox）方法，以寻找相对于正类的最大同质超盒。Friedman 和 Fisher [10] 提出了患者规则归纳方法（PRIM），用于在特征空间中寻找结果均值较高的盒。这两种方法均不要求$\mathbf{x}'$ 位于盒内。

如前所述，IRD 也可以被视为一种用于总结大量 SFEs 的方法。大多数已提出的 SFE 方法仅返回一个点作为解释 [3,14,19]。相比之下，Guidotti 等人提出的方法 [12,13] 使用代理树返回一组 SFEs。该方法通过追踪通往感兴趣点的路径，揭示了哪些特征值对于得出预测最为重要。代理树的可靠性取决于一个假设，即树能够充分复现底层模型，但这一假设往往并不成立。此外，IRD 要求盒具有同质性，而这只有通过过拟合/深度生长树才能实现。因此，只有当底层模型是基于树的模型时，树结构才适合用于推导 SFEs [8,25]。

一种利用超盒的可解释机器学习方法是 Anchors 方法 [22]。返回的超盒表明必须固定或锚定哪些特征，才能防止模型改变数据点的分类。Anchors 最初被提出时，旨在寻找也能部分覆盖其他类别观测值的超盒；在其实现中，精确率 0.95 是默认值 [23]。尽管可以将精确率改为 1，但由于搜索空间有限，Anchors 仍不适合用于生成可解释区域描述符：特征的盒边界要么被设为数据中观测到的完整特征范围，要么被设为值 $\mathbf{x}$。这带来了生成“过度具体的锚点”的风险，其覆盖率较低[22]。为了生成覆盖率更高的盒，可以预先对特征进行分箱。

然而，目前还不存在针对 Anchors 的成熟离散化技术，而且 Anchors 所依赖的优化过程不允许在优化期间调整分箱。

为了解决离散化问题，Sharma 等人 [24] 提出了与模型无关的可解释规则提取（MAIRE）流程。MAIRE 通过基于梯度的优化，为连续特征寻找更优的边界。但它仍无法对类别特征进行更精细的选择：要么盒不允许更改某个特征，要么就涵盖该特征的所有可能取值。

Dash 等人 [2] 提出了一种基于一组超盒的分类器。该方法利用列生成，着重寻找超盒的最优组合，以便为来自整个特征空间的输入构建准确的模型。因此，该方法并不关注局部性，按我们的定义也不具备可解释性。

表 2 总结了所讨论的方法是否适用于生成可解释区域描述符。总体而言，这些方法都未能满足式 (5) 中的所有目标以及第 2.2 节提出的要求。具体来说，没有一种方法处理稀疏性约束，且只有少数方法与模型无关。在第 4.4 节中，我们对 MaxBox、PRIM 和 MAIRE 进行了修改，使其满足我们的所有要求，从而将其转化为实用的 IRD 方法。其他方法都无法进行所需程度的修改，因为其底层且不可替代的优化方法要么针对多个盒，要么面向不同的搜索空间。后者尤其适用于 Anchors。不过，在第 6 节的基准研究中，该方法可作为基线方法。

4 生成 IRD

现在，我们提出一个用于生成 IRD 的统一框架，该框架包含四个步骤：限制、选择、初始化和优化。还可以选择执行后处理步骤（第 4.5 节）。

4.1 搜索空间的限制

为了限制 B的初始搜索空间，我们提出一种简单的方法，用于找到$\mathbf{x}'$ 的最大局部盒 B_{-} ，使得 $B \subset \mathrm{B}_{-}$。对于连续型特征X_j，我们在等距网格上改变$\mathbf{x}'$ 的值 x'_j 。B_{-j}的上界和下界设为 x'_j中导致 Y'之外的预测所需的最小变化量。这种方法类似于个体条件期望（ICE）值 [11]。对于类别型特征 X_j，B_{-j} 包含$\mathcal{X}_j$ 的所有类别；在调整$\mathbf{x}'$的 x'_j 后，这些类别仍会产生 $\in Y'$ 预测。如果用户设置稀疏性约束，将特征 X_j 设为不可变，则必须满足 $\mathrm{B}_{-j} = x'_j$ 。我们在附录 D 中证明如下结论。

定理 3。 对于任意一个求解式 *(5)* 优化问题的盒 B ，都有 $B \subseteq B$ 。

4.2 底层数据集的选择

所有方法都需要一个由$\mathbf{x} \in \mathcal{X}$构成的数据集 $\mathbf{X}_{-}$ 作为输入。该数据集用于根据覆盖率和精确率的经验版本（式 (2) 和式 (3)）评估（相互竞争的）盒。对于某些方法，该数据集还提供一组待评估的潜在框边界。训练数据是合适的数据集。由于只有实例 $\in \mathbf{B}_{-}$ 相关（定理 3），我们从$\mathbf{X}_{-}$ 中移除所有实例 $\notin \mathbf{B}_{-}$。因此，对于所有不可变特征 X_j，$x_j = x'_j \, \forall \mathbf{x} \in \mathbf{X}_{-}$ 均成立。更多特征和稀疏性约束会增加$\mathbf{X}_{-}$ 在$\mathbf{x}'$附近仅稀疏分布的风险。由于我们的目标是获得忠实于模型而非数据生成过程（DGP）的 IRD，因此可以通过从 $\mathbf{B}_{-}$的可行特征范围内均匀采样来生成新数据。在第 6 节中，我们考察在 $\mathbf{B}_{-}$ [8] 内使用样本量加倍的采样数据，相较于使用训练数据，会如何影响 IRD 和 IRD 方法的质量。

4.3 盒的初始化

所有方法都需要将初始框 B 作为输入，该框可以设为覆盖所有 $\mathbf{X}_{-}$ 的最大局部盒$\mathbf{B}_{-}$，也可以设为仅包含 $\mathbf{x}'$ 的最小可能盒。我们将从最大局部盒开始的方法定义为自上而下的 IRD 方法，将从最小可能盒开始的方法定义为自下而上的方法。

4.4 框边界优化

最后一步包括对框边界进行优化。自顶向下方法通过迭代收缩最大局部盒的框边界来提高该盒的精确率（保持 $\mathbf{x}' \in B$），而自底向上方法则通过迭代扩展最小盒的框边界来提高该盒的覆盖率（保持精确率为 1）。在本节中，我们将介绍 MaxBox、MAIRE 和 PRIM 方法，以及我们提出的扩展方法，使这些方法能够优化式 (5) 并满足第 2.2 节中的设计要求。扩展方法的伪代码及其内部工作原理示意图见附录 E。所有方法均以数据集 $\mathbf{X}_{-}$ 和一个初始框 B作为输入。

MaxBox – 自顶向下方法 MaxBox 最初是为二元分类问题提出的，此类问题包含正类和负类。该方法从覆盖所有数据的最大盒开始。分支定界（BnB）算法 [17] 检查缩小盒子的各种方案，以优化其针对正类的精确率。分支规则通过排除一个负类样本$\mathbf{x}$ 来创建新盒，使盒在至少一个特征维度上缩小到$\mathbf{x}$ 的取值之上或之下（类别特征采用独热编码）。盒覆盖率上界的估计值决定了接下来对哪个不精确的盒进行分支、使用哪个样本进行分支，以及丢弃哪些上界未超过覆盖率的盒

[8] 扩大一倍指的是训练数据的大小，而不是 $\mathbf{X}_{-}$的大小。

当前最大同质盒的覆盖率。如果没有剩余可缩小的盒，则返回最大同质盒作为 IRD。

扩展 将带有预测值 $\in Y'$ 的观测标记为正类后，该方法便与模型无关。由于原始算法未考虑相应的盒是否仍包含$\mathbf{x}'$, ，我们调整了该方法，舍弃不包含$\mathbf{x}'$的盒，以保证局部性。

PRIM – 自上而下的方法 该方法最初旨在寻找平均结果较高的盒。该过程从一个包含所有点的盒开始。在剥离阶段，PRIM 会迭代识别一组符合条件的子框（对于数值特征，由 α 分位数和 (1-α) 分位数定义；对于类别特征，则由每个存在的类别定义），并剥离排除后平均结果最高的子框。重复此步骤，直到盒中包含的点数低于总点数的一定比例。在粘贴阶段，通过添加使结果均值提升最多的子框，迭代扩大该盒。这些子框至少包含 α 个观测，其在一个维度上具有最近的较低值或较高值（数值型 X_j），或具有一个新类别（类别型 X_j）。

扩展 我们调整了该方法以针对公式（5）：在每次剥离迭代中，排除一个子框，使得生成的盒具有最高精确率（覆盖率作为破平局规则）；在每次粘贴迭代中，则添加最大的同质子框。如果精确率和覆盖率不足以选出用于剥离或粘贴的最佳盒，则从最佳候选项中随机选择一个子框。当生成的盒变为同质时，剥离即停止；当不存在可添加的同质盒时，粘贴即停止。此外，仅剥离不覆盖$\mathbf{x}'$ 的子框。按照作者的建议，我们在基准研究（第 6 节）中使用 $\alpha = 0.05$ 。

MAIRE – 自底向上方法 该方法从一个覆盖$\mathbf{x}'$的盒开始。每次迭代中，通过优化覆盖率度量的可微近似，使用 ADAM [15] 调整框边界。如果精确率低于某个阈值，或$\mathbf{x}'$ 不在盒内，该方法还会分别优化公式（3）和公式（4）的可微版本。MAIRE 在达到指定的迭代次数后停止。最后，该方法返回迭代过程中最大的同质盒。

扩展该方法要求特征缩放到 0–1。为解决类别特征的一对多问题（第 3 节），我们对类别特征进行独热编码。为与其他（收敛型）方法进行公平比较，我们实现了一种收敛判据：让 MAIRE 不断扩大框边界，直到精确率降至 1 以下，然后 MAIRE 最多再运行 100 次迭代。第 6 节实验所用的实现基于作者的实现 [24] ，并作了上述修改。超参数按照作者的建议设置。我们仅将精度阈值设为 1，而非 0.95。

4.5 后处理

上一节所述的所有方法都根据$\mathbf{X}_{-}$中有限数量的数据点来确定框边界$^{-}$。有限的数据访问可能导致$\mathbf{X}_{-}$未能覆盖特征空间中的某些区域$^{-}$，从而使生成的 B 的边界并非最优： B 中可能存在预测为 $\notin Y'$的区域，也可能在 B 之外存在相邻区域，其预测同样为 $\in Y'$。为了改进给定盒 B的框边界，我们利用新采样的数据开发了以下后处理方法。该流程包含与 prim 相同的剥离和粘贴步骤。

首先，基于新采样的数据测量 B 的精确率。如果 $\exists \mathbf{x} \in B$ with $\hat{} f(\mathbf{x}) \notin Y'$，则根据新采样数据在子框内的分布，迭代剥离大小占比精确率最低的子框。如果所有待剥离的子框均为同质，则停止剥离。在随后的粘贴步骤中，添加已被证明为同质的最大子框（根据该子框内的新采样数据判断）。如果无法明确确定最佳盒（因为多个盒具有相同的精确率和覆盖率），则随机选择一个子框。该方法具有三个超参数：用于评估的样本数量；对于连续特征，剥离或粘贴盒时的相对盒大小（相对于$\mathcal{X}$的大小）；以及最小盒大小的阈值。后者用作粘贴的停止准则。如果无法添加任何同质子框，则将连续特征待添加的相对盒大小减半，只要该相对盒大小不低于阈值即可。我们方法的伪代码见附录 F。

第 6 节考察我们的后处理方法是否能够改进可解释区域描述符。对于这些实验，我们将用于评估盒子的样本数设为 100，将相对盒大小设为 0.1，并将最小盒大小的阈值设为 0.05。

5 质量度量

现在，我们介绍一组用于评估生成的可解释区域描述符 和*IRD* 方法的质量度量。这些度量适用于要解释的单个实例$\mathbf{x}'$，其中 B 是$\mathbf{x}'$ 的 IRD 方法 G所返回的 IRD。评估需要评估数据$\mathbf{E}$，其中包含$\mathbf{x} \in \mathcal{X}$；对于第 6 节中的基准研究，我们使用训练$^{-}$数据以及从 B_{-}中均匀采样的新数据。训练数据有助于评估这些方法在生成可解释区域描述符时是否恰当地使用了训练数据（例如，精确率应为 1），而大量新增生成的数据$_{-}\in \mathrm{B}_{-}$ 则会带来相对于模型而非数据生成过程（DGP）更精确的评估。

局部性 IRD 应覆盖 $\mathbf{x}'$。如果 $locality(B) = \mathbb{I}(\mathbf{x}' \in B)$ 等于 1，则满足此属性。

覆盖率 给定两个精确率相同的 IRD，我们更倾向于覆盖率较高的那个（公式 (2)）。为评估覆盖率，我们使用来自覆盖 $\mathbf{x}'$ 的连通凸水平集 $\mathcal{L}$ 的样本 $\mathbf{x} \in \mathbf{E}$ 。

定义 2. 若存在一条连接 $\mathbf{x}$ 与 $\mathbf{x}'$ 的路径，且所有中间点的预测均为 $\in Y'$，则数据点 $\mathbf{x}$ *with* $\hat{} f(\mathbf{x}) \in Y'$ 属于 $\mathbf{x}'$ 的 L。

路径通过 Kuratomi 等人的识别算法确定。 [16],详情见附录 G。

精确率 对于覆盖率相同的两个 IRD，优先选择精确率较高的 IRD（式（3））。

极大性 根据定义 1，盒应基于$\mathbf{x} \in \mathbf{E}$而非$\mathbf{x} \in \mathcal{X}$达到极大。

$\hat{}$调用次数 优先选择调用 f 次数较少的 IRD 方法。

鲁棒性 如果我们使用相同的$\mathbf{x}'$ 和 $\hat{} f R$，并使用相同的$\bar{}\mathbf{X}_{-}$重新运行方法 G，那么生成的可解释区域描述符 $B_1, ..., B_R$ 应与最初生成的B重叠，使得 $robustness(G) = \min_{k \in \{1,...,R\}} \frac{\sum_{\mathbf{x} \in \mathbf{E}} \mathbb{I}(\mathbf{x} \in B \cap B_k)}{\sum_{\mathbf{x} \in \mathbf{E}} \mathbb{I}(\mathbf{x} \in B \cup B_k)}$ 具有较高的值。

6 性能评估

在一项基准研究中，我们探讨以下研究问题（RQ）：

1. 根据第 5 节中列出的质量度量，在将训练数据用作$\mathbf{X}_{-}$（不进行后处理）时，第 4.4 节中的不同方法与基线方法$\bar{}$相比表现如何？$\bar{}$2. 使用源自 B_{-}、大小加倍的采样数据，相较于使用训练数据，会对可解释区域描述符和方法的质量产生什么影响？3. 后处理（第 4.5 节）会对 IRD 方法的质量产生什么影响?

作为基线方法，我们采用 Anchors 方法 [22]，其精确率为 1，数值特征使用 20 个基于分位数的分箱（详情见第 3 节）。

6.1 设置

为回答这些研究问题，我们使用了 OpenML 平台上的六个数据集[26], ，其目标变量为二元、多类别或连续型。表3总结了这些数据集的维度，以及目标变量和特征的类型。在训练模型之前，从数据集中随机抽取五个数据点并将其排除，以作为$\mathbf{x}'$。对于每个数据集，我们训练四种模型：超盒模型、逻辑回归/多项逻辑回归/线性模型（取决于结果变量）、具有一个隐藏层的神经网络，以及随机森林模型。我们对随机森林模型的树木数量和隐藏层中的神经元数量进行调优（详情见附录H）。每个$\mathbf{x}'$ 的超盒模型均由分类与回归树（CART）模型推导而来。对于一个

表3：基准数据集概览。编号：OpenML 编号；类型：目标类型；观测数：行数；连续型/类别型：连续型/类别型特征数量。

Name	编号	类型	观测数	连续	类别
糖尿病	37	二元	768	8	0
井字棋_ _	50	二元	958	0	9
cmc	23	三分类	1473	2	7
车辆	54	四分类	846	18	0
no2	886	回归	500	7	0
plasma_retinol	511	回归	315	10	3

给定的$\mathbf{x}'$，，后处理后的模型会在某个点落入与$\mathbf{x}'$相同的终端节点时预测为1，否则预测为0。[9]

对于分类模型，预测函数返回 $\mathbf{x}$ 属于概率最高类别的概率。对于二元目标，我们设置 $[0.5, 1]$。对于回归和多分类目标，设置为 $\hat{}(\mathbf{x})$，$\hat{}(\mathbf{x})\hat{}$，其中 δ 为训练数据预测值的标准差 f。对于多分类，区间还会被限制在 0 到 1 之间。对于每个数据集、模型和 $\mathbf{x}'$，，我们使用 MaxBox、PRIM 和 MAIRE 生成 IRD，并将 Anchors 作为基线方法。方法的超参数根据第4节进行设置。方法要么在训练数据上运行，要么在从 B_ 均匀采样的数据上运行（RQ 2）；要么不进行后处理，要么进行后处理（RQ 3）。对于稳健性评估，我们重复了实验 $R = 5$ 次。

这些方法及其生成的 IRD 根据第 6 节中的性能指标进行了评估——评估数据要么是训练数据，要么是从 B_ 中均匀采样的 1000 个新实例。我们还通过执行 Wilcoxon秩和检验，对这些方法进行了统计比较，以检验以下假设：两种（IRD）方法的覆盖率和精确率值的分布是否存在差异（研究问题1）；同一种方法使用训练数据与采样数据时是否存在差异（研究问题2）；以及同一种方法在未进行后处理与进行后处理时是否存在差异（研究问题3）。实验在一台配备 2.60 GHz Intel(R) Xeon(R)处理器和 32 个 CPU 的计算机上进行。总体而言，生成这些盒耗时 63 小时，期间使用了 20 个 CPU。稳健性评估中的五次重复实验还额外耗时 316 小时。

6.2 结果

图1以可视化方式比较了各方法的覆盖率和精确率值。表4展示了各方法满足极大性的频率以及对$\hat{}f$的调用次数。每个数据集和模型的单独结果、统计分析以及鲁棒性结果见附录I。由于所有返回的可解释区域描述符都覆盖$\mathbf{x}'$，我们省略了局部性度量的结果。

研究问题1 - 方法比较 在不进行后处理且训练数据为$\mathbf{X}_$的情况下（图1第一行），根据

[9] CART模型的真实超盒可能大于终端节点诱导的超盒（见附录中的图 6）。

ˆ 表4：各方法在极大性和 f 调用次数方面的比较，结果按所有数据集、模型和x′取平均值。每种方法均在¯训练数据或来自 B_ 的均匀采样数据上运行或评估，并且不进行 (0) 或进行 (1) 后处理。极大性越高、调用次数越少越好。

	训练数据						**采样数据**					
	最大值$_{train}$		最大值$_{samp}$		对 ˆ f Max$_{train}$ 的调用次数		Max$_{samp}$		对 ˆ f 的调用次数	
	0	1	0	1	0	1	0	1	0	1
MaxBox	**0.60**	**0.42**	0.06	**0.41**	**184**	55769	0.23	**0.45**	0.24	**0.43**	**1621**	**37627**
PRIM	0.42	0.37	**0.18**	0.39	**184**	**46070**	0.20	0.42	**0.25**	0.39	**1621**	42958
MAIRE	0.18	0.41	0.04	**0.41**	**184**	68126	0.06	0.41	0.11	0.35	**1621**	92976
Anchors	0.27	**0.42**	0.16	0.40	26402	94448	**0.31**	0.42	0.18	0.36	77818	129276

在训练数据和新采样数据上的评估结果，MaxBox 的精确率最高，其次是 MAIRE。PRIM 的 IRD 平均覆盖率最大，但它们也覆盖了预测结果不在 Y' 之内的采样数据。由于出现并列时会随机选择一个子框，按照我们的鲁棒性度量，PRIM 不具备鲁棒性。在极大性方面，没有一种方法优于其他方法。总体而言，根据覆盖率、精确率以及对 f 的调用次数，所有方法都优于基线方法 Anchors。后者是因为竞争盒是在观测数据的按列置换数据上进行评估的。其他所有方法的调用次数均仅为 $f\,|\,\bar{\mathbf{X}}\_\,|$ 次。

RQ 2 - 训练数据与采样数据 平均而言，来自 B_ 的采样数据¯大小翻倍后，其覆盖率、精确率和极大性指标略有提高（相对于新采样数据而言），但相对于训练数据则没有提高。由于 ¯**X**_的大小增加，因此需要更多次调用 ˆ f 。[10]

RQ 3 - 无后处理与有后处理对比 对于所有方法，后处理都提高了可解释区域描述符（IRDs）的覆盖率和精确率。不同方法之间以及不同底层数据方案（训练数据与采样数据）之间的可解释区域描述符（IRDs）质量差异有所减小。质量提升是以效率和鲁棒性为代价的；平均而言，后处理导致额外增加了 57,000 次对 f 的调用，而新数据的采样降低了鲁棒性。maire 平均需要最多的后处理迭代次数，其次是 Anchors。

7 结论、局限性与展望

结论 我们介绍了可解释区域描述符（IRD），它们以超盒的形式描述特征空间中的区域，这些区域不会影响某个实例的预测。这些超盒提供了一组半事实解释，用于证明预测的合理性，并指出哪些特征会影响预测，以及是否可能存在逐点偏差或不合理性。我们形式化了对可解释区域描述符的搜索，并为 IRD 方法提出了期望目标、统一框架和质量度量。我们

[10] 大小相较于训练数据增长了十倍，而非两倍，因为并非所有训练数据都是 $\in$ ¯B_ ，因此并非都在 ¯**X**_中。

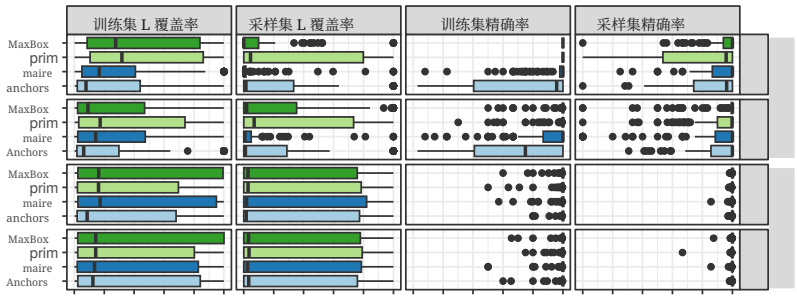


图1：IRD 方法在覆盖率和精确率方面的比较，评估基于训练数据或 B_ 内新采样的数据¯。附加项L 表示，在评估覆盖率时，仅考虑 $\mathcal{L}$ 内的训练点或采样点。箱线图中的每个点反映了一个实验设置（数据集、模型和 $\mathbf{x}'$）所生成的一个 IRD 的性能。每种方法均在训练数据（traindata）或从¯B_ 中均匀采样的数据（sampled）上运行或评估，并且方法的运行方式分为不使用后处理（postproc）和使用后处理两种。精确率和覆盖率的值越高越好。

详细讨论了三种现有的超盒方法，并对其进行了调整，使其能够搜索可解释区域描述符。基准研究中缺乏一种“能够一统天下”的方法，强调了建立包含多种方法的统一框架的必要性。研究还表明，获取更大的均匀采样数据集，或使用我们提出的后处理方法，可以进一步提升可解释区域描述符的质量。

局限性 我们的工作为进一步研究提供了可能，例如研究方法超参数的敏感性、采样规模的影响，或方法相对于$\mathbf{x}'$ 或底层数据发生轻微变化时的鲁棒性。虽然我们在基准研究中仅考虑了低维数据集，但对于高维数据集，我们提出了两种限制搜索空间的策略：一是让用户决定哪些特征可以更改、哪些特征不可更改¯（第2.2节）；二是基于个体条件期望曲线推导最大的局部盒 $B \subset$B_（第4.1节）。未来研究可以探索：（1）使用其他可解释机器学习方法，例如特征重要性方法，来选择需要研究其变化的特征（所有其他特征均设为其允许值范围）；（2）考虑特征相关性或因果关系来生成可解释区域描述符，这不仅能自然地限制搜索空间，还能使可解释区域描述符忠实于数据生成过程。对于将可解释区域描述符应用于表格数据以外的数据，考虑特征相关性同样十分重要。虽然本文介绍的所有方法都是与模型无关的，但我们将针对图像数据和文本数据的具体研究留待未来开展。

展望 我们相信，我们的工作也可以成为研究 IRD 在其他领域应用的起点，例如用于超参数（HP）调优：如果调优方法为某个机器学习模型确定了一个有潜力的超参数集，IRD 可以

揭示其敏感性，以及是否存在其他同样出色但效率更高的超参数设置。IRD 还可以为可解释的局部替代模型（如 LIME [21]）识别高保真区域。LIME 使用一个（正则化）线性模型 $\hat{}g(\mathbf{x})$ 来近似黑盒模型 $\hat{}\ f(\mathbf{x})$ 在观测数 $\mathbf{x}'$ 附近的预测。在此情况下，了解线性模型在哪个区域 B 近似黑盒模型（高保真区域）可能会很有用；$\hat{}g$ 仅在 $\mathbf{x}'$ 周围的区域 B 提供有价值的洞见，其中 $\forall \mathbf{x} \in B: \hat{}h(\mathbf{x}) := |\hat{}f(\mathbf{x}) - \hat{}g(\mathbf{x})| \leq \epsilon$ 对于用户定义的 $\epsilon > 0$。以 $\hat{}h$ 作为预测模型并结合 $Y' = [0, \epsilon]$，IRD 方法可能以可解释的方式识别此类高保真区域 B。</output>

致谢

本研究部分得到了德国联邦统计局的支持。

参考文献

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.：SMOTE：合成少数类过采样技术。《人工智能研究杂志》**16**，321 357 2002101613– ( )。https://doi.org/./jair.953 2. Dash, S., Günlük, O., Wei, D.：通过列生成实现布尔决策规则。载于：第 32 届国际神经信息处理系统会议论文集。第 4660–4670 页。NIPS'18, Curran Associates Inc.，美国纽约州雷德胡克（2018）3. Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K., Das, P.：基于缺失信息的解释：迈向具有相关负例的对比解释。载于：第 32 届国际神经信息处理系统会议论文集。第 590–601 页。NIPS'18，Curran Associates Inc.，美国纽约州雷德胡克（2018）4. Doshi-Velez, F., Kim, B.：迈向严谨的可解释机器学习科学。arXiv 1702.08608 v2, arXiv.org 电子印本档案（2017）。https://doi.org/10. 4855008608 51702/arXiv.. Dua, D., Graff, C.：UCI 机器学习资料库（2017），https://archive.ics.uci.edu/ml/datasets/statlog+(德语+信用+数据）6. Eckstein, J., Hammer, P.L., Liu, Y., Nediak, M., Simeone, B.：最大盒问题及其在数据分析中的应用。《计算优化与应用》**23**1020546910706 73285298 2002101023( )，– ( )。https://doi.org/./a:. Emmerich, M.T.M., Deutz, A.H., Kruisselbrink, J.W.：黑盒水平集近似的质量指标。载于：Tantar, E., Tantar, A.A., Bouvry, P., Del Moral, P., Legrand, P., Coello Coello, C.A., Schütze, O.（编）《EVOLVE——概率、集合导向数值方法与进化计算之间的桥梁》，第 157–185 页。Springer Berlin Heidelberg，德国柏林海德堡（2013）。https://doi.org/ 10 1007 978 3 642 32726 1_4 ./----. Fernandez, G., Aledo, J.A., Gamez, J.A., Puerta, J.M.：模糊分类树中的事实解释与反事实解释。《IEEE 模糊系统汇刊》**30**(12)，5484–5495（2022）。https://doi.org/10.1109/tfuzz.2022.3179582

9. Ferreira, L.：德国信用风险（2018），https://www.kaggle.com/datasets/kabure/ german-credit-data-with-risk，最后访问日期：2023年10月1日。Friedman, J.H., Fisher, N.I.：高维数据中的峰值搜寻。Statistics and Computing 9(2)，123–143（1999）。https://doi.org/10.1023/A:1008894516817Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.：窥探黑箱内部：通过个体条件期望图可视化统计学习。Journal of Computational and Graphical Statistics 24(1)，44–65（2015）。https://doi.org/10.1080/10618600.2014.907095Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F.：用于黑箱决策的事实解释与反事实解释。IEEE Intelligent Systems 34(6)，14–23（2019）。https://doi.org/10.1109/MIS.2019.2957223Guidotti, R., Monreale, A., Ruggieri, S., Naretto, F., Turini, F., Pedreschi, D., Giannotti, F.：通过事实规则与反事实规则，为黑箱模型提供稳定且可操作的解释。Data Mining and Knowledge Discovery（2022）。https://doi.org/10.1007/s10618-022-00878-5Kenny, E.M., Keane, M.T.：关于为深度学习生成合理的反事实解释与半事实解释。Proceedings of the AAAI Conference on Artificial Intelligence 35(13)，11575–11585（2021）。https://doi.org/10.1609/aaai.v35i13.1737715. Kingma, D.P., Ba, J.：Adam：一种随机优化方法。arXiv 1412.6980 v9，arXiv.org电子预印本库（2017）。https://doi.org/10.48550/arXiv.1412.698016. Kuratomi, A., Miliou, I., Lee, Z., Lindgren, T., Papapetrou, P.：JUICE：有依据的反事实解释。载于：Pascal, P., Ienco, D.（编）Discovery Science。第493–508页。Lecture Notes in Computer Science，Springer Nature Switzerland，Cham（ 2022 10 1007 978 3 031 18840 4_35 ）。https://doi.org/ . / - - - - .Land, A.H., Doig, A.G.：一种求解离散规划问题的自动方法。Econometrica 28(3)，497–520（1960）。https://doi.org/10.2307/1910129Lemhadri, I., Li, H.H., Hastie, T.：RbX：预测模型的基于区域的解释。arXiv 2210.08721，arXiv.org电子预印本库（2022）。https://doi.org/10.48550/arXiv.2210.0872119. Nugent, C., Doyle, D., Cunningham, P.：通过基于案例的解释获得洞见。Journal of Intelligent Information Systems 32(3)，267–295（2009）。https://doi.org/10.1007/s10844-008-0069-3Pfisterer, F., Poon, J., Lang, M.：mlr3keras：mlr3的Keras扩展。GitHub仓库。网址：https://github.com/mlr-org/mlr3keras（2022），提交版本：bad8434b7898b51b2143fc680594057c00dc708021. Ribeiro, M.T., Singh, S., Guestrin, C.："我为什么应该相信你？"解释任意分类器的预测结果。载于：第22届ACM SIGKDD国际知识发现与数据挖掘会议论文集。第1135–1144页（2016）。22. Ribeiro, M.T., Singh, S., Guestrin, C.：Anchors：高精度与模型无关的解释。Proceedings of the AAAI Conference on Artificial Intelligence 32(1)（2018）。https://doi.org/10.1609/aaai.v32i1.1149123. Ribeiro, M.T., Singh, S., Guestrin, C.：Anchor。GitHub仓库。网址：https://github.com/marcotcr/anchor（2022），提交版本：b1f5e6ca37428613723597e85c38558e8cd21c2e24. Sharma, R., Reddy, N., Kamakshi, V., Krishnan, N.C., Jain, S.：MAIRE——一种与模型无关的可解释规则提取流程，用于解释分类器。载于：Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E.（编）Machine Learning and Knowledge Extraction，第12844卷，第329–349页。Springer International

出版，Cham（2021）。https://doi.org/10.1007/978-3-030-84060-0_21，丛书名称：Lecture Notes in Computer Science 25. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.：决策树和基于模糊规则的分类器的事实解释与反事实解释的生成和评估。载于：2020 年 IEEE 国际模糊系统会议（FUZZ-IEEE）。第 1–8 页。IEEE，英国格拉斯哥（2020）。https://doi.org/10.1109/ FUZZ48607.2020.9177629 26. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.：OpenML：机器学习中的网络化科学 **15** 2641198 272496020141011452641190(), –(). https://doi.org/./.. Wachter, S., Mittelstadt, B., Russell, C.：无需打开黑箱的反事实解释：自动化决策与 GDPR。《哈佛法律与技术期刊》**31**(2), 841–887（2018）28. Zabinsky, Z.B., Huang, H.：用于水平集近似的基于分区的优化方法：概率分支定界。载于：Smith, A.E.（编）《工业与系统工程中的女性：新兴主题的关键进展与观点》，第 113–155 页。Springer International Publishing，Cham（2020）。https://doi. org/ 10 1007 978 3 030 11866 2_6 ./----. Zabinsky, Z.B., Wang, W., Prasetio, Y., 29 Ghate, A., Yen, J.W.：用于水平集近似的自适应概率分支定界。载于：2011 年冬季仿真会议论文集（WSC）。第 4146–4157 页。IEEE，美国亚利桑那州凤凰城（2011）。https://doi.org/10.1109/WSC.2011.6148103

A 应用示例

除了第 1 节中的信贷申请案例外，下面我们还展示一个医疗和司法管辖权方面的应用。

医疗 考虑一个用于预测某人未来是否会患糖尿病的机器学习模型。（为简单起见，我们假设该模型能准确近似现实世界中的关系。）下面我们讨论两种情况：

（1）某人被预测将会患糖尿病，想知道原因，以及可以采取哪些措施来预防。可采取的行动有很多：增加运动、少吃红肉、使用顺势疗法药物等。IRD 可以指出哪些行动不太可行，例如，当各种现实可行的运动量都处于盒内时，运动就不是一个有希望的选择。不过，改变饮食可能是一种选择，因为仅仅通过每周有一天吃肉来改变饮食，并不属于盒内（具体的预防策略可以揭示反事实解释）。

（2）某人被预测不会患糖尿病，想知道在不改变预测结果的情况下，自己的生活方式有多大的调整空间。只要不改变饮食、增加红肉摄入，体重增加或许也不会使其患糖尿病的风险上升。

司法管辖权 考虑一个预测某人在未来两年内是否会犯罪的机器学习模型。得分较高的人希望了解原因。不包含所有受保护属性组（例如性别）的可解释区域描述符可能表明针对这些群体的不公平歧视。因此，可解释区域描述符可以启动对机器学习模型公平性和偏见的进一步调查。

B 定理 1 的证明

证明。给定一个未参与预测模型 $\hat{f}$ 的特征 X_j，使得 $\forall \tilde{\mathbf{x}} \in \mathcal{X} \wedge \forall x_j \in \mathcal{X}_j$:

$$\hat{f}(\tilde{x}_1, ..., \tilde{x}_{j-1}, \tilde{x}_j, \tilde{x}_{j+1}, ..., \tilde{x}_p) = \hat{f}(\tilde{x}_1, ..., \tilde{x}_{j-1}, x_j, \tilde{x}_{j+1}, ..., \tilde{x}_p), \tag{6}$$

并且给定一个关于定义 1 极大的盒 B，其对应于$\mathbf{x}'$。现在假设定理 1 不成立，因此 $B_j = [l_j, u_j] \subset \mathcal{X}_j$。然而，由于公式 (6) 成立，对于数值型 X_j，以下条件之一成立：$(\exists\, x_j \in \mathcal{X}_j \wedge x_j < l_j\colon\ precision(B \cup [x_j, l_j]) = 1)$，或 $(\exists\, x_j \in \mathcal{X}_j \wedge x_j > u_j\colon\ precision(B \cup [u_j, x_j]) = 1)$；对于类别型 X_j，则有 $(\exists\, x_j \in \mathcal{X}_j \setminus B_j\colon\ precision(B \cup x_j) = 1)$ 成立，这与 B的极大性假设相矛盾。

C 定理 2 的证明

证明。给定一个盒 B，满足 $precision(B) = 1$，且对应于$\mathbf{x}' \in B$；再给定一个特征 X_j，该特征对于 $\hat{f}(x')$ 是相关的，使得 $\exists x_j \in \mathcal{X}_j \setminus B_j\colon \hat{f}(x'_1, ..., x'_{j-1}, x_j, x'_{j+1}, ..., x'_p) \notin Y'$。现在假设定理 2 不成立，因此 $B_j = \mathcal{X}_j$。这与 $precision(B) = 1$ 的表述相矛盾，因为 x_j 这会导致一个预测 $\notin Y'$，而该预测对应的$\mathbf{x}'$ 也被该盒覆盖。

D 定理 3 的证明

证明。不失一般性，我们假设只有数值特征。假设我们计算得到了 $\bar{B}_- = \bigcup_{j=1}^{p} [l_j, u_j]$，使得 $\forall j \in \{1, ...p\}$：

$$\hat{f}(\underbrace{x'_1, .., x'_{j-1}, l_j, x'_{j+1}, ..., x'_p}_{:=\mathbf{x}'_l}) \notin Y' \wedge \hat{f}(\underbrace{x'_1, .., x'_{j-1}, u_j, x'_{j+1}, ..., x'_p}_{:=\mathbf{x}'_u}) \notin Y'.$$

我们暂且假设 $B \subset \bar{B}_-$ 不成立，因此存在一个同质的B，满足 $min(B_j) < l_j$ 或 $max(B_j) > u_j$，且 $\mathbf{x}' \in B$。然而，此时$\mathbf{x}'_l$ 或$\mathbf{x}'_u$ 也会属于 B，但对于这两者，$\hat{f}(\mathbf{x}'_u) \notin Y'$ 或 $\hat{f}(\mathbf{x}'_l) \notin Y'$ 成立，这与 B 是同质的相矛盾。

E IRD方法的伪代码与示意图

E.1 伪代码

算法 1 改编的 MaxBox 方法 [6]

输入： 目标实例 $\mathbf{x}'$, 期望范围 Y', 预测模型 $\hat{f}:\mathcal{X} \to \mathbb{R}$, 输入数据集 $\bar{\mathbf{X}}$, 初始框 B

初始化候选项 $=[]$, 上_界_覆盖率_最佳值 $=$-Inf, 当前值_最佳值 $=[]$

如果 $\exists \mathbf{x} \in \bar{\mathbf{X}} \wedge \mathbf{x} \in B:\_\ \hat{f} \notin Y'$ **然后**

候选项 $=$候选项.$append(B)$

当 $length$(候选项) > 0 **执行**

$B^{best} = choose\_best$(候选项)

▷如果上界_界限_覆盖率_最佳值 < 0, B^{best} 则对应于具有的盒之前执行的收缩步骤数最多（以上界为准的覆盖率作为破平局规则），否则， B^{best} 对应于使其最大化的盒

$$\text{<escape>}\left(\frac{|\{\mathbf{x} \in B|f \in Y'|}{(\mathbf{x})\}\{\mathbf{x}\hat{}(\mathbf{x})\}|}\right).\text{</escape>}$$

候选项 $=$ 候选项。$remove(B^{best})$

子框 $= create\_new\_candidates(B^{best})$ 。 在图2中，C 和 D 是新的由初始框生成的候选项

对于 $B \in$ 子框 **执行**

若 $\forall \mathbf{x} \in B:\hat{f}$（$\mathbf{x}$） $\in Y'$ **则**

覆盖率 $= upper\_bound\_coverage(B)$

如果覆盖率 $>$ 上界_覆盖率_最佳值 **则**

当前值_最佳值 $= B$

上界_覆盖率_覆盖率_最佳值 $=$覆盖率

结束条件判断

else

如果 $upper\_bound\_coverage(B) >$ 上界_覆盖率_最佳值_**则**

候选项 $=$ 候选项.$append(B)$

结束条件判断

结束条件判断

结束循环

结束循环

else

当前值_最佳值 $=$ B

结束条件判断

返回 当前值 _best

算法 2 改编的 PRIM 方法 [10]输入：目标实例 x′, 期望范围 Y'，预测模型 $\hat{f}: \mathbb{X} \to \mathbb{R}$, 输入 数据集 $\mathbb{X}_{-}$，初始框 B当 $\exists \mathbf{x} \in \bar{\mathbb{X}}_{-} \wedge \mathbf{x} \in B: \hat{f} \notin Y'$时，执行：对于 $j \in \{1, ..., p\}$，执行$C_j = []$ ▷ 创建剥离候选项 如果 X_j 为数值型，则 $C_j = C_j.append(B_j^-, B_j^+)$，其中 $B_j^- = [l_j, min(X_{j(\alpha)}, x_j')]$ 且$B_j^+ = [max(X_{j(1-\alpha)}, x_j') u_j]$ 分别为当前盒中 X_j 的 $x_{j(\alpha)}$ 分位数和 $x_{j(1-\alpha)}$ 分位数 α和 $(1-\alpha)$ B否则，如果 X_j 为类别型，则$C_j = \{s \in B_j \mid s \neq x_j'\}$ 结束如果 结束循环$b^{best} = \arg\max_{b \in C_j, j \in \{1,...,p\}} precision(B \setminus b)B = B \setminus b^{best}$结束循环 当同质时 =TRUE 当同质时执行：对于 $j \in \{1, ..., p\}$，执行$C_j = []$ ▷ 创建粘贴候选项 如果 X_j 为数值型，则在框内 $= \{\mathbf{x} \in \bar{\mathbb{X}}_{-} \mid x_k \notin B_k\}$，对于 $k \in \{1, ..., j-1\}$， $j+1$， $...p$ 个数$C_j = C_j.append\ B_j^-\ B_j^+$ 已添加 $= |\{\mathbf{x} \in \bar{\mathbb{X}}_{-} \mid \mathbf{x} \in B\}| \cdot (,)$，其中 $B_j^- = [x_j^l, l_j]$和 $B_j^+ = [u_j, x_j^u]$ 分别为第 j个特征值，来自框内第（已添加_个）观测 $\mathbf{x} \in$ ，其值低于 l_j ；而x_j^u 为第 j个特征值，来自框内第（已添加_个）观测 $\mathbf{x} \in$ ，其值高于 u_j x_j 否则，如果 X_j 为类别型，则$C_j = \{s \in X_j \mid s \notin B_j\}$ 结束如果$C_j = \{b \in C_j \mid precision(B \cup b) = 1\}$ 结束循环 如果 $\exists j \in \{1, ..., p\}$: $|C_j| > 0$ 则$b^{best} = \arg\max_{b \in C_j, j \in \{1,...,p\}} coverage(B \setminus b)B = B \cup b$否则 同质 =FALSE 结束如果 结束循环 返回 B

算法 3 改进的 MAIRE 方法 [24]

输入： 目标实例 $\mathbf{x}'$, 期望范围 Y', 预测模型 $\hat{f}: \mathcal{X} \to \mathbb{R}$, 输入数据集 $\bar{\mathbf{X}}$_, 初始框 B, 精度阈值 τ （默认值为 1），最大数量
最大迭代次数_迭代（默认 100）
缩放 $\mathbf{x} \in$ 的所有特征值 将 $\mathbf{X}$_ 和 $\mathbf{x}'$ 缩放至 0-1 范围
最佳值_覆盖率 $= 0$
已收敛 $=$ 否
best_候选项 $= B$
$i = 0$
当 $i \leq max\_iterations$ **时执行**
  $B = optimize\_with\_adam(B)$
    ▷ 优化覆盖率、精确率和局部性的可微版本
  如果 $precision(B) \geq \tau \wedge coverage(B) \geq$ 最佳值_覆盖率 **那么**
    best_候选项 $= B$
  否则，如果 $precision(B) < \tau$ **那么**
    已收敛 $=$ TRUE
  结束 if
  如果 收敛 $=$ TRUE **则**
    i $=$i $+ 1$
  结束 if
结束循环
返回 最佳值_候选项

E.2 插图

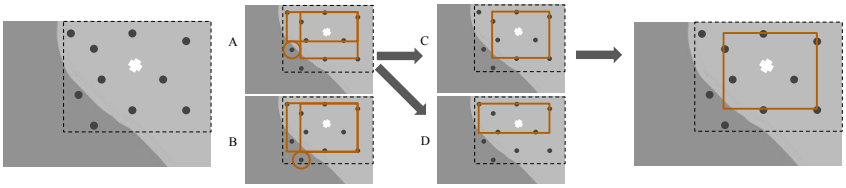


图2：改进版 MaxBox 算法示意图。算法从 $\bar{\mathbf{B}}$_（虚线框）开始。框中有两个预测为 $\notin Y'$ 的数据点（称为负样本），因此需要进一步优化该框。首先，选择一个负样本——A 或 B 中的一个。然后，检查在一个特征维度上排除这些点后，预测为 $\in Y'$ 的样本数量。两个负样本各自得到的框最多覆盖七个样本。我们选择 A 中的负样本（选择 B 也可以）。由此得到的框成为新的子问题/候选项（C 和 D）。C 和 D 中的两个框都只包含预测为 $\in Y'$的样本，但选择 C 中的框作为最优解，因为它包含更多预测为 $\in Y'$的样本。D 因样本数量较少而被舍弃。由于两个框中都没有负样本，C 和 D 无法继续拆分，因此 MaxBox 返回的框是 C 中的框。

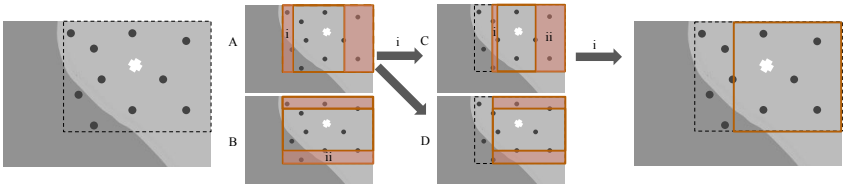


图3：改编后的 PRIM 算法示意图。算法从¯B_开始。在第一次迭代中，有四个可能被移除的子框（A 与 B 两个特征维度中各有两个）。选择子框 i，是因为它的精确率最高，但与 ii 相比，其大小更小。在下一步（C 和 D）中，同样有四个子框可能被移除。出于同样的原因，我们再次选择 i。移除它之后，得到的盒同时也是最终盒，因为在粘贴步骤中只有一个子框可以添加——仍然是 i。所有其他维度均为极大值。

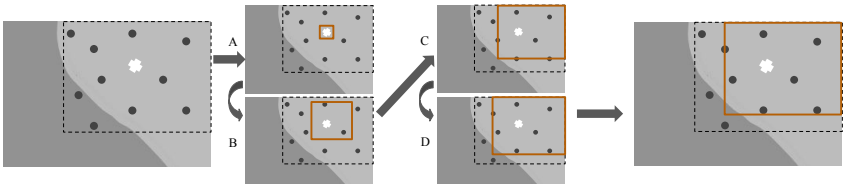


图4：改编后的 MAIRE 算法示意图。算法从可能的最小盒开始。随后迭代扩大框边界（A-D）。只有当新盒的精确率 = 1时，才会更新框边界。

F 后处理方法伪代码

算法 4 后处理算法 - 剥离（灵感源自 [10]）

输入： 目标实例 $\mathbf{x}'$, 期望范围 Y', 预测模型 $\hat{f}: \mathcal{X} \to \mathbb{R}$, 初始盒 B, 评估样本数量 M （默认 100）, 相对小盒大小 of 连续特征 α (默认 0.1)

对于 $j \in \{1,...,p\}$ **执行**

如果 X_j 数值型**则**

$s_j = (max(\mathcal{X}j) - min(\mathcal{X}j)) \cdot \alpha$ ▷推导数值特征的子盒大小特征基于$\mathcal{X}$

如果 X_j 整数**则**

$s_j = round(s_j)$

结束如果

结束如果

结束循环

$\bar{\mathbf{X}}\_ = sample\_uniformly(B, n = M \cdot 5)$ 。采样新数据以进行检查如果 B 同质

如果 $\exists \mathbf{x} \in \bar{\mathbf{X}} \wedge \mathbf{x} \in B :\_ \hat{f} \notin Y'$ **那么**

非_同质 = TRUE ▷开始剥离

当 非_同质**时**

对 $j \in \{1, ..., p\}$ **执行**

$C_j = [\,]$ ▷ 创建剥离候选项

如果 X_j 数值型，**则**

$C_j = C_j.append(B_j^-, B_j^+)$

其中 $B_j^- = [l_j, min(l_j + s_j, x'_j)]$ 和 $B_j^+ = [max(u_j - s_j, x'_j) u_j]$

否则如果 X_j 类别型 **则**

$C_j = \{s \in B_j \mid s \neq x'_j\}$

结束如果

$C_j = \{b \in C_j \mid precision(B_j^b) < 1\}$ 与 $B_j^b = (B_1 \times ... \times B_{j-1} \times b \times B_{j+1} \times ... \times B_p)$

结束循环

如果 $\exists j \in \{1,...,p\} : |C_j| > 0$ **那么**

$b^{best} = \underset{b \in C_j,\, j \in \{1,...,p\}}{\arg\max} precision\_to\_boxsize(B_j^b)$ 。在 M 上进行评估新的在 B_j^b 内采样的实例

$B^{best} = (B_1 \times ... \times B_{j-1} \times b^{best} \times B_{j+1} \times ... \times B_p)$ ▷选择最低的那个相对于大小的精确率

$B = B^{best}$

else

not_同质 = FALSE

结束 if

结束循环

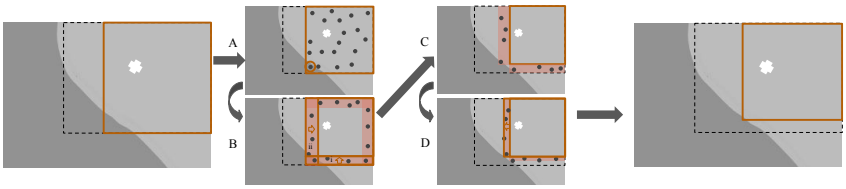
结束条件判断

返回 B, s $= \{s_j \mid X_j$ 数值型$\}$

算法 5 后处理算法——粘贴（灵感来自 [10]）

输入： 目标实例 $\mathbf{x}'$, 期望范围 Y', 预测模型 $\hat{f}: \mathcal{X} \to \mathbb{R}$, 初始框 B （可能已剥离）, 评估样本数 M （默认 100）, 相对连续特征子框大小 α （默认 0.1）, 相对下限阈值子框大小 α_0 （默认值 0.05）, 数值特征的子盒大小 $\mathbf{s}$
同质 =TRUE ▷ 开始粘贴
步长 $= 1$
当同质 **时**
  对于 $j \in \{1,...,p\}$ **执行**
    $C_j = [\,]$ ▷ 创建用于粘贴的候选项/子框
    如果 X_j 数值型**则**
      $C_j = C_j.append(B_j^-,\ B_j^+)$
其中 $B_j^- = [l_j - \text{步长} \cdot\ s_j,\ \ l_j]$ 以及 $B_j^+ = [u_j, u_j + \text{步长} \cdot\ s_j]$
    否则如果 X_j 类别型 **则**
      $C_j = \{s \in X_j \mid s \notin B_j\}$
    结束 if
    $C_j = \{b \in C_j \mid precision(B_j^b) = 1\}$ 与 $B_j^b = (B_1 \times ... \times B_{j-1} \times b \times B_{j+1} \times ... \times B_p)$
  结束 for
  如果 $\exists j \in \{1,...,p\} : |C_j| > 0$ **则**
    $b^{best} = \underset{b \in C_j,\, j \in \{1,...,p\}}{\text{参数最大值}} size(B_j^b)$ ▷ 在 M 内采样的新实例上评估 B_j^b
    $B = B \cup b$ ▷ 选择精确率为 1 的最大项
  else
    如果步长 $\geq \alpha_0$ **则**
      步长 $=$ 步长$/2$ ▷ 如果不存在精确率为 1 的盒，考虑减小子盒大小
    else
      同质 =FALSE
    结束 if
  结束 if
结束 while
返回 B

图5：后处理算法示意图。算法从另一种方法生成的盒（棕色实线盒，是¯虚线盒 B_的子框）开始。首先，对新点进行采样，并评估该盒是否同质（A）。如果不是，则根据相对于大小的精确率，迭代剥离精确率最低的子框（B）。精确率根据子框内新采样的点进行评估。首先剥离子框 i，然后剥离子框 ii（两者都包含一个预测为 $\notin Y'$的样本）。如果不存在精确率为 < 1 的子框，则评估该盒能否进一步扩大（C）。如果所有考察的子框都具有 < 1精确率，则将子框大小减半（D），只要相对小盒大小不低于阈值，就继续此过程。



G 水平集识别

算法从 $\mathbf{x}'$ 开始，并尝试在$\mathbf{x}$ 和$\mathbf{x}'$的名义型、继而有序型、最后连续型特征之间寻找连接 $\in Y'$ 。如果找到路径，则$\mathbf{x}$ 属于$\mathcal{L}$。对于类别特征，将检查特征顺序的所有排列。[11] 对于连续特征，则检查给定等距步数下最短的线性路径。Kuratomi 等人 [16] 使用了 DBSCAN，但其最大距离阈值的选择存在歧义。识别算法的复杂度为 $O(c! \cdot c + o! \cdot \sum_{j=1}^{o} k_j + q)$，其中$c$和 o 分别表示名义型和有序型特征的数量， k_j 表示有序型特征 X_j 的可能取值数量， q 表示连续特征的检查步数。

可以通过尝试在未连接点与已连接点之间建立联系，进一步丰富水平集。不过，在比较 IRD 方法时，凸水平集已经足够，因为超盒本身就是凸的。

模型的超参数调优

对于超参数调优，我们采用了随机搜索（进行 15 次评估）和 5 折交叉验证（CV），并使用误分类误差（分类任务）或均方误差（回归任务）作为性能评估指标。表 5 展示了各个模型的调优搜索空间。相对有限的调优设置应当是……

[11] 如果排列数超过 100，则随机选择 100 种特征顺序。

足以完成我们解释预测模型的任务——准确度较低的模型并不会构成障碍。对于tic_tac_toe、糖尿病 和 避孕方法选择 等不平衡数据集，使用 SMOTE 算法 [1]进行了平衡。对于 SMOTE，数值特征经过标准化，类别型特征则进行了独热编码。神经网络的优化器为 ADAM [15]，训练 500 个 epoch。所有其他超参数均采用 mlr3keras R 包的默认值 [20]（层单元数除外，见表 5）。表 6 展示了使用嵌套重采样评估各模型的准确率，内层和外层循环均采用 5 折交叉验证。

表 5：各模型的调优搜索空间。树的数量的超参数值经过了对数转换。

模型	超参数范围	
随机森林	树的数量	[1, 1000]
逻辑回归	-	-
线性模型	-	-
多项式模型	-	-
超盒/rpart	-	-
神经网络	层_单元	[1, 20]

表 6：各模型在各数据集上的分类误差或均方误差（回归）。性能通过嵌套重采样计算，内层和外层循环均采用 5 折交叉验证。我们未测量（终端节点）超盒模型的性能，因为每个$\mathbf{x}'$对应的模型都不同。

	随机森林	线性模型	神经网络	超盒
糖尿病	0.233	0.224	0.229	-
tic_tac_toe	0.036	0.019	0.094	-
cmc	0.466	0.495	0.389	-
车辆	0.256	0.201	0.254	-
no2	33502.856	37678.319	77866.331	-
血浆_视黄醇	45391.218	59224.452	297481.249	-

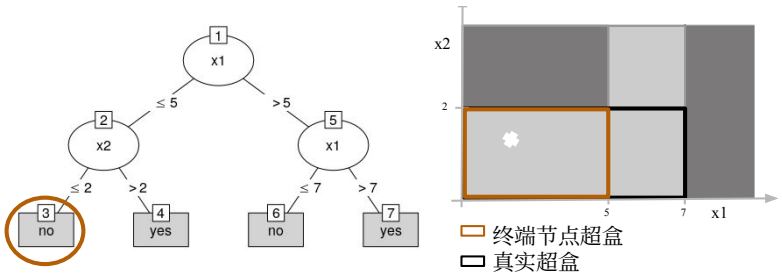


图6：CART树的真实超盒与终端节点超盒。白色十字标记对应$\mathbf{x}'$。

I 基准测试 - 附加结果

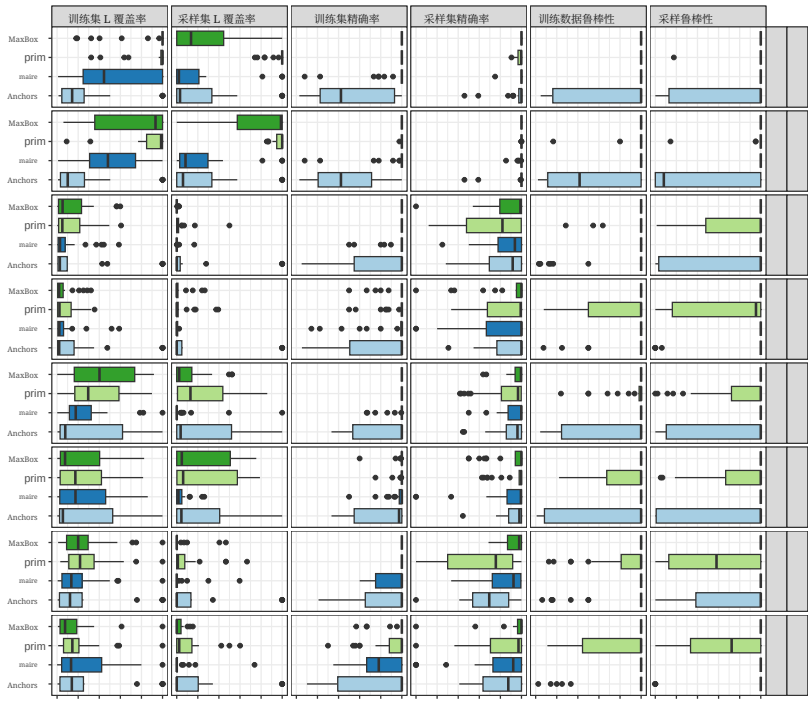


图7：分别比较每个模型的 MaxBox、PRIM、Anchors 和 MAIRE 在覆盖率与精确率方面的表现。每种方法均在¯训练数据（traindata）或从 B_ 均匀采样的数据（sampled）上运行或评估，未进行后处理。精确率和覆盖率越高越好。

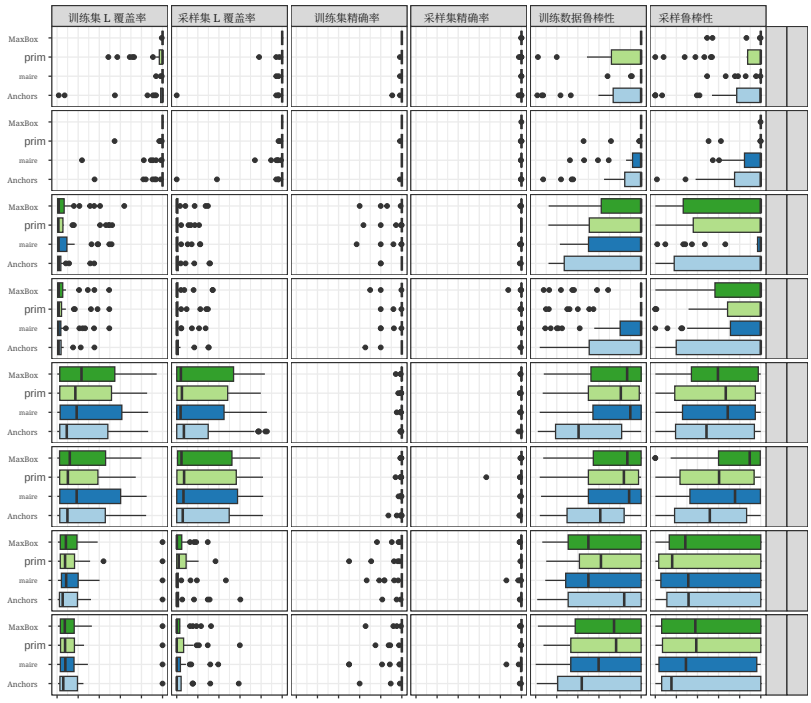


图 8：分别比较每个模型的 MaxBox、PRIM、Anchors 和 MAIRE 在覆盖率和精确率方面的表现。每种方法要么在训练数据（traindata）上运行或评估¯，要么在来自 B_（采样）的均匀采样数据上运行或评估并进行后处理。精确率和覆盖率越高越好。

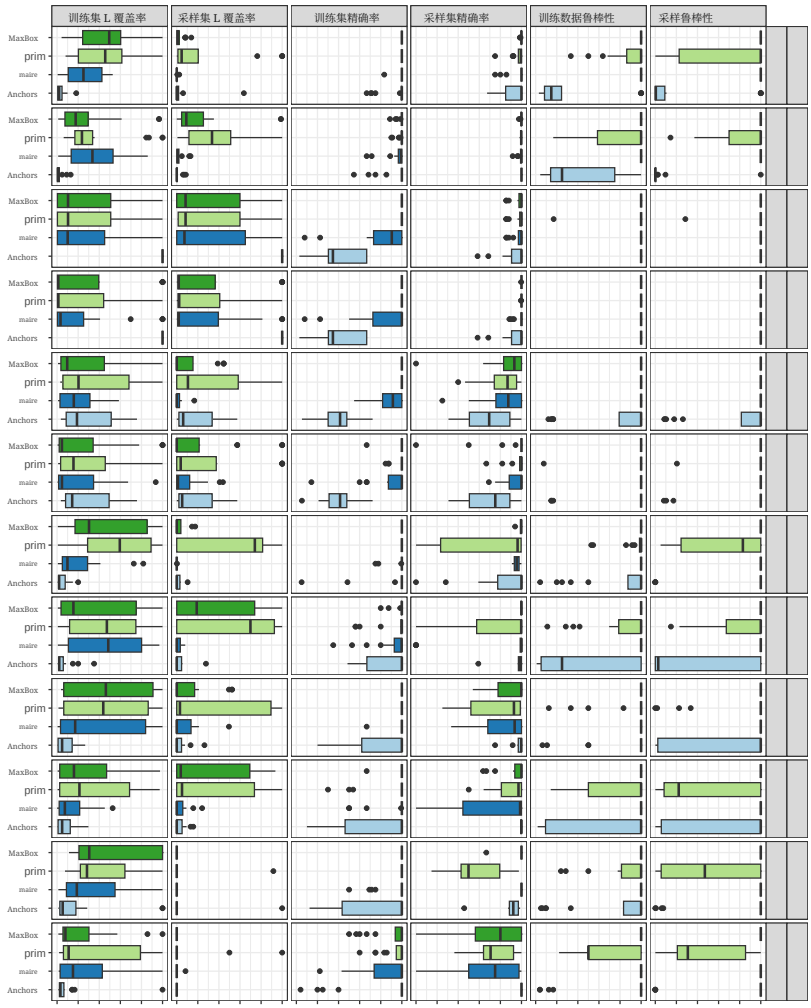


图9：分别比较各个数据集上的 MaxBox、PRIM、Anchors 和 MAIRE 的覆盖率与精确率。每种方法要么在¯训练数据（训练数据）上运行或评估，要么在来自 B_ 的均匀采样数据（采样数据）上运行或评估，不进行后处理。精确率和覆盖率越高越好。

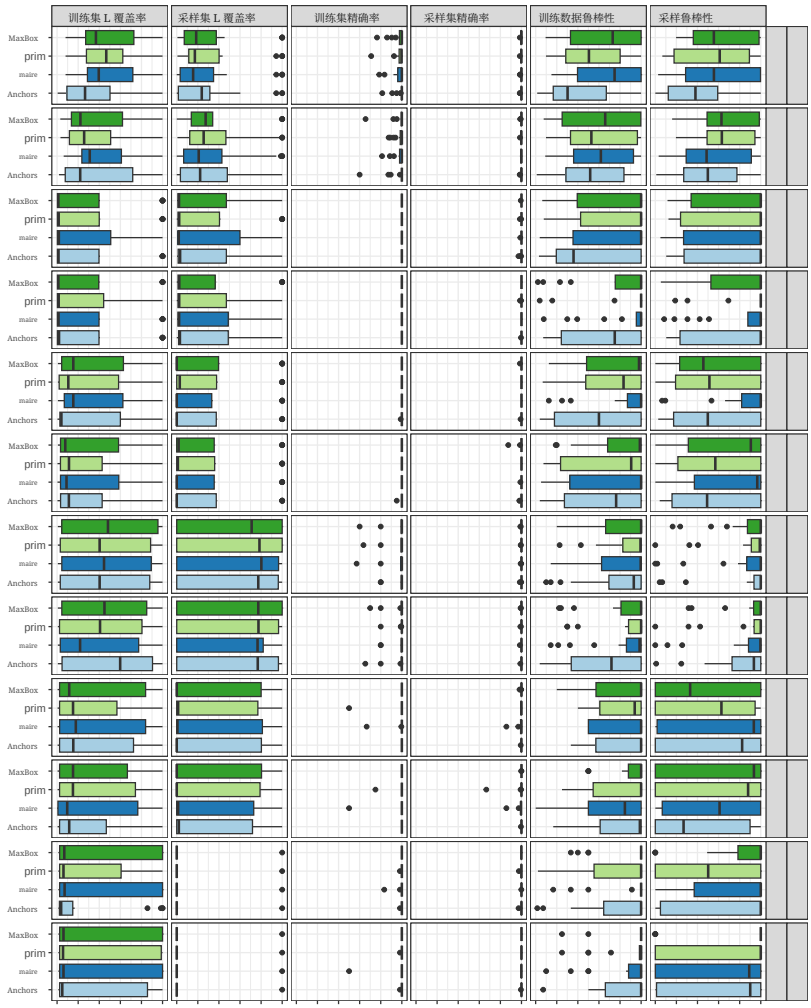


图10：分别比较各数据集上 MaxBox、PRIM、Anchors 和 MAIRE 的覆盖率与精确率。每种方法均在训练数据（traindata）或从 B_ (sampled) 均匀采样的数据上运行或评估¯并进行后处理。精确率和覆盖率越高越好。

Table 7: Statistical analysis of RQ 1. Pairwise comparison of MaxBox, PRIM, Anchors, and MAIRE w.r.t. coverage and precision. Each value corresponds to the p-value obtained for the Wilcoxon rank-sum test with H_0 that the performances do not differ. Cells printed in bold font correspond to p-values that are lower than $\alpha = 0.05/36$ (Bonferroni-adjustment) and indicate that one method outperforms the other. Only methods run on the training data without post-processing were compared.

measure	MaxBox = PRIM	MaxBox = Anchors	MaxBox = MAIRE	PRIM = Anchors	PRIM = MAIRE	Anchors = MAIRE
coverage_train	0.761	0.618	**0**	0.579	**0**	0.473
coverage_sampled	**0**	0.044	**0**	**0**	**0**	**0**
coverage_$\mathcal{L}$_train	0.431	**0.001**	**0**	**0**	**0**	0.127
coverage_$\mathcal{L}$_sampled	**0**	0.035	0.004	0.059	**0**	**0**
precision_train	1	**0**	**0**	**0**	**0**	**0**
precision_sampled	0.025	**0**	0.623	0.042	0.104	0.004

Table 8: Statistical analysis of RQ 2. Pairwise comparison of using training data vs. sampled data for $\bar{\mathbf{X}}$. Each value corresponds to the p-value obtained for the Wilcoxon rank-sum test with H_0 that the performance of methods using training data is better than the performance of methods using sampled data. Cells printed in bold font correspond to p-values that are lower than $\alpha = 0.05/30$ (Bonferroni-adjustment) and indicate a preference towards using sampled data. Comparisons were only conducted for the methods run without post-processing.

measure	overall	MaxBox	PRIM	Anchors	MAIRE
coverage_train	1.00	1	0.999	0.445	0.744
coverage_sampled	0.00	**0**	0.987	0.402	0.003
coverage_$\mathcal{L}$_train	1.00	1	1	0.781	0.896
coverage_$\mathcal{L}$_sampled	0.00	**0**	0.236	0.476	0.172
precision_train	1.00	0.995	0.998	0.782	0.993
precision_sampled	0.00	0.011	**0**	**0.001**	0.381

