可解释人工智能逻辑

Adnan Darwiche加利
福尼亚大学计算机科学系美
国洛杉矶
darwiche@cs.ucla.edu

arXiv:2305.05172v1 [cs.AI] 9 May 2023

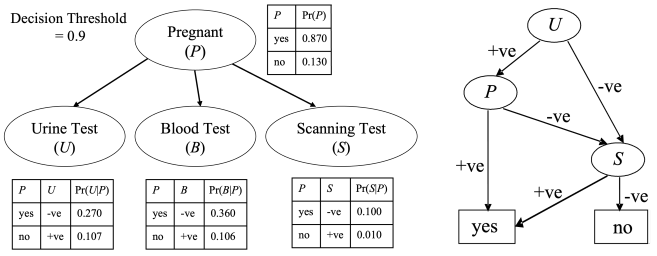
摘要——可解释人工智能的一项核心追求，是理解（学习得到的）分类器所做出的决策。近年来，人们对这一理解的三个维度给予了极大关注。第一个维度是刻画实例上对于决策而言既必要又充分的条件，从而提供实例抽象，可将其视为"决策背后的理由"。下一个维度是刻画对于某项决策而言充分的极小条件，从而识别实例中与该决策无关的最大方面。最后一个维度是刻画对于某项决策而言必要的极小条件，从而识别能够产生其他决策的实例极小扰动。本教程将讨论一种涵盖这些维度的、全面的可解释性语义与计算理论，该理论基于符号逻辑领域近期的一些进展。本教程还将讨论该理论如何特别适用于非符号分类器，例如基于贝叶斯网络、决策树、随机森林以及某些类型神经网络的分类器。

索引词—**可解释人工智能、符号逻辑、分类器、主蕴含项、主蕴含项、量化逻辑**

I. 引言

解释（学习得到的）分类器的决策，或许是可解释人工智能领域研究最多的任务。分类器可以有不同形式——例如决策树、随机森林、贝叶斯网络或神经网络——但本质上是一个将实例 映射到有限数量类别。的函数。当分类器将一个实例映射到某个类别时，我们称它对该实例做出了决策 。每个分类器都有一组特征 ，这些特征是离散或数值变量。通过为每个特征赋值，就可以生成一个实例。请看图1(a)中的朴素贝叶斯分类器，它有三个二元特征（U、 B、 S），代表医学检测，共生成八个实例。该分类器有两个类别（P=YES、 P=NO），分别对应患者是否怀孕。给定一个代表患者检测结果的实例，该分类器首先根据这些结果计算变量 P 的分布，然后仅当该概率不低于0.90时，才将类别 P=YES 赋予该实例（此值称为分类阈值）。该分类器可能是从数据中学习得到的，并通过概率推理做出决策，但从本质上说，它是一个将有限数量实例映射到类别的函数。图2(a)展示了另一种以决策树形式呈现的分类器。它有两个数值特征。

本研究部分由美国国家科学基金会资助（资助编号：ISS-1910317）。
发表于 38th 2023年美国计算机协会/电气电子工程师学会计算机科学逻辑研讨会



(a) 朴素贝叶斯分类器　　(b) 决策图

图 1. 朴素贝叶斯分类器，取自 [1],，具有二值特征 U， B， S 以及类别 P=是、P=否。该决策图表示同一个分类器，因为它对每个实例都会作出完全相同的决策。

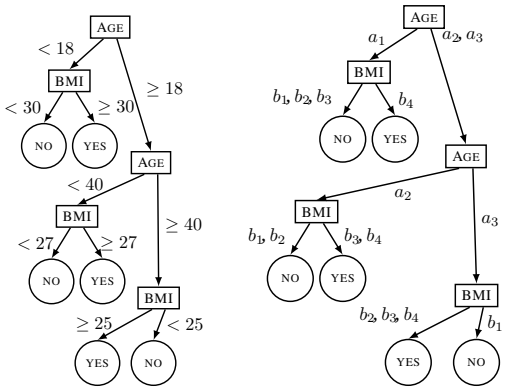
（AGE、身体质量指数）以及由此产生的无限多个实例，这些实例被分类器映射到两个类别之一（YES、否）。

本教程旨在探讨一种基于符号逻辑的理论，用于解释此类分类器的决策。该理论针对三个基本问题：

问题一：针对某个实例作出决策的原因是什么？问题二：实例具备哪些最少的特征，就能保证针对该实例作出相应的决策？问题三：对实例进行哪些最少的改变，会导致针对该实例作出不同的决策？

第一个问题的答案将是关于该决策的充要条件，以逻辑公式的形式表达，揭示导致该决策的实例本质。第二个问题的答案也将是一个逻辑公式，用于刻画特征的最小集合，以及这些特征状态所需满足的最弱条件；在这些条件下，必然会触发该决策。这将立即识别出与决策无关的特征，也会识别出相关特征中无关的状态。第三个问题的答案同样是一个逻辑公式，用于识别特征的最小集合，以及如何改变这些特征，从而将决策转变为某个指定的或任意的其他决策。正如我们稍后将看到的，第一个问题的答案将构成回答第二和第三个问题的基础。

所讨论的理论基于使用类别公式 表示分类器，这些逻辑公式用于描述每个类别中的实例。这里会产生两个问题。首先，这些类别公式从何而来？其次，是否总能以这种方式表示分类器？当考虑如下分类器时，第一个问题可能会浮现。



(a) 决策树 (b) 离散化决策树

图2。来自 [2], 的决策树，包含数值特征年龄、身体质量指数以及类别是、否。年龄被离散化为三个区间 $[0, 18), [18, 40)$ 和 $[40, \infty)$，因此可以将其视为一个离散变量，其对应取值为 a_1，a_2，a_3。身体质量指数被离散化为四个区间 $[0, 25), [25, 27)$, $[27, 30)$ $[30, \infty)$，因此可以将其视为一个离散变量，其对应取值为 b_1，b_2，b_3，b_4。

如图 1(a) 所示，它本质上是数值型的，并基于概率推理。该分类器的特征是离散的，因此每个类别中的实例数量有限，可以用逻辑公式表示。该公式通过一个编译 过程得到，我们将在本教程最后的第六节讨论这一过程。第二个问题出现在图 2(a) 所示的决策树这类分类器中，因为其中包含数值特征，所以涉及无限多个实例。众所周知，这类决策树可以轻松离散化，如图 2(b) 所示。这里，特征 AGE 的数值被划分为三个等价区间（标记为 a_1, a_2, a_3），而身体质量指数的数值被划分为四个等价区间（标记为b_1, b_2, b_3, b_4）。从该分类器的角度来看，特征取值分别落在相同对应区间内的两个实例是同构的，因此每个类别中的实例都只有有限个等价类，也可以用逻辑公式表示。同样的情况也适用于包括决策图和采用多数投票的随机森林在内的许多分类器。同样，这类分类器的类别公式的编译将在第六节讨论。

本教程首先在第二节介绍一些技术预备知识，接着在第三、四、五节探讨三个可解释性问题。随后，我们将在第六节讨论如何将分类器编译为类别公式，最后在第七节作一些总结。

II. 离散逻辑

即使存在通过“二值化”处理的离散变量，文献中仍普遍采用布尔逻辑。例如，具有 n个取值的离散变量可以用 n 个布尔变量进行编码；有关若干编码方案的综述，请参见 [3]。本教程不采用这一传统做法，而是使用一种可称为离散逻辑的方法，如 [2], [4]中所述。在这种逻辑中，离散变量是基本对象，布尔变量只是其特例。从语义上说，这对于本文讨论的可解释性理论至关重要，对于计算也可能至关重要。[1] 之所以偏向布尔逻辑，是因为经过数十年发展，它拥有强大而完备的计算工具，但这些工具无法直接应用于离散变量。不过，我们将看到，如今我们已有强大的工具，能够在可解释人工智能领域处理离散变量。

A. 离散逻辑的语法与语义

我们假设存在一个有限的离散变量集合 Σ。每个变量 $X \in \Sigma$ 都有有限数量的状态 $x_1, \ldots, x_n$, $n > 1$。变量 X的文字 ℓ 称为 X- 文字，是满足 $\emptyset \subset \ell \subset \{x_1, \ldots, x_n\}$ 的一组状态。我们通常将形如 $\{x_1, x_3, x_4\}$ 的文字表示为 x_{134}，其含义是：变量 X的状态为 x_1 、 x_3 或 x_4。当且仅当一个文字只包含一个状态时，它才是简单文字。因此， x_3 是一个简单文字，而 x_{134} 则不是。由于简单文字与一个状态相对应，因此我们交替使用这两个概念。如果变量 X 具有两个状态，则称其为布尔变量或二元变量，其状态表示为x 和 x。这样的变量只能具有两个简单文字，即$\{x\}$ 和 $\{x\}$，为方便起见，分别记为 x 和 x 。公式可以是常量 $\top$、 $\bot$、文字 $\overline{\ell}$、否定 α、合取$\alpha \cdot \beta$或析取 $\alpha + \beta$，其中 α、 β均为公式。公式 Δ 中出现的变量集合记为 vars(Δ)。

离散逻辑的语义直接推广了布尔逻辑的语义。世界，记作 ω，将 Σ中的每个变量映射到其某个状态。若公式 α 满足于世界 ω，则称 ω 为公式 α 的模型，记作 $\omega \models \alpha$（即， α 在 ω中为真）。常量 $\top$ 表示有效公式（在每个世界中都成立），常量 $\bot$ 表示不可满足公式（没有模型）。若公式 β 的每个模型也是公式 α 的模型，则公式 α 蕴含公式 β，记作 $\alpha \models \beta$。在这种情况下，我们称 α 比 β 更强，也称 β 比 α 更弱。

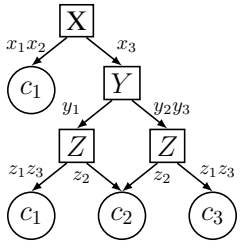
B. 使用类别公式表示分类器

离散变量 $X \in \Sigma$ 称为特征，状态 x_i（或简单文字 $\{x_i\}$）称为特征。实例是特征 Σ 中每个特征各取一个特征值后构成的合取（实例与世界一一对应）。

定义 *1*：一个具有类别 c_1、.....、c_n 的分类器，是由一组互斥且完备的公式 Δ_1、.....、Δ_n 构成的集合，称为类别公式。当且仅当 $\mathcal{I} \models \Delta_i$ 时，实例 $\mathcal{I}$ 属于类别 c_i 。

下面的决策图是一个分类器，包含三个三值特征 X， Y， Z 和三个类别 c_1， c_2， c_3。其类别公式为 $\Delta_1 = x_{12} + x_3 \cdot y_1 \cdot z_{13}$、 $\Delta_2 = x_3 \cdot z_2$ 和 $\Delta_3 = x_3 \cdot y_{23} \cdot z_{13}$。

这个分类器有 27 个实例：A型有 20 个实例，AB型有 3 个实例，O型有 4 个实例。例如，实例 $\mathcal{I} = x_3 \cdot y_2 \cdot z_2$ 属于 c_2 类，因为 $\mathcal{I} \models \Delta_2$。同样，尽管分类器可能有不同的形式，本文提出的可解释性理论将其视为一组类别公式，因为解释决策只需要这些内容。



[1]例如，我们见过离散化会导致变量具有数千个取值的随机森林。在这种情况下，我们希望避免二值化。

C. 作为范式的解释

问题一、问题二和问题三的答案将用具有范式的逻辑公式表示，接下来将对此进行讨论。项 是由不同变量的文字合取而成。子句 是由不同变量的文字析取而成。析取范式（*DNF*） 是若干项的析取。合取范式（*CNF*） 是若干子句的合取。否定范式（*NNF*） 是不含否定的公式。这些定义意味着，项不能自相矛盾，子句不能恒真，析取范式和合取范式都是否定范式，并且这些范式都不包含否定。若公式（项/子句）只包含简单文字，我们就称其为简单的 。简单公式和非简单公式将根据其传达决策信息的多少，划分出回答问题一、问题二和问题三的两种方法。

D. 使用条件化构造抽象

在构造实例的抽象以解释决策时，条件化的概念起着核心作用。首先请注意，像 $x_1 \cdot y_3 \cdot z_2$ 这样的简单项表示一组状态。公式 Δ 在简单项 τ 上的条件化记作 $\Delta|\tau$，并按如下方式得到：对于简单项 τ 中出现的每个状态 x_i，如果 $x_i \in \ell$，则将 Δ 中的每个 X文字 ℓ 替换为 $\top$；否则，将 ℓ 替换为 $\bot$。考虑公式$\Delta = x_{12} + x_3 \cdot y_1 \cdot z_{13}$和简单项 $\tau = x_3 \cdot z_1$。那么$\Delta|\tau = \bot + \top \cdot y_1 \cdot \top = y_1$。一般而言，条件化后的公式$\Delta|\tau$ 不包含项 τ 中出现的任何变量。

E. 使用主蕴含项和蕴含子句定义极小性

问题二和问题三涉及一种极小性概念，可通过主蕴含项和蕴含子句这两个经典概念来刻画。若项 τ 蕴含公式 Δ（ Δ （ $\tau \models \Delta$）），则称其为该公式的蕴含项 。若不存在比 τ 更弱的 Δ 的其他蕴含项（即，对于任何项 $\tau' \neq \tau$， $\tau \models \tau' \models \Delta$ 都不成立），则称其为主蕴含项 。若子句 σ被公式 Δ 蕴含（ Δ （$\Delta \models \sigma$）），则称其为该公式的蕴含子句。若不存在比σ 更强的 Δ 的其他蕴含子句（即，对于任何子句 $\sigma' \neq \sigma$， $\Delta \models \sigma' \models \sigma$ 都不成立），则称其为主蕴含子句 。若不存在其他主蕴含项 τ'满足 vars(τ') $\subset$vars(τ)，则主蕴含项 τ 称为变量极小 。变量极小素蕴含式的定义类似。

离散公式的主蕴含项和蕴含子句，其行为方式可能会让习惯于布尔语境下这些概念的人感到意外。此外，这些行为对问题二和问题三有着有趣的影响。考虑布尔公式 $\Delta_b = (x + z) \cdot (x \cdot z + y)$。项 $x \cdot y \cdot z$ 是蕴含项，但不是主蕴含项。弱化该项使其成为主蕴含项的唯一方法，是删除其中的一些变量_($y \cdot z$ 是一个主蕴含项）。在离散逻辑中，我们可以通过向某些文字添加状态，在不删除任何变量的情况下弱化一个项。考虑离散公式 $\Delta_d = (x \cdot z_1 + z_2) \cdot (x \cdot z_3 + y)$，其中变量 Z 是三值变量，而变量 X、Y 是二值变量。项 $x \cdot y \cdot z_1$ 是该公式的蕴含项，但不是主蕴含项。我们可以通过向文字 z_1 添加状态 z_2 来弱化该项，从而得到主蕴含项 $x \cdot y \cdot z_{12}$。对于主蕴含子句：强化布尔子句的唯一方法是删除其中的一些变量，而我们可以通过从离散子句的文字中移除状态来强化离散子句。

III. 决策背后的原因

现在我们转而回答问题一：对一个实例作出决策的原因是什么？我们将介绍两种不同的答案。第一种答案是过去几年可解释人工智能领域广泛发展的基础。第二种答案则是最近提出的，信息量更大，并且进一步涵盖了第一种答案。不过，如果分类器的所有特征都是二元的（即布尔值），这两种答案是等价的。

考虑一个由类别公式 $\Delta_1, \ldots, \Delta_n$ 指定的分类器，并假设它判定实例 $\mathcal{I}$ 属于类别 c_i，$\mathcal{I} \models \Delta_i$。我们需要知道其中的原因。这个问题的答案必须是实例上的一个条件，该条件能够推出公式 Δ_i。由于这个条件旨在表示实例的一种抽象，因此我们希望它尽可能弱，从而获得最一般的抽象。问题的有趣之处就在这里。如果我们不对这个条件施加额外要求，那么问题一 Q1 就会得到一个平凡的答案，因为类别公式Δ_i 满足我们提出的要求。原因如下：根据假设， $\mathcal{I} \models \Delta_i$，因此 Δ_i 确实是实例所满足的一个条件。此外，不存在其他关于实例 $\mathcal{I}$的条件，其弱于 Δ_i 且能够推出 Δ_i。这个答案是平凡的，因为在解释类别 c_i中任何实例的决策时，都会得到相同的答案。因此，我们需要对所寻求的条件提出一项额外要求，接下来将回顾为此目的而提出的两种方案。

A型. 完整理由

第一个提案要求，所寻求的条件——也就是决策的理由——必须由实例中出现的状态构成；这些状态可以任意组合，但只能使用合取和析取。严格来说，它要求该条件是一个简单的否定范式公式，其中所有文字都出现在待解释的实例中。由此便引出了 [5] 中提出的完整理由概念。[2]

定义*2*：假设实例 $\mathcal{I}$ 属于类别 c_i。对实例 $\mathcal{I}$ 所作决策的完整理由，是文字均取自 $\mathcal{I}$ 且满足 $\mathcal{I} \models \Gamma \models \Delta_i$ 的最弱否定范式 Γ。

这里我们强调，完整理由是一个简单的否定范式公式，也就是说，其中包含的每个文字都对应一个状态。

考虑一个具有三个三值特征 X， Y， Z 的分类器。实例 $\mathcal{I} = x_2 \cdot y_2 \cdot z_1$ 被判定属于类别 c_2，其公式为 $\Delta_2 = x_{23} \cdot (x_2 + y_{23}) \cdot (y_{23} + z_1)$。这一决策的完整理由是 $x_2 \cdot (y_2 + z_1)$。它表示，实例 $\mathcal{I}$ 属于类别 c_1，因为它具有特征 x_2，并且具有两个特征 y_2， z_1 之一。这确实是关于该实例的最一般条件，它蕴含类别公式，并且由对实例特征 x_2、 y_2 和 z_1 应用合取和析取得到。

现在有两个问题变得十分重要。第一，我们如何计算完整理由？第二，当完整理由过于复杂、无法被人类解读时，我们该如何使用它们

[2]参考文献 [5]给出了不同但等价的定义，并适用于二元变量。有关更一般的论述，请参见 [4] 。

？接下来我们将讨论第一个问题，而把第二个问题留到本教程稍后再处理。

参考文献 [5] 指出，如果类别公式得到了适当的表示（例如，表示为有序二叉决策图 OBDD [6]），就可以在线性时间内计算出完整理由。随后， [7] 取得了一项重大进展，提出了以下量化算子，并证明可以利用它来表达完整理由。

定义 *3*：对于具有状态 x_1，……，x_n 的变量 X，公式 Δ 中状态 x_i 的全称文字量化 定义为 $\forall x_i \cdot \Delta = \Delta|x_i \cdot \prod_{j \neq i}(x_i + \Delta|x_j)$。

这个算子具有交换性，因此可以按照任意顺序对 $\mathcal{I}$ 的状态进行量化，从而将实例 $\mathcal{I}$ 从其类别公式 Δ_i，（写作 $\forall \mathcal{I} \cdot \Delta_i$）中量化出去。此外，如 [7] 所示，对于布尔变量， $\forall \mathcal{I} \cdot \Delta_i$ 对应完整理由；如 [4] 所示，对于离散变量， $\forall \mathcal{I} \cdot \Delta_i$ 也对应完整理由。这个结果的一个重要意义是，该量化算子在多种逻辑形式上的计算表现良好。例如，可以在线性时间内从合取范式中量化任意一组状态。更一般地，如 [7] 所示，该算子可分配于合取；当各析取项不共享变量时，也可分配于析取（较弱的条件见 [4]）。这使得以下类别公式形式受到关注，因为它可以在线性时间内计算完整理由（以及第三节 B 部分讨论的一般理由）。

定义 *4*： NNF 当且仅当对于 NNF 中的任意析取项 α 和 β，$\text{vars}(\alpha) \cap \text{vars}(\beta) = \emptyset$。[3]

由于这种形式及其某些弱化形式， [4] 为基于决策图的某些类别分类器提供了闭式完整理由。

B. 一般理由

现在我们转向一项近期进展： [2] 该进展确定了关于实例抽象的一项较弱要求，并且能够在解释决策时产生更多信息。[4]

定义 *5*：假设实例 $\mathcal{I}$ 属于类别 c_i。实例 $\mathcal{I}$ 上该决策的一般理由 是最弱的 NNF Γ，其文字由 $\mathcal{I}$ 蕴含，并且满足 $\mathcal{I} \models \Gamma \models \Delta_i$。

完整理由与一般理由之间的唯一差异在于，前者要求文字出现在实例中，而后者只要求这些文字由实例蕴含。因此，一般理由不再像完整理由那样是简单公式，因为其中可能包含非简单文字。如果所有特征都是二值特征，那么每个文字都是简单的，完整理由与一般理由也就相等。然而，在存在非二值特征时，一般理由能够提供更多关于为何作出某项决策的信息。让我们通过一个例子来说明这一点。

考虑图 3 中的分类器，以及一位患者 Rob，

$$(\textsc{Diabetes}{=}\textsc{y}) \cdot (\textsc{Weight}{=}\textsc{under}) \cdot (\textsc{Btype}{=}\textsc{A}),$$

[3]这与与可分解 NNF（称为可分解 NNF（DNNF））形成对比，后者已得到广泛研究。 [8], 定义 5 实际上是 [2] 中的一个命题。我们在此反转了表述，以配合本教程的叙述脉络。[4]

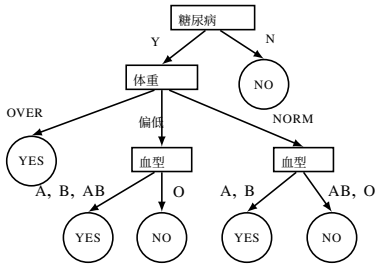


图 3. 以决策树形式表示的某种疾病分类器 [2].

其决策为 是。该决策的完整理由是

$$(\textsc{Diabetes}{=}\textsc{y}) \cdot (\textsc{Btype}{=}\textsc{A})$$

而一般理由是

$$(\textsc{Diabetes}{=}\textsc{y}) \cdot (\textsc{Btype} \in \{\textsc{A}, \textsc{B}, \textsc{AB}\}) \cdot (\textsc{Weight} \in \{\textsc{under}, \textsc{over}\} + \textsc{Btype} \in \{\textsc{A}, \textsc{B}\}).$$

完整理由通过 Rob 患有糖尿病且其血型为 A型来论证决策。但一般理由提供了较弱的论据。其中之一是 Rob 患有糖尿病，且其血型不是 AB型或 O型。另一个论据是，他患有糖尿病，体重不正常，且血型不是 O型。

一般理由所提供的附加信息在离散化分类器的背景下尤为关键。再次考虑图2中的决策树，该决策树具有两个数值特征 AGE、身体质量指数。特征 AGE 被离散化为三个区间 $[0, 18), [18, 40)$ 和 $[40, \infty)$，对应标签为 a_1、a_2、a_3。特征 BMI 被离散化为四个区间 $[0, 25), [25, 27), [27, 30)$ $[30, \infty)$，对应标签为 b_1、b_2、b_3、b_4。因此，AGE 可以被视为具有状态 a_1、a_2、a_3 的离散变量，而 BMI 可以被视为具有状态 b_1、b_2、b_3、b_4 的离散变量。现在考虑实例 AGE$=42 \cdot$ BMI$=28$，该实例被离散化为 $a_3 \cdot b_3$（AGE ≥ 40 和 $27 \leq$ BMI < 30）。该实例的决策为 是，其类别公式 Δ_{YES} 为 $a_1 \cdot b_4 + a_2 \cdot b_{34} + a_3 \cdot b_{234}$（通过追踪通向类别 是 的路径获得）。该决策的完整理由为 $a_3 \cdot b_3$，即实例本身，因此没有发生抽象。一般理由为 $a_{23} \cdot b_{34} + a_3 \cdot b_{234}$，其内容是

$$(\textsc{Age} \geq 18) \cdot (\textsc{BMI} \geq 27) + (\textsc{Age} \geq 40) \cdot (\textsc{BMI} \geq 25).$$

这是实例的一种较弱性质，但仍然能够推出该决策。例如，第一部分表示，Age 不低于 18 且身体质量指数不低于 27 足以论证该决策。这类论据无法通过完整理由获得，因为完整理由只能引用实例特征，也就是 Age ≥ 40 和 $27 \leq$ 身体质量指数 < 30。

一般理由也可以使用 [2] 中引入的量化算子来表示。

定义*6*：对于变量 X，其状态为 x_1、……、x_n，且公式为 Δ，$\forall x_i \cdot \Delta$ 定义为 $\Delta|x_i \cdot \Delta$。

这个运算符也具有交换性，因此我们可以通过以任意顺序量化 $\mathcal{I}$ 的状态，根据其类别公式 Δ_i，$\forall \mathcal{I} \cdot \Delta_i$ 对实例 $\mathcal{I}$ 进行量化。此外， $\forall \mathcal{I} \cdot \Delta_i$

对应于如 [2]所示的实例$\mathcal{I}$ 决策的一般理由。就合取范式和或可分解否定范式等形式的分配性和易处理性而言，算子 $\forall$ 和 $\forall$ 具有相似的计算性质。因此，与完整理由类似，对于某些类别的分类器 [2]，我们也有闭式一般理由。

C. 固定性与单调性

完整理由和一般理由具有一种称为固定性的性质，它具有重要意义，包括对计算的影响，我们将在后文讨论。 [2]中也指出了这一性质。

定义 7：当一个否定范式在实例 $\mathcal{I}$ 上的文字与 $\mathcal{I}$ 一致时，称其在该实例上局部固定。如果一个公式等价于一个在实例 $\mathcal{I}$ 上局部固定的否定范式，则称该公式在实例 $\mathcal{I}$ 上固定。

公式 $x_{12}\cdot y_{13}+z_{23}$ 在实例 $\mathcal{I}=x_1\cdot y_1\cdot z_2$ 上局部固定，但公式 $x_{12}\cdot y_{23}+z_{23}$ 并非如此，因为文字 y_{23} 与实例 $\mathcal{I}$ 不一致。完整理由的固定性直接由定义 2 得出，而一般理由的固定性则直接由定义 5 得出。

如果一个简单公式具有固定性，那么它也是单调的。

定义 *8*：对于每个变量 X，公式中出现的所有 X-文字均为简单且相同，当且仅当该公式单调。

例如，公式 $(x_1+y_2)\cdot(x_1+z_3)\cdot(y_2+z_3)$ 是单调的，但公式 $(x_1+y_2)\cdot(x_{12}+z_3)\cdot(y_2+z_3)$ 不是单调的，因为它包含一个非简单文字 x_{12}。公式 $(x_1+y_2)\cdot(x_1+z_3)\cdot(y_3+z_3)$ 也不是单调的，因为对于变量 Y，它包含不同的文字 y_2 和 y_3。众所周知，完整理由是单调的；在计算完整理由的主蕴含项和蕴含子句时， [4], [5] 利用了这一性质。

D. 选择语义

对于实例 $\mathcal{I}$ 及其类别公式 Δ_i，我们有 [2]:

$$\mathcal{I}\models\forall\mathcal{I}\cdot\Delta_i\models\overline{\forall}\mathcal{I}\cdot\Delta_i\models\Delta_i$$

因此，一般理由总是弱于完整理由。上述关系还表明，无论是完整理由还是一般理由，都强于类别公式。因此，算子 $\forall$ 和 $\forall$ 可以被视为<覆选择算子，因为它们会从类别 c_i [2]、[7]中的实例中选择子集。具体而言， $\forall\mathcal{I}\cdot\Delta_i$ 会选择类别 c_i 中的所有实例，只要它们属于该类别这一点不取决于与实例 $\mathcal{I}$不一致的特征。例如，假设实例 $\mathcal{I}'$ 属于类别 c_i ，并且仅在特征 X、Y、Z的状态上与实例 $\mathcal{I}$存在差异。那么，当改变其特征X、Y、Z 的任意子集的状态后，所得实例也属于类别 c_i时， $\mathcal{I}'$才会被 $\forall\mathcal{I}\cdot\Delta_i$ 选中。相比之下，$\forall\mathcal{I}\cdot\Delta_i$ 会选择所有实例，只要将其任意特征改变为与实例 $\mathcal{I}$一致后，它们仍属于类别 c_i （其中包括由 $\forall\mathcal{I}\cdot\Delta_i$选中的实例）。在前面的例子中，当将实例 $\mathcal{I}'$的特征 X、Y、Z的任意子集的状态设为它们在 $\mathcal{I}$ 中的状态后，所得实例也属于类别 c_i时，实例 $\mathcal{I}'$才会被 $\forall\mathcal{I}\cdot\Delta_i$ 选中。

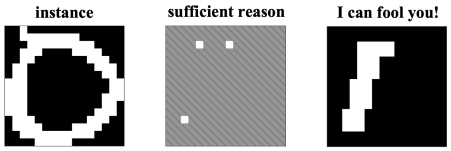


图 4. 一个 来自分类数字域的充分理由 字 [11]。

IV. 决策的充分理由

现在我们来回答问题二：一个实例的哪些最小方面能够保证对该实例作出的决策？我们将给出两种答案，这取决于我们如何定义“方面”。第一种答案基于以下概念。

定义*9*：假设实例 $\mathcal{I}$ 属于类别 c_i。针对实例 $\mathcal{I}$ 的决策的充分理由（*SR*）是完整理由 $\forall\mathcal{I}\cdot\Delta_i$ 的主蕴含项。

充分理由是一个项 $\tau=\ell_1\cdot\ldots\cdot\ell_n$，其中每个文字 ℓ_i 都是简单的（即一个状态），且出现在实例 $\mathcal{I}$ 中。[5]这会得出关系 $\mathcal{I}\models\tau\models\forall\mathcal{I}\cdot\Delta_i\models\Delta_i$，进而意味着，任何包含状态 ℓ_1、.....、 ℓ_n 的实例，其决策都将与实例 $\mathcal{I}$ 相同。此外，这些状态的任何真子集都不满足这一点；否则， τ 就不可能是 $\forall\mathcal{I}\cdot\Delta_i$ 的主蕴含项。换言之，充分理由 τ 是实例 $\mathcal{I}$ 中一组最小特征，用以证明对该实例作出的决策合理；因此，实例的其他特征可视为与决策无关，因为无论如何改变这些特征，都不会影响该决策。

充分理由这一概念最初在 [9]中以另一种但等价的定义提出，当时称为*PI*解释 。“充分理由”一词最早在 [5]中使用，该文引入了完整理由，并指出其主蕴含项就是PI解释。充分理由有时也称为溯因解释 [10]。

以一个具体的充分理由为例，考虑图 4 所示的 [11]，图像 16×16 。左侧的图像被输入一个卷积神经网络（CNN），该网络的任务是对数字 0,1 进行分类，而 CNN 正确地将该图像分类为数字 0。图 4 中间展示了这一决策的一个充分理由。这个充分理由包含 3 个白色像素（共 256 个像素），因此，任何包含这 3 个白色像素的图像都会被该 CNN 分类为数字 0，如图 4 右侧所示。

再举一个具体例子，考虑图 3 中的分类器以及以下患者 Lara，

$$\mathcal{I}\colon\ (\textsc{Diabetes}{=}\textsc{y})\cdot(\textsc{Weight}{=}\textsc{over})\cdot(\textsc{Btype}{=}\textsc{a}). \tag{1}$$

关于 Lara 的决策是 是，类别公式 Δ_{YES} 是

$$\begin{aligned}&(\textsc{Diabetes}{=}\textsc{y})\cdot\\&\quad[(\textsc{Weight}{=}\textsc{over})+\\&\quad(\textsc{Weight}{=}\textsc{norm})\cdot(\textsc{Btype}\in\{\textsc{a},\textsc{b}\})+\\&\quad(\textsc{Weight}{=}\textsc{under})\cdot(\textsc{Btype}\in\{\textsc{a},\textsc{b},\textsc{ab}\})].\end{aligned} \tag{2}$$

[5]由于完整理由简单且聚焦于该实例 $\mathcal{I}$。

此决策的完整理由是 $\forall \mathcal{I} \cdot \Delta_{\text{YES}}$

$$(\text{DIABETES}=\text{Y}) \cdot [(\text{WEIGHT}=\text{OVER}) + (\text{BTYPE}=\text{A})]. \quad (3)$$

它有两个主蕴含项。

$$(\text{DIABETES}=\text{Y}) \cdot (\text{WEIGHT}=\text{OVER}) \quad (4)$$

$$(\text{DIABETES}=\text{Y}) \cdot (\text{BTYPE}=\text{A}) \quad (5)$$

这些构成了该决策的充分理由。因此，对于 Lara 而言，鉴于她的特征（$\text{D}_{\text{型}}=\text{Y}$）·（$\text{W}_{\text{型}}=\text{OVER}$），特征 $\text{B}_{\text{型}}$ 与该决策无关；而在给定（$\text{D}_{\text{型}}=\text{Y}$）·（$\text{B}_{\text{型}}\text{A}$）的情况下，特征 $\text{W}_{\text{型}}$ 与该决策无关。

通过采用近期在 [2] 中所展示的决策背后的一般理由，可以得到较弱的充分理由，从而找出实例中更多的无关方面。

定义*10*：假设实例 $\mathcal{I}$ 属于类别 c_i。关于 $\mathcal{I}$ 的决策的一般充分理由（*GSR*）是一般理由 $\forall \mathcal{I} \cdot \Delta_i$ 的变量极小主蕴含项。[6]

Lara决策的一般理由是

$$\begin{aligned}&(\text{DIABETES}=\text{Y})\cdot\\&\quad[(\text{WEIGHT}=\text{OVER}) + (\text{BTYPE} \in \{\text{A}, \text{B}\}) +\\&\quad(\text{WEIGHT} \in \{\text{UNDER}, \text{OVER}\}) \cdot (\text{BTYPE} \in \{\text{A}, \text{B}, \text{AB}\})]\end{aligned} \quad (6)$$

并有三个主蕴含项

$$(\text{DIABETES}=\text{Y}) \cdot (\text{WEIGHT}=\text{OVER}) \quad (7)$$

$$(\text{DIABETES}=\text{Y}) \cdot (\text{BTYPE} \in \{\text{A}, \text{B}\}) \quad (8)$$

$$(\text{糖尿病}=\text{Y}) \cdot (\text{体重} \in \{\text{UNDER}, \text{高于}\}) \cdot (\text{血型} \in \{\text{A型}、\ \text{B}、\ \text{AB型}\}) \quad (9)$$

只有前两个是变量极小的，因此它们是该决策的一般充分理由。我们已在（4）中看到（7）是一个充分理由。然而，（8）中的一般充分理由弱于（5）中的充分理由，并且传达了有关该决策的更多信息。具体而言，（5）中的充分理由指出，Lara 的体重与该决策无关，因为她患有糖尿病，且她的血型为 A型；而（8）中的一般充分理由则指出，这是因为她患有糖尿病，且她的血型不是 AB型或 O型。

每个一般充分理由都与该实例相容，因为一般理由以该实例为固定依据。将实例 $\mathcal{I}$ 与一般充分理由 τ 的交集 定义为 $\mathcal{I}$ 中蕴含 τ 的最小子集。那么，该交集必定是一个充分理由。此外，当且仅当满足定义 10 中的变量极小性条件时，这一保证才成立。例如，(1) 和 (8) 的交集是 $(\text{DIABETES}=\text{Y}) \cdot (\text{BTYPE}=\text{A})$，它是一个充分理由；而 (1) 和 (9) 的交集是 $(\text{DIABETES}=\text{Y}) \cdot (\text{WEIGHT}=\text{OVER}) \cdot (\text{BTYPE}=\text{A})$，它不是充分理由。[7] 更一般地，某项决策的充分理由恰好是一般充分理由与该实例的交集。因此，一般充分理由涵盖其经典对应项，并为决策及其背后的原因提供更多信息。

[6]这个定义出现在 [2] 一篇文章中，该文提出了一种不同的一般充分理由定义，且未引用一般理由。[7]参考文献 [3] 建议使用非变量极小的主蕴含项。分类器。最后我们再次指出，如果所有特征都是二元的，则完整理由与一般理由等价，因此一般充分理由可简化为经典充分理由。此前所有观察均可由 [2]中的结果推出。

V. 决策的必要理由

现在我们转而回答问题三：对一个实例进行哪些最小改动，才能改变针对该实例的决策？回答这个问题将找出一些最小特征集，恰当地改变这些特征后即可翻转决策。这个问题有两个相互独立的变化维度。第一，确定是否也要明确指出对这些特征进行哪些具体改动；这取决于我们为此采用完整理由还是一般理由。第二，涉及我们是要将决策翻转为某个特定的替代决策，还是任意不同的决策。

A. 削弱当前决策

我们在本节中假设，我们的目标是推翻当前的决策，而不针对某个特定的替代决策。我们可以通过以下概念实现这一目标。

定义 *11*：假设实例 $\mathcal{I}$ 属于类别 c_i。对实例 $\mathcal{I}$ 作出决策的必要理由（*NR*）是完整理由 $\forall \mathcal{I} \cdot \Delta_i$ 的主蕴含子句。

每个必要理由都是一个子句 $\ell_1 + \ldots + \ell_n$，其中每个文字 ℓ_i 都是简单的（一个状态），并且出现在实例 $\mathcal{I}$中。[8] 因此，每个必要理由都由该实例满足。然而，如果我们以违反某个必要理由的方式改变实例，就不再能保证决策得以维持。假设状态 ℓ_i 对应于特征 X_i。为了违反必要理由 $\ell_1 + \ldots + \ell_n$，我们必须改变实例 X_i 中每个特征的状态。必要理由所提供的保证是：至少有一种这样的改变会使决策发生反转，并且不存在能够使决策发生反转的该改变的严格子集。

必要理由的概念最初在 [12] 中以另一种但等价的定义形式化，并称为对比解释 [13]。必要理由有时也称为反事实解释 [14]。[9]“必要理由”一词最早在 [4] 中使用；该文指出，对比解释是完整理由的主蕴含子句，并给出了此前讨论过的必要性语义（即若要保留决策，就必须保留的一种性质）。现在，让我们通过一个具体示例来说明必要理由。

我们回到（1）中的患者 Lara，分类器在图 3 中将其判定为$_{\text{是}}$。类别公式见（2），该决策的完整理由见（3）。这个完整理由有两个主蕴含子句，分别是（$\text{D}_{\text{糖尿病}}=\text{Y}$）和（$\text{W}_{\text{体重}}=\text{OVER}$）+（$\text{B}_{\text{血型}}=\text{A}$）。违反第一个理由只有一种方法，即将 $\text{D}_{\text{糖尿病}}$ 设为$_{\text{N}}$，这确实会使决策翻转为$_{\text{否}}$，如图 3 所示。

[8]由于完整理由简单且固着于该实例。9

哲学、社会科学和人工智能领域有大量研究探讨对比解释和反事实解释；例如参见 [15]–[24]。虽然这些概念的定义有时只是彼此的变体或改进版本，但它们并不总是兼容的。

第二个必要理由可以通过六种不同方式被违反：将 WEIGHT 设为 {UNDER, NORM} 中的一个状态，并将 BTYPE设为 {B、AB、O} 中的一个状态。这些改变／违反并非都会改变决策，但至少有一种会。将 (WEIGHT=UNDER) · (BTYPE=O) 改变会改变决策，但将 (WEIGHT=UNDER) ·(BTYPE=AB) 改变则不会（始终假设 DIABETES 保持不变）。我们再次强调，第一个改变的任何严格子集都不会改变决策，因为必要理由所建议的改变是最小的。

现在我们将看到，如果采用 [2] 中最近提出的决策一般理由，就能做得比这更好。

定义 *12*：假设实例 $\mathcal{I}$ 属于类别 c_i。关于 $\mathcal{I}$ 的决策的 一般必要理由（*GNRs*）是一般理由 $\forall \mathcal{I} \cdot \Delta_i$ 的变量极小素蕴含式。[10]

与必要理由一样，违反一般必要理由会削弱该决策。不过，我们现在还有一项额外保证：每一次 对一般必要理由的违反都会翻转该决策。如果我们不坚持使用变量极小素蕴含式，这项保证就不会成立；而如果某个并非变量极小的素蕴含式满足这项保证，那么它所建议的某些变更就不可能是极小的（也就是说，我们可以通过对实例进行更少的变更来翻转该决策）。此外，任何能够翻转决策且由某个必要理由提出的变更，也会由某个一般必要理由提出。换言之，后者蕴含前者，并且传达了关于决策及其底层分类器的更多信息。[11] 让我们来看一个具体例子。

Larais 决策的一般理由见（6），并有三个素蕴含子

$$
\begin{aligned}
&(\text{DIABETES}=\text{Y})\\
&(\text{WEIGHT}=\text{OVER}) + (\text{BTYPE} \in \{\text{A}, \text{B}, \text{AB}\})\\
&(\text{WEIGHT} \in \{\text{UNDER}, \text{OVER}\}) + (\text{BTYPE} \in \{\text{A}, \text{B}\}).
\end{aligned}
$$

它们都具有变量极小性，因此都是一般必要理由。第一条只能通过一种方式违反：将 DIABETES 设为N ，这会改变决策。第二条可以通过两种方式违反：将 WEIGHT 设为 {UNDER, NORM} 中的某个状态，并将 BTYPE 设为状态 O。正如使用图 3 中的分类器可以验证的那样，这两种违反方式都会改变决策。第三条一般必要理由可以通过两种方式违反：将 WEIGHT 设为状态NORM ，并将 BTYPE设为 {AB,O} 中的某个状态。同样，这两种违反方式都会改变决策。因此，违反这些一般必要理由中的任意一条都必定会改变决策。此外，必要理由所建议的每一种会改变决策的变更，也都由某条一般必要理由所建议。

我们以以下几点结束本节。布尔归结从子句 $x+\alpha$ 和 $x+\beta$ 推导出子句 $\alpha+\beta$。此外，计算布尔合取范式的主蕴含子句的一种经典方法是，对合取范式进行归结闭包，并在每个归结步骤后删除被包含的子句；例如参见 [25], [26]。正如 [2], 最近所示，这也可以用于离散合取范式，但需使用广义归结规则：该规则从子句 $\ell+\alpha$ 和 $\ell'+\beta$ 推导出子句 $\ell\cdot\ell'+\alpha+\beta$，其中 ℓ 和 ℓ' 是同一变量的文字。尤其值得注意的是：如果合取范式是局部固定的（定义 7），一般理由正是如此，那么每个归结步骤后，只需丢弃非变量极小的子句，就能计算出一般必要理由 [2]。

B. 针对新的决策

前面的讨论展示了如何利用（一般）必要理由来翻转决策，但无法保证新的决策是什么。当分类器有两个以上类别，因而能够做出两个以上决策时，这就成了一个问题。设有一个分类器，用于决定是否批准申请人的大额贷款（c_1）、批准小额贷款（c_2）或拒绝申请（c_3）。令 Δ_1, Δ_2, Δ_3为相应的类别公式，并假设有一位申请人 $\mathcal{I}$ 获批了小额贷款，即$\mathcal{I} \models \Delta_2$。（一般）必要理由确实指出了能够翻转该决策的最小申请变更，但无法保证新的决策会批准更大额度的贷款，还是拒绝该申请。假设我们希望翻转该决策，使申请人获批更大额度的贷款（c_1）。我们可以合并类别 c_2和 c_3 ，从而得到一个具有两个决策 c_1, c_{23} 的分类器，以及相应的类别公式 Δ_1 和 $\Delta_{23} = \Delta_2 + \Delta_3$。在原分类器 $\mathcal{I} \models \Delta_2$中，批准小额贷款的决策（$c_2$）现在变成了新分类器 $\mathcal{I} \models \Delta_{23}$中的一项决策，即批准小额贷款或拒绝申请（$c_{23}$）。因此，翻转这一决策就能保证申请人获批大额贷款（c_1）。

更一般地，假设我们有一个分类器，其类别为c_1、...、c_n，类别公式为 Δ_1、.....、Δ_n，并且有一个属于类别 c_i $\mathcal{I} \models \Delta_i$的实例 $\mathcal{I}$。如果我们希望以最小幅度改变该实例，使其转入其他某个类别 c_j，就需要计算完整理由$\forall \mathcal{I} \cdot \bigvee_{k\neq j} \Delta_k$ 或一般理由 $\forall \mathcal{I} \cdot \bigvee_{k\neq j} \Delta_k$。然后，可以按照前文所述，计算这些理由的（变量极小）主蕴含子句。[4] 提出了这种合并类别公式的技术，由此得到的必要理由后来被称为目标对比解释 [12]。

VI. 编译类别公式

全面讨论如何为各种类型的分类器编译类别公式，超出了本教程的范围。[12] 不过，我们会分享一些关于这一编译过程的重要见解，让它不再那么神秘。首先要指出的是，该过程所依据的技术取决于分类器的类型，而其难度则取决于分类器的类型以及所需编译公式的形式。许多研究致力于将公式编译为易处理电路 [27]，但本文基本上会忽略这一方面，因为将公式编译为易处理电路是一个已被充分理解、并在知识编译 [28]领域得到广泛研究的问题。

[10]这个定义是 [2] 中的一个命题，该命题为一般必要理由提供了另一种定义，且未提及一般理由。[11]一般必要理由在实例 $\mathcal{I}$ 上被固定（定义 7）。

[12]“类别公式”一词最早见于 [4]。编译分类器的“决策函数”或“输入输出行为”是更常见的说法。

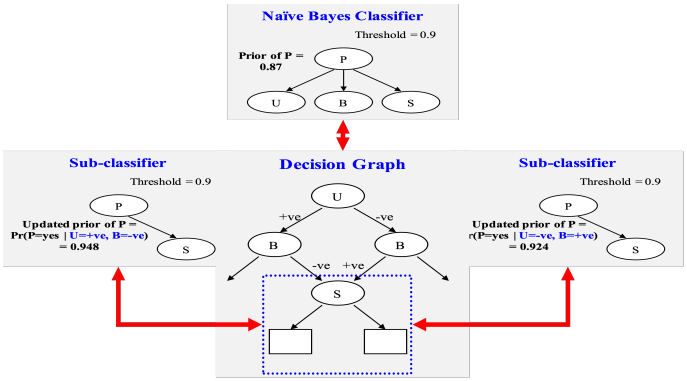


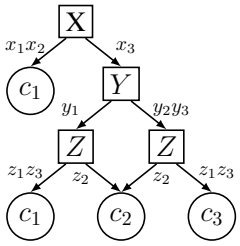
图 5. 将朴素贝叶斯分类器编译为决策图 [29]。

A. 贝叶斯网络

编译贝叶斯网络分类器的类别公式，概念上或许是最微妙的，因为要判断某个实例属于哪个类别，需要进行概率推理。这一编译过程背后有两个基本见解。第一个是 [29] 中引入的等价区间概念。考虑图 1(a) 中的朴素贝叶斯分类器，其类别分布为 $\Pr(P{=}\text{YES})$、$\Pr(P{=}\text{NO}) = (0.87, 0.13)$。只要 $\Pr(P{=}\text{YES}) \in [0.684, 0970]$，我们就可以显著改变这一分布，而不改变该分类器作出的任何决策；这一区间称为等价区间。第二个基本见解是 [29] 中同样引入的等价子分类器概念。继续看同一个分类器，假设将特征 U、B 分别设为 +VE、–VE。这会得到一个关于单个特征 S 的子分类器，其类别分布为 (0.948, 0.052)，因为 $\Pr(P{=}\text{YES}|U{=}\text{+VE}\ B{=}\text{–VE}) = 0.948$。同样，如果将这两个特征设为 –VE +VE，我们会得到另一个关于相同特征 U 的子分类器，但其类别分布不同，为 (0.924, 0.076)，因为 $\Pr(P{=}\text{YES}|U{=}\text{–VE}\ B{=}\text{+VE}) = 0.924$。利用等价区间的概念，我们可以得出结论：这两个子分类器是等价的，也就是说，它们对各自的子实例 $S{=}$+VE 和 $S{=}$–VE 作出相同的决策。结合这两个见解，我们可以通过对特征赋值空间进行深度优先搜索，将贝叶斯网络分类器编译为一个（符号）决策图；每当遇到一个与先前遇到（已编译）的子分类器等价的子分类器时，就剪枝搜索；见图 5。这一方法曾用于编译 [29], 中的朴素贝叶斯分类器、[9] 中的树结构分类器，以及 [30] 中的图结构分类器。计算等价区间和识别等价子分类器的技术取决于分类器的结构，这也解释了这一发展顺序。图 1(b) 展示了采用此方法编译图 1(a) 中朴素贝叶斯分类器得到的决策图。接下来将讨论如何从决策图中提取类别公式。

B. 决策图与随机森林

接下来我们对决策图的讨论也适用于决策树，因为决策树是决策图的一种特殊情况。众所周知，决策图的类别公式可以直接写成析取范式。要构造类别 c_i 的析取范式，我们只需为从根节点到标记为 c_i 的叶节点的每条路径构造一个项，然后将这些项析取起来。例如，在下方的决策图中，从根节点到类别 c_1 有两条路径，它们生成析取范式 $\Delta_1 = x_{12} + x_3 \cdot y_1 \cdot z_{13}$。



对于决策图（非树），这可能会生成一个比图大得多（呈指数级）的析取范式。我们可以通过构造一个大小保证与决策图大小成线性关系的 NNF 电路来缓解这一问题。 [4] 接下来给出的构造会生成满足更强性质的电路，因此在条件较弱的情况下，这些电路可以让我们在与电路大小成线性关系的时间内，计算完整的一般理由。

设 c_1、……、c_n 是决策图的类别（即叶节点的标签），并假设我们希望为某个类别 c_k 的公式构造一个 NNF 电路。我们首先定义函数 $\text{nnf}(N)$，它将决策图中的节点 N 映射为一个 NNF 片段，定义如下。若节点 N 有出边 $\xrightarrow{\ell_1} C_1$、……、$\xrightarrow{\ell_m} C_m$，则 $\text{nnf}(N) = \sum_{i=1}^{m}(\ell_i \cdot \text{nnf}(C_i))$。对于叶节点，若 $c_i = c_k$，则 $\text{nnf}(c_i) = \top$；若 $c_i \neq c_k$，则 $\text{nnf}(c_i) = \bot$。现在，我们可以通过调用 $\text{nnf}(R)$ 将决策图转换为 NNF 电路，其中 R 是图的根节点。这是标准方法，但 [4] 以不同方式定义了函数 nnf(.)，使 NNF 电路满足一些理想性质。具体而言，对于内部节点 N，它使用 $\text{nnf}(N) = \prod_{i=1}^{m}(\ell_i' + \text{nnf}(C_i))$，其中文字 ℓ_i' 是文字 ℓ_i 的补。叶节点保持不变：若 $c_i = c_k$，则 $\text{nnf}(c_i) = \top$；若 $c_i \neq c_k$，则 $\text{nnf}(c_i) = \bot$。新方法等价于先使用第一种方法为类别 c_1、……、c_{k-1}、c_{k+1}、……、c_n 的并集构造 NNF 电路，然后根据德摩根定律对所得电路取反。

如果决策图满足至多测试一次性质 （任意路径上一个特征最多被测试一次），则可以保证通过这种新方法构造的 NNF 电路是或可分解的（定义 4）。请回想一下，如果类别公式是或可分解否定范式，那么可以在线性时间内计算完整理由和一般理由，因为算子 $\forall$ （定义 3）和 $\overline{\forall}$ （定义 6）会在此类否定范式中对析取和合取进行分配。若决策图满足 [4] 中讨论的弱至多测试一次性质 ，也可以得到能够在线性时间内计算完整理由和一般理由的 NNF 电路。离散化决策图满足这一性质（例如图2(b)中的决策树）。上述构造及其变体是 [4] 中提出的闭式完整理由和 [2] 中闭式一般理由的基础。

对于采用多数投票的随机森林，可以通过使用多数电路组合森林中各棵树对应的 NNF 电路，轻松构造出类别公式的 NNF 电路。但即使各棵树对应的电路是析取可分解的，最终得到的电路也不一定是析取可分解的。

C. 神经网络

将神经网络分类器编译为类别公式的过程较为复杂，因为相关技术和假设多种多样，例如激活函数的类型，以及网络是二元、二值化还是量化的。因此，我们将仅讨论一种适用于二值神经网络的方法，并介绍其他方法的相关资料。

我们将介绍的研究 [11] 假设神经网络具有二值输入和阶跃激活函数。该编译技术基于若干观察。首先，具有阶跃激活的神经元输出为二值，因此，如果其输入也为二值，那么该神经元就表示一个布尔函数。因此，具有二值输入和阶跃激活函数的神经网络必定表示一个布尔函数（即，其输入、内部线路和输出上的信号都必须属于 $\{0,1\}$）。因此，如果能将神经元编译为布尔电路（即具有二值输入和阶跃激活函数的神经元），就可以将神经网络转换为布尔电路（从中可以轻松提取类别公式）。下一个观察是，具有阶跃激活的神经元是一种线性分类器，与朴素贝叶斯分类器类似。因此，我们先前讨论过的、将朴素贝叶斯分类器编译为决策图的技术，可以从 [29], 中采用，用于将这类神经元编译为决策图，进而编译为 NNF 电路。成功编译这类神经元后，就能立即将神经网络表示为布尔电路，并由此轻松获得类别公式。这项研究更进一步，采用了 [29] 中编译方法的一种变体，该方法假设神经元权重 $w_1, \ldots, w_n$ 和阈值 T 均为整数。这样便可以在 $O(nW)$ 时间内完成编译，其中 $W = |T| + \sum_{i=1}^{n} |w_i|$ 是绝对值之和。这种伪多项式时间编译算法也可用于实值权重：只需将权重乘以一个常数后再截断即可（即权重具有固定精度）。该技术使得编译具有数百个输入（特征）的神经网络成为可能。图 4 中的示例以及 [11]中的许多其他示例，都是通过这种方法生成的。

关于以符号方式编码输入输出行为的更多技术，请参见 [31]–[37] 中针对二值化神经网络的研究，以及 [38]–[41]中针对量化神经网络的研究。不过，这些研究大多着眼于验证任务，而非解释行为。

VII. 总结 remarks

我们以三点评论结束本教程。第一点评论涉及可解释人工智能领域的两类不同研究。一类研究假设我们掌握了足够多的分类器信息，从而能够构建类别公式，或者更一般地，刻画哪些实例属于哪些类别。这显然是本教程采用的假设，由此产生的方法称为基于模型的。 这些方法通常旨在提供具有严格保证的解释，例如我们讨论过的那些保证。另一类研究则假设我们只能查询分类器，也就是说，让它对实例进行分类。这些方法称为模型无关的 ，并且主要是通过早期系统而广为人知，这些系统见于 [42], [43]。这些方法往往更具可扩展性，但不提供严格保证，可以视为是在计算近似解释 [44]。

第二点评论涉及过去几年中针对可解释性所开展的大量研究，尤其是关于计算解释的复杂性。以下是一些例子。对于朴素贝叶斯（以及线性）分类器，研究表明，可以在对数线性时间内生成一个充分理由，并以多项式延迟生成所有充分理由 [45]。对于决策树，研究表明，生成一个充分理由的复杂度为多项式时间 [46]。后续研究表明，决策图 [47] 和某些类别的易处理电路 [14], [48]也具有相同的复杂度。文献 [49],还研究了决策树充分理由的生成，包括最短充分理由的生成；研究表明，即使只生成一个最短充分理由也是困难的。文献 [4]在更广泛的背景下研究了最短充分理由的生成，其中涵盖决策图和 SDD。[13] 文献 [55] 研究了最短充分理由的复杂度，涉及对应于决策图的布尔分类器，以及具有 ReLU 激活函数的神经网络。研究还进一步表明，必要理由的数量与决策树的大小呈线性关系 [4], [47],，所有此类理由都可在多项式时间内计算得出 [4], [49],；如果分类器满足文献 [56]所述的某些条件，则可以通过多项式延迟枚举最短必要理由。文献 [14], [48], 还给出了更多复杂性结果，其中根据表示分类器的易处理电路 [48] 或其允许在多项式时间内执行的处理类型 [14], 对分类器进行了分类。文献 [57] 还针对大量解释查询和多类布尔分类器，对复杂性进行了全面研究。此外，研究者还提出了基于 SAT、MaxSAT 或部分 MaxSAT 的计算方法，用于随机森林（例如 [58], [59], ）、树集成（例如 [60]）和提升树（例如 [61]）。可解释性的形式化结果甚至被用于质疑一些普遍看法，例如与决策树可解释性相关的看法 [62]。

最后一点与本教程采用的叙述方式有关：本教程将决策背后的完整且一般的理由视为可解释性理论的核心概念，并用它们来描述其他概念，包括那些在这些理由被提出之前就已出现的概念。这源于我们坚信，“实例抽象”这一概念必须成为任何全面且基础扎实的可解释性理论的核心。我们希望读者会认同，这种处理方式还促成了一种极简的表述，既阐明了语义，也便于计算。

致谢

谨此感谢 Yizuo Chen、Haiying Huang 和 Albert Ji 提供的宝贵反馈并参与讨论。

[13]SDD（句子决策图）是一种决策图，其分支依据是公式（句子），而不是变量 [50]。SDD 是 [51], OBDD 的超集，表达也简洁得多，但在简洁性方面与其他某些类型的决策图不可比 [52]–[54]。

参考文献

[1] H. Chan 和 A. Darwiche，“关于贝叶斯网络分类器的推理”，载于 UAI。Morgan Kaufmann，2003 年，第 107–115 页。[2] C. Ji 和 A. Darwiche，“面向含非二元变量分类器的一类新型解释”，CoRR，第 abs/2304.14760 卷，2020 年。[3] A. Choi、A. Shih、A. Goyanka 和 A. Darwiche，“关于随机森林行为的符号编码”，CoRR，第 abs/2007 01493 卷，2020 年。[4] A. Darwiche 和 C. Ji，“关于必要且充分解释的计算”，载于 AAAI。AAAI Press，2022 年。[5] A. Darwiche 和 A. Hirth，“决策背后的原因”，载于 ECAI，《人工智能与应用前沿》系列，第 325 卷。IOS Press，2020 年，第 712–720 页。[6] R. E. Bryant，“用于布尔函数操作的基于图的算法”，《IEEE 计算机汇刊》，第 35 卷，第 8 期，第 677–691 页，1986 年。[7] A. Darwiche 和 P. Marquis，“关于布尔逻辑中文字的量化及其在可解释人工智能中的应用”，《人工智能研究杂志》，第 72 卷，第 285–328 页，2021 年。[8] A. Darwiche，“可分解否定范式”，《美国计算机协会会刊》，第 48 卷，第 4 期，第 608–647 页，2001 年。[9] A. Shih、A. Choi 和 A. Darwiche，“解释贝叶斯网络分类器的符号方法”，载于 IJCAI。ijcai.org，2018 年，第 5103–5111 页。[10] A. Ignatiev、N. Narodytska 和 J. Marques-Silva，“基于溯因的机器学习模型解释”，载于第三十三届人工智能会议论文集（AAAI），2019 年，第 1511–1519 页。[11] W. Shi、A. Shih、A. Darwiche 和 A. Choi，“关于二值神经网络的易处理表示”，载于 KR，2020 年，第 882–892 页。[12] A. Ignatiev、N. Narodytska、N. Asher 和 J. Marques-Silva，“从对比解释到溯因解释，再回到对比解释”，载于 AI*IA，《计算机科学讲义》系列，第 12414 卷。Springer，2020 年，第 335–355 页。[13] P. Lipton，“对比解释”，《英国皇家哲学研究所增刊》，第 27 卷，第 247–266 页，1990 年。[14] G. Audemard、F. Koriche 和 P. Marquis，“基于编译表示的易处理 XAI查询”，载于 KR，2020 年，第 838–849 页。[15] A. Garfinkel，“解释的形式：重新思考社会理论中的问题”，《英国科学哲学杂志》，第 33 卷，第 4 期，第 438–441 页，1982 年。[16] D. Lewis，“因果解释”，载于《哲学论文集》第二卷，D. Lewis 编。牛津大学出版社，1986 年，第 214–240 页。[17] D. Temple，“为什么问题的对比理论”，《科学哲学》，第 55 卷，第 1 期，第 141–151 页，1988 年。[18] S. Wachter、B. D. Mittelstadt 和 C. Russell，“无需揭开黑箱的反事实解释：自动化决策与 GDPR”，《哈佛法律与技术杂志》，第 31 卷，第 2 期，2018 年。[19] J. van der Waa、M. Robeer、J. van Diggelen、M. Brinkhuis 和 M. Neerincx，“基于局部反例树的对比解释”，arXiv 预印本 arXiv:1806.07470，2018 年。[20] T. Miller，“人工智能中的解释：来自社会科学的启示”，《人工智能》，第 267 卷，第 1–38 页，2019 年。[21] B. Mittelstadt、C. Russell 和 S. Wachter，“解释人工智能中的解释”，公平性、问责制与透明度会议论文集，2019 年 1 月。[22] Y. Goyal、Z. Wu、J. Ernst、D. Batra、D. Parikh 和 S. Lee，“反事实视觉解释”，载于第 36 届国际机器学习会议论文集，《机器学习研究论文集》系列，K. Chaudhuri 和 R. Salakhutdinov 编，第 97 卷。PMLR，2019 年 6 月 9–15 日，第 2376–2384 页。[23] S. Verma、V. Boonsanong、M. Hoang、K. E. Hines、J. P. Dickerson 和 C. Shah，“机器学习中的反事实解释与算法补救：综述”，2022 年。[24] R. K. Mothilal、A. Sharma 和 C. Tan，“通过多样化反事实解释来解释机器学习分类器”，2020 年公平性、问责制与透明度会议论文集，2020 年 1 月。[25] V. Gurvich 和 L. Khachiyan，“关于生成单调布尔函数的无冗余合取范式和析取范式”，《离散应用数学》，第 96 卷，第 363–373 页，1999 年。[26] Y. Crama 和 P. L. Hammer，“布尔函数——理论、算法与应用”，载于《数学及其应用百科全书》，2011 年。[27] A. Darwiche，“逻辑在人工智能中的三种现代作用”，载于 PODS。美国计算机协会，2020 年，第 229–243 页。[28] A. Darwiche 和 P. Marquis，“知识编译图谱”，《人工智能研究杂志》，第 17 卷，第 229–264 页，2002 年。[29] H. Chan 和 A. Darwiche，“贝叶斯网络分类器推理”，载于 UAI。Morgan Kaufmann，2003 年，第 107–115 页。[30] A. Shih、A. Choi 和 A. Darwiche，“将贝叶斯网络分类器编译为决策图”，载于 AAAI。AAAI Press，2019 年，第 7966–7974 页。[31] N. Narodytska、S. P. Kasiviswanathan、L. Ryzhyk、M. Sagiv 和 T. Walsh，“验证二值化深度神经网络的性质”，载于 AAAI。AAAI Press，2018 年，第 6615–6624 页。[32] N. Narodytska、H. Zhang、A. Gupta 和 T. Walsh，“寻找适用于 SAT 的二值化神经网络架构”，载于 ICLR。OpenReview.net，2020 年。[33] N. Narodytska、A. A. Shrotri、K. S. Meel、A. Ignatiev 和 J. Marques-Silva，“使用模型计数评估启发式机器学习解释”，载于 SAT，《计算机科学讲义》系列，第 11628 卷。Springer，2019 年，第 267–278 页。[34] T. Baluta、S. Shen、S. Shinde、K. S. Meel 和 P. Saxena，“神经网络的定量验证及其安全应用”，载于 CCS。美国计算机协会，2019 年，第 1249–1264 页。[35] A. Shih、A. Darwiche 和 A. Choi，“通过 Angluin 风格学习验证二值化神经网络”，载于 SAT，《计算机科学讲义》系列，第 11628 卷。Springer，2019 年，第 354–370 页。[36] K. Jia 和 M. Rinard，“二值化神经网络的高效精确验证”，载于《第 34 届神经信息处理系统国际会议论文集》，NIPS’20 系列。美国纽约州雷德胡克：Curran Associates Inc.，2020 年。[37] Y. Zhang、Z. Zhao、G. Chen、F. Song 和 T. Chen，“BDD4BNN：基于 BDD 的二值化神经网络定量分析框架”，载于《计算机辅助验证》，A. Silva 和 K. R. M. Leino 编。Cham：Springer International Publishing，2021 年，第 175–200 页。[38] M. Giacobbe、T. A. Henzinger 和 M. Lechner，“量化神经网络需要多少位？”，载于《系统构建与分析的工具和算法》，A. Biere 和 D. Parker 编。Cham：Springer International Publishing，2020 年，第 79–97 页。[39] T. A. Henzinger、M. Lechner 和 D. Zikelic，“量化神经网络的可扩展验证”，载于 AAAI。AAAI Press，2021 年，第 3787–3795 页。[40] Y. Zhang、Z. Zhao、G. Chen、F. Song、M. Zhang、T. Chen 和 J. Sun，“QVIP：一种基于整数线性规划的量化神经网络形式化验证方法”，ASE ’22 系列。美国纽约州纽约：美国计算机协会，2023 年。[41] Y. Zhang、F. Song 和 J. Sun，“QEBVerif：神经网络量化误差界验证”，CoRR，第 abs/2212.02781 卷，2022 年。[42] M. T. Ribeiro、S. Singh 和 C. Guestrin，“‘我为什么要相信你？’：解释任意分类器的预测”，载于 KDD。美国计算机协会，2016 年，第 1135–1144 页。[43] ——，“Anchors：高精度的模型无关解释”，载于 AAAI。AAAI Press，2018 年，第 1527–1535 页。[44] A. Ignatiev、N. Narodytska 和 J. Marques-Silva，“关于启发式机器学习解释的验证、修复与改进”，CoRR，第 abs/1907 02509 卷，2019 年。[45] J. Marques-Silva、T. Gerspacher、M. C. Cooper、A. Ignatiev 和 N. Narodytska，“以多项式时间和延迟解释朴素贝叶斯及其他线性分类器”，载于 NeurIPS，2020 年。[46] Y. Izza、A. Ignatiev 和 J. Marques-Silva，“关于解释决策树”，CoRR，第 abs/2010.11034 卷，2020 年。[47] X. Huang、Y. Izza、A. Ignatiev 和 J. Marques-Silva，“关于高效解释基于图的分类器”，载于 KR，2021 年，第 356–367 页。[48] X. Huang、Y. Izza、A. Ignatiev、M. C. Cooper、N. Asher 和 J. Marques-Silva，“知识编译语言的高效解释”，CoRR，第 abs/2107.01654 卷，2021 年。[49] G. Audemard、S. Bellart、L. Bounia、F. Koriche、J.-M. Lagniez 和 P. Marquis，“布尔决策树的解释能力”，《数据与知识工程》，第 142 卷，第 102088 页，2022 年。[50] A. Darwiche，“SDD：命题知识库的一种新型规范表示”，载于 IJCAI。IJCAI/AAAI，2011 年，第 819–826 页。[51] S. Bova，“SDD 比有序二叉决策图简洁得多”，载于 AAAI。AAAI Press，2016 年，第 929–935 页。[52] B. Bollig 和 M. Buttkus，“关于语句决策图的相对简洁性”，《计算机系统理论》，第 63 卷，第 6 期，第 1250–1277 页，2019 年。

[53] P. Beame 和 V. Liew，“知识编译的新限制及其在精确模型计数中的应用”，载于 *UAI*。AUAI出版社，2015年，第131–140页。[54] P. Beame、J. Li、S. Roy 和 D. Suciu，“精确模型计数的下界及其在概率数据库中的应用”，载于 *UAI*。AUAI出版社，2013年。 [55] P. Barcel´o、M. Monet、J. P´érez 和 B. Subercaseaux，“通过计算复杂性的视角理解模型可解释性”，载于 *NeurIPS*，2020年。[56] G. Audemard、F. Koriche 和 P. Marquis，“基于编译表示的可处理XAI查询”，载于 第*17*届知识表示与推理原理国际会议论文集，2020年9月，第838–849页。[57] G. Audemard 、 S. Bellart 、 L. Bounia 、 F. Koriche 、 J. Lagniez 和 P. Marquis，“布尔分类器的计算可理解性”，载于 *KR*，2021年，第74–86页。 [58] Y. Izza 和 J. Marques-Silva，“使用SAT解释随机森林”，载于 *IJCAI*。ijcai.org，2021年，第2584–2591页。 [59] G. Audemard、S. Bellart、L. Bounia、F. Koriche、J. Lagniez 和 P. Marquis，“以稀疏性换取随机森林解释的复杂性”，载于 *AAAI*。AAAI出版社，2022年，第5461–5469页。 [60] A. Ignatiev、Y. Izza、P. J. Stuckey 和 J. Marques-Silva，“使用最大可满足性高效解释树集成”，载于 *AAAI*。AAAI出版社，2022年，第3776–3785页。 [61] G. Audemard、J. Lagniez、P. Marquis 和 N. Szczepanski，“计算提升树的溯因解释”，丛书：第26届人工智能与统计国际会议论文集（AISTATS），第206卷。PMLR，2023年。 [62] Y. Izza、A. Ignatiev 和 J. Marques-Silva，“应对决策树中的解释冗余”，*J. Artif. Intell. Res.*，第75卷，第261–321页，2022年。