学习 最大化动态特征选择的互信息

Ian Covert [1] **Wei Qiu** [1] **Mingyu Lu** [1] **Nayoon Kim** [1] **Nathan White** [2] **Su-In Lee** [1]

arXiv:2301.00557v2 [cs.LG] 8 Jun 2023

摘要

特征选择有助于降低机器学习中的数据获取成本，但标准做法是使用静态特征子集训练模型。本文考虑动态特征选择 （DFS）问题，在该问题中，模型根据当前可用信息依次查询特征。DFS 通常通过强化学习来解决，但我们探索了一种更简单的方法：根据条件互信息贪心地选择特征。这种方法在理论上颇具吸引力，但需要访问数据分布的预言机，因此我们开发了一种基于摊销优化的学习方法。研究表明，所提出的方法在训练至最优时能够恢复贪心策略；在实验中，该方法的表现优于众多现有特征选择方法，从而验证了它是解决这一问题的一种简单而强大的方法。

1. 引言

机器学习模型的输入可能需要付出高昂成本才能获取，因此人们常用特征选择来降低数据获取成本。在信息按序收集的应用中，一种自然的做法是根据当前可用的信息自适应地选择特征，而不是使用固定的特征集。这种设置称为动态特征选择 （DFS），[1] 过去十年间已有多项研究探讨了这一问题（Saar-Tsechansky 等人，2009年；Dulac-Arnold 等人，2011；Chen 等人，2015b；Early 等人，2016a；He 等人，2016a；Kachuee 等人，2018）。

与使用固定特征集的静态特征选择相比（Li 等，2017年；Cai 等，2018），DFS 在固定预算下能够提供更好的性能。这很容易理解，因为当信息量最大的特征因人而异时，为所有样本选择相同的特征（例如，所有前往医院急诊室就诊的患者）并非最优。尽管从理论上讲，DFS 应能带来更好的性能，但它也带来了更具挑战性的学习问题，因为它既要求学习（i）特征选择策略，也要求学习如何根据可变的特征集进行预测。

以往研究以多种方式处理动态特征选择（DFS），但通常采用强化学习（RL）（Dulac-Arnold 等人，2011；Shim 等人，2018；Kachuee 等人，2018；Janisch 等人，2019；Li & Oliva，2021）。强化学习是解决序贯决策问题的一种自然方法，但现有方法难以训练，而且无法稳定地优于静态特征选择方法（Henderson 等人，2018；Erion 等人，2021）。因此，我们的工作探索了一种更简单的方法：根据特征与响应变量之间的条件互信息（CMI）贪心地选择特征。

贪心方法已在以往研究中得到应用（Chen 等人，2015b；Ma 等人，2019），但由于计算条件互信息需要对数据分布进行预言机访问，因此在实践中难以使用（Cover & Thomas，2012）。因此，我们的重点是开发一种实用的近似方法。以往研究要么对数据作出较强假设（Geman & Jedynak，1996），要么使用生成模型近似数据分布（Ma 等人，2019）；而我们开发了一种灵活的方法，能够直接预测每一步的最优选择。我们的方法基于对贪心条件互信息策略的变分视角，并采用一种称为摊销优化 （Amos，2022）的技术，从而仅使用标准带标签数据集即可进行训练。值得注意的是，当模型训练至最优时，其目标函数能够恢复出贪心策略。

本研究的贡献如下：

1. 我们从变分角度（或基于优化的角度）推导了贪心条件互信息策略，表明该策略等价于在给定最优分类器的情况下最小化一步预测损失。

2. 我们提出一种基于摊销优化的训练方法，通过训练策略网络，直接预测每一步的最优选择。我们的方法不需要包含正确选择信息的数据集，而是

[1]华盛顿大学保罗·G·艾伦计算机科学与工程学院 [2]华盛顿大学急诊医学系。通讯作者：Ian Covert< icovert@cs.uw.edu>。

40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023，作者所有。[1]该问题也被称为序列特征选择、主动特征获取和信息追寻。

基于标准带标签数据集和一个目标函数，其全局优化器即为贪心条件互信息策略。

3. 我们针对本质上离散的学习目标提出了一种连续松弛方法，从而能够进行高效且与架构无关的基于梯度的优化。

我们的实验在众多数据集上评估了所提出的方法，结果表明，该方法优于许多近期的动态和静态特征选择方法。总的来说，我们的工作表明，只要经过适当学习，贪心条件互信息策略就是一种简单而强大的动态特征选择方法。

2. 问题表述

本节将介绍动态特征选择问题，并说明全文使用的符号。

2.1. 符号

令$\mathbf{x}$ 表示输入特征向量，$\mathbf{y}$ 表示监督学习任务中的响应变量。输入包含 d 个不同的特征，即$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$。我们用记号 $s \subseteq [d] \equiv \{1, \ldots, d\}$ 表示索引子集，用$\mathbf{x}_s = \{\mathbf{x}_i$ 表示特征子集。粗体符号$\mathbf{x}$,$\mathbf{y}$表示随机变量，符号 x, y 表示可能的取值，而 $p(\mathbf{x},\mathbf{y})$ 表示数据分布。

我们的目标是设计一种策略，根据当前可用的信息控制选择哪些特征。选择策略可视为函数 $\pi(x_s) \in [d]$，即接收一个特征子集作为输入，并输出下一个要查询的特征索引。该策略配有一个预测器 $f(x_s)$，可根据可用特征集合进行预测；例如，如果$\mathbf{y}$是离散的，那么预测值位于概率单纯形中；对于 K 个类别，则为 $f(x_s) \in \Delta^{K-1}$。记号 $f(x_s \cup x_i)$ 表示基于组合特征得到的预测。我们首先考虑作用于特征子集的策略函数和预测器函数；第 4 节将提出一种使用掩码变量 $m \in [0,1]^d$ 的实现方式，其中这些函数作用于 $x \odot m$。

2.2. 动态特征选择

动态特征选择的目标是以尽可能少的预算选择特征，同时实现最高的预测准确率。通常，能够使用更多特征会让预测更容易，因此挑战在于选出少量信息丰富的特征。这个问题有多种形式，包括具有非均匀特征成本以及针对每个样本设置不同预算的情形（Kachuee 等人，2018），但我们关注固定预算且成本均匀的设定。我们的目标是在推理时处理预测：从不使用任何特征开始，依次选择特征 x_s 使得 $|s| = k$ 在固定预算 $k < d$下，最终准确预测响应变量 y。

给定一个用于衡量预测结果与标签之间差异的损失函数 $\ell(\hat{y},\ y)$，一种自然的评分标准是在选择 k 特征后的期望损失。该评分应用于策略-预测器对（π，f），对于固定预算 k，我们将评分定义如下：

$$v_k(\pi, f) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\big[\ell\big(f\big(\{\mathbf{x}_{i_t}\}_{t=1}^{k}\big), \mathbf{y}\big)\big], \qquad (1)$$

其中，对于每个（$\mathbf{x}$，$\mathbf{y}$），根据 $i_n = \pi(\{\mathbf{x}_{i_t}\}_{t=1}^{n-1})$依次选择特征索引。目标是最小化$v_k(\pi,\ f)$，或者等价地，最大化最终的预测准确率。

一种方法是将其构建为马尔可夫决策过程（MDP），并使用标准强化学习技术求解，使得π 和 f 经过训练后能够优化基于公式（1）的奖励函数。近期已有多项研究设计了此类公式（Shim 等人，2018；Kachuee 等人，2018；Janisch 等人，2019年；Li 和 Oliva，2021年）。然而，这些方法难以进行有效训练，因此我们的工作聚焦于一种更易学习且更易解释的贪心方法。

3. 贪心信息最大化

本节首先定义贪心条件互信息策略，然后介绍一种依赖生成模型的现有近似策略。

3.1. 贪心选择策略

作为动态特征选择的一种理想化方法，我们关注这样一种贪心算法：在每一步都选择信息量最大的特征。特征的信息量可以通过多种方式定义，但我们采用信息论视角，认为最有用的特征是与响应变量具有最大条件互信息（Cover & Thomas，2012）的特征。条件互信息记作$I(\mathbf{x}_i;\mathbf{y} \mid x_s)$，用于量化在考虑当前特征 x_s的情况下，未知特征$\mathbf{x}_i$ 能够提供多少关于响应变量$\mathbf{y}$ 的信息，其定义为联合分布与因子分解分布之间的 KL 散度：

$$I(\mathbf{x}_i; \mathbf{y} \mid x_s) = D_{\mathrm{KL}}\big(p(\mathbf{x}_i, \mathbf{y} \mid x_s) \,||\, p(\mathbf{x}_i \mid x_s) p(\mathbf{y} \mid x_s)\big).$$

基于此，我们将贪心条件互信息策略定义为 $\pi^*(x_s) \equiv \arg\max_i I(\mathbf{x}_i;\mathbf{y} \mid x_s)$，以便依次选择特征，使我们对响应变量的信息量最大化。我们也可以将该策略理解为执行贪心不确定性最小化，因为这等价于在每一步最小化$\mathbf{y}$的条件熵，即$\pi^*(x_s) = \arg\min_i H(\mathbf{y} \mid \mathbf{x}_i, x_s)$（Cover & Thomas，2012年）。为了完整刻画这一理想化方法，我们还考虑将该策略与贝叶斯分类器配合作为预测器，即 $f^*(x_s) = p(\mathbf{y} \mid x_s)$。

在每一步最大化关于$\mathbf{y}$的信息量直观上合理，并且应能在许多问题中取得良好效果。以往研究也考虑过同一思路，但视角与我们不同。首先，Chen 等人（2015b）从理论角度出发，证明贪心算法的性能与最优策略相比，仅相差一个乘法因子；该证明需要特定的分布假设，但我们发现，贪心算法在许多真实世界数据集上表现良好（第 6 节）。其次，从实现角度看，有两项工作旨在提供实用的近似方法；然而，这些方法存在若干局限，因此我们的工作旨在开发一种简单灵活的替代方案（第 4 节）。在这些工作中，Ma 等人（2019）和 Chattopadhyay 等人（2022）都需要数据分布的条件生成模型，我们将在下文讨论这一点。

3.2. 估计条件互信息

如果能够直接计算条件互信息，那么实现贪心策略就轻而易举，但在实践中这种情况很少见。另一种做法是对其进行估计。下面介绍一种针对每个特征迭代执行的估计过程；暂且假设我们可以预言机访问所有 $s \subseteq [d]$ 对应的响应分布 $p(\mathbf{y} \mid \mathbf{x}_s)$，以及所有 $s \subseteq [d]$ 和 $i \in [d]$ 对应的特征分布 $p(\mathbf{x}_i \mid \mathbf{x}_s)$。

在选择过程的任意时刻，给定当前特征 x_s，我们可以按如下方式估计特征 $\mathbf{x}_i$ 的条件互信息，其中$i \notin s$。首先，我们可以从其条件分布中为 $\mathbf{x}_i$ 抽取多个值；对于$j \in [n]$，也可以从 $x_i^j \sim p(\mathbf{x}_i \mid x_s)$中抽取。接下来，我们可以为每个抽样值生成贝叶斯最优预测，即 $p(\mathbf{y} \mid x_s, x_i^j)$。最后，我们可以计算平均预测，以及相对于该预测的平均 KL 散度，从而得到以下条件互信息估计量：

$$I_i^n = \frac{1}{n}\sum_{j=1}^{n} D_{\mathrm{KL}}\Big(p(\mathbf{y} \mid x_s, x_i^j) \;||\; \frac{1}{n}\sum_{l=1}^{n} p(\mathbf{y} \mid x_s, x_i^l)\Big). \tag{2}$$

该评分衡量预测之间的变化性，并捕捉不同的$\mathbf{x}_i$值是否会显著影响$\mathbf{y}$的条件分布。该估计量可用于选择特征；或者，由于以下极限结果（见附录 A），我们可以令 $\pi(x_s) = \arg\max_i I_i^n$：

$$\lim_{n\to\infty} I_i^n = I(\mathbf{y}; \mathbf{x}_i \mid x_s). \tag{3}$$

因此，该方法通过估计条件互信息，提供了一种识别正确贪心选择的途径。先前的研究曾探索过类似思路，即根据采样得到的预测为特征评分（Saar-Tsechansky 等人，2009年；Chen 等人，2015a；Early 等人，2016a；b），但这些研究中的实现选择使其无法进行贪心信息最大化。在公式（2）中，我们的估计器使用贝叶斯分类器、从条件分布 $p(\mathbf{x}_i \mid x_s)$ 中采样特征，并使用 KL 散度来衡量预测变异性，这些都很重要。然而，这种估计器并不实用，因为我们通常无法同时获取 $p(\mathbf{y} \mid \mathbf{x}_s)$ 和 $p(\mathbf{x}_i \mid \mathbf{x}_s)$。

在实践中，我们需要为每个分布学习相应的替代模型。例如，我们可以使用一个分类器来近似 $f(x_s) \approx p(\mathbf{y} \mid x_s)$，并使用一个生成模型来近似从 $p(\mathbf{x}_i \mid \mathbf{x}_s)$ 中采样。类似地，Ma 等人（2019）提出联合建模 $(\mathbf{x},\mathbf{y})$，并通过改进的变分自编码器（Kingma 等人，2015）实现条件生成模型。由于多种原因，这种方法存在局限，包括：(i) 难以训练准确的条件生成模型；(ii) 难以对混合的连续/类别特征建模（Ma 等人，2020；Nazabal 等人，2020）；以及 (iii) 条件互信息估计过程缓慢。在接下来的方法中，我们通过直接预测每一步的最佳选择，避开了这三项挑战。

4. 所提方法

现在我们介绍所提出的方法：一种通过摊销优化训练的贪心策略实用近似。不同于以往将条件互信息作为中间步骤进行估计的工作，我们从变分视角研究贪心策略，并以此训练一个网络，使其能够根据当前特征直接预测最优选择。

4.1. 条件互信息的变分视角

对我们的目的而言，认识到贪心策略可以视为某个优化问题的解很有帮助。第 3 节将条件互信息常规定义为 KL 散度，但这很难融入端到端学习方法。相反，我们现在考虑由策略 π 和预测器 f实现的一步前瞻预测，并确定使其损失最小化的行为。给定当前特征 x_s 和一个选择$i = \pi(x_s)$，期望的一步前瞻损失为：

$$\mathbb{E}_{\mathbf{y},\mathbf{x}_i \mid x_s}\Big[\ell\big(f(x_s \cup \mathbf{x}_i), \mathbf{y}\big)\Big]. \tag{4}$$

我们在此提出的变分视角包含关于这一期望损失的两个主要结果。第一个结果与预测器有关，我们证明，最小化损失的预测器可以独立于策略 π定义。我们在以下命题中针对分类任务对此进行形式化说明；我们的结果也可以推广到回归任务（证明见附录 A）。

命题 1. 当$\mathbf{y}$ 为离散变量且 ℓ 为交叉熵损失时，对于任意策略 π，式 (4) 均由贝叶斯分类器最小化，即 $f^*(x_s) = p(\mathbf{y} \mid x_s)$。

这一性质要求在选择特征时，不掌握其余特征或响应变量的信息。这一假设适用于动态特征选择（DFS），但不适用于依据完整特征集进行选择的情形（Chen 等人，2018；Yoon 等人，2018；Jethani 等人，2021）。现在，假设我们使用贝叶斯分类器 f^* 作为预测器，我们的第二个结果涉及选择策略。如下文所示，损失最小化策略等价于基于条件互信息进行选择。

命题 2。当 $\mathbf{y}$ 为离散变量时，ℓ 为交叉熵损失，且预测器为贝叶斯分类器 f^*，则贪心条件互信息策略可使式(4)达到最小值，即 $\pi^*(x_s) \in \arg\max_i I(\mathbf{y};\mathbf{x}_i \mid x_s)$。

由此可见，第 3 节中定义的贪心条件互信息策略等价于最小化一步预测损失。接下来，我们利用这一变分视角，开发一种联合学习流程，用于学习策略和预测器网络。

4.2. 摊销优化方法

我们不再通过估计每个特征的条件互信息来确定下一步的选择，而是开发一种能够直接预测每一步最佳选择的方法。贪心策略隐含地要求针对每次选择求解一个优化问题，即 $\arg\max_i I(\mathbf{y}, \mathbf{x}_i; x_s)$，但由于无法访问这个目标函数，我们现在提出一种直接预测解的方法。遵循一种称为摊销优化的方法（Amos，2022 年），我们将第 4.1 节中关于条件互信息的变分视角构造成一个由可学习网络优化的目标函数。

首先，由于这有助于基于梯度的优化，我们现在假设策略输出一个关于特征索引的分布。稍微滥用一下记号，本节将策略视为函数 $\pi(x_s) \in \Delta^{d-1}$，这是对先前定义 $\pi(x_s) \in [d]$ 的推广。采用这一随机策略版本，我们现在可以将目标函数表述如下。

令选择策略由神经网络 $\pi(\mathbf{x}_s;\phi)$ 参数化，预测器由神经网络 $f(\mathbf{x}_s;\theta)$ 参数化。令 $p(\mathbf{s})$ 表示一个支撑集包含所有子集的分布，或者对所有 $|s| < d$，均有 $p(s) > 0$。那么，目标函数 $\mathcal{L}(\theta,\phi)$ 定义为

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\mathbb{E}_{p(\mathbf{s})}\Big[\mathbb{E}_{i\sim\pi(\mathbf{x_s};\phi)}\big[\ell\big(f(\mathbf{x_s}\cup\mathbf{x}_i;\theta),\mathbf{y}\big)\big]\Big]. \tag{5}$$

直观地说，式 (5) 表示生成一个随机特征集 $\mathbf{x}_s$，根据 $i \sim \pi(\mathbf{x}_s;\phi)$ 对一个特征索引进行采样，然后衡量预测 $f(\mathbf{x}_s \cup \mathbf{x}_i;\theta)$ 的损失。因此，我们的目标优化的是单独的选择和预测，而非整个轨迹，这使我们能够基于命题 1-2 进行构建。我们将其描述为一种贪心方法的实现，因为当训练至最优时，它会恢复贪心条件互信息选择。在分类情形下，我们在一个温和的假设下证明了以下结果：最优选择是唯一的。

定理 1. 当 $\mathbf{y}$ 为离散变量且 ℓ 为交叉熵损失时，式 (5) 的全局最优解是满足 $f(x_s;\theta^*) = p(\mathbf{y} \mid x_s)$ 的预测器，以及将全部概率质量置于 $i^* = \arg\max_i I(\mathbf{y};\mathbf{x}_i \mid x_s)$ 上的策略 $\pi(x_s;\phi^*)$。

如果放宽最优选择唯一这一假设，最优策略 $\pi(\mathbf{x}_s;\phi^*)$ 只需在最佳索引之间分配概率质量。回归情形也有类似结果，此时可以将贪心策略理解为执行条件方差最小化。

定理 2. 当 $\mathbf{y}$ 为连续变量且 ℓ 为平方误差损失时，式(5) 的全局最优解是一个满足 $f(x_s;\theta^*) = \mathbb{E}[\mathbf{y} \mid x_s]$ 的预测器，以及一个将全部概率质量分配给 $i^* = \arg\min_i \mathbb{E}_{\mathbf{x}_i|x_s}[\mathrm{Var}(\mathbf{y} \mid \mathbf{x}_i, x_s)]$ 的策略 $\pi(x_s;\phi^*)$。

这些结果的证明见附录 A。需要注意的是，每个模型的函数类都必须具有足够的表达能力，以包含各自的最优解；并且该结果对所有支持集覆盖全部子集的 $p(\mathbf{s})$ 均成立。

与第 3.2 节中的条件互信息估计流程相比，这种方法有两个主要优点。首先，我们无需对所有 (s,i) 建模特征条件分布 $p(\mathbf{x}_i \mid \mathbf{x}_s)$。对这些分布建模是一个困难的中间步骤，而我们的方法则旨在直接输出最优索引。其次，我们的方法速度更快，因为每次选择只需进行一次前向传播：使用 Ma 等人（2019）的方法选择 k 个特征需要 $\mathcal{O}(dk)$ 个评分步骤，而我们的方法只需通过策略网络 $\pi(\mathbf{x}_s;\phi)$ 进行 k 次前向传播。

此外，与通过强化学习训练的策略相比，贪心方法更容易学习。我们的训练过程可以视为一种奖励塑形（Sutton 等，1998年；Randløv 和 Alstrøm，1998年），其中奖励会考虑每一步之后的损失，并为每次选择是否有益提供强烈信号。相比之下，只有在选择 k 特征后才观察到奖励，这向策略网络提供的信号相对较弱（见式（1））。强化学习方法通常面临颇具挑战的探索-利用权衡，但学习贪心策略更简单，因为它只需在每一步找到局部最优选择。

4.3. 使用连续松弛进行训练

我们的公式 (5) 中的目标在得到完美优化时会产生正确的贪心策略，但 $\mathcal{L}(\theta, \phi)$ 很难通过梯度下降进行优化。特别是，给定一个采-

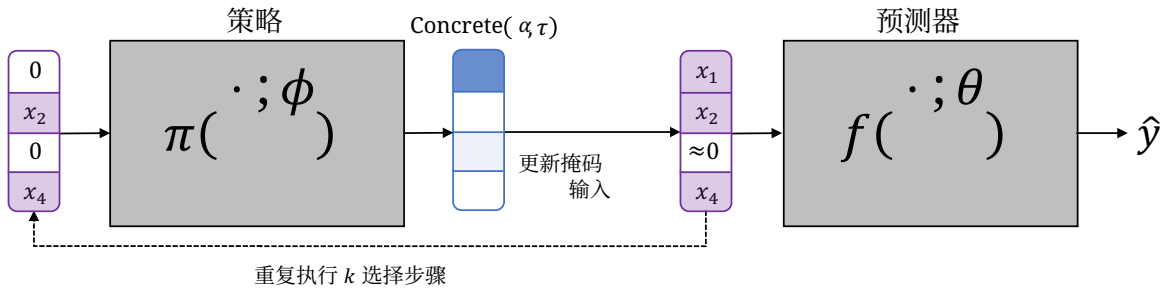

图 *1.* 我们的训练方法示意图。左：通过使用掩码输入反复调用策略网络来选择特征。右：每次选择后，使用预测器网络进行预测。执行梯度下降时，只有实线参与反向传播。

样索引 $i \sim \pi(\mathbf{x}_S; \phi)$ 时，很难通过策略网络传播梯度。REINFORCE 技巧（Williams，1992年）是获得随机梯度的一种方法，但较大的梯度方差可能使其在许多问题中效果不佳。针对如何减少这一设置下梯度方差的问题，已有大量成熟的研究（Tucker 等，2017年；Grathwohl 等，2018），但我们提出使用一种简单的替代方案：Concrete 分布（Maddison 等，2016）。

根据 $i \sim \pi(x_S; \phi)$ 采样的索引，可以用一个指示所选索引的独热向量 $m \in \{0,1\}^d$ 表示；而使用 Concrete 分布时，我们采样的是概率单纯形中的一个近似独热向量，或 $m \in \Delta^{d-1}$。这种连续松弛使我们能够使用重参数化技巧计算梯度（Maddison 等，2016；Jang 等，2016）。将子集 $s \subseteq [d]$ 松弛为连续向量，也需要对策略函数和预测器函数进行松弛，因此我们让它们作用于掩码输入 x，即逐元素乘积 $x \odot m$。为避免混淆特征值为零还是被掩码的情况，我们还可以将掩码作为模型输入传入。

使用 Concrete 分布进行训练时，需要指定温度参数 $\tau > 0$，以控制样本的离散程度。以往的研究通常采用固定温度进行训练，或在预先设定的训练轮次数内逐步退火（Chang 等人，2017年；Chen 等人，2018；Balın 等人，2019年），而我们则使用一系列 τ 值进行训练，并在每个步骤执行提前停止。这样一来，温度和训练轮次数便不再是需要调整的重要超参数。我们的训练流程总结于图 1，并由算法 1 进行了更详细的说明。

此外，我们还发现，以下几个可选步骤有助于改进优化：

- 预测器和策略网络可以共享参数 $f(\mathbf{x}; \theta)$ $\pi(\mathbf{x}\ \phi)$。这不会使二者的联合优化变得复杂；在早期层学习共享表示，有时还能帮助网络更快地完成优化（例如，处理图像数据时）。
- 我们不使用随机子集分布进行训练 $p(\mathbf{s})$，而是使用当前策略 $\pi(\mathbf{x}; \phi)$ 选出的特征来生成子集。这样，模型便能专注于推理时可能遇到的子集，同时不会影响全局最优策略/预测器：选择之间不会传播梯度，因此，公式 (5) 和这种采样方法都将每个特征集视为独立的优化问题，只是权重不同（见附录 D）。
- 在联合训练策略-预测器对之前，我们先使用随机子集对预测器 $f(\mathbf{x}; \theta)$ 进行预训练。与从随机初始化开始优化 $\mathcal{L}(\theta, \phi)$ 相比，这种做法效果更好，因为随机预测器 $f(\mathbf{x}; \theta)$ 无法向 $\pi(\mathbf{x}; \phi)$ 提供有关哪些特征有用的信号。

5. 相关工作

以往研究常使用强化学习来解决动态特征选择问题。例如，Dulac-Arnold 等人（2011）；Shim 等人（2018）；Janisch 等人（2019）；Li 和 Oliva（2021）优化基于最终预测准确率的奖励，而 Kachuee 等人（2018）则使用一种考虑预测不确定性的奖励。强化学习是解决序贯决策问题的自然方法，但在实践中可能难以优化：强化学习需要复杂的训练流程、收敛速度慢，而且对初始化高度敏感（Henderson 等人，2018）。因此，基于强化学习的动态特征选择并不能可靠地优于静态特征选择，这一点已由 Erion 等人（2021）展示，并在我们的实验中得到证实。

其他方法包括模仿学习（He 等人，2012；2016a）和迭代特征评分方法（Melville 等人，2004；Saar-Tsechansky 等人，2009；Chen 等人，2015a；Early 等人，2016b；a）。模仿学习将动态特征选择视为监督分类，而我们的训练方法无需预言机策略。大多数现有的特征评分技术都是贪心方法，与我们的方法类似，但它们采用的评分启发式与最大化条件互信息无关（见第 3.2 节）。有两种特征评分方法专门用于计算条件互信息，但它们存在重要的实际局限：Ma 等人（2019）和

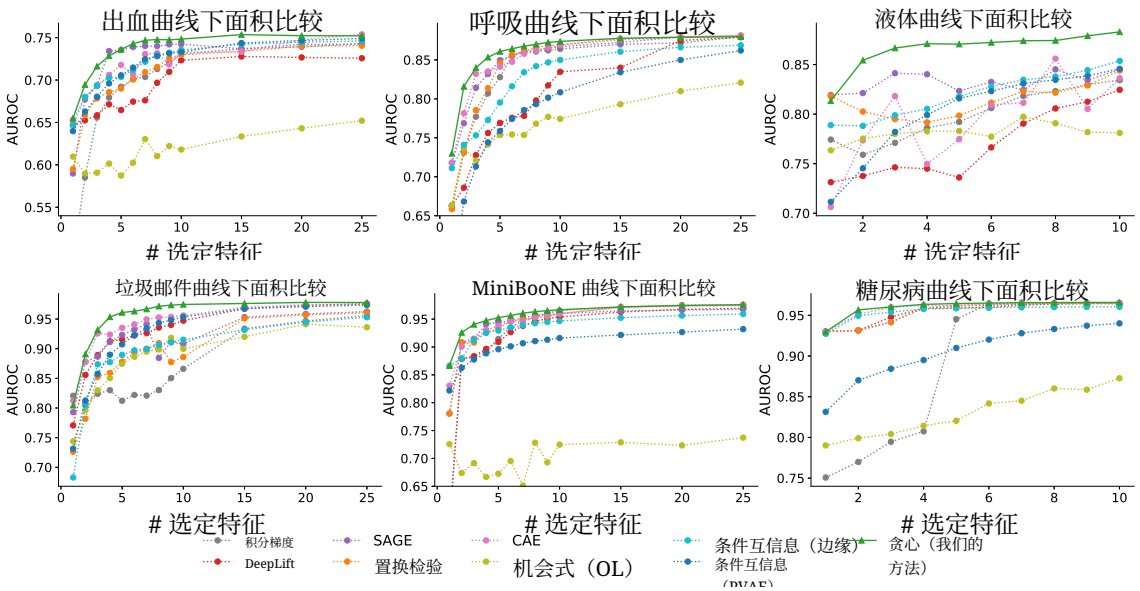

图 2。评估　　在六个表格数据集上评估贪心方法。每种方法的结果均为五次运行的平均值　　s.

Chattopadhyay 等人（2022）都依赖难以训练的生成模型，这可能导致条件互信息估计不准确。我们的方法更简单、更快速，也更灵活，因为选择逻辑由策略网络承载，无需进行生成建模。[2]

静态特征选择是一项长期存在的问题（Guyon 和 Elisseeff, 2003年；Cai 等人，2018）。神经网络尚无默认方法，但一种选择是根据局部或全局重要性分数对特征进行排序（Breiman，2001；Shrikumar 等人，2017年；Sundararajan 等人，2017年；Covert 等人，2020年）。此外，此前的若干研究利用连续松弛，通过梯度下降学习特征选择策略：例如，Chang 等人（2017年）；Balın 等人（2019年）；Yamada 等人（2020年）；Lee 等人（2021年）；Covert 等人（2022年）进行静态特征选择，而 Chen 等人（2018）和 Jethani 等人（2021年）则在能够访问所有特征的前提下进行实例级特征选择。我们的工作采用类似的连续松弛进行优化，但应用于动态特征选择的场景；在该场景中，我们的方法学习的是选择策略，而非静态选择层。

最后，一些研究从理论角度考察了贪心特征选择算法。例如，Das & Kempe (2011)；Elenberg 等人 (2018) 表明，弱次模性意味着静态特征选择能够取得近乎最优的性能。与我们的工作更相关的是，Chen 等人（2015b）发现，相关概念自适应次模性（Golovin & Krause, 2011）在动态特征选择场景中并不成立，但作者在特定的分布假设下给出了性能保证。

[2]与此同时，Chattopadhyay 等人（2023）提出了一种类似的方法，用于预测每一步的最优选择。

6. 实验

现在，我们展示如何在多个数据集上使用贪心方法。我们首先考察不同规模的表格数据集，包括四项医学诊断任务，随后研究两个图像分类数据集。其中一些任务天然适合动态特征选择，其余任务则可用于检验我们方法的有效性。复现我们实验的代码可在线获取：

h
ttps://github.com/iancovert/dynamic-selection。

我们通过与动态和静态特征选择方法进行比较来评估我们的方法。为确保比较一致，我们也只使用适用于神经网络的方法。作为静态基线，我们使用置换检验（Breiman，2001）和 SAGE（Covert 等人，2020），根据特征对模型准确率的重要性为特征排序；此外，还使用 DeepLift（Shrikumar 等人，2017）和积分梯度（Sundararajan 等人，2017）对每次预测计算的分数，并在整个数据集上汇总。随后，我们使用监督式 Concrete 自编码器（CAE，Balın 等人，2019），这是一种先进的静态特征选择方法。作为动态基线，我们使用第 3.2 节所述条件互信息估计程序的两个版本。首先，我们使用 Ma 等人（2019）的 PVAE生成模型对未知特征进行采样；其次，改为从这些特征的边缘分布中采样。在这两种情况下，我们都使用通过随机特征子集训练的分类器进行预测。最后，我们还使用基于强化学习的机会式学习（OL）方法（Kachuee 等人，2018）。附录 C提供了有关基线方法的更多信息。

表 *1.* 预算为 1–10 个特征时的曲线下面积平均值（95% 置信区间）。

		Spam	MiniBooNE	糖尿病	出血	呼吸	液体
Static	积分梯度	82.84 ± 0.68	89.10 ± 0.33	88.91 ± 0.24	66.70 ± 0.27	81.10 ± 0.04	79.94 ± 0.94
	DeepLift	90.16 ± 1.24	88.62 ± 0.30	95.42 ± 0.13	67.75 ± 0.49	76.05 ± 0.35	76.96 ± 0.56
	SAGE	89.70 ± 1.10	92.64 ± 0.03	95.43 ± 0.01	71.34 ± 0.19	82.92 ± 0.26	83.27 ± 0.53
	置换检验	85.64 ± 3.58	92.19 ± 0.15	95.46 ± 0.02	68.89 ± 1.06	81.56 ± 0.28	81.35 ± 1.04
	CAE	92.28 ± 0.27	92.76 ± 0.41	95.91 ± 0.07	70.69 ± 0.57	83.10 ± 0.45	79.40 ± 0.86
Dynamic	机会式（OL）	85.94 ± 0.20	69.23 ± 0.64	83.07 ± 0.82	60.63 ± 0.55	74.44 ± 0.42	78.13 ± 0.31
	条件互信息（边缘）	86.57 ± 1.54	92.21 ± 0.40	95.48 ± 0.05	70.57 ± 0.46	79.62 ± 0.62	81.97 ± 0.93
	条件互信息 (PVAE)	89.01 ± 1.40	88.94 ± 1.25	90.50 ± 5.16	70.17 ± 0.74	74.12 ± 3.50	80.27 ± 1.02
	贪心（我们的方法）	**93.91 ± 0.17**	**94.46 ± 0.12**	**96.03 ± 0.02**	**72.64 ± 0.31**	**84.48 ± 0.08**	**86.59 ± 0.25**

6.1. 表格数据集

我们首先将方法应用于源自急诊医学场景的三项医学诊断任务。这些任务包括：根据纤维蛋白原浓度偏低预测患者的出血风险（出血）；判断患者是否需要气管插管以获得呼吸支持（呼吸）；以及判断患者是否会对液体复苏产生反应（液体）。有关数据集的更多详情，请参阅附录B。在每种情形下，由于时间和资源限制，在推理时收集所有可能的输入都很困难，因此动态特征选择是自然的解决方案。

所有方法均使用全连接网络，并通过丢弃来减少过拟合（Srivastava 等人，2014年）。图 2（上）展示了各方法在不同特征预算下的结果。分类准确率以曲线下面积衡量；在全部三项任务中，贪心方法在几乎所有特征预算下都取得了最佳结果。在基线方法中，有几种静态方法有时表现接近，但条件互信息估计方法很少具有竞争力（Ma 等人，2019年）。此外，在线学习的结果不稳定且表现较弱。选择少量特征时，贪心方法的优势通常最大；准确率趋于饱和后，这一优势通常会缩小。

接下来，我们使用三个公开的表格数据集开展实验：垃圾邮件分类（Dua 和 Graff，2017年）、粒子识别（MiniBooNE）（Roe 等人，2005年）和糖尿病诊断（Miller，1973年）。糖尿病任务是动态特征选择的一项自然应用，并已用于先前的研究（Kachuee 等人，2018）。我们再次测试了不同数量的特征，图 2（下）展示了各特征预算对应的曲线下面积图。在这些任务中，贪心方法在几乎所有特征数量下再次取得最高准确率。表 1 汇总了特征数量从 $k = 1$到10 时的平均曲线下面积结果，进一步凸显了贪心方法在全部六个数据集上的优势。附录 E 展示了更大版本的曲线下面积曲线（图 4 和图 5），以及展示各数据集中选择结果变化情况的图。

这些数据集上的结果表明，或许出乎意料的是，动态方法可能不如静态方法。有趣的是，以往的研究并未强调这一点，因为它们没有测试强有力的静态基线（Kachuee 等人，2018；Janisch 等人，2019）。例如，在线学习在这些数据集上缺乏竞争力，而条件互信息估计方法的两个版本也并非始终位列基线方法前列。原则上，动态方法有能力取得更好的表现，因此这些方法结果不佳，凸显了同时学习选择策略和适用于多个特征集的预测函数的难度。在这些实验中，我们的方法是唯一成功做到这两点的动态方法。

6.2. 图像分类数据集

接下来，我们考虑了两个标准的图像分类数据集：MNIST（LeCun 等人，1998年）和 CIFAR-10（Krizhevsky 等人，2009年）。我们的目标是从一幅空白图像开始，依次显示多个像素或图像块，最终仅使用图像的一小部分进行分类。尽管这并不是 DFS 的明显应用场景，但它对我们的方法而言是一个具有挑战性的问题，并且此前已有几项工作研究了类似任务（Karayevet al.，2012年；Mnih 等人，2014年；Early etal.，2016a；Janisch 等人，2019年）。

对于 MNIST，我们对策略和预测器均采用全连接架构，并将像素视为单独的特征；因此，我们有 $d = 784$。对于 CIFAR-10，我们为策略和预测器网络采用共享的 ResNet 骨干网络（He 等人，2016b），并且每个网络使用各自的输出头。这些 32×32 图像被粗化为$d = 64$ 大小为 4×4 的图像块，因此选择器头会生成与每个图像块对应的 logits，而预测器头会生成每个类别的概率。

图 3 展示了我们的方法在不同特征预算下的准确率。对于 MNIST，我们采用之前的基线方法，但由于计算成本较高，排除了条件互信息估计方法：评估大量候选特征时，该方法会变得很慢。我们观察到，我们的方法带来了显著提升，尤其是在只进行少量选择时。仅用 10 个像素，我们的贪心方法就达到了近 90% 的准确率，比表现最好的基线方法高出约 10%，并且远高于先前的工作（Balın 等人，2019年；Yamada 等人，2020年；Covert 等人，2020年），

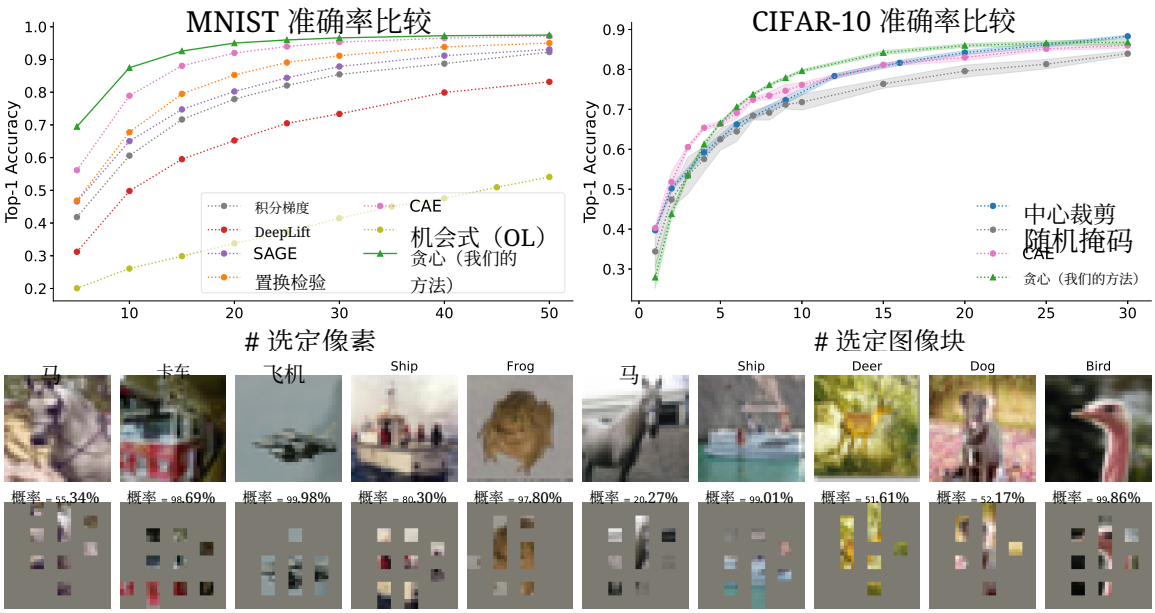

图 *3*。 用于图像分类的贪心特征选择。左上：MNIST 上的准确率比较，结果取五次运行的平均值。右上：CIFAR-10 上的准确率比较，附有 95% 置信区间。下方：针对 CIFAR-10 图像，使用贪心方法从 64 个图像块中选取 10 个后的示例选择结果和预测。

在线学习（OL）的结果最差，而且由于状态数量庞大，训练速度也很慢。

对于 CIFAR-10，由于计算成本较高，我们省略了若干基线方法的比较。我们采用 CAE（这是我们最具竞争力的静态基线），以及两种简单的基线方法：中心裁剪和不同尺寸的随机掩码。对于每种方法，我们绘制五次试验所得的均值和 95% 置信区间。使用极少量图像块时，我们的贪心方法准确率略低；但使用 6–20 个图像块时，其准确率显著更高。最后，图 3（下方）还展示了我们的方法从 64 个图像块中选取 10 个后的预测示例，附录 E 则展示了选取不同数量图像块时的类似图表。

7. 结论

在这项工作中，我们探讨了一种用于动态特征选择（DFS）的贪心算法，该算法根据特征与响应变量之间的条件互信息（CMI）来选择特征。我们提出了一种方法，通过直接预测每一步的最优选择来近似这一策略；实验结果表明，我们的方法优于多种现有的特征选择方法，包括动态和静态基线方法。未来的研究可以考虑纳入不均等的特征成本，或针对每个样本确定理想的特征预算；从理论角度来看，在我们所研究的固定预算情形之外，刻画贪心算法的性能也是未来研究的一个有趣方向（Chen 等人，2015b）。最后，未来的工作还可以探索适用于处理部分输入的架构，尤其是针对图像等结构化数据。

致谢

感谢 Samuel Ainsworth、Kevin Jamieson、Mukund Sudarshan 和 Lee 实验室提供的宝贵讨论。本研究由美国国家科学基金会（NSF）DBI-1552309 和 DBI-1759487 项目，以及美国国立卫生研究院 R35-GM-128638 和 R01-NIA-AG-061132 项目资助。

参考文献

美国国家健康与营养调查，2018。网址：https://www.cdc.gov/nchs/nhanes。

Amos, B. 关于通过摊销优化学习在连续域上进行优化的教程。*arXiv* 预印本 *arXiv:2202.00665*，2022年。

Balın, M. F., Abid, A., 和 Zou, J. Concrete自编码器：可微特征选择与重构。载于国际机器学习会议，第444–453 页。PMLR，2019年。

Breiman, L. 随机森林。《机器学习》，45(1): 5–32, 2001.

Cai, J., Luo, J., Wang, S., 和 Yang, S. 机器学习中的特征选择：一种新视角。Neurocomputing，300:70–79, 2018.

Chang, C.-H.、Rampasek, L. 和 Goldenberg, A. 面向深度学习模型的 Dropout特征排序。*arXiv*预印本 *arXiv:1712.08645*，2017年。

Chattopadhyay, A.、Slocum, S.、Haeffele, B. D.、Vidal, R. 和 Geman, D. 可解释性设计：通过组合可解释查询来学习预测器。*IEEE*模式分析与机器智能汇刊，2022年。

Chattopadhyay, A.、Chan, K. H. R.、Haeffele, B. D.、Geman, D. 和 Vidal, R. 面向可解释预测的变分信息追求。*arXiv*预印本 *arXiv:2302.02876*，2023。

Chen, J.、Song, L.、Wainwright, M. 和 Jordan, M. 学会解释：模型解释的信息论视角。发表于国际机器学习会议，第883–892页。PMLR，2018。

Chen, S.、Choi, A. 和 Darwiche, A. 基于决策稳健性的信息价值。载于*AAAI*人工智能会议论文集，第29卷，2015a。

Chen, Y.、Hassani, S. H.、Karbasi, A. 和 Krause, A. 序贯信息最大化：贪心策略何时近似最优？载于学习理论会议，第338–363页。PMLR，2015b。

Cover, T. 和 Thomas, J. 信息论基础。Wiley，2012年。ISBN9781118585771。

Covert, I.、Lundberg, S. M. 和 Lee, S.-I. 使用加性重要性度量理解全局特征贡献。神经信息处理系统进展，33:17212–17223，2020年。

Covert, I.、Lundberg, S. M. 和 Lee, S.-I. 通过移除进行解释：一个统一的模型解释框架。《机器学习研究杂志》，22:209–1，2021年。

Covert, I., Gala, R., Wang, T., Svoboda, K., Lee, S.-I. 等。用于空间转录组学的预测性和稳健基因选择。*bioRxiv*，2022年。

Das, A. 和 Kempe, D. 次模与谱相遇：用于子集选择、稀疏逼近和字典选择的贪心算法。*arXiv* 预印本 *arXiv:1102.3975*，2011。

Dua, D. 和 Graff, C. UCI机器学习资料库，2017年。网址：http://archive.ics.uci.edu/ml。

Dulac-Arnold, G.、Denoyer, L.、Preux, P. 和 Gallinari, P. 逐数据分类：一种针对稀疏性的序贯方法。载于欧洲机器学习与数据库知识发现联合会议，第 375–390 页。施普林格，2011年。Early, K.、Fienberg, S. E. 和 Mankoff, J. 使用 FOCUS 进行测试时特征排序：以最少的用户负担实现交互式预测。载于*2016* 年 *ACM* 普适与泛在计算国际联合会议论文集，第 992–1003 页，2016a。Early, K.、Mankoff, J. 和 Fienberg, S. E. 在线调查中的动态问题排序。*arXiv* 预印本 *arXiv:1607.04209*，2016b。

Elenberg, E. R., Khanna, R., Dimakis, A. G. 和 Negahban, S. 受限强凸性意味着弱次模性。统计年刊，46(6B)：3539–3568，2018。

Erion, G., Janizek, J. D., Hudelson, C., Utarnachitt, R. B., McCoy, A.M., Sayre,M. R.,White, N.J. 和Lee, S.-I. CoAI：面向医疗保健的成本感知人工智能。*medRxiv*，2021年。

Feng, J. 和 Simon, N. 用于高维非参数回归和分类的稀疏输入神经网络。*arXiv* 预印本 *arXiv:1711.07592*，2017年。

Gal, Y. 和 Ghahramani, Z. 将丢弃作为贝叶斯近似：在深度学习中表示模型不确定性。载于国际机器学习会议，第 1050–1059 页。PMLR，2016。

Geman, D. 和 Jedynak, B. 一种用于卫星图像道路跟踪的主动测试模型。*IEEE*模式分析与机器智能汇刊，18(1):1–14，1996。

Golovin, D. 和 Krause, A. 自适应次模性：主动学习和随机优化中的理论与应用。人工智能研究杂志，42:427–486，2011。

Grathwohl, W.、Choi, D.、Wu, Y.、Roeder, G. 和 Duvenaud, D. 穿过虚空的反向传播：优化用于黑箱梯度估计的控制变量。载于国际学习表征会议，2018。

Guyon, I. 和 Elisseeff, A. 变量与特征选择简介。《机器学习研究杂志》，3（3月）：1157–1182，2003年。

He, H.、Daum´e III, H. 和 Eisner, J. 成本敏感动态特征选择。发表于*ICML Inferning*研讨会，2012年。

He, H.、Mineiro, P. 和 Karampatziakis, N. 主动信息获取。*arXiv* 预印本 *arXiv:1602.02181*，2016a。

He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016b。

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. 重要的深度强化学习。载于*AAAI*人工智能会议论文集，第 32 卷，2018。

Jang, E., Gu, S., and Poole, B. 使用 Gumbel-softmax 的类别重参数化。*arXiv* 预印本 *arXiv:1611.01144*，2016。

Janisch, J., Pevn`y, T., and Lis`y, V. 使用深度强化学习进行高成本特征分类。载于*AAAI*人工智能会议论文集，第 33 卷，第 3959–3966 页，2019年。

Jethani, N., Sudarshan, M., Aphinyanaphongs, Y., 和 Ranganath, R. 我们学会解释了吗？：解释方法如何学会在其解释结果中编码预测。在人工智能与统计国际会议，第 1459–1467 页。PMLR，2021年。Kachuee, M., Goldstein, O., K¨arkk¨ainen, K., Darabi, S., 和 Sarrafzadeh, M. 机会式学习：从数据流中进行预算受限的成本敏感学习。在国际学习表征会议，2018。Kachuee, M., Karkkainen, K., Goldstein, O., Zamanzadeh, D., 和 Sarrafzadeh, M. 利用公共卫生数据进行成本敏感诊断与学习。*arXiv* 预印本 *arXiv:1902.07102*，2019年。Karayev, S., Baumgartner, T., Fritz, M., 和 Darrell, T. 及时目标识别。神经信息处理系统进展，25，2012年。Kingma, D. P., Salimans, T., 和 Welling, M. 变分 dropout 与局部重参数化技巧。神经信息处理系统进展，28，2015年。Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., 等。Captum：面向 PyTorch 的统一通用模型可解释性库。*arXiv* 预印本 *arXiv:2009.07896*，2020年。Krizhevsky, A., Hinton, G., 等。从小尺寸图像中学习多层特征。2009年。

LeCun, Y., Bottou, L., Bengio, Y., 和 Haffner, P. 将基于梯度的学习应用于文档识别。*IEEE* 论文集，86(11):2278–2324，1998年。

Lee, C.、Imrie, F. 和 van der Schaar, M. 利用相关门控增强自监督的特征选择。载于国际学习表征会议，2021年。

Lemhadri, I.、Ruan, F. 和 Tibshirani, R. LassoNet：具有特征稀疏性的神经网络。载于国际人工智能与统计学会议，第 10–18 页。PMLR，2021年。

Li, J.、Cheng, K.、Wang, S.、Morstatter, F.、Trevino, R. P.、Tang, J. 和 Liu, H. 特征选择：数据视角。*ACM* 计算调查（CSUR），50(6):1–45，2017年。

Li, Y. 和 Oliva, J. 利用生成式代理模型进行主动特征获取。载于国际机器学习会议，第 6450–6459 页。PMLR，2021年。

Ma, C., Tschiatschek, S., Palla, K., Hernandez-Lobato, J. M., Nowozin, S. 和 Zhang, C. EDDI：利用部分变分自编码器高效动态发现高价值信息。发表于国际机器学习会议，第 4234–4243 页。PMLR，2019 年。

Ma, C., Tschiatschek, S., Turner, R., Hern´andez-Lobato, J. M., and Zhang, C. VAEM：一种用于异构混合类型数据的深度生成模型。神经信息处理系统进展，33:11237–11247，2020年。Maddison, C. J., Mnih, A., and Teh, Y. W. Concrete 分布：离散随机变量的连续松弛。*arXiv* 预印本 *arXiv:1611.00712*，2016。Melville, P., Saar-Tsechansky, M., Provost, F., and Mooney, R. 用于分类器归纳的主动特征值获取。载于第四届 *IEEE* 国际数据挖掘会议 *(ICDM'04)*，第 483–486 页。IEEE，2004年。Miller, H. W. 美国健康与营养调查的计划与实施，1971–1973年。美国卫生、教育与福利部出版物编号 *(PHS)*-卫生、教育与福利部（美国），1973年。Mnih, V., Heess, N., Graves, A., 等。视觉注意力的循环模型。神经信息处理系统进展，27，2014年。Mosesson, M. W. 纤维蛋白原与纤维蛋白的结构和功能。血栓与止血杂志，3(8): 1894–1904，2005年。Nazabal, A., Olmos, P. M., Ghahra mani, Z., and Valera, I. 使用变分自编码器处理不完整异构数据。模式识别，107:107501，2020年。

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. PyTorch 中的自动微分。2017年。Randløv, J. and Alstrøm, P. 使用强化学习和塑形学习骑自行车。载于国际机器学习大会，第 98 卷，第 463–471 页。Citeseer，1998年。Roe, B., Yand, H., Zhu, J., Lui, Y., Stancu, I., 等。提升决策树：人工神经网络的替代方案。核仪器与方法*A*辑，543:577–584，2005年。Saar-Tsechansky, M., Melville, P., and Provost, F. 主动特征值获取。管理科学，55(4): 664–684，2009年。Shim, H., Hwang, S. J., and Yang, E. 通过可变大小集合编码实现联合主动特征获取与分类。神经信息处理系统进展，31，2018。Shrikumar, A., Greenside, P., and Kundaje, A. 通过传播激活差异来学习重要特征。载于国际机器学习会议，第 3145–3153 页。PMLR，2017年。Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 丢弃：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》，15(1):1929–1958，2014年。Subco mmittee, A., Group, I. A. W., 等。高级创伤生命支持（ATLS®）：第九版。《创伤与急诊外科杂志》，74(5):1363–1366，2013年。Sundararajan, M., Taly, A., and Yan, Q. 深度网络的公理化归因。载于国际机器学习会议，第 3319–3328 页。PMLR，2017年。Sutton, R. S., Barto, A. G., 等。《强化学习导论》。1998年。Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. 神经格兰杰因果关系。*IEEE*模式分析与机器智能汇刊，44(8):4267–4279，2021年。Tucker, G., Mnih, A., Maddison, C. J., Lawson, J., and Sohl-Dickstein, J. Rebar：针对离散潜变量模型的低方差、无偏梯度估计。神经信息处理系统进展，30，2017年。Williams, R. J. 用于联结主义强化学习的简单统计梯度跟踪算法。《机器学习》，8(3):229–256，1992年。Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. 使用随机门进行特征选择。载于国际机器学习会议。PMLR，2020年。

Yoon, J., Jordon, J. 和 van der Schaar, M. INVASE：使用神经网络进行实例级变量选择。发表于国际学习表征会议，2018。

A. 证明

本节中，我们重新陈述并证明主要理论结果。首先，我们给出关于任意策略 π的最优预测器的命题。

命题 1. 当$\mathbf{y}$ 为离散变量且 ℓ 为交叉熵损失时，对于任意策略 π，贝叶斯分类器都能使式(4) 达到最小值，即 $f^*(x_s) = p(\mathbf{y} \mid x_s)$。

证明。 给定预测器的输入 x_s，我们的目标是确定使期望损失最小的预测。由于特征由 π 依次选择，且选择时并不知道未选特征的取值，因此没有其他信息可供条件化；对于预测器而言，我们甚至无需区分特征的选择顺序。因此，对于离散响应变量$\mathbf{y} \in [K]$，最优预测 $\hat{} y \in \Delta^{K-1}$ 可推导如下：

$$\begin{aligned} f^*(x_s) &= \arg\min_{\hat{y}} \ \mathbb{E}_{\mathbf{y}|x_s}\big[\ell(\hat{y}, \mathbf{y})\big] \\ &= \arg\min_{\hat{y}} \ \sum_{i \in \mathcal{Y}} p(\mathbf{y} = i \mid x_s) \log \hat{y}_i \\ &= \arg\min_{\hat{y}} \ D_{\mathrm{KL}}\big(p(\mathbf{y} \mid x_s) \mid\mid \hat{y}\big) + H(\mathbf{y} \mid x_s) \\ &= p(\mathbf{y} \mid x_s). \end{aligned}$$

对于平方误差损失下的连续响应变量$\mathbf{y} \in \mathbb{R}$，我们也有一个涉及响应变量条件期望的类似结果：

$$\begin{aligned} f^*(x_s) &= \arg\min_{\hat{y}} \ \mathbb{E}_{\mathbf{y}|x_s}\big[(\hat{y} - \mathbf{y})^2\big] \\ &= \arg\min_{\hat{y}} \ \mathbb{E}_{\mathbf{y}|x_s}\big[(\hat{y} - \mathbb{E}[\mathbf{y} \mid x_s])^2\big] + \mathrm{Var}(\mathbf{y} \mid x_s) \\ &= \mathbb{E}[\mathbf{y} \mid x_s]. \end{aligned}$$

□

命题 2. 当$\mathbf{y}$ 为离散变量、ℓ 为交叉熵损失，且预测器为贝叶斯分类器 f^*时，贪心条件互信息策略可使式(4) 达到最小值，即 $\pi^*(x_s) = \arg\max_i \ I(\mathbf{y}; \mathbf{x}_i \mid x_s)$。

证明。 根据公式（4），策略网络的选择 $i = \pi(x_s)$ 相对于分布 $p(\mathbf{y}, \mathbf{x}_i \mid x_s)$ 会产生以下期望损失：

$$\begin{aligned} \mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\ell(f^*(x_s \cup \mathbf{x}_i), \mathbf{y})\big] &= \mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\ell(p(\mathbf{y} \mid \mathbf{x}_i, x_s), \mathbf{y})\big] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}\Big[\mathbb{E}_{\mathbf{y}|\mathbf{x}_i,x_s}[\ell(p(\mathbf{y} \mid \mathbf{x}_i, x_s), \mathbf{y})]\Big] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}\big[H(\mathbf{y} \mid \mathbf{x}_i, x_s)\big] \\ &= H(\mathbf{y} \mid x_s) - I(\mathbf{y}; \mathbf{x}_i \mid x_s). \end{aligned}$$

请注意， $H(\mathbf{y} \mid x_s)$ 是一个与 i 无关的常数。因此，在确定使期望损失最小化的索引时，我们有以下结果：

$$\arg\min_i \ \mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\ell(f^*(x_s \cup \mathbf{x}_i), \mathbf{y})\big] = \arg\max_i \ I(\mathbf{y}; \mathbf{x}_i \mid x_s).$$

对于具有平方误差损失的连续响应变量，若最优预测器为 $f^*(x_s) = \mathbb{E}[\mathbf{y} \mid x_s]$，则有类似的结果：

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\left[(f^*(x_s \cup \mathbf{x}_i) - \mathbf{y})^2\right] &= \mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\left[(\mathbb{E}[\mathbf{y} \mid \mathbf{x}_i, x_s] - \mathbf{y})^2\right] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}\Big[\mathbb{E}_{\mathbf{y}|\mathbf{x}_i,x_s}[(\mathbb{E}[\mathbf{y} \mid \mathbf{x}_i, x_s] - \mathbf{y})^2]\Big] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}[\mathrm{Var}(\mathbf{y} \mid \mathbf{x}_i, x_s)].\end{aligned}$$

当我们的目标是最小化期望损失时，因此应选择能使条件方差的期望最低的索引：

$$\arg\min_i \ \mathbb{E}_{\mathbf{x}_i|x_s}[\mathrm{Var}(\mathbf{y} \mid \mathbf{x}_i, x_s)].$$

□

N接下来，我们还证明了式 (3) 中给出的极限结果，该结果表明 $I_i^n \to I(\mathbf{y};\mathbf{x}_i \mid x_s)$.

证明。条件互信息 $I(\mathbf{y};\mathbf{x}_i \mid x_s)$ 定义如下（Cover & Thomas，2012年 ）：

$$\begin{aligned}I(\mathbf{y};\mathbf{x}_i \mid x_s) &= D_{\mathrm{KL}}\big(p(\mathbf{x}_i, \mathbf{y} \mid x_s) \,||\, p(\mathbf{x}_i \mid x_s)p(\mathbf{y} \mid x_s)\big) \\ &= \mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\Big[\log \frac{p(\mathbf{y}, \mathbf{x}_i \mid x_s)}{p(\mathbf{x}_i \mid x_s)p(\mathbf{y} \mid x_s)}\Big].\end{aligned}$$

重新排列各项后，可将其写成关于$\mathbf{x}_i$ 的期望 KL 散度：

$$\begin{aligned}I(\mathbf{y};\mathbf{x}_i \mid x_s) &= \mathbb{E}_{\mathbf{x}_i|x_s}\mathbb{E}_{\mathbf{y}|x_s,\mathbf{x}_i}\Big[\log \frac{p(\mathbf{y}, \mathbf{x}_i \mid x_s)}{p(\mathbf{x}_i \mid x_s)p(\mathbf{y} \mid x_s)}\Big] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}\mathbb{E}_{\mathbf{y}|x_s,\mathbf{x}_i}\Big[\log \frac{p(\mathbf{y} \mid \mathbf{x}_i, x_s)}{p(\mathbf{y} \mid x_s)}\Big] \\ &= \mathbb{E}_{\mathbf{x}_i|x_s}\Big[D_{\mathrm{KL}}\big(p(\mathbf{y} \mid \mathbf{x}_i, x_s) \,||\, p(\mathbf{y} \mid x_s)\big)\Big]\end{aligned}$$

现在，当我们抽取多个值 $x_i^1, \ldots, x_i^n \sim p(\mathbf{x}_i \mid x_s)$ 并使用贝叶斯分类器进行预测时，随着 n 变大，我们得到以下平均预测：

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} p(\mathbf{y} \mid x_s, x_i^j) = \mathbb{E}_{\mathbf{x}_i|x_s}\left[p(\mathbf{y} \mid \mathbf{x}_i, x_s)\right] = p(\mathbf{y} \mid x_s).$$

C计算相对于该预测的平均 KL 散度后，我们得到以下结果 ：

$$\lim_{n\to\infty} I_i^n = \mathbb{E}_{\mathbf{x}_i|x_s}\Big[D_{\mathrm{KL}}\big(p(\mathbf{y} \mid \mathbf{x}_i, x_s) \,||\, p(\mathbf{y} \mid x_s)\big)\Big] = I(\mathbf{y};\mathbf{x}_i \mid x_s).$$

□

定理 1. 当$\mathbf{y}$ 是离散变量且 ℓ 是交叉熵损失时，式 (5) 的全局最优解是一个满足$f(x_s;\, \theta^*) = p(\mathbf{y} \mid x_s)$ 的预测器，以及一个将全部概率质量分配给 $i^* = \arg\max_i I(\mathbf{y};\mathbf{x}_i \mid x_s)$ 的策略 $\pi(x_s;\, \phi^*)$。

证明。 我们首先考虑预测器网络 $f(\mathbf{x}_s;\,\theta)$。当向预测器提供特征值 x_s 时，这意味着策略根据 $\pi(x_{s\setminus i};\,\phi)$ 选取了一个索引 $i \in s$，其余索引 $s \setminus i$ 则从 $p(\mathbf{s})$ 中采样。由于$\mathbf{s}$是独立于 $(\mathbf{x},\mathbf{y})$ 采样的，并且 $\pi(x_{s\setminus i};\,\phi)$ 无法访问 $(\mathbf{x}^{[d]\setminus s}, \mathbf{x}_i, \mathbf{y})$，因此必须针对分布 $\mathbf{y} \mid x_s$ 来考虑预测器的期望损失。由此定义全局最优预测器 $f(x_s;\,\theta^*)$，无论选择策略 $\pi(x_s;\phi)$ 如何，也无论最后选中了哪个索引 i：

$$f(x_s;\theta^*) = \arg\min_{\hat{y}} \; \mathbb{E}_{\mathbf{y}|x_s}\big[\ell(\hat{y},\mathbf{y})\big] = p(\mathbf{y} \mid x_s).$$

上述结果可由我们对命题 1 的证明推出。现在，给定最优预测器 $f(x_s;\,\theta^*)$，我们可以通过针对固定输入 x_s 最小化期望损失，来定义全局最优策略。将分配给每个索引 $i \in [d]$ 的概率质量记为 $\pi_i(x_s;\phi)$，其中 $\pi(x_s;\phi) \in \Delta^{d-1}$，则期望损失如下：

$$\begin{aligned}\mathbb{E}_{i\sim\pi(x_s;\phi)}\mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\ell(f(x_s\cup\mathbf{x}_i;\theta^*),\mathbf{y})\big] &= \sum_{i\in[d]}\pi_i(x_s;\phi)\mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\ell\big(f(x_s\cup\mathbf{x}_i;\theta^*),\mathbf{y}\big)\big]\\ &= \sum_{i\in[d]}\pi_i(x_s;\phi)\mathbb{E}_{\mathbf{x}_i|x_s}[H(\mathbf{y}\mid\mathbf{x}_i,x_s)].\end{aligned}$$

上述结果可由我们对命题 2 的证明推出。如果存在唯一的索引 $i^* \in [d]$ 使得条件熵的期望最小，或者

$$\mathbb{E}_{\mathbf{x}_{i^*}|x_s}[H(\mathbf{y}\mid\mathbf{x}_{i^*},x_s)] < \mathbb{E}_{\mathbf{x}_i|x_s}[H(\mathbf{y}\mid\mathbf{x}_i,x_s)] \quad \forall i \neq i^*,$$

那么最优预测器必须将全部概率质量分配给 i^*，即 $\pi_{i^*}(x_s;\phi^*) = 1$。注意，相应的特征$\mathbf{x}_{i^*}$ 与$\mathbf{y}$具有最大的条件互信息，因为我们有

$$I(\mathbf{y};\mathbf{x}_{i^*}\mid x_s) = \underbrace{H(\mathbf{y}\mid x_s)}_{\text{Constant}} - \mathbb{E}_{\mathbf{x}_{i^*}|x_s}[H(\mathbf{y}\mid\mathbf{x}_{i^*},x_s)].$$

总而言之，我们首先考虑最优预测器 $f(\mathbf{x}_s;\,\theta^*)$，然后在假设使用最优预测器的情况下，考虑最优策略 $\pi(\mathbf{x}_s;\phi^*)$，由此推导出目标函数 $\mathcal{L}(\theta,\phi)$ 的全局最优解。 □

定理 2. 当$\mathbf{y}$ 是连续变量且 ℓ 为平方误差损失时，式(5) 的全局最优解是满足以下条件的预测器：$f(x_s;\,\theta^*) = \mathbb{E}[\mathbf{y} \mid x_s]$；以及一个策略 $\pi(x_s;\,\phi^*)$，它将全部概率质量赋予 $i^* = \arg\min_i \mathbb{E}_{\mathbf{x}_i|x_s}[\mathrm{Var}(\mathbf{y} \mid \mathbf{x}_i, x_s)]$。

证明. 我们的证明思路与定理 1 的证明相同。对于给定任意策略时的最优预测器，我们有：

$$f(x_s;\theta^*) = \arg\min_{\hat{y}} \; \mathbb{E}_{\mathbf{y}|x_s}\big[(\hat{y}-\mathbf{y})^2\big] = \mathbb{E}[\mathbf{y}\mid x_s].$$

接下来，对于该策略的期望损失，我们有：

$$\mathbb{E}_{i\sim\pi(x_s;\phi)}\mathbb{E}_{\mathbf{y},\mathbf{x}_i|x_s}\big[\big(f(x_s\cup\mathbf{x}_i;\theta^*)-\mathbf{y}\big)^2\big] = \sum_{i\in[d]}\pi_i(x_s;\phi)\mathbb{E}_{\mathbf{x}_i|x_s}[\mathrm{Var}(\mathbf{y}\mid\mathbf{x}_i,x_s)].$$

如果存在一个索引 $i^* \in [d]$，使其对应的条件方差期望最低，那么最优策略必须将全部概率质量分配给 i^*，或者 $\pi_{i^*}(x_s;\phi^*) = 1$。 □

B. 数据集

我们实验所用的数据集汇总于表2。三个表格数据集和两个图像分类数据集均可公开获取，而三个急诊医学任务的数据则由 Harborview Medical Center Trauma Registry 私下整理。

表2. 我们实验中使用的数据集概要。

数据集	# 特征	# 特征组	# 类别	# 样本
液体	224	162	2	2,770
呼吸	112	35	2	65,515
出血	121	44	2	6,496
Spam	58	–	2	4,601
MiniBooNE	51	–	2	130,064
糖尿病	45	–	3	92,062
MNIST	784	–	10	60,000
CIFAR-10	1,024	64	10	60,000

B.1. MiniBooNE 与垃圾邮件分类

垃圾邮件数据集包含从电子邮件中提取的特征，用于预测邮件是否为垃圾邮件。其中，三个特征描述了电子邮件中大写字母的使用情况，其余 54 个特征描述了特定关键词或字符的出现频率。MiniBooNE 粒子识别数据集根据各种连续特征区分电子中微子与 μ子中微子（Roe 等人，2005年）。这两个数据集均取自 UCI 数据库（Dua 和 Graff，2017年）。

B.2. 糖尿病分类

糖尿病数据集取自美国国家健康与营养检查调查（NHANES）（NHA，2018），这是一项持续进行的调查，旨在评估美国成人和儿童的健康状况。我们使用了 Kachuee 等人（2018；2019年）预处理的数据版本，其中包含 1999 年至 2016 年收集的数据。输入特征包括人口统计信息（年龄、性别、族裔等）、实验室检查结果（总胆固醇、甘油三酯等）、体格检查数据（体重、身高等）以及问卷答案（吸烟、饮酒、睡眠习惯等）。我们还请一位专家根据经济负担、患者隐私和患者不便程度为每项特征建议成本，但在实验中假设所有特征成本相同。最后，我们使用空腹血糖值，根据标准阈值将样本划分为三类：正常、糖尿病前期和糖尿病。

B.3. 图像分类数据集

MNIST 和 CIFAR-10 数据集通过 PyTorch（Paszke 等人，2017年）下载。我们采用标准的训练集与测试集划分，并从训练集中划分出一个验证集，其大小与测试集相同（10,000 个样本）。

B.4. 急诊医学数据集

本研究使用的急诊医学数据集收集于 13 年期间（2007-2020年），涵盖 14,463 例急诊科入院病例。我们排除了年龄未满 18 岁的患者，并整理出 3 个在住院前场景中常见的临床队列，分别为：1）院前液体复苏；2）急诊科呼吸支持；3）创伤后出血。出于患者隐私方面的考虑，这些数据集不公开。

院前液体复苏 我们选取了院前环境中可获取的 224 个变量，包括调度信息（受伤日期、时间、原因和地点）、人口统计信息（年龄、性别）以及院前生命体征（血压、心率、呼吸频率）。结局指标为每位患者对液体复苏的响应，定义依据为高级创伤生命支持（ATLS）的定义（Subcommittee 等人，2013年）。

急诊科呼吸支持 在该队列中，我们旨在预测哪些患者抵达急诊科时需要呼吸支持。与前一个数据集类似，我们选取了 112 个院前临床特征，包括调度信息（受伤日期、时间、原因和地点）、人口统计信息（年龄、性别）以及院前生命体征（血压、心率、呼吸频率）。结局根据患者是否接受了呼吸支持来定义，包括侵入式（插管）和非侵入式（双水平气道正压通气）两种方式。

出血 在该队列中，我们仅纳入测量过纤维蛋白原水平的患者，因为该指标可用于判断是否存在出血或纤维蛋白溶解（Mosesson，2005年）。与之前的数据集一样，我们使用人口统计信息、调度信息和院前观察数据作为输入特征。根据专家意见，我们将结局定义为个体的纤维蛋白原水平是否低于 200 毫克/分升，这代表创伤后出血风险较高。

C. 基线方法

本节将进一步介绍我们在实验（第 6 节）中使用的基线方法。

C.1. 全局特征重要性方法

我们采用的两种静态特征选择基线方法——置换检验和 SAGE——都是根据特征在提升模型准确率方面的作用对特征进行排序的全局特征重要性方法 （Covert 等人，2021年）。在实验中，我们使用在整个数据集上训练的单个分类器运行每种方法，然后根据预算选择排名前 k 的特征。

运行置换检验时，我们将相应特征列中的值替换为从训练集中随机抽取的值，并计算验证集 AUROC。运行 SAGE 时，我们使用了作者提供的实现，并采用自动收敛检测（Covert 等人，2020年）。为处理留出特征，对于六个表格数据集，我们对 128 个采样值取平均；对于 MNIST，我们使用零值基线以加快收敛。

C.2. 局部特征重要性方法

我们的静态特征选择基线方法中，有两种——DeepLift 和积分梯度——属于局部特征重要性方法，它们根据特征对单个预测的重要性对特征进行排序。在实验中，我们使用验证集中的所有样本，为真实类别生成特征重要性分数。然后根据平均绝对重要性选取排名靠前的 k 个特征。对于积分梯度，我们使用均值基线（Sundararajan 等人，2017年）；两种方法均使用 Captum 软件包运行（Kokhlikyan 等人，2020年）。

C.3. 可微特征选择

我们最后一个静态特征选择基线是 Balın 等人（2019）提出的 Concrete 自编码器（CAE）。该方法最初通过重建完整的输入向量来进行无监督特征选择，但我们将预测目标进行了修改，使其能够用于监督式任务。作者建议在手动调定的训练轮次数内，以指数衰减的温度进行训练；而我们采用了类似于自身方法的做法：使用一系列温度值进行训练，对每个温度值根据验证损失执行提前停止，并返回使用最终温度训练后选出的特征。

我们尝试了 Yamada 等人（2020）提出的一种类似方法，但该方法需要调整惩罚参数才能达到所需的特征数量；我们发现，在 MNIST 上的实验中，其性能与我们的方法相近。在通过神经网络学习选择特征的方法中，还有几种方法使用组稀疏惩罚（Feng & Simon，2017年；Tank 等人，2021年；Lemhadri 等人，2021年）；我们测试了 Lemhadri 等人（2021年）提出的 LassoNet 方法，发现它在 MNIST 上的表现不具竞争力。为求简洁，我们仅展示监督式 CAE 的结果。

C.4. 条件互信息估计

我们的实验采用了第 3.2 节所述条件互信息估计方法的两个版本。两者均受 Ma 等人（2019）提出的 EDDI 方法启发，但一个关键区别是，我们不会在同一个条件生成模型中联合建模 $(\mathbf{x},\mathbf{y})$：相反，我们使用分类器 $f(\mathbf{x}_s) \approx p(\mathbf{y} \mid \mathbf{x}_s)$ 单独对响应变量建模，并使用生成模型 $p(\mathbf{x}_i \mid \mathbf{x}_s)$ 对特征建模。这在一定程度上缓解了该方法面临的一项挑战，即处理混合连续/类别数据（也就是说，我们无需联合建模类别型响应变量）。

对于该方法的第一个版本，我们训练 PVAE 作为生成模型（Ma 等人，2019）。编码器和解码器均包含两个隐藏层，潜在维度设为 16，并使用来自潜在后验分布的 128 个样本来近似$p(\mathbf{x}_i \mid x_s) = \int p(\mathbf{x}_i \mid \mathbf{z})p(\mathbf{z} \mid x_s)$。潜在空间和解码器空间均采用高斯分布，并且与原始方法（Ma 等人，2019）类似，我们使用解码器均值生成样本。第二个版本则通过一种简单的近似绕过对生成模型的需求：从特征的边缘分布中采样，这等同于假设特征独立。

C.5. 机会式学习

Kachuee 等人（2018）提出了机会式学习（OL），这是一种使用强化学习解决动态特征选择的方法。该模型由两个网络组成，分别类似于我们的策略和预测器：一个 Q 网络，用于估计每个动作对应的价值，其中动作对应于特征；另一个 P 网络，负责进行预测。使用 OL 时，我们采用与自身方法相同的架构；此外，OL 的 P 网络和 Q 网络共享网络参数。

作者在式（6）中引入了一个效用函数来计算奖励，该函数计算预测不确定性的差异，并使用蒙特卡洛 dropout 对其进行近似（Gal 和 Ghahramani，2016）。奖励还考虑了特征成本，但我们将所有特征成本设为 $c_i = 1$:

$$r_i = \frac{||Cert(x_s) - Cert(x_s \cup x_i)||}{c_i} \tag{6}$$

为了与其他方法进行公平比较，我们对作者的实现进行了几项修改，包括：1）在达到预先指定的预算之前禁止执行预测操作；2）将所有特征成本设为相同；3）支持附录 D.4 中所述的预定义特征组。训练时，我们每经过 $1 + \frac{d}{100}$ 次经验就更新 P 网络、Q 网络和目标 Q 网络，其中 d 是数据集中的特征数量。此外，经验回放缓冲区设置为存储最近的 $1000d$ 次经验，并逐渐降低随机探索概率，使其最终达到 0.1。

D. 训练方法与超参数

本节将进一步介绍我们的训练方法和超参数选择。

D.1. 训练伪代码

算法 1 总结了我们的训练方法。简而言之，我们使用策略网络的 logits 抽取一个 Concrete 样本来选择特征，根据后续预测计算损失，然后使用来自策略分布的离散样本更新下一步的掩码。我们使用 PyTorch（Paszke 等人，2017年）和 PyTorch Lightning 实现了这一方法。[3]

算法 1:训练伪代码

输入：数据分布 $p(\mathbf{x}, \mathbf{y})$，预算 $k > 0$，学习率 $\gamma > 0$，温度 $\tau > 0$输出：预测器模型 $f(\mathbf{x}; \theta)$，策略模型 $\pi(\mathbf{x}; \phi)$ 初始化 $f(\mathbf{x}; \theta)$ $\pi(\mathbf{x}; \phi)$ 当尚未收敛时 do 抽取 $x, y \sim p(\mathbf{x}, \mathbf{y})$ 初始化 $\mathcal{L} = 0, m = [0, \dots, 0]$for $j = 1$to k do 计算 logits $\alpha = \pi(x \odot m; \phi)$，抽取 $G_i \sim$ Gumbel for $i \in [d]$set $\tilde{m} =$max $\big(m,$ softmax$(G + \alpha, \tau)\big)$ // 使用 Concrete 更新 set $m =$max $\big(m,$ softmax$(G + \alpha, 0)\big)$ // 使用独热更新 update $\mathcal{L} \leftarrow \mathcal{L} + \ell\big(f(x \odot \tilde{m};\ \theta),\ y\big)$end update $\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}$, $\phi \leftarrow \phi - \gamma \nabla_\phi \mathcal{L}$end

返回 $f(\mathbf{x}; \theta)$ $\pi(\mathbf{x}; \phi)$

算法 1 与正文中的目标 $\mathcal{L}(\theta,\ \phi)$ 之间一个显著差异在于，前者使用策略 $\pi(\mathbf{x};\ \phi)$ 来生成特征子集。这不同于公式 (5)，后者使用子集分布 $p(\mathbf{s})$ 来生成特征子集。这两种方法的主要共同点是：它们都针对每个特征集分别设置了优化问题，并将这些问题视为相互独立。对于每个特征集 x_s，问题的目标是一步前瞻损失，并按如下方式同时纳入策略和预测器：

$$\mathbb{E}_{i \sim \pi(\mathbf{x_s}; \phi)} \big[\ell\big(f(\mathbf{x_s} \cup \mathbf{x}_i; \theta), \mathbf{y}\big) \big]. \tag{7}$$

各子集对应的问题彼此不相互作用：在优化过程中，给定 x_s 的选择仅依据损失的即时变化；与基于强化学习的解决方案不同，梯度不会通过多次选择进行反向传播。在求解这些问题时，区别仅在于，公式 (5) 根据 $p(\mathbf{s})$ 对它们加权，而算法 1 则根据当前策略 $\pi(\mathbf{x}\ \phi)$ 对它们加权。

我们发现，在生成特征集时纳入当前策略，对于获得良好性能至关重要。作为一项消融实验，我们测试了改为随机生成训练样本 $(\mathbf{x_s},\ \mathbf{y})$，而非使用当前策略时，方法的性能会有多大变化：在 MNIST 数据集上，我们发现，使用随机子集会导致性能显著下降（表 3）。

D.2. 模型选择

算法 1 中未展示、但我们在正文中提到的一处细节，是我们对 Concrete 分布的温度参数 τ 进行衰减的方法。我们使用一系列数量较少的温度值进行训练，并针对每个温度值根据验证损失执行提前停止。为了进行模型选择，我们另行计算温度值为零时的验证损失，因为这能更准确地反映模型在推理时的使用情况；最终，我们在所有训练温度值对应的模型中，选取零温度损失表现最佳的版本。

[3]https://www.pytorchlightning.ai

表 *3.* 使用 MNIST 进行的消融实验。

# 特征	5	10	15	20	25	30	40	50
Ours	0.695	0.875	0.926	0.950	0.960	0.966	0.973	0.975
消融实验	0.578	0.757	0.807	0.819	0.838	0.850	0.869	0.883

D.3. 超参数

我们在六个表格数据集上的实验使用了全连接架构，并在所有层中应用了丢弃（Srivastava 等人，2014年）。丢弃概率设为 0.3，网络包含两个宽度为 128 的隐藏层，我们根据验证损失执行提前停止。对于我们的方法，预测器和策略是两个架构相同的独立网络。使用静态方法选出的特征训练模型时，我们根据验证损失报告多次训练中表现最佳的模型结果。由于训练的模型数量较多，我们没有进行额外的超参数调优。

对于 MNIST，我们使用了全连接架构，包含两个宽度为 512 的层，并将丢弃概率设为 0.3。同样，我们的方法使用了架构相同但彼此独立的网络。对于 CIFAR-10，我们使用了共享的 ResNet 骨干网络（He 等人，2016b），该网络由若干残差连接的卷积层组成。分类头由全局平均池化和一个线性层构成；选择头则由一个转置卷积层和一个 1×1卷积层组成，输出尺寸为 8×8 的 logits 网格。我们的 CIFAR-10 网络通过随机裁剪和随机水平翻转进行数据增强训练。

D.4. 特征分组

我们在实验中使用的所有方法都是为选择单个特征而设计的，但在处理采用独热编码的类别特征时，这并不理想。我们的三个急诊医学任务都包含此类特征，因此我们扩展了每种方法，使其支持特征分组。

将 SAGE 和置换检验扩展到特征组十分简单：计算重要性分数时，我们只需移除特征组，而非单个特征。对于 DeepLift 和积分梯度，我们使用每个组内的重要性总和，这样可以保留各方法的可加性。对于基于 Concrete 自编码器的方法，我们实现了一个经过泛化的选择层，使其能够以特征组为单位进行操作。我们还扩展了 OL，使其能够以特征组为单位进行操作，具体做法是让动作映射到特征组，而非单个特征。

最后，对于我们的方法，我们对策略网络 $\pi(\mathbf{x}; \phi)$ 进行参数化，使其输出数量等于特征组的数量g，而非特征总数 d（其中 $g < d$）。应用掩码时，我们首先生成一个二值掩码$m \in [0,1]^g$，然后使用二值分组矩阵 $G \in \{0,1\}^{d \times g}$将该掩码投影到 $[0,1]^d$中；若特征i 属于特征组 j，则 $G_{ij} = 1$，否则为 $G_{ij} = 0$。因此，我们的掩码输入向量为 $x \odot (Gm)$。

E. 补充结果

本节给出若干补充实验结果。首先，图 4 和图 5 展示了与图 2 相同的结果，但尺寸更大，以便更清晰地查看。接下来，图 6 至图 11 展示了使用贪心方法时，各表格数据集的特征选择频率。每幅图中的热图显示，在特定特征预算下，某个特征（或特征组）被选中的比例。这些图表表明，我们的方法确实会针对不同样本选择不同的特征。

最后，图 12 展示了 CIFAR-10 的预测示例，其中显示的图像块数量各不相同。通常，只需显示少量图像块，预测结果就会变得相对准确，这与图 3 中的结果类似。从定性角度来看，我们可以看到策略网络学会了选择垂直条纹，但填充每条条纹的顺序取决于它预测重要信息可能所在的位置。

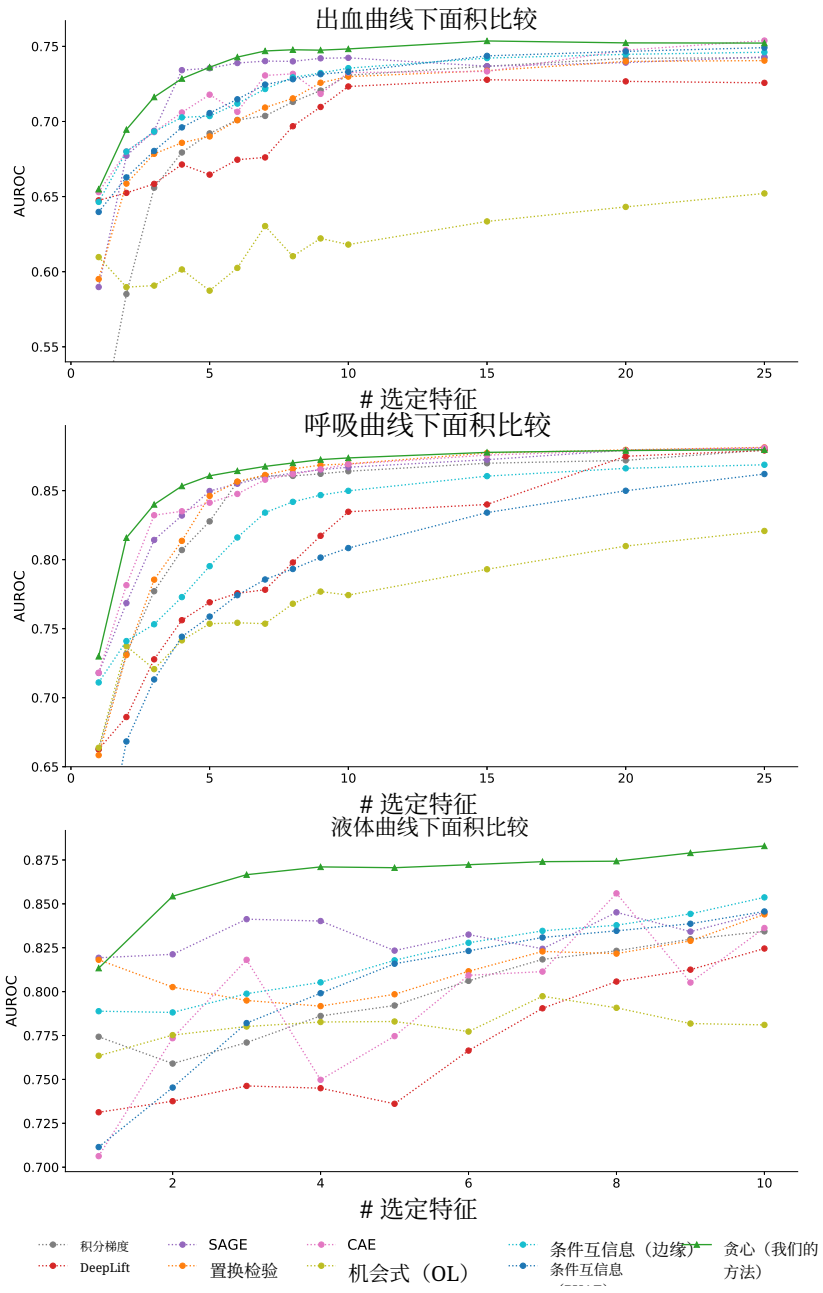

图 *4.* 三个急诊医学诊断任务的曲线下面积比较。

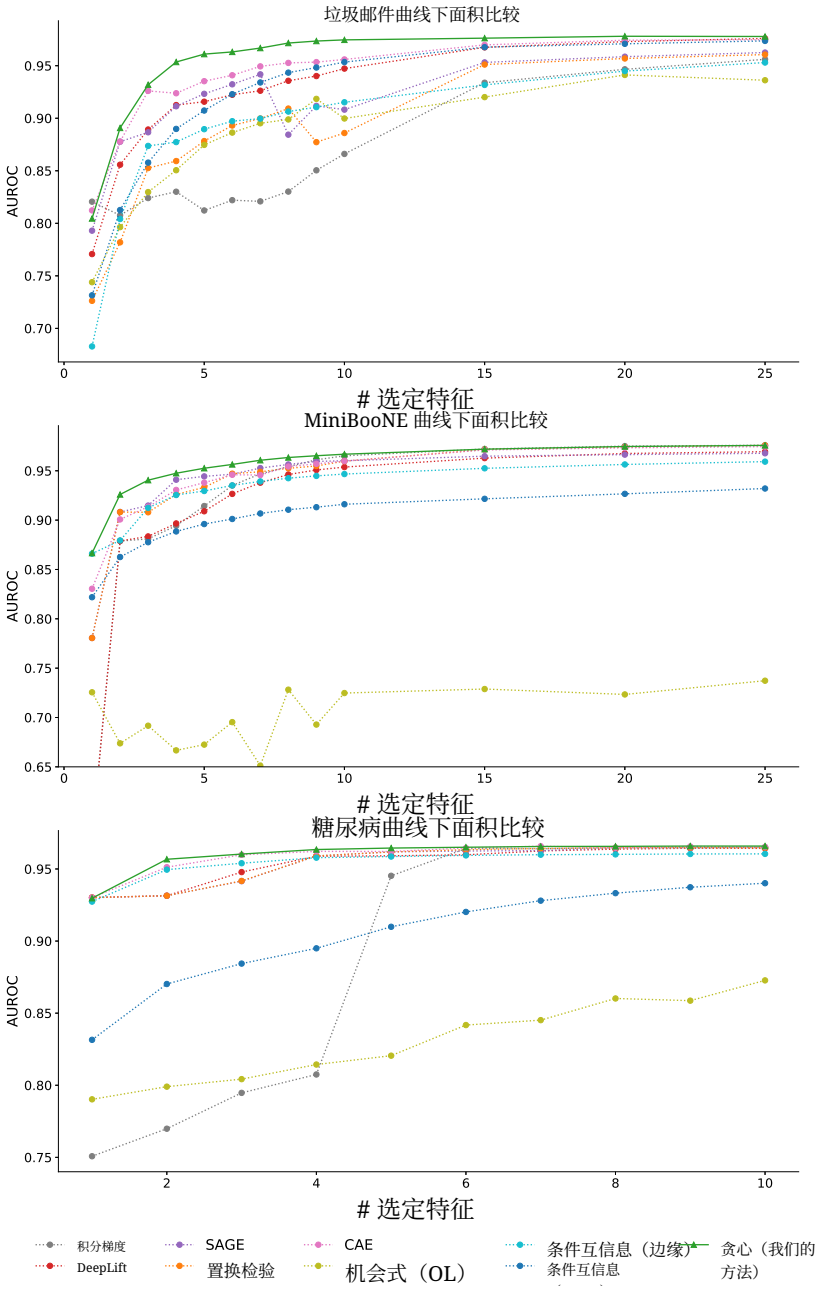

图 *5*。 三个公开表格数据集上的曲线下面积比较。

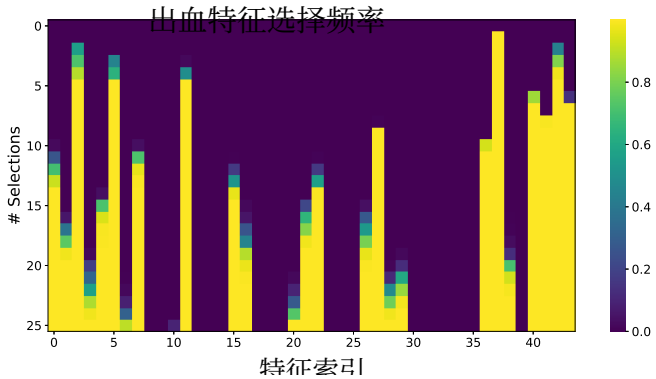

图 *6.* 贪心方法在出血数据集上的特征选择频率。

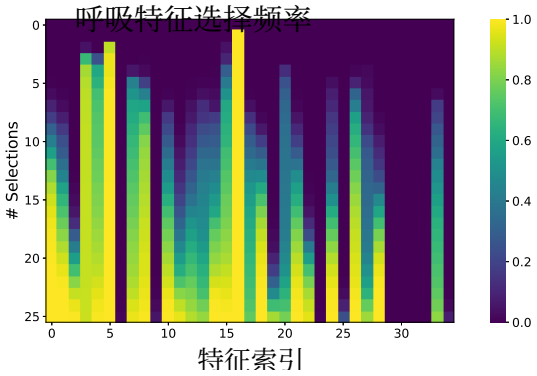

图*7.* 贪心方法在呼吸数据集上的特征选择频率。

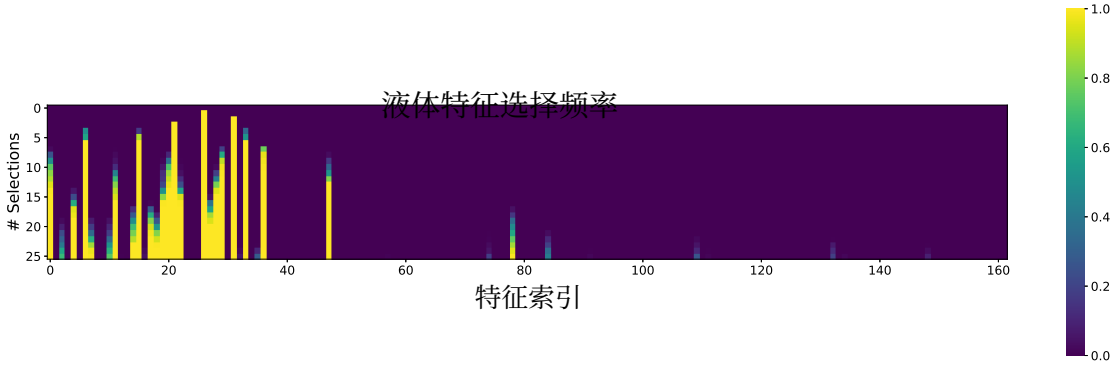

图*8.* 贪心方法在液体数据集上的特征选择频率。

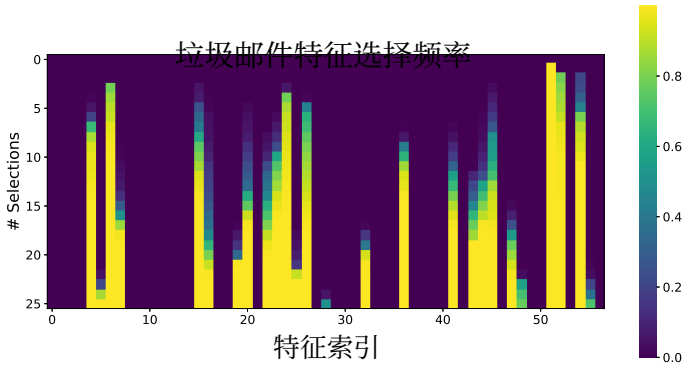

图 *9.* 我们在垃圾邮件数据集上采用贪心方法时的特征选择频率。

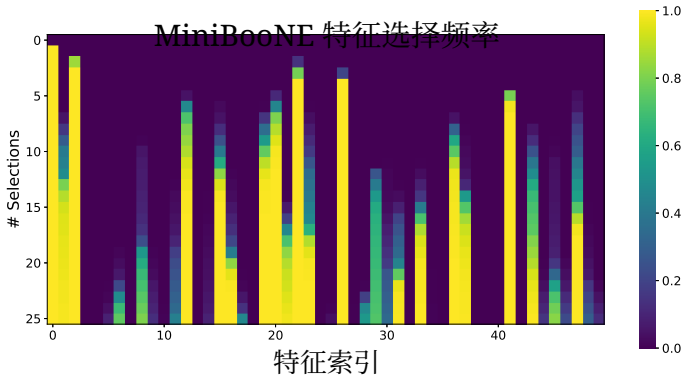

图 *10.* 我们在 MiniBooNE 数据集上采用贪心方法时的特征选择频率。

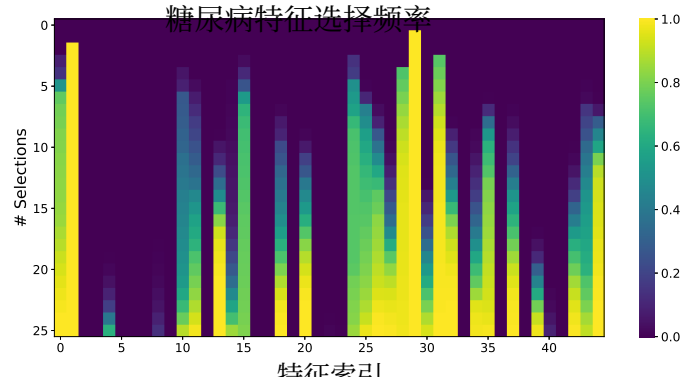

图 *11.* 我们在糖尿病数据集上采用贪心方法时的特征选择频率。

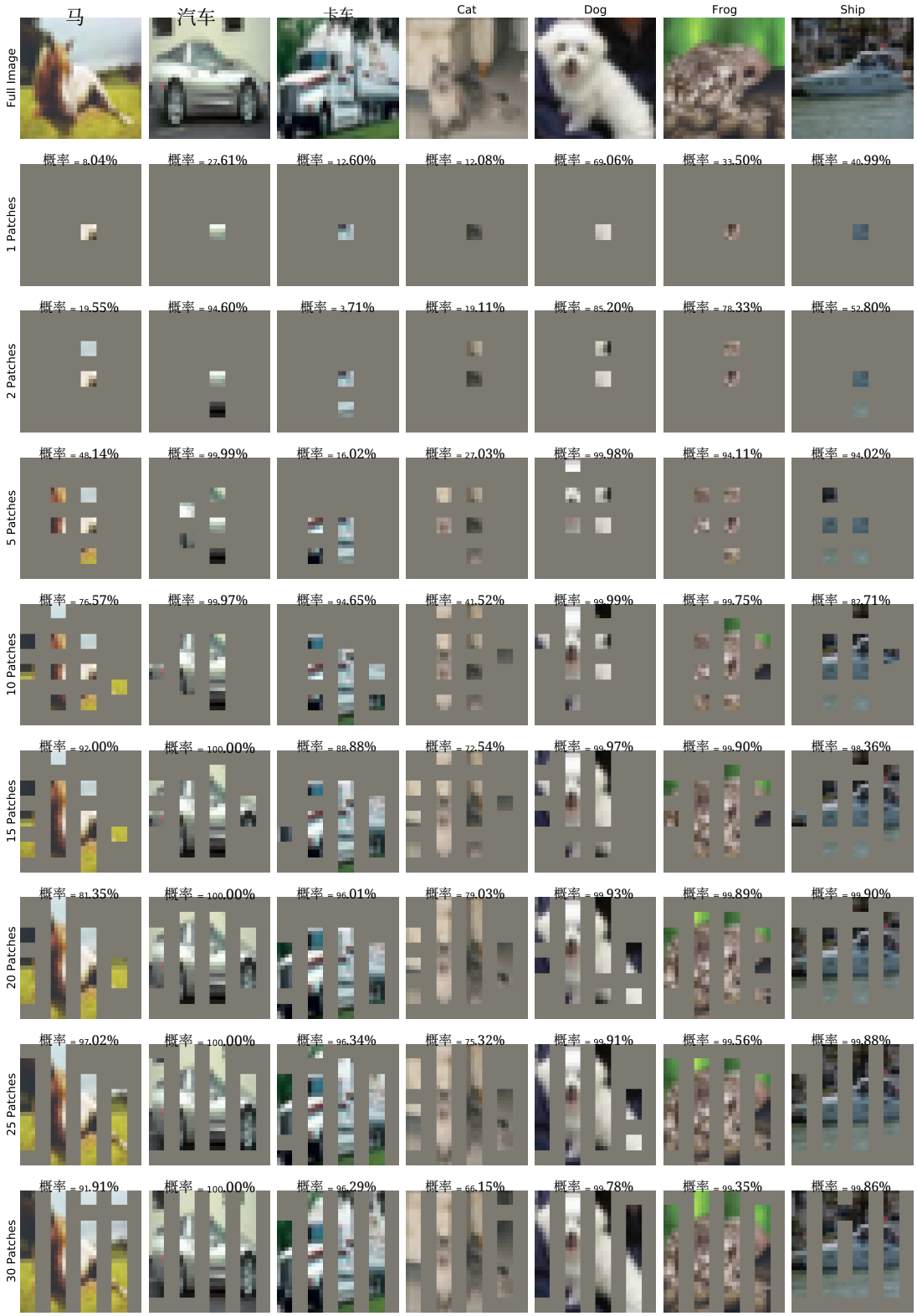

图 *12.* 我们的方法揭示的不同数量图像块对应的 CIFAR-10 预测。