arXiv:2303.01378v1 [cs.AI] 2 Mar 2023

语义增强数据科学的愿景

Udayan Khurana[1], Kavitha Srinivas[1], Sainyam Galhotra[2], Horst Samulowitz[1]IBM Research, [2]芝加哥大学
ukhurana@us.ibm.com,kavitha.srinivas@ibm.com,sainyam@uchicago.edu,samulowitz@us.ibm.com

摘要

近年来，机器学习或数据科学自动化方面的努力已在超参数优化、模型选择等各种任务中取得成功。然而，利用领域知识和数据语义等关键领域的自动化程度仍然很低。数据科学家长期以来一直利用常识推理和领域知识来理解和丰富数据，以构建预测模型。本文讨论当前数据科学和机器学习解决方案的重要不足。随后，我们展望如何将对数据的语义理解和推理与新型数据科学自动化工具相结合，从而实现一致且可解释的数据增强与转换。此外，我们还讨论语义如何以新的方式帮助数据科学家应对机器学习中与信任、偏差和可解释性相关的挑战。语义标注还可以帮助更好地探索和组织大型数据源。

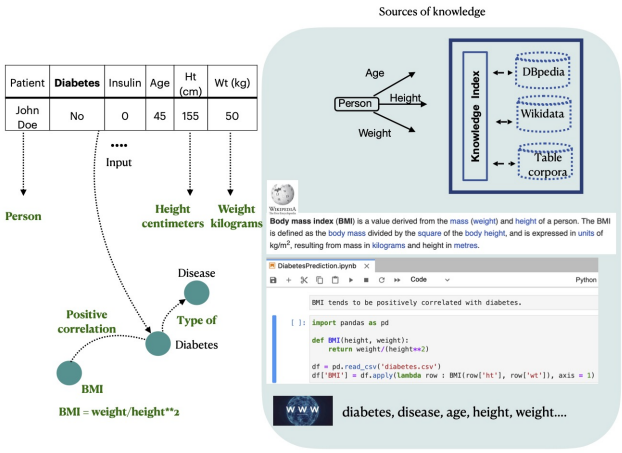


图 1：使用语义进行糖尿病建模

1 引言

机器学习（自动化机器学习）或数据科学流程的自动化近年来受到了广泛关注[1]。推动自动化机器学习发展的动力在于减少人工干预，从而节省时间和成本，并提高建模过程的效能。近来，人们已通过强化学习、多臂老虎机以及贝叶斯优化等数值优化方法实现了此类自动化。尽管迄今提出的自动化工作已提高了数据科学（DS）流程不同环节的效率，但其中一个相对未被触及的量纲是数据中的语义。传统上，在高度依赖人类数据科学家或领域专家的数据科学中，数据的语义发挥着至关重要的作用。数据科学家研究给定的问题或数据，将其与现实世界中的概念联系起来，并揭示它们之间的关系。随后，通常由目标变量定义的回归或分类任务会与由特征表示的概念相联系。最后，数据科学家利用领域知识或关于世界的一般知识，执行能够增强给定数据建模能力的操作，例如特征增强。在迄今的自动化技术中，只有少数在数据科学的语义方面取得了进展。这一领域的大部分内容仍有待进一步研究；它不仅为业界所亟需，也受到学术研究的关注。本文汇集了与语义数据科学相关的工作。我们识别了机遇，并提出了一些构想，以扩展自动化数据科学技术和方法，从而利用语义。

图 1 展示了一个将语义注入数据科学流程的动机示例。输入表格用于糖尿病建模作为年龄、身高、体重等各种参数的函数。随着近期技术的进步，如今可以为这些数据赋予语义，并以绿色标签显示，帮助理解数据。右侧蓝色面板展示了可用于添加语义的知识来源。DBpedia、维基百科等知识图谱，或网络上的海量表格语料库，都有助于理解各列。例如，列到概念的映射技术[20, 26]可以将“患者”等命名实体列映射到“人物”概念。知识图谱中的“人物”概念通常包含一组属性，可用于通过列到属性的匹配方法，将“Ht”匹配到“身高”。单位则可以根据知识图谱中各种属性分布的特征进行近似推断。除知识图谱外，维基百科、数据科学笔记本的文本单元格和代码等非结构化来源中，也蕴藏着大量信息。挖掘技术可用于处理这些语料，为当前的数据科学问题自动创建特征。例如，可以从维基百科的代码和文本中挖掘出计算身体质量指数的公式，并将身体质量指数作为基于语义的特征，以改进糖尿病预测。此外，从代码笔记本中挖掘出的公式既可以用于提取表达式以进行特征增强，也可以提示何时进行此类增强可能会产生有意义的信号，例如某个表达式经常出现在处理类似数据的笔记本中。请注意，这些操作对经验丰富的数据科学家而言可能是常规工作，但要让软件自动完成这些映射和推理却颇具挑战。本文旨在重点介绍相关的语义技术研究方向，并展望下一步的重要工作。

[1]https://www.tinyurl.com/forbes-automated-data-science/

数据准备[2] 是任何成功模型构建任务的核心。它是调整特征空间表示的过程，目的是更好地适应模型拟合或提高目标变量 [25]的可预测性。这一环节也需要数据科学家大量运用语义映射和推理。图 1 中的身体质量指数推导示例展示了这一过程如何通过维基百科实现自动化。为说明语义在数据准备中的普适性，再来看两个来自不同领域的示例。首先，考虑 Kaggle[3] 上的贷款审批预测任务，目标是预测给定的贷款申请是否会获批。提供的属性包括贷款金额、申请人的性别、婚姻状况、教育程度、申请人收入、共同申请人收入等。熟悉贷款领域的数据科学家会将 $\frac{applicant_income+coapplicant_income}{loan\ \ amount}$ 确定为预测变量之一。~~完成这一步数据准备需要运用不同层次的语义操作。首先，将给定的列映射到现实世界概念；然后，找出这些特征概念之间的关系；最后，确定哪些关系与目标变量的概念相符，等等。~~

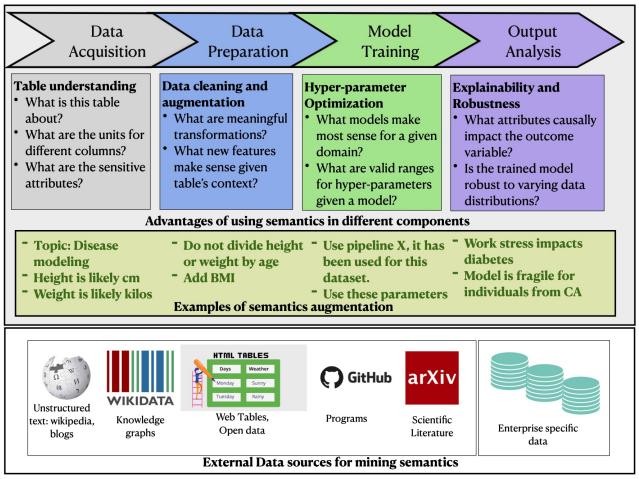


图 2：数据科学流程中的语义。

我们的第二个示例取自 Kaggle 上有关 COVID-19 的数据集。用于疾病建模的一个常见比率是病死率（CFR）。CFR 是一种将死亡人数与确诊病例数关联起来的公式：病死率% = ~~$\frac{\text{疾病死亡人数}}{\text{疾病确诊病例数}} \times 100$~~。此类公式既可见于用于处理 COVID 数据的代码笔记本中，也可见于维基百科；同样，它们还可用于为模型构建创建新特征。

除了数据准备，机器学习（ML）和数据科学的其他方面也依赖于对数据和模型的语义解读。例如，模型可解释性部分基于对数据中概念的解读，以及对在这些概念上执行的操作之含义的理解。为通过去偏确保公平性，我们需要识别传统的偏见点 [31]，并确保今后不再重蹈覆辙。要自动化这一过程，语义概念识别以及对已识别概念之间因果依赖关系的理解，对于能够识别并消除偏差的可信系统至关重要。同样，通过符号人工智能 [5]和语义来执行业务规则，也进一步彰显了语义驱动的自动化数据科学的价值。此外，早期研究还致力于理解超参数并提取应用这些参数时的约束 [2],，以及利用语义进行数据清洗 [47]。

本文首先探讨与表格数据（第 2.1 节）、代码（第 2.3.1 节）和文本（第 2.3.2 节）的语义理解与标注相关的研究，这些研究旨在构建面向语义的数据科学基础设施。随后，我们将重点关注自动化数据准备（第 3 节），介绍广受欢迎的非语义研究，以及面向语义的数据准备领域的最新进展和新机遇。接下来讨论机器学习，尤其是自动化机器学习（第 4 节）。最后，我们将从语义维度拓展这些理念，以实现可信机器学习（第 5 节）。

[2]有时也称为特征工程或特征增强[3]
https://www.tinyurl.com/yonatanrabinovich-loan [4]
https://www.kaggle.com/imdevskp/corona-virus-report [5]
https://tinyurl.com/inrule-business-rules-ml

2 语义理解

数据科学的首要步骤是理解数据。如图 2 所示，理解表有助于指导语义在特征生成中的应用，并选择适合该领域的模型。了解领域中关键变量之间的因果关系，可以增强模型的可解释性。该图说明，语义能够显著提升数据科学工作流中的每个组成部分。近期研究已展示了理解数据语义方面的进展，而这正是语义数据科学至关重要的第一步。总的来说，该领域的研究可分为以下几类：（a）表格理解——将单元格值、列乃至整张表映射到一组定义明确的概念，这些概念取自本体或知识图谱；（b）表间数据发现——例如，基于更大规模的表索引，发现其他可进行连接或合并的表；（c）理解代码、文本或其他较少结构化数据源中的语义。

2.1 表格理解

表格理解有助于数据科学中的可解释性、信任、偏差和特征工程，因为它有助于整体上选择更合适的特征来构建模型。表格理解还可以指导相关数据的搜索，包括适合进行连接或合并的表。本文将介绍部分有关表格理解的文献，以说明尽管这项任务具有挑战性，但其可行程度足以支持语义数据科学。

表格理解包括识别列类型标注（CTA），找出包含感兴趣实体的核心列、映射到这些实体属性的列（CPA），以及将表本身映射到更广泛的类别。其目标是自动识别数据中所提供列的正确概念。此外，我们还希望以恰当的层级关系表达列之间的关系，从而体现数据集的意图。下面将介绍不同的表格理解方法。

2.1.1 基于知识图谱的方法。 大多数表格理解方法都假设存在一个或多个知识图谱或本体，用来定义理解表格单元格或列所需的概念全集。主要包含数值、日期、社会保障号码等内容的列不太可能出现在知识图谱中，通常称为 L 列。对于包含可能映射到实体的字符串（称为 NE 列）的列，表格理解可分为单元格实体标注（CEA）、列类型标注（CTA）和列属性标注（CPA）；许多系统通常会利用 CEA 来辅助 CTA 和 CPA。许多系统还会考虑另一项任务，即检测主题列（S），以识别核心实体列，而表中的其他列则表示该核心实体的属性。目前已有众多系统着眼于这些任务，这主要得益于针对这些任务的挑战赛 [21], ，其基础知识图谱为 DBpedia 或维基数据。能够执行一项或多项 NE 任务的系统包括 MantisTable [6], ColNet [4], TableMiner+ [38]和 Sherlock [20]。这些系统大多基于列向量嵌入进行神经网络元学习。当有大量优质训练样例时，这些方法在某些情况下表现出色。然而，在许多情况下，它们难以扩展到大量列概念，尤其是在处理数值数据以及训练数据稀缺或存在噪声时。 C^2[26] 提出了一种基于多个数据源（如知识库和带噪声的网页表格）的最大似然集成方法，该方法在概念数量增加时更具可扩展性。一些系统（如 Memei [43] ）则联合训练 CTA 任务和列间关系，以便为 L 列添加标注。

虽然各系统采用的方法和执行的任务各不相同（例如监督式或半监督式），但在几乎所有情况下，对于现实中的表而言，将列与知识图谱进行匹配都颇具挑战。充其量，这只能为完全自动化的语义特征工程提供有限支持。这在一定程度上是由于基于知识的方法至少存在两项关键局限：（a）知识库的召回率严重不足。据估计，在普通爬网网络语料库的 35 亿个 HTML 页面中，只有 3% 的表格能够与 DBpedia 匹配 [39]。（b）对于 L 列，这种方法尤其具有挑战性，除非主题列（S 列）的检测完全准确，且列标题能够通过 CPA 映射到知识图谱属性。下文将介绍的知识无关技术，能够绕过基于知识图谱的方法在召回率较低时面临的一些局限。

2.1.2 知识图谱无关方法。知识图谱无关方法将大量公开可用表格作为知识，并使用模式匹配技术映射语料库中的表和列。这里的挑战在于如何将这项计算扩展到数百万条标注，从而实现文献中通常所称的整体模式匹配。针对这一问题，人们采用了多种方法，包括二元集成策略，通过迭代匹配创建单一词汇表 [1], ；以及对列值进行向量化，并使用局部敏感哈希（ LSH）等技术减少成对比较的数量（[12]），等等。

近期，有研究尝试对大型语料库中的表进行向量化。这一思路受到语言模型领域相关工作的启发，例如 BERT [8], 能够有效地封装句子中词语之间的相关性。一些近期研究通过构建基于列值的基于 BERT 的无监督语言模型，来解决列类型分配问题 [44], ，并添加其他中间列标签以提供上下文，从而改进标注。在 [44], 中，他们使用 WikiTables 领域的 160 万张表来构建模型。将语料库用作参考知识库的方法显然很有吸引力，因为它可以大幅缓解知识图谱的召回率问题。网络上收集的大型表格语料库（例如 VizNet7、Data.gov[8] ）的可用性，使这一方法前景可期且切实可行。

2.2 表间发现

虽然通常可以基于单个表构建特征，但在数据科学中，科学家更常通过连接或合并多个表中的数据来计算模型。表格相似性检测针对的是这一数据整理领域。基于多种指标进行表格相似性检测，例如匹配行的数量、匹配列的数量，以及组合相关表格可能带来的信息增益等；Juneau 等系统利用这些指标来增加特征工程中可用特征的数量 [49]。JOSIE [50], 解决了在数据湖中查找有趣连接的问题，它能在数百万个表中高效地执行 top-k 重叠集合相似性搜索。可连接表格按其与查询键列的重叠规模排序。JOSIE 适用于等值连接。PEXESO [10]使用词嵌入对字符串值列进行连接，因为可连接列的字符串表示通常存在细微差异。PEXESO 也经过数据丰富化评估；在机器学习任务中，该系统的 micro-F1 分数提高了 1.9%，均方误差降低了 10%。KAFE [16] 和 S3D [15] 通过识别给定数据的语义概念与数据库表列中的语义概念，利用知识图谱和大型数据湖中的表来创建有效的连接和并集。

2.3 从代码和文本中挖掘语义

假设我们拥有能够找到有用数据源、以对目标变量建模的技术，自动化特征工程仍然需要挖掘公式，以便针对给定目标构建有意义的特征，如身体质量指数、CFR 或贷款示例所示。我们介绍两个挖掘领域：一个是从代码中挖掘，另一个是从文本中挖掘。

2.3.1 通过代码挖掘开展数据科学：近期研究致力于挖掘代码仓库，以帮助数据科学家。Wranglesearch[9] 是一个通过动态分析挖掘 Kaggle 笔记本的系统，可提取有助于数据整理的函数和表达式；它或许是与自动化特征工程最相关的研究。其他研究则着眼于更广泛的数据科学开发领域。例如， [32] 会挖掘 Kaggle 笔记本和脚本，以找出其所称的决策点——在这些决策点上，人们可能会考虑采用不同的模型或参数；该系统会向数据科学用户提供这些备选方案，帮助他们构建更稳健的模型。

[6]http://webdatacommons.org/webtables/ [7] https://github.com/mitmedialab/viznet [8]https://www.data.gov/ [9] https://www.josecambronero.com/pdf/wranglesearch-draft.pdf

SOAR [35] 将 API 调用文档的嵌入技术与示例编程方法相结合，建议将数据科学代码从一个 API 重构为另一个 API（例如，从 TensorFlow 迁移到 PyTorch）。

2.3.2 从文本中挖掘表达式：大量有助于数据科学的领域知识也埋藏在文本之中。已有研究致力于提取数学方程的主题，通过构建联合神经模型，同时处理方程周围的文本上下文和方程本身。在数据科学领域，如果能成功将方程中的术语映射到列，那么搜索相关表达式并将其呈现给数据科学家，或提取特征，就会更有用。 [18]介绍了如何在一个包含数百万个方程的提取数据库上使用经过改进的信息检索技术；例如，他们能够找到与雅可比多项式相关的方程，或补全方程。同样，从数据科学的角度来看， [18] 将方程与文本主题联系起来，但其在完全自动化的特征工程机制中的效用仍有待探讨。

另一个相关研究领域是从文本中挖掘因果关系；因果关系往往决定哪些特征在语义上可能对目标变量建模有意义。关于从文本中自动抽取因果关系的研究成果丰富（例如， [29], [13]）。从笔记本和领域专用语料库中提取此类信息，有助于进行特征选择。

2.4 利用大型语言模型中的语义

受基于 Transformer 的模型在自然语言处理（NLP）领域取得成功的启发，目前已有许多不同的研究致力于将语言模型应用于表格数据（参见 [9] 以获取综述）。这些语言模型面向多种下游任务，包括以文本形式总结表格、列属性标注、SQL 生成或实体链接。尽管这一领域的初步结果颇为有趣，但仍存在许多悬而未决的问题：（a）目前缺乏一套规模庞大的下游任务集合，而正是这类任务推动了自然语言处理领域的显著进步；近期的自然语言处理模型会在数百项任务上进行训练（例如 BigBench [10]）。表格任务尚不存在这样的标准任务套件，而且每项研究针对的下游任务各不相同。（b）现有公共语料库往往是 Web表格，其特征通常与企业数据库存在较大差异；而且，即使使用电子表格，这些电子表格也不会向社区公开；（c）对于包含数百万行和列的表格数据而言，模型的序列长度限制似乎尤其具有挑战性；（d）数值列经常被当作文本处理，或采用专门的表示方式（例如表示数量级、首位数字和末位数字），但这些方式可能无法恰当地捕捉列的特征。

3 数据准备

数据科学流程中的数据准备组件旨在整合从多个来源获取的数据，并进行数据清洗和格式化，以统一现有数据集。数据准备通常包含许多组件。我们在第 3.1 节概述这些组件，并讨论它们对下游分析的影响。随后，我们介绍初步的语义应用工作，并指出忽视语义带来了错失良机。接下来，我们提出具体思路，为各种未考虑语义的数据准备方法融入语义。

3.1 数据准备任务

数据准备通常包含许多不同的步骤，用于集成、清理、转换和验证可用的数据集。我们介绍三类具有代表性的数据准备任务。

数据发现与集成： 这些组件旨在利用外部数据源，识别可用于下游分析的新数据集。这些组件面临的主要挑战，是如何从数百万个可用选项中搜索有用的数据，并集成从异构数据源汇集的信息。以往的数据发现与集成技术并未考虑使用语义来识别有用的数据集。这些技术往往会产生许多假阳性，即与目标无关或语义上不连贯的数据集。

数据验证与转换： 数据转换技术会使用转换函数修改数据值并使其标准化。例如，它们可以将重量单位从千克转换为磅，或通过组合已有属性生成新属性，例如利用身高和体重属性计算身体质量指数。许多不同的方法都致力于生成新属性。FICUS [30] 会在可能的特征空间中执行束搜索，并通过应用构造函数来构建新特征。ExploreKit [24] 会逐个显式扩展特征空间，每次添加一个特征。AlphaD3M [11]和 Cognito [28] 利用强化学习，通过智能体优化特征转换。LFE [34] 则通过感知机，根据历史数据草图表示上的转换效果，直接预测每个特征最有用的转换方式。有关此类方法的更多详情，请参阅综述文章 [25], [51] 和 [40]。

以下工作在一定程度上考虑了用于数据增强的语义。Friedman等人 [14] 和Galhotra等人 [16] 分别利用知识库，为文本分类和Web表格递归生成新特征。FeGeLOD [37] 通过查询从链接开放数据中提取特征。然而，这些方法未能充分挖掘来自多个数据源的可用信息。S3D [15] 能够引入来自网页表格和多个知识图谱等多个来源的数据，但其添加新列时并不考虑目标变量。语义特征发现 [41]通过将给定数据问题中的概念与历史代码中的概念关联起来，利用历史数据科学代码文件和笔记本代码库中的公式。尽管这些方法很有前景，但由于知识库的可用性和丰富性有限，以及底层概念匹配算法的质量受限，其影响力也受到限制。此外，如何在考虑不同背景和领域的情况下建立数据与表达式之间的联系，仍是一个开放问题。

数据筛选和清洗技术旨在去除噪声值、填补缺失值并替换不准确记录。以往的数据清洗技术会处理各种数据质量问题，例如重复项、违反函数依赖、异常值检测等。大多数此类技术并未探索如何利用

[10]https://github.com/google/BIG-bench

语义以及外部数据源中的信息来改进这些环节。

3.1.1 数据准备中缺乏语义导致的机会错失。上述最先进的特征工程系统依赖反复试验，通过对给定特征应用变换函数来生成新特征。它们力求最大化准确率，却对数据和变换函数的实际语义视而不见。这会带来若干弊端。首先，它们无法利用外部知识对数据执行特定变换，而这类变换往往很难通过通用探索技术发现。例如，身体质量指数（图 1）涉及一个复杂公式，这些技术中的任何一种都需要耗费很长时间并进行大量试验才能发现它。其次，它们可能违反物理定律，例如将身高与体重相加；这在这些方法中可能是一个合理的探索选项，但凭借常识我们知道，这在物理上毫无意义，应该避免。第三，这些方法无法引入问题中未提供、但可从外部获取的额外数据，以提升建模能力。知识库、数字图书、维基百科、博客以及数据科学代码等形式中蕴藏着海量知识，而自动化数据准备或特征工程却错失了利用这些知识的机会。

3.2 在自动化特征工程中利用语义的提案

Cherrier 等 [39] 采用遗传编程的方式构造特征，但在物理领域中施加约束。例如，有些约束规定能量和量纲不能相加，因为这在物理上没有意义；平方根之后再平方也没有太大意义，等等。我们认为，将这种方法推广到现有的非语义方法中，能够显著提升后者的效能。无论是按照 [5] 的建议，通过遗传编程进行特征工程，还是采用第3.1节所述的其他优化方法，例如 Cognito [27],DSM [22],ExploreKit [24] ，抑或采用 LFE [34], 这样的元学习方法，通过识别语义类型并施加某些约束（其中一些是全局性的，另一些则针对具体问题），并将这些约束表示在适当的语法中，都将带来益处。只需检查语义或约束一致性，就能限制这些方法执行（无益的）操作。为此，需要一种通用语法来表示这类约束，还需要一个有效的引擎来强制执行这些约束。该领域的研究人员或许可以开发一种语言：既具有足够的表达能力，能够描述不同领域的约束，又足够简单，便于指定约束，或将现有的业务规则或科学规则转换为这种新语言。

4 语义机器学习

当前机器学习（包括深度学习）在实际应用中大多缺乏语义。对于特征、值、状态、对象或标签究竟“意味着”什么，并没有明确的概念（例如，将其作为输入的一部分）。因此，语义并未被明确用于构建、调优或部署模型。这甚至包括用于语言（例如 GPT3 [3]）、语音、图像（例如 ImageNet [7]）或视频识别的强大基础模型；它们在特定任务上展现出令人瞩目的性能。尽管这类模型确实能够生成有效的回答，有时看起来似乎具备语义意识，但其底层模型几乎没有任何内在的语义理解。这些机器学习模型既没有获得输入数据的语义，也不会推断其语义。这些模型大多基于数据的统计和数学属性。虽然人们希望通过数据驱动的方式获取洞见，但数据本身的含义却被忽视了，未能超越统计层面的理解；由此产生的问题众所周知，包括偏差和可解释性 11、伪相关 12，以及不恰当或毫无意义的回答 [13]。

将语义融入其中颇具挑战，因为人工智能尚未实现常识推理。不过，我们认为，即使只融入有限的语义，也有助于构建更好的人工智能模型。例如，在其他方面不依赖模型的强化学习中， [42], 如果模型预先知道某个状态-动作对在现实中无效，就没有必要探索这种情形。这既能避免不必要的计算，也能避免模型可能出现的次优表现。同样，如果已知特征之间不存在语义关系，就可以考虑不在神经网络中连接这些特征，从而有望让模型的解释更加清晰。如图 1 所示，基于数据转换、由语义驱动的特征工程，对数据科学家、监管机构、审计员和公众而言都是透明的。例如，了解计算模型为何拒绝或接受一项贷款申请，具有社会、法律和伦理价值。

值得注意的是，人工智能中的许多问题领域实际上要求在输入中对任务作出更明确的定义。例如，人工智能规划和约束编程都以用户提供的定义为基础（如变量、约束和目标）。在这些情形下，整体数据语义至少有一部分是作为输入提供的。然而，众所周知，以这种方式对问题建模会给用户带来沉重负担，因为问题的表述往往一点也不简单（例如， [36, 46]）。因此，通用的语义自动化也将使这些领域受益。

4.1 自动化机器学习

语义的影响也可以融入自动化机器学习优化本身。超参数搜索是自动化机器学习中已开始应用语义的一个方面。寻找正确模型的搜索空间很大，这里的重点是利用机器学习领域现有的工件，在模型的超参数空间中进行更有针对性的搜索。[32] 例如，有一种方法会挖掘 Kaggle 笔记本和脚本，以找出他们所说的决策点——在这些地方，可能会考虑其他模型或参数——并向数据科学用户提供这些备选方案，帮助他们构建更稳健的模型。另一种方法则探究能否挖掘数据科学代码，以便针对未见过的数据集更好地进行流水线搜索 [19]。

我们进一步认为，在超参数优化过程中，可以分析代码和 Python 文档字符串，不仅自动确定可用的超参数设置 [2],，还可以提取超参数配置中的约束，例如

[11]https://venturebeat.com/datadecisionmakers/bias-in-ai-is-spreading-and-its-time-to-fix-the-problem/ [12]http://www.tylervigen.com/spurious-correlations [13] https://medium.datadriveninvestor.com/chatbot-conversations-gone-wrong-78f7c1af4401

互斥设置。这样至少可以自动将搜索空间缩小到有效选项。在流水线配置层面，可以观察现有代码中的模式，例如主成分分析几乎总是在归一化之后执行，而不是反过来，并利用这些模式，例如在自动化流水线探索中优先考虑特定流水线。

5 使用语义进行可信输出分析

使用机器学习技术做出高风险决策，使信任的重要性日益凸显，而信任通常体现为公平性、可解释性和鲁棒性等方面。以往大多数技术通过探索数据集属性之间的相关性来研究机器学习的社会影响，却没有理解这些属性的真正含义。因此，公平性感知学习技术容易受到数据集噪声的影响，难以推广到现实场景；在这些场景中，数据集可能存在选择偏差，甚至包含伪相关。可解释人工智能技术生成的解释缺乏语义一致性 [45]，也难以理解。

5.1 用因果依赖关系推断语义

实现语义一致学习的初步步骤之一，是利用属性之间的因果依赖关系。这些研究以两种不同方式探索了因果信息的使用。(i) 语义约束：这些技术对输入属性施加约束，以确保所研究的属性值组合是可行的 [33]。例如，一条年龄为 < 18、婚姻状况为 $=$ 已婚的记录是不可能存在的。(ii) 依赖信息：这些技术将属性之间的依赖关系作为外部知识进行评估 [17, 23]。例如，薪水增加通常意味着储蓄增加。

因果关系并不能体现某一特定特征的含义，而是描述特征之间的依赖关系。考虑因果依赖关系，是利用语义开展分析的一步。这些技术展现出将因果信息纳入机器学习的巨大潜力，但也带来了诸多挑战。

如何识别属性依赖关系？ 属性之间的因果依赖关系是根据观察数据估计得出的，但由于可用数据规模有限，这种估计往往并不准确。领域科学家需要通过事后分析，手动验证语义约束/因果依赖关系，以确保分析在因果关系上保持一致。

如何纳入语义约束和依赖关系？ 因果关系需要检验属性变化所产生的效应，这通常被称为处理或干预。将干预纳入传统的基于相关性的框架，需要开发新的技术。近期的一项技术研究了因果公平性概念的应用，并探索了在没有任何外部信息的情况下估计因果效应。

外部信息能发挥什么作用？ 大多数技术都依赖领域科学家施加因果约束，或以因果图的形式输入依赖关系信息。如何减少此类技术对人工提供外部信息的依赖，是一项亟待深入研究的重要挑战。数据湖、知识图谱等外部信息源包含有助于推断依赖关系的有用信息。

6 个开放性研究问题

在本文讨论的语义数据科学的各种研究机遇中，我们总结了一些最有前景的研究方向。

研究问题1： 从数值数据中推断语义：(a) 找到合适的嵌入，用于编码数值列所指的概念、单位和数量。研究结合草图化思路与摘要方法的技术是否必要。(b) 使用数值嵌入将列映射到其概念，这种做法常用于文本数据。(c) 为列到概念的映射创建真实可信的人工标注基准，尤其针对数值数据。

研究问题2： 通过语义对齐和分类，开发理解和组织大量数据（数据湖）的技术。这包括语义搜索和数据浏览，帮助用户理解海量读时模式、未经组织的数据。（图3中的 DLE 和 SemSearch）

研究问题3： 挖掘大型代码库中的数据转换，并为这些转换中的变量关联概念。

研究问题4： 挖掘维基百科、电子书、博客等大型专家文本库，从中提取可付诸行动的信息，并将其关联到知识图谱所表示的概念。

RP5: 通过识别概念并查找这些概念的历史值，改进数据插补技术。（图3中的 SDI）。**RP6:** 根据数据和任务规范的上~~下文~~，发现反映模型构建中偏差的概念。调整模型训练以消除偏差。

参考文献

[1] C. Batini、M. Lenzerini 和 S. B. Navathe。1986。数据库模式集成方法的比较分析。*ACM Comput. Surv.* (1986)。[2] G Baudart、P Kirchner、M Hirzel 和 K Kate。2020。挖掘文档以提取超参数模式。*CoRR* abs/2006.16984 (2020)。[3] T Brown 等。2020。语言模型是少样本学习者。*CoRR* abs/2005.14165 (2020)。[4] J. Chen、E. Jiménez-Ruiz、I. Horrocks 和 C. Sutton。2019。ColNet：嵌入 Web 表格的语义以预测列类型。*AAAI* 33（2019年7月）。[5] Noëlie Cherrier、Jean-Philippe Poli、Maxime Defurne 和 Franck Sabatié。2019。利用受约束遗传编程在实验物理中进行一致的特征构造。载于*IEEE Congress on Evolutionary Computation*。[6] M Cremaschi、F Paoli、A Rula 和 B Spahiu。2020。一种完全自动化的完整语义表解释方法。*Future Gener. Comput. Syst.* (2020)。[7] J Deng、W Dong、R Socher、L Li、K Li 和 L Fei-Fei。2009。ImageNet：一个大规模分层图像数据库。载于*IEEE conference on computer vision and pattern recognition*。[8] J Devlin、M Chang、K Lee 和 K Toutanova。2018。BERT：用于语言理解的深度双向变换器预训练。*arXiv preprint arXiv:1810.04805* (2018)。[9] Haoyu Dong、Zhoujun Cheng、Xinyi He、Mengyu Zhou、Anda Zhou、Fan Zhou、Ao Liu、Shi Han 和 Dongmei Zhang。2022。表预训练：关于模型架构、预训练目标和下游任务的综述。载于*Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*，Lud De Raedt（编）。International Joint Conferences on Artificial Intelligence Organization，5426–5435。https://doi.org/10.24963/ijcai.2022/761 综述专题。[10] Y. Dong、K. Takeoka、C. Xiao 和 M. Oyamada。2021。数据湖中的高效可连接表发现：一种基于高维相似性的方法。*IEEE ICDE* (2021)。[11] I Drori、Y Krishnamurthy、R Rampin、R Lourenco、J Ono、K Cho、C Silva 和 J Freire。2018。AlphaD3M。*International Workshop on AutoML* (2018)。[12] S Duan、A Fokoue、O Hassanzadeh、A Kementsietsidis、K Srinivas 和 M Ward。2012。使用局部敏感哈希进行大规模本体的基于实例匹配。载于*ISWC*。

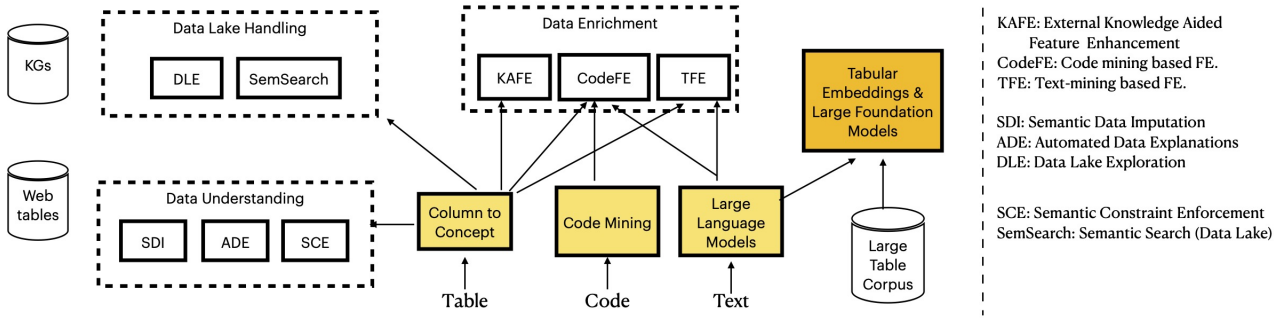


图3：面向数据科学的语义机制构成。浅黄色填充的模块是核心语义构建模块，而深黄色模块是潜在的核心构建模块（基础表格模型）。

[13] Hassanzadeh 等。2019。通过大规模文本挖掘回答二元因果问题：使用人类专家提供的因果对进行评估。载于 IJCAI。[14] L Friedman 和 S Markovitch。2018。面向知识的学习的递归特征生成。arXiv 预印本 arXiv:1802.00050 (2018)。[15] S Galhotra 和 U Khurana。2020。结构化数据上的语义搜索。载于 ACM CIKM。[16] S Galhotra、U Khurana、O Hassanzadeh、K Srinivas、H Samulowitz 和 M Qi。2019。利用外部知识进行预测建模的自动化特征增强。载于 IEEE ICDM。[17] S Galhotra、R Pradhan 和 B Salimi。2021。使用概率对比反事实解释黑盒算法。载于 ACM SIGMOD。[18] A Greiner-Petter、M Schubotz、F Müller、C Breitinger、H Cohl、A Aizawa 和 B Gipp。2020。发现感兴趣的数学对象——数学记法研究。[19] M Helali、E Mansour、I Abdelaziz、J Dolby 和 K Srinivas。2022。一种基于图神经网络的可扩展自动化机器学习方法。载于 VLDB。[20] Madelon Hulsebos 等。2019。Sherlock：一种用于语义数据类型检测的深度学习方法。载于 ACM SIGKDD。[21] E Jiménez-Ruiz、O Hassanzadeh、V Efthymiou、J Chen 和 K Srinivas。2020。SemTab 2019：用于对表格数据与知识图谱匹配系统进行基准测试的资源。载于语义网。[22] J Kanter 和 K Veeramachaneni。2015。深度特征合成：迈向数据科学工作的自动化。载于 IEEE数据科学与高级分析国际会议。[23] A Karimi、G Barthe、B Balle 和 I Valera。2020。面向重大决策的与模型无关的反事实解释。载于人工智能与统计国际会议。PMLR。[24] Gilad Katz、Eui Chul、Richard Shin 和 Dawn Song。2016。ExploreKit：自动特征生成与选择。载于 IEEE ICDM。[25] U Khurana。2018。监督学习中的基于变换的特征工程：自动化策略。载于《机器学习与数据分析的特征工程》。[26] U Khurana 和 S Galhotra。2021。表格数据的语义标注。ACM CIKM (2021)。[27] U Khurana、H Samulowitz 和 D Turaga。2018。利用强化学习进行预测建模的特征工程。载于 AAAI。[28] U Khurana、D Turaga、H Samulowitz 和 S Parthasarathy。2016。Cognito：面向监督学习的自动化特征工程。载于 IEEE ICDM。[29] P. Li 和 K. Mao。2019。面向知识的卷积神经网络：从自然语言文本中抽取因果关系。Expert Systems with Applications (2019)。[30] Shaul Markovitch 和 Dan Rosenstein。2002。使用通用构造函数进行特征生成。Machine Learning (2002)。[31] N Mehrabi、F Morstatter、N Saxena、K Lerman 和 A Galstyan。2021。机器学习中的偏见与公平性综述。ACM Computing Surveys (CSUR) (2021)。[32] M Merrill、G Zhang 和 T Althoff。2021。MULTIVERSE：从网络代码中挖掘集体数据科学知识，以推荐替代分析方法。[33] R Mothilal、A Sharma 和 C Tan。2020。通过多样化反事实解释来解释机器学习分类器。载于公平性、问责制与透明度会议。[34] F Nargesian、H Samulowitz、U Khurana、E Khalil 和 D Turaga。2017。面向分类的特征工程学习。载于 IJCAI。[35] A Ni、D Ramos、A Yang、I Lynce、V Manquinho、R Martins 和 C Le Goues。2021。SOAR：一种用于数据科学 API 重构的合成方法。载于 IEEE/ACM ICSE。

[36] Barry O'Sullivan。2010。约束编程中的自动建模与求解。载于第二十四届 *AAAI*人工智能大会 （佐治亚州亚特兰大）（*AAAI'10*）。AAAI Press，1493–1497。[37] H Paulheim 和 J Fümkranz。2012。基于链接开放数据的数据挖掘特征的无监督生成。*WIMS*（2012）。[38] S Ramnandan、A Mittal、C Knoblock 和 P Szekely。2015。为数据源分配语义标签。载于《语义网：最新进展与新领域》。[39] D Ritze、O Lehmberg、Y Oulabi 和 C Bizer。2016。分析Web表格增强跨领域知识库的潜力。载于*WWW*。[40] P Sondhi。2009。特征构造方法：综述。*UIUC*技术报告 （2009）。[41] K Srinivas、T Tateishi、D Weidele、U Khurana、H Samulowitz、T Takahashi、D Wang 和 L Amini。2022。通过代码挖掘和语义类型检测进行语义特征发现。载于*AAAI*。[42] R Sutton 和 A Barto。2018。强化学习：导论 （第二版）。麻省理工学院出版社。http://incompleteideas.net/book/the-book-2nd.html [43] K Takeoka、M Oyamada、S Nakadai 和 T Okadome。2019。Meimei：一种用于表格语义标注的高效概率方法。*AAAI*（2019）。[44] Mohamed Trabelsi、Jin Cao 和 Jeff Heflin。2020。使用深度上下文化语言模型进行语义标注。arXiv: 2010 16037 。B Ustun、A [45] 2019 Spangher 和 Y Liu。。线性分类中的可操作的补救措施。载于公平、问责与透明度会议论文集。[46] J Vielma。2015。混合整数线性规划公式化技术。*SIAM Rev.*（2015）。[47] J Wang、Z Song、Q Li、J Yu 和 F Chen。2014。面向大数据的基于语义的智能数据清理框架。载于*IEEE SPAC*。[48] M Yasunaga 和 J Lafferty。2019。TopicEq：面向科学文本的主题与数学方程联合模型。*AAAI*（2019）。[49] Y Zhang 和 Z Ives。2020。在数据湖中查找相关表格以实现交互式数据科学。载于*ACM SIGMOD*。[50] E Zhu、D Deng、F Nargesian 和 R Miller。2019。JOSIE：用于在数据湖中查找可连接表格的重叠集相似性搜索。载于*ACM SIGMOD*。[51] M Zöller 和 M Huber。2021。自动化机器学习框架的基准测试与综述。《人工智能研究杂志》（2021）。