评人工智能如何走向新阶段

（兼谈国内外跟帖评论）

陆首群

2022.1.4
（2022.3.16补充）

国内外AI跟帖留言

（780条~837条跟帖）

第十一集

中国开源软件推进联盟
China OSS Promotion Union

评人工智能如何走向新阶段?

陆首群

2022.1.4

（2022.3.16 补充三篇通用人用智能的论述）

1、当前人工智能发展的热点是如何将弱人工智能转化为强人工智能，而机器学习（包括深度学习）是弱人工智能的代表。机器学习／深度学习是一种强大的数据统计分析工具，但它也是有缺陷的，它本质上是一项暗箱操作（黑盒子）技术，其训练过程不可解释。如何打破机器学习内在的黑盒子痼疾，实现可解释的人工智能，使弱人工智能的机器学习转变为强人工智能的可解释机器学习，如今已成为全球人工智能研发的亮点。

早在 2019 年 8 月，COPU 就提出研发可解释性人工智能（XAI）的任务,并予 2020.6 在开源高峰论坛上邀请 IBM 作“可信任人工智能(反欺炸、可解释、公平性）”的报告。

在国内，2020.12，沈向洋教授提出“拥抱开源，我们现在最重要的事情是要做可解释的人工智能”，2021.1，姚期智院士提出“机器学习算法缺乏可解释性，很多算法处于黑盒子状态，这项人工智能的技术瓶颈亟待突破”。

迄今 COPU 收到全球研发可解释性人工智能的跟帖 70 多件。

由于全球人工智能技术(XAI)尚未完全成熟，在研发 XAI 算法时，专家对各道演算程序的理解和操作尚具不确定性，最后评估还只能靠

人工，XAI 演算结果不够精准，致使可解释性机器学习推广应用存在一定困难，为此 COPU 要求 IBM 专家对其列举的案例的各道演算程序进行解析和说明。

经过一年半的讨论和研究，2021.12.9，IBM 的专家们终于按要求提出了《可信任的人工智能》的报告（分析了三个可解释性人工智能案例：银行信贷业务、个人医疗开支、皮肤镜检查应用）。

我们对 IBM 应 COPU 之请三番五次对可解释机器学习案例作业程序进一步作出说明，表示感谢！

2、研发基于异步脉冲神经网络的神经拟态计算系统，期望走上强人工智能之路。

在这方面表现突出的有英特尔的 Loihi 芯片、Pohoiki Springs 神经拟态网络（由 1 亿神经元组成）及神经拟态计算系统，目前已制成原型机并正在开发应用生态；还有曼彻斯特大学等研究集成的由 10 亿个神经元组成的神经拟态网络及制成 Spin Naker 类脑计算机原型机；浙江大学自主研发达尔文-2 芯片，集成 1.2 亿个神经元组成的神经拟态网络，正在跟进。

3、采用数据、知识双驱动，立足于知识工程，研发大规模语义网络（知识图谱）以支持实现认知智能，这条通向强人工智能之路难度较大，主要原因在于目前研发的语义网络缺乏逻辑推理机制，机器尚难识别常识等短板。

4、探索如何构建通用人工智能（或第三代人工智能）之路，还处于早期构思阶段，尚未探索到核心理论，可说是路漫漫！

过去在本刊的国内外人工智能跟帖留言中，论述未来通用人工智能的主要有北京邮电大学的钟义信教授、清华大学人工智能研究院的张钹院士的团队，在本集中我们还发表了东北大学科学技术哲学研究中心（陈凡、吴怡等）、中国社会科学院（刘方喜）、复旦大学哲学学院（徐英瑾）等学者对通用人工智能的有关论述。

国内外 AI 跟帖留言（780-837）

780，空间机器学习模型诊断：一种基于模型不可知距离的方法

Alexander Brenning，Friedrich Schiller 大学，2021.11.13

虽然在解释黑箱机器学习（ML）模型方面已经取得了重大进展，但仍然缺乏诊断工具来阐明 ML 模型在预测技术和可变重要性方面的空间行为。

本文贡献提出了空间预测误差剖面（SPEP）和空间变量重要性剖面（SVIP）作为新的模型不可知评估和解释工具，用于空间预测模型，重点是预测距离。它们的适用性在两个案例研究中得到了证明，这两个案例研究分别代表了环境科学背景下的区域化任务和遥感土地覆盖分类的分类任务。

新的诊断工具丰富了空间数据科学的工具包，并可能改进 ML 模型的解释、选择和设计。

781，生物启发的语音情感识别

Reza Lotfidereshgi 等，Sherbrooke 大学，2021.11.15

传统的基于特征的分类方法不能很好地应用于语音情感的自动识别，主要是因为识别说话人情感状态所需的精确的光谱和韵律特征集尚未确定。

本文提出了一种直接对语音信号进行处理的方法，从而避免了特征提取的困难步骤。此外，该方法结合了人类语音产生的经典源滤波器模型和最近引入的液态机（LSM）的优点，后者是一种受生物启发的尖峰神经网络（SNN）。首先分离语音信号的源和声道分量，并将其转换为感知相关的频谱表示。然后由两个神经元库分别处理这些表示。该方法在柏林情感语音数据库（Emo-

DB）上具有很好的分类性能。

782，人工智能找到全新抗生素，可杀死超级耐药菌

MIT，载于 2020.2.21《细胞》杂志《Cell》

通过一种机器学习 / 深度学习模型（弱人工智能）发现了一种潜在的药物的抗菌潜力。在动物实验中，这种全新的抗生素能有效杀死一种对已知所有抗生素都耐药的超级细菌。

机器学习模型能够自动学习药物分子里的结构，不但可以掌握这些分子的不同位置是否存在特定的化学基因，还能预测这些分子的特性。随后研究人员给这种模型提供了 2335 个用于“学习”的不同分子，这些分子中有美国 FDA 已经批准的药物，也有不少具有广泛生物特性的天然分子。研究人员希望一任训练之后，这种模型能够学会识别能有效杀死大肠杆菌的药物。

训练完后是检验这套机器学习模型学习能力的时候，研究人员使用 Broad 研究所的一个化合物库，让这套模型从其中 6111 个分子里寻找具有潜在抗菌潜力的分子，这种模型认为一个分子具有很强的抗菌活性，这种分子原先是作为一种糖尿病药物而开发出来的，在结构上和已有的任何一种抗生素都明显不同，后续研究也表明该分子对人类细胞的毒性较低。

随后，他们在培养皿里测试了 halicin 对多种耐药菌的杀菌效果，令人欣喜！除了铜绿假单胞菌（Pseudomonasaeruginosa，一种难治的肺部病原体）之外 halicin 对所有测试的耐药菌都有杀伤作用，显示良好的广谱抗菌活性。

783，通过数据驱动仿真学习交互式驾驶策略

Tsun-Hsuan Wang 等，MIT（美），TRI（日），2021.11.23

论文地址：https://arxiv.org/pdf/2111.12137.pdf

论文内容：

数据驱动的模拟器为驱动策略学习提供了高效的数据效率。但当用于交互建模时，这时数据效率成为一个瓶颈：小的底层数据集，通常缺乏学习交互驾驶的关键且具有挑战性的边缘案例。

在本文中，提出了一个端到端的框架，用于在静态和动态代理交互的情况下对自治代理进行真实世界场景的模拟和训练。本文中的训练环境是使用真实数据的，并且支持多个代理的高保真渲染，这样自我代理学习的控制策略可以直接应用在在真实世界中的全尺寸自动驾驶车辆上，而不需要任何程度的域随机化、增强或微调方法。通过模拟任意代理交互，不需要大量的专家培训数据和密集的监控信号，而这是现有模仿学习方法的两个常见缺陷。通过这种支持交互环境建模和仿真的方法，采用最先进的策略学习技术学习交互式系统的驾驶任务（如跟车和超车），智能代理可以在丰富的多代理交互以及不同的照明和环境条件下，实现复杂的连续控制和决策。同时在全尺寸自动驾驶车辆上，进行零次策略转移的实验，包括在具有感知挑战性的测试跑道上成功超车的测试任务。实验表明模型具有高性能和通用性，所学习到的策略可以直接转移到全尺寸的自动驾驶车辆上，而无需使用任何传统的 sim-to-real 转移技术，如域随机化。

总结来说，本文提出了一种新的方法，使用多智能体数据驱动的仿真来学习一个端到端控制器，实时决策和自动驾驶。同时提出了几个多代理任务，随

着复杂程度的增加，需要进行广泛的模拟和现实世界中的实证分析，将学到的策略部署在全尺寸自动驾驶汽车上。而未来的研究方向包括但不限于多模态，模拟更复杂的任务，包括情景模拟，代理交互的感知和边缘案例生成。

784，堆叠式深层卷积神经网络来预测涡扇发动机剩余使用寿命

David Solis-Martin 等，2021.11.24

论文地址：https://arxiv.org/abs/2111.12689

内容：

介绍了用于预测飞机发动机剩余使用寿命（RUL）的基于数据的技术和方法。解决方案基于两个堆叠在两个级别的深度卷积神经网络 (DCNN)。第一个 DCNN 用于提取低维特征，并使用归一化的原始数据作为输入的向量。 第二个 DCNN 摄取了从以前的 DCNN 中提取的向量列表并估计 RUL。模型选择是通过使用重复随机子采样验证方法的贝叶斯优化进行的。在这项工作中，获得最终解决方案的过程分为两个学习阶段。在第一个中，学习原始数据的编码并将其用作第二个学习阶段的输入以获得能够估计 RUL 的最终模型。

训练模型的源码： https://github.com/DatrikIntelligence/Stacked-DCNN-RUL-PHM21.

785，用于皮肤病变诊断的可解释性深度图像分类器

Carlo Metta 等，2021.11.22

论文地址：https://arxiv.org/pdf/2111.11863.pdf

内容：决策系统中采用的深度学习模型的可解释性是医疗诊断等重要环境

中的关键问题，可解释人工智能 (XAI) 的研究正在试图解决医疗诊断的问题，然而，现在的 XAI 方法通常只在一般分类器上进行测试，并不代表诸如医学诊断等现实问题。本文的目的是研究解释方法在实际医疗环境中的可用性，在本文中，作者们分析了一个关于皮肤病变图像的案例研究，定制了一个 XAI 方法来解释能够识别不同类型皮肤病变的深度学习模型。作者使用 ResNet 分类器对 ISIC 数据集（该数据集由 25,331 张皮肤病变及其类别（标签）图像的训练集组成；一个包含 8,238 张图像的测试集）进行分类，ABELE 解释器对 ISIC 数据集进行解释，研究发现，解释者采用的潜在空间分析揭示了一些最常见的皮肤病变类别是明显分开的，而这种现象可能源于每个类别的内在特征，经过证明，通过针对性的训练，abele 能够做出有意义的解释，真正能够帮助从业者。

786，一种通过 FPGA 上新兴的神经编码加速脉冲神经网络的端到端框架

Daniel Gerlinghoff 等，香港中文大学（深圳），2021.11.19

论文地址： https://arxiv.org/abs/2111.10027

内容：

编译器框架对于广泛使用基于 FPGA 的深度学习加速器至关重要。它们允许不熟悉硬件工程的研究人员和开发人员利用特定领域逻辑获得的性能。传统人工神经网络存在多种框架。然而，在创建针对脉冲神经网络 (SNN)优化的框架方面并未投入太多研究工作。这种新一代神经网络对于在具有严格功率和资源限制的边缘设备上部署 AI 变得越来越有趣。我们的端到端框架 E3NE 可自动为 FPGA 生成高效的 SNN 推理逻辑。它基于 PyTorch 模型和

用户参数，应用各种优化并评估基于脉冲的加速器的固有优缺点。多级并行和新兴神经编码方案的使用导致效率优于以前的 SNN 硬件实现。对于类似的模型，E3NE 使用不到 50%的硬件资源和 20%的功耗，同时将延迟降低了一个数量级。此外，可扩展性和通用性允许部署大规模 SNN 模型 AlexNet 和 VGG。

787，使用 Loihi-2 进行高效的神经形态信号处理

Garrick Orchard 等，英特尔，2021.11.5

论文地址：https://arxiv.org/pdf/2111.03746.pdf

内容：神经形态计算中使用的受生物启发的尖峰神经元是带有动态状态变量的非线性滤波器——与深度学习中使用的无状态神经元模型非常不同。英特尔神经形态研究处理器 Loihi 2 的下一个版本支持多种具有完全可编程动态的状态脉冲神经元模型。在这里，我们展示了先进的尖峰神经元模型，可用于在仿真 Loihi 2 硬件上的模拟实验中有效地处理流数据。在一个例子中，共振和激发（RF）神经元用于计算短时傅里叶变换（STFT），其计算复杂度与传统 STFT 相似，但输出带宽比传统 STFT 少 47 倍。在另一个例子中，我们描述了一种使用时空 RF 神经元的光流估计算法，与传统的基于 DNN 的解决方案相比，该算法需要的运算量少 90 倍以上。我们还展示了利用反向传播训练用于音频分类任务的 RF 神经元的有希望的初步结果。

最后，我们证明了 Hopf 谐振器（RF 神经元的一种变体）的级联复制了耳蜗的新特性，并激发了一种有效的基于尖峰的频谱图编码器。

788，A survey on knowledge graphs: representation, acquisition and applications

Shaoxiong Ji 等，2021.11.28

论文地址：https://arxiv.org/pdf/2002.00388.pdf

内容：读后感

这属于一篇比较新的知识图谱类的综述文章，主要贡献是对近期前沿的图谱研究工作有一个全局性的总览，并且调整了图谱研究工作的分类，并以分类为纲对前沿的主要研安技术进行总结，对未来知识相关的应用趋势进行了前瞻预测。

本文按照知识图谱的调研和总结、按照新的分类来对知识图谱领域的研究内容进行划分、前沿技术的分析、对未来的研究方向领域的预测这四个部分来进行论述。本文按照当前知识图谱的重点研究工作来对知识图谱进行分类，分别是知识图谱表示学习（KRL，也称为图谱嵌入）、知识获取和补全、时序知识图谱、知识感知应用。

知识图谱表示学习

- 表示空间：主要是四类包括实数点向空间、复杂空间、高斯分布空间、流形和群。
- 评分函数：包括基于距离和基于语义两种。
- 编码空间：主要是论述了线性/非线性模型、张量模型、神经网络三种。
- 辅助信息：图谱嵌入表示过程中，为了最小化语义损失，会将更多的图谱信息进行嵌入，这些信息包括文本描述、实体/关系类别信息、可视信息（如图片等）、实体/关系的属性、关系路径、逻辑规则。

知识获取和补全

- 图谱补全
- 实体发现：包括实体识别、实体分类、实体消歧、实体对齐。
- 关系抽取：主要是 DNN、GCN、CNN 的结合。

时序知识图谱

- 总结了相关的几篇文献。

知识感知应用

- 自然语言理解
- 问题回答
- 推荐系统

789，图神经网络中的问题

Xiang Song（AWS AI），2021.11.23

论文地址：https://arxiv.org/pdf/2111.11638.pdf

内容：

图神经网络(GNN)在从包含节点/边缘特征信息的图结构数据中学习方面取得了成功，并应用于社交网络、推荐、欺诈检测和知识图推理。在这方面，过去已经提出了各种策略来提高 GNN 的表达能力。例如，一个直接的选择是通过扩展隐藏维度或增加 GNN 层数来简单地增加参数大小。然而，更宽的隐藏层很容易导致过度拟合，并且逐渐增加更多的 GNN 层可能会导致这个 http URL 在本文中，我们提出了一种与模型无关的方法，即图神经网络中的网络(NGNN)，它允许任意 GNN 模型通过使模型更深来增加他们的模型

容量。然而，NGNN 不是添加或加宽 GNN 层，而是通过在每个 GNN 层中插入非线性前馈神经网络层来加深 GNN 模型。 NGNN 在 ogbn-products 数据上应用于基于 Graph Sage 的 GNN 的分析表明，它可以使模型保持稳定，不受节点特征或图结构扰动的影响。此外，节点分类和链接预测任务的广泛评估结果表明 NGNN 在不同的 GNN 架构上可靠地工作。例如，它将 Graph Sage 在 ogbn-products 上的测试准确率提高了 1.6%，并提高了 hits@100 分数 SEAL 在 ogbl-ppa 上的得分提高了 7.08%，Graph Sage+Edge-Attr 在 ogbl-ppi 上的 hits@20 得分提高了 6.22%。并且在本次提交时，它在 OGB 链接预测排行榜上获得了两个第一名。

790，一种用于脑机接口 P300 信号识别的新型神经网络

Jingrou Xu 等，电子科技大学 2021.11.18

论文地址：https://ieeexplore.ieee.org/document/9604420

内容:P300 事件相关电位(ERP)对脑机接口的研究具有重要意义。由于 P300 脑电图（EEG）信号的信噪比低，特征提取困难，目前对 P300 ERP 信号识别精度的优化有限。为了解决这些问题，我们提出了一种用于 P300 信号识别的新型神经网络。首先，通过实例标准化，对脑电信号各通道内部数据进行归一化，消除分布差异，有利于 P300 信号的特征提取。其次，我们设计了 MultFeat 模块来提取样本的全局特征和通道特征，并融合它们以增强特征表示能力。此外，我们还提出了一种脑电信号数据预处理方法，可有效提高脑电信号的信噪比，抑制样本不平衡。我们进行了大量实验，所提出的网络在 BCI 竞赛 III 的数据库 II 上达到了 95% 的准确率，优于传统的支持

向量机机器学习方法。结果表明,本文提出的方法能够有效识别 P300 信号。

791，关于两种 XAI 文化：已部署 AI 系统中非技术解释的案例研究

Helen Jiang 等，佐治亚理工学院，2021.12.2

链接：https://arxiv.org/pdf/2112.01016.pdf

简介：可解释人工智能（XAI）的研究已经蓬勃发展，但“我们让人工智能对谁解释？”这个问题还没有得到足够的重视。非人工智能专家对 XAI 的理解不多,尽管如此,他们是实际部署的人工智能系统的主要受众和主要利益相关者。差距是显而易见的：人工智能专家和非专家认为“解释”的内容在实际场景中非常不同。因此，这一差距在现实 AI 部署中产生了两种截然不同的期望、目标和 XAI 形式文化。

我们主张为非技术受众开发 XAI 方法至关重要。然后，我们介绍了一个现实案例研究，其中人工智能专家向非技术利益相关者提供了人工智能决策的非技术性解释，并在高度监管的行业中成功部署。然后，我们综合从案例中吸取的经验教训，并分享 AI 专家建议，当解释 AI 决策的非技术利益相关者的意见。

792，人工智能驱动的移动网络：从认知到决策

Guiyang Luo 等，2021.12.8，

链接：https://arxiv.org/pdf/2112.04263.pdf

简介:移动网络(MN)预计将提供前所未有的机会,实现互联体验的新世界,并从根本上改变人们与一切事物互动的方式。在日益复杂的配置问题和不

断增长的新服务需求的推动下，MN 变得越来越复杂。这种复杂性给部署、管理、操作、优化和维护带来了重大挑战，因为它们需要对 MN 有全面的理解和认知。人工智能（AI）处理计算机中智能行为的模拟，在许多应用领域已显示出巨大的成功，表明其在认知 MN 状态和做出智能决策方面的潜力。在本文中，我们首先提出了一种人工智能驱动的移动网络体系结构，并讨论了认知复杂性、高维行动空间决策和系统动态自适应方面的挑战。然后，讨论了与人工智能相关的潜在解决方案。最后，我们提出了一种深度学习方法，将认知与决策相结合，直接将 MN 的状态映射到感知的 QoS。我们提出的方法有助于运营商做出更智能的决策，以保证 QoS。同时，我们提出的方法的有效性和优势在一个真实的数据集上得到了验证，该数据集在 5 天内涉及 77 个站点的 31261 个用户。

793，适用于图像分类模型研究的认知心理学外推框架

Roozbeh Yousefzadeh 等，耶鲁大学，2021.12.6

链接：https://arxiv.org/pdf/2112.03411.pdf

简介：我们研究了深度学习图像分类模型的功能任务，并表明图像分类需要外推能力。这表明，为了理解深度学习，必须开发新的理论，因为当前的理论假设模型只是插值，留下了许多关于它们的问题没有答案。我们研究了通过训练模型从图像中提取的像素空间和特征空间（在其隐藏层中，包括预训练残差神经网络最后一个隐藏层中的 64 维特征空间），以及通过小波/剪切波提取的特征空间。在所有这些领域中，测试样本都大大超出了训练集的凸包，图像分类需要外推。与深度学习文献相反，在认知科学、心理学和神经

科学中，外推和学习通常是同时进行的。此外，据报道，人类视觉认知和行为的许多方面都涉及外推。我们提出了一个新的外推框架，用于深入学习模型的数学研究。在我们的框架中，我们使用术语外推，在训练集的凸包外（在像素空间或特征空间中），但在训练数据定义的特定范围内，以这种特定的方式外推，认知科学的许多研究中定义了相同的外推方式。我们解释说，我们的外推框架可以为深度学习的开放性研究问题提供新的答案，包括其过度参数化、训练机制、分布外检测，等等。我们还发现，在学习任务中，外推的程度可以忽略不计，据报道，深度学习没有简单模型的优势。

794，基于不同原型赋值的可解释图像分类

雅盖隆大学数学与计算机科学学院，阿尔迪根公司，2021.12.7

论文地址：https://arxiv.org/pdf/2112.02902.pdf

内容：本文介绍了 ProtoPool，一种自解释的细粒度原型模型。图像分类。在 ProtoPool 中，作者们实现了一些主要的新颖元素，与之前的模型（如 ProtoPNet、ProtoPShare 和 ProtoTree）相比，这些元素大大减少了原型的数量，同时获得了更高的可解释性和更容易的训练。作者们没有将原型硬分配给类，而是实现了作为原型集分布的软原型。该分布在训练期间使用 Gumbel-Softmax 技巧随机初始化和二值化。这种机制通过删除 ProtoPNet、ProtoPShare 和 ProtoTree 中所需的修剪步骤简化了训练过程。第二个新颖之处是焦点相似度函数，它将模型集中在罕见的前景特征上。此外，引入了一个新的焦点相似度函数来将模型集中在罕见的前景特征上。作者们表明 ProtoPool 在 CUB-200-2011 和斯坦福汽车数据集上获得了最先进的准

确性，大大减少了原型的数量。在文章中，作者分析 了 ProtoPool 模型的可解释性。首先，作者们表明 ProtoPool 模型既可以用作局部解释，也可以用作全局解释。然后，作者们讨论 ProtoPool 和其他基于原型的方法之间的区别，调查两个类在 ProtoPool 训练的每次运行中是否共享相似数量的原型。然后，作者们对 ProtoPNet、ProtoTree 和 ProtoPool 使用的相似性函数进行用户研究，以评估人类对它们的可理解性。最后，作者们从认知心理学的角度考虑 ProtoPool。在文章中，作者们提供了该方法的理论分析和用户研究并提供了代码，以表明作者们的原型比使用竞争方法获得的原型更具特色。

795，对比和反事实调查 可解释人工智能的解释生成方法

IlIA Stepin 等，圣地亚哥-德孔波特斯拉大学，2021.3

论文地址：https://ieeexplore.ieee.org/abstract/document/9321372

内容：

a.对比解释

在人文社会科学中积累的解释结果表明，它具有内在的反差。对比性的属性预设了一个解释，即在假设的非发生选项(“为什么 P 发生了而不是 Q?”)的前提下，回答了关于事件原因的“为什么”问题(“为什么 P 发生了?”)因此，实用主义解释方法的支持者认为，正是这种能力将解释性问题的答案从一组对比的假设备选方案中区分出来，为被解释者提供了关于问题背后推理的充分全面的信息。这种方法也被认为设定了一个解释必须满足的最低标准:它必须有利于观察到的事件 P 的概率，而不是所有假设的选项(A。

问 2， .，问 n)。对比解释是认知科学中最具影响力的话题之一。因此，对比解释被认为是人类认知与生俱来的。事实上，我们习.惯于质疑我们曾经作出的那些决定，特别是如果这些决定)或巧合的情况导致了悲剧事件。

此外，对比推理是外展推理的基础。即推断某些事实，使某些观察结果可信的过程。换句话说，一个给定的观察结.果可以在一堆相互竞争的假说的基础上得到解释。

b.反事实的解释

鉴于对比性的性质，我们可以想象，如果在某一时刻做出了不同的决定，事情会如何发展，那么我们就有可能做出解释性的选择。它们可以用来解释这种不同的未采取的替代决定的潜在后果。在这种情况下，大脑被假设为构建并比较一个实际发生的事件的心理表征和它的一-些替代事件。认知科学家把这种对过去事件的替代方案的心理表征称为反事实(“与事实相反”)。“思考过去的可能性和过去或现在的不可能”的过程因此被称为反事实思维。另外，想象与实际发生的情况相关的另一种情况，并探索其后果的组合被称为反事实推理。此外，反事实推理被认为是解释变化环境中适应性行为的关键机制。

796，基于信息瓶颈的 Hebbian 学习规则自然地将工作记忆和突触更新联系起来

Kyle Daruwalla，威斯康辛大学，2021.11.24，

论文地址：https://arxiv.org/pdf/2111.13188.pdf

内容：人工神经网络通过反向传播训练极深的网络，成功地解决了各种各样

的问题。反向传播对尖峰神经网络的直接应用包含生物学上不可信的组件，例如重量传输问题或单独的推理和学习阶段。各种方法分别针对不同的组件，但完整的解决方案仍然是无形的。在这里，作者采用了一种替代方法，完全避免了反向传播及其相关问题。最近在深度学习方面的工作提出通过信息瓶颈 (IB) 独立训练网络的每一层。随后的研究指出，这种逐层方法避免了跨层的错误传播，从而形成了生物学上合理的范式。很遗憾，IB 是使用一批样本计算的。先前的工作通过仅使用两个样本（当前和上一个样本）的权重更新解决了这个问题。作者的工作采用不同的方法，将权重更新分解为局部和全局分量。本地组件是 Hebbian，仅取决于当前样本。全局组件计算依赖于一批样本的逐层调制信号。作者表明，这种调制信号可以通过具有工作记忆 (WM) 的辅助电路（如水库）来学习。因此，作者可以使用大于 2 的批量大小，批量大小决定了 WM 所需的容量。据作者所知，作者的规则是第一个将突触更新与任务的 WM 直接耦合的生物学上合理的机制。作者在合成数据集和图像分类数据集（如 MNIST）上评估作者的规则，并探索 WM 容量对学习性能的影响。作者希望他们的工作是了解记忆在学习中的机制作用的第一步。

797，面向节能嵌入式神经拟态计算多核 BigLittleuBrain 设计

M.L. Varshika，比利时德雷塞尔大学，2021.11.23

论文地址： https://arxiv.org/abs/2111.11838

内容：随着嵌入式系统中基于脉冲的深度学习推理应用的增加，这些系统倾向于集成神经拟态加速器，例如 μ 提高能源效率的大脑。我们提出一个 μ

基于大脑的可扩展多核神经拟态硬件设计，可加速脉冲深度卷积神经网络(SDCNN) 的计算。为了提高能源效率，内核在其神经元和突触容量方面被设计为异构（大内核比小内核具有更高的容量），并且它们使用并行分段总线互连进行互连，与相比，这导致更低的延迟和能量到传统的基于网格的片上网络(NoC)。我们提出了一个名为 SentryOS 的系统软件框架，将 SDCNN 推理应用程序映射到所提出的设计。SentryOS 由一个编译器和一个运行时管理器组成。编译器利用 big 和 little 的内部架构将 SDCNN 应用程序编译成子网络 μ 脑核。运行时管理器将这些子网络调度到核心上并流水线化它们的执行以提高吞吐量。我们使用五个常用的 SDCNN 推理应用程序评估了所提出的大、小众核神经拟态设计和系统软件框架，并表明所提出的解决方案降低了能量（37%到 98% 之间），减少了延迟（9%到 25%之间），并增加应用程序吞吐量（在 20%到 36%之间）。我们还表明，SentryOS 可以很容易地扩展到其他脉冲神经拟态加速器。

798，使用图摘要技术改进知识图谱上的问答系统

Sirui Li，澳大利亚默多克大学，2021.12.5

论文地址：https://link.springer.com/content/pdf/10.1007%2F978-3-030-92273-3_40.pdf

内容：知识图谱 (KG) (KGQA) 上的问答 (QA) 系统使用 KG 中包含的三元组自动回答自然语言问题。关键思想是将 KG 的问题和实体表示为低维嵌入。以前的 KGQA 曾尝试使用知识图嵌入 (KGE) 和深度学习 (DL) 方法来表示实体。然而，KGEs 太浅，无法捕捉表达特征，DL 方法独立处理每个三

元组。最近，图卷积网络 (GCN) 在提供实体嵌入方面表现出色。然而，将 GCN 用于 KGQA 是低效的，因为 GCN 在聚合邻域时平等对待所有关系。此外，在使用以前的 KGQA 时可能会出现一个问题：在大多数情况下，问题的答案通常是不确定的。为了解决上述问题，我们提出了一种使用循环卷积神经网络 (RCNN) 和 GCN 的图摘要技术。 GCN 和 RCNN 的组合确保嵌入与与问题相关的关系一起传播，从而获得更好的答案。所提出的图摘要技术可用于解决 KGQA 无法回答答案数量不确定的问题。在本文中，我们在最常见的问题类型（单关系问题）上演示了所提出的技术。实验表明，与 GCN 相比，所提出的使用 RCNN 和 GCN 的图摘要技术可以提供更好的结果。当问题的答案数量不确定时，所提出的图摘要技术显着提高了对实际答案的回忆。

799，为什么通用人工智能不会实现

Ragnar Fjelland，挪威卑尔根大学，2020.6

内容：创造类人人工智能(AI)的现代项目始于第二次世界大战后，当时人们发现电子计算机不仅是处理数字的机器，而且还能处理符号。追求这一目标是有可能的，无需假设机器智能与人类智能是相同的。这就是所谓的弱 AI。然而，许多人工智能研究人员追求的目标是开发原则上与人类智能相同的人工智能，称为强人工智能。弱 AI 不如强 AI 雄心勃勃，因此争议较少。然而，也存在着与弱 AI 相关的重要争议。本文重点讨论了人工智能(AGI)和人工窄智能(ANI)的区别。虽然 AGI 可能被归类为弱 AI，但它接近于强 AI，因为人类智能的一个主要特征是它的普遍性。尽管 AGI 没有强大的 AI 那么雄

心勃勃，但几乎从一开始就有批评者。主要的批评者之一是哲学家休伯特·德雷福斯(Hubert Dreyfus)，他认为，没有身体、没有童年、没有文化实践的计算机根本不可能获得智力。德雷福斯的主要论点之一是，人类的知识在一定程度上是隐性的，因此无法在计算机程序中表达和整合。然而，今天有人可能会说，人工智能研究的新方法已经使他的观点过时了。深度学习和大数据是最新的方法，倡导者认为它们将能够实现 AGI。仔细观察就会发现，尽管用于特定目的的人工智能(ANI)的发展令人印象深刻，但我们还没有在开发人工通用智能(AGI)方面走得更近。这篇文章进一步指出，这在原则上是不可能的，它恢复了休伯特·德雷福斯的观点，即计算机并不存在于这个世界上。

800，机器智能的限度：尽管机器智能取得了进步，但人工智能一般智能仍然是一大挑战

Henry Shevlin 等，剑桥大学，帝国理工学院，2019.9

内容：尽管最近机器学习取得了突破性进展，但目前的人工智能系统缺乏生物智能的关键特征。当前的局限性能否克服是一个开放性的问题，但鉴于其对社会的影响，回答这一问题至关重要。在这篇论文中的目标是提出一般智能的含义，以及神经科学将如何变得至关重要，以及人类-人工智能混合系统的使用。无论如何，正在进行的机器学习研究的命运肯定会影响到认知科学中关于大脑的结构和功能，也许还有智能本身的未来的长期争论。

801，通过频率和通道神经注意力进行基于 EEG 的听觉注意力检测

Siqi Cai 等，华南理工大学、香港中文大学、新加坡国立大学，2021.12.2

论文地址：https://ieeexplore.ieee.org/abstract/document/9633231

内容：人类可以集中注意力在多声源环境中的单个声源上。听觉注意力检测(AAD) 旨在从一个人的大脑信号中检测出特定的声源，这将使许多创新的人机系统成为可能。然而，脑电图（EEG）信号的有效表征学习仍然是一个挑战。在本文中，我们提出了一种神经注意力机制，该机制动态地为 EEG 信号的子带和通道分配不同的权重，以推导出 AAD 的判别式表示。简而言之，我们想建立一个计算注意力机制，即神经注意力，来模拟人脑中的听觉注意力。我们将提出的神经注意力整合到 AAD 系统中，并通过对两个公开可用数据集的综合实验验证神经注意力机制。实验结果表明，所提出的系统显着优于最先进的参考基线。

802，用发型、妆容和面部形态解释人脸识别准确性的性别差异

作者：V í tor Albiero 等，IEEE、圣母大学（诺特丹大学）、佛罗里达科技大学，2021.12.29

链接：https://arxiv.org/pdf/2112.14656.pdf

简介：媒体报道指责人脸识别有“偏见”、“性别歧视”和“种族主义”。研究文献一致认为，女性的人脸识别准确率较低，她们通常具有较高的错误匹配率和较高的错误不匹配率。然而，很少有公开发表的研究旨在确定女性准确率较低的原因。例如，2019 年的人脸识别供应商测试记录了一系列算法和数据汇总女性准确率较低的情况，该测试还在“我们没有做的事情”标题下列出了“分析原因和影响”。我们提出了第一个实验分析，以确定研

究中观察到这一结果的数据集中女性人脸识别准确率较低的主要原因。控制测试图像中相同数量的可见人脸可以降低女性的错误不匹配率。其他分析表明，化妆平衡数据集进一步提高了女性的假不匹配率。最后，一项聚类实验表明，两个不同女性的图像本质上比两个不同男性的图像更相似，这可能解释了错误匹配率的差异。

803，在 CARLA 实施的基于交叉口情况覆盖的自动驾驶车辆验证和确认框架，Zaid Tahir，Rob Alexander，约克大学、波士顿大学，2021.12.24

链接：https://arxiv.org/pdf/2112.14706.pdf

简介：自动驾驶汽车（AVs）在安全关键领域运行，因为自动驾驶软件中的错误可能导致巨大损失。据统计，道路交叉口是 AVs 领域驱动设计（ODD）的一部分，其事故率最高。因此，对道路交叉口的 AVs 进行测试并确保其在道路交叉口的安全性是重要的，也是本文的重点。我们提出了一个基于情景覆盖（SitCov）的 AV 测试框架，用于 AVs 的验证与确认（V&V）以及安全保证，该框架是在一个名为 CARLA 的开源 AV 模拟器中开发的。SitCov AV 测试框架侧重于不同环境和交叉口配置情况下道路交叉口的行车交互，使用自动生成 AVs 安全保证测试套件的情况覆盖标准。我们开发了一个用于交叉点情况的本体，并使用它生成一个情况超空间，即由该本体产生的所有可能情况的空间。为了评估我们的 SitCov AV 测试框架，我们在 ego AV 中植入了多个故障，并比较了基于情景覆盖和随机情景生成。我们发现，两种生成方法触发的种子故障数量大致相同，但基于情境覆盖的生成告诉我们更多关于 ego AV 自主驱动算法的弱点，尤其是在边缘情况下。我们的代码是

在线公开的，任何人都可以使用我们的 SitCov AV 测试框架，或者在它的基础上进一步构建。本文旨在为 AVs 的 V&V 和开发领域做出贡献，不仅从理论角度，而且从开源软件贡献的角度，发布一个灵活/有效的 AVs V&V 和开发工具。

804，迈向医学同行影响的 Shapley 价值图框架

Jamie Duell,Monika Seisenberger 等，斯旺西大学、普利茅斯大学，2021.12.29

链接：https://arxiv.org/pdf/2112.14624.pdf

简介可解释人工智能（XAI）是人工智能（AI）的一个子领域，处于 AI 研究的前沿。在 XAI 中，特征属性方法以特征重要性的形式产生解释。现有特征归因方法的一个局限性是缺乏对干预后果的解释。虽然强调了对某一预测的贡献，但并未说明特征与干预结果之间的影响。本文的目的是引入一个新的框架，以更深入地研究使用图形表示的特征对特征交互的解释，从而提高黑箱机器学习（ML）模型的可解释性，并为干预提供信息。

805，可信任的人工智能——

人工智能可解释性方法总结、案例分析及前景展望

署名 IBM 程海旭 吴婧 董琳 马小明 南驰 张红兵

我们在深度信息技术第四集介绍了 IBM 有关 AI 可解释性，健壮性及公平性的方法论。IBM 在这些方法论的基础上在 Linux 基金会开源了可解释性

工具套件 AIX360[1]，健壮性工具套件 ART[2] 和公平性工具套件 AIF360[3]。我们在第四集主要集中讨论了 AI 可解释性的技术背景及一个银行案例。

陆首群教授非常关注我们有关可信 AI 的技术，特别是 AI 可解释性的方法和案例。陆教授基于我们在第四集的文章提出了有关 AI 可解释性的八个问题，并邀请我们在这篇文章里详细阐述那八个问题是怎样在案例里解决的。陆教授的热情激励我们在这篇文章里总结了人工智能可解释性方法，分析了 AI 可解释性怎样帮助银行贷款，个人医疗支出预测和皮肤镜检查等三个案例，并展望了 AI 可解释性及可信 AI 的前景。

一、人工智能可解释性方法总结

AI 模型解释的背景：

1) 随着人工智能广泛应用，越来越多的 AI 模型应用落地，人们对于模型需要有比较清晰的认知，以便在模型使用时，更加确定，使用更加合适；

2) 复杂的机器学习模型，像深度学习（deep learning），集合模型（ensemble model，如 XGBoost）预测精度，效果好，但其结构庞大复杂，不像传统的机器学习模型（如线性回归，决策树）其结构明确，内涵清晰，容易解释。对于这些复杂模型，用户希望除了预测效果之外，希望进一步的了解；

3) 自动 AI 技术近年来发展迅速，其利用自动技术，在基本的属性特征基础上，做一系列的变化，操作和选择，进而生成新的特征，和基本特征一起建模，生成模型准确度高，效果好。 这个过程作为整体模型，如何解

[1] https://github.com/Trusted-AI/AIX360
[2] https://github.com/Trusted-AI/adversarial-robustness-toolbox
[3] https://github.com/Trusted-AI/AIF360

释和理解，也很重要；

4) 当使用模型作出业务预测之后，往往只有预测结果，而用户于各个影响因素所起的作用需要进一步了解，有助于增强用户对结果的信任和后续的决策。

综上，从应用的广泛性和技术发展复杂度，解释性变得日益必要与迫切。

1、选择演译方法（如决策树：树干指向演译目标，树枝指向特征），

- 用简单的，结构清晰的模型来解释复杂模型

模型表达的是:影响因素或特征与目标之间的映射关系，如果映射关系结构是清晰的，明确的，就是可以解释的。

该种演绎方法除了决策树外，典型演绎方法还有线性回归模型（特征是自变量，目标是因变量，目标的取值是多个自变量的线性组合，一个自变量贡献一部分,其中系数表达了自变量的重要程度）。

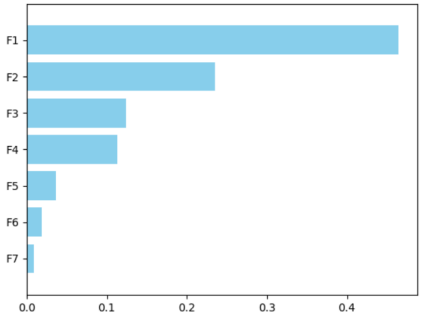
在特定场景，如果典型的，结构明确的算法模型（决策树或者线性回归）的预测或者识别的效果和复杂模型近似，就可以这些算法来解释复杂模型的算法。比较成熟和应用广泛的就是对于一条实例的预测结果解释。 用复杂模型（比如深度学习，xgboost）对图像，或者文字，或者一个贷款记录作出了识别或者预测。那么围绕着这条实例数据建立一个典型的线性回归模型（建立这个线性回归模型的数据由该 1）条实例数据，2)该复杂模型和 3）建立复杂模型数据共同生成， 以此来保证在这个实例上线性回归和复杂模型的等效性），用线性回归模型来解释这条记录的预测。0penScala 对于实例的解释采用该技术

- 从特征与目标之间的关系来理解和解释模型

当模型的结构特别复杂，或者其结构很难解释时。 从宏观上看多个特征与目标之间的关系，有助于理性模型，对模型有宏观的，整体的认知。宏观的认知包括

特征重要性（Feature Importance）

在模型众多的特征中，计算出每一个模型的重要度值。 从这些值的排序中可以看到哪些特征重要，哪些特征不太重要。典型的特征重要度如下所示



（横轴是重要度的值， 纵轴是各个特征，最上面的重要度最高的，往下依次降低）

2、选择特征

特征的选择基于既有的数据（客观存在）和一些主观经验，可以使用一下方法

- **相关关系（correlation）法**

通过特征值和目标的观测值，计算相关系值，常用*皮尔逊*相关系数（ Pearson correlation coefficient），如果值大于某阈值，一般是0.5，说明特征与目标有较强的关系，可以作为模型的预测特征

- **模型选择法**

将所有可特征作为预测变量，使用通用准确好的模型，比如 XGBoost。

然后逐个从模型中去掉某特征变量，再建模，比较两次模型准确度变化，判断特征是否有用。

- **经验判断**

业务人员根据主观经验，预判哪些特征变量影响目标。

- **自动建模技术(Auto AI)**

使用基于既有特征，自动选择和生成新的特征作为模型预测特征，今年来该技术发展较快（IBM 有相应的产品研发）

3、依据特征和数据建模

当数据比较充足和完整时，使用 **2（选择特征）**中方法，使用现有各种建模算法（包括传统机器学习算法， XGboost， 深度学习等，结合具体业务开发模型。典型的 AI 模型算法众多，具体选择算法，根据业务需求而定。例如是否放贷，属于典型的分类问题，XGBoost 常用典型算法。模型开发除了选择算法，一般还包括特征变量选择和参数调优，从这两个方面调高模型的精度。近年来随着 AutoAI 技术发展，建模开发的难度和周期开始降低和缩短。与此同时，模型解释的要求变高。

4、根据模型求解算法

一般而言，当选择了特定的 AI 模型，该模型的求解算法就已经存在。通常业务逻辑比较复杂，需要再 AI 模型结果的基础上，基于业务需求，二次加工。

5、在计算基础上进行评估（人工或机器）

对于 AI 模型的预测结果评估有通用的评估方法，数据一般会随机分为训练数据和测试数据两部分，训练数据主要用来训练模型，学习数据中的规

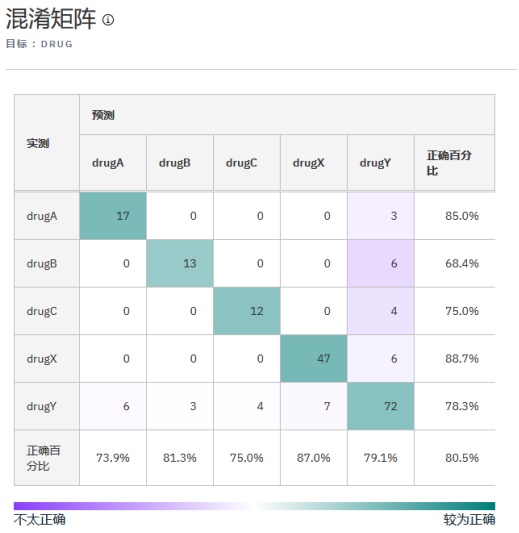
律；测试数据对学习的结果评估，主要是从准确度角度，通过目标的观测值和预测值比较评估。区分回归模型（Regression）和分类模型（Classification）

- **回归模型（regression）评估指标**

常用的评价指标有，均方误差（Mean Squared Error），均方根误差（Root Mean Squared Error），平均绝对误差（Mean Absolute Error）和 R Squared

- **分类模型评估**

分别统计每一类别中正确预测与错误预测的个数和占比，叫混淆矩阵（Confusion Matrix），如下所示 有 5 类药品，它们正确预测和错误预测个数



6、进一步研究是否达到公平、公正、可信？！

AI 模型的结果是从训练数据中学习到的，当测试准确度达到要求的指标够，首先说明模型是准确的，完成了从数据中学习规律的任务，是基于提供的数据是“可信的”。但训练数据可能并不完整，训练的模型可能数据没有出现以偏概全或偏向。

公正，公平是主观认知（基于法律，业务等），例如，根据法律或公司政策，不同性别的工资不应该有显著的差别，以保证公正、公平。 因此，建立一个以工资为目标，其他如年龄、学历、工龄和性别等为特征的模型，需要检测模型在性别方面是否有公平（Fairness）。 公正、公平的内容需要根据业务需求明确。IBM 相关产品（如 OpenScale ）提供公平的检测能力。

二、人工智能可解释性案例分析

■ 案例分析一：AI 可解释性在银行贷款业务中的应用

1、背景

随着机器学习使用的不断普及，有时会被用来支持银行信用贷款审批流程，即针对用户贷款申请，通过机器学习模型来预测申请是被接受还是被拒绝。我们使用来自 FICO 可解释机器学习挑战赛的数据来讲述该场景，同时针对该场景中不同用户期望的解释来说明 AI Explainability 360 Toolkit （AIX360）的使用。此场景中涉及的三种类型的用户是：数据科学家，他在部署之前评估机器学习模型；信贷员，根据模型的输出做出最终决定；以及银行客户，他想了解申请结果的原因。

对于数据科学家来说，他更期望从模型的整体上来理解模型的推断过程，而不是某个具体的贷款申请。信贷员是最终决定用户申请批准与否的人，他们期望理解机器学习模型推断的具体原理，以此来做错正确且理由充分的审批。银行客户作为贷款申请人，他们期望知道申请被通过和拒绝的原因，特别是在被拒绝的情况下。

2、数据说明

FICO 挑战赛数据集包含有关真实房主提出的房屋净值信贷额度 (Home Equity Line of Credit, HELOC) 申请的匿名信息。我们正在考虑的机器学习任务是使用申请人信用报告中的信息来预测他们是否会在两年内及时付款。然后可以使用机器学习预测来决定房主是否有资格获得信贷额度。

下表列出了训练样本的主要特征，包括预测变量和目标变量。例如，NumSatisfactoryTrades 是一个预测变量，它计算过去与申请人签订的信用协议的数量，这些协议导致按时付款。要预测的目标变量是一个称为 RiskPerformance 的二元变量。 “差”值表示申请人在信用账户开立后的 24 个月内至少逾期 90 天或更糟一次。值“良好”表示他们已付款，逾期未超过 90 天。预测变量和目标之间的关系为表中的最后一列。如果预测变量相对于坏的概率 = 1 单调递减，则意味着随着变量值的增加，贷款申请为“坏”的概率降低，即变得更“好”。例如，External Risk Estimate 和 Num Satisfactory Trades 显示为单调递减。单调递增则相反。

特征	含义	单调性约束（对“坏”结果的影响）
ExternalRiskEstimate	综合风险标记	单调递减
MSinceOldestTradeOpen	最早账目的时长（以月为单位）	单调递减
MSinceMostRecentTradeOpen	最新账目的时长（以月为单位）	单调递减
AverageMInFile	账目的平均时长（以月为单位）	单调递减
NumSatisfactoryTrades	合规账目数量	单调递减
NumTrades60Ever2DerogPubRec	拖欠超过 60 天以上的账目数量	单调递减
NumTrades90Ever2DerogPubRec	拖欠超过 90 天以上的账目数量	单调递减
PercentTradesNeverDelq	未拖欠账目占比	单调递减
MSinceMostRecentDelq	最近一次拖欠账目距今的月数	单调递减
MaxDelq2PublicRecLast12M	过去 12 个月内最差拖欠分数	取值为 0-7 时单调递减
MaxDelqEver	最差拖欠分数	取值为 2-8 时单调递减
NumTotalTrades	总账目数量	无约束
NumTradesOpeninLast12M	过去 12 月账目数量	单调递增
PercentInstallTrades	分期付款账目占比	无约束
MSinceMostRecentInqexcl7days	距离 7 天前最近一次信用查询的月数	单调递减
NumInqLast6M	近 6 月信用查询次数	单调递增
NumInqLast6Mexcl7days	近 6 月信用查询次数（不包含最近七天）	单调递增
NetFractionRevolvingBurden	循环债务余额占信用额度的百分比	单调递增
NetFractionInstallBurden	分期付款债务余额占原始贷款金额的百分比	单调递增
NumRevolvingTradesWBalance	含余额循环债务账目数量	无约束
NumInstallTradesWBalance	含余额分期付款债务账目数量	无约束
NumBank2NatlTradesWHighUtilization	高利用率账目数量	单调递增
PercentTradesWBalance	含余额债务账目比例	无约束
RiskPerformance	风险表现	目标

3、数据科学家

在评估用于部署的机器学习模型时，理想情况下，数据科学家希望了解模型的整体行为，而不仅仅是在特定情况下的行为。在可能需要更高标准的可解释性的银行业等受监管行业尤其如此。数据科学家可能必须将模型呈现给：1）技术和业务经理在部署前进行审查，2）贷款专家将模型与专家的

知识进行比较，或 3) 监管机构检查合规性。此外，将模型部署在与其训练的地理区域不同的地理区域是很常见的。在部署之前，模型的全局视图可能会帮助发现过度拟合和对其他地区的泛化能力差的问题。

	8960	8403	1949	4886	4998
ExternalRiskEstimate	64.0	57.0	59.0	65.0	65.0
MsinceOldestTradeOpen	175.0	47.0	168.0	228.0	117.0
MsinceMostRecentTradeOpen	6.0	9.0	3.0	5.0	7.0
AverageMInFile	97.0	35.0	38.0	69.0	48.0
NumSatisfactoryTrades	29.0	5.0	21.0	24.0	7.0
NumTrades60Ever2DerogPubRec	9.0	1.0	0.0	3.0	1.0
NumTrades90Ever2DerogPubRec	9.0	0.0	0.0	2.0	1.0
PercentTradesNeverDelq	63.0	50.0	100.0	85.0	78.0
MSinceMostRecentDelq	2.0	16.0	NaN	3.0	36.0
MaxDelq2PublicRecLast12M	4.0	6.0	7.0	0.0	6.0
MaxDelqEver	4.0	5.0	8.0	2.0	4.0
NumTotalTrades	41.0	10.0	21.0	27.0	9.0
NumTradesOpeninLast12M	1.0	1.0	12.0	1.0	2.0
PercentInstallTrades	63.0	30.0	38.0	31.0	56.0
MSinceMostRecentInqexcl7days	0.0	0.0	0.0	7.0	7.0
NumInqLast6M	1.0	2.0	1.0	0.0	0.0
NumInqLast6Mexcl7days	1.0	2.0	1.0	0.0	0.0
NetFractionRevolvingBurden	16.0	66.0	85.0	13.0	54.0
NetFractionInstallBurden	94.0	70.0	90.0	66.0	69.0
NumRevolvingTradesWBalance	1.0	2.0	10.0	3.0	2.0
NumInstallTradesWBalance	1.0	2.0	5.0	2.0	3.0
NumBank2NatlTradesWHighUtilization	NaN	0.0	4.0	0.0	1.0
PercentTradesWBalance	50.0	57.0	94.0	46.0	83.0

可直接解释的模型可以提供这样的全局理解，它们具有足够简单的形式，因此它们的工作模式是透明的。下面我们通过 AIX360 提供的基于 Boolean Rule (BR)的 Boolean Rule Column Generation (BRCG)算法构建可直接解释的模型。

为了让 BRCG 可以更好的处理数据，可以将训练数据特征中的某些特殊值（如负数）转化为 NaN，而不是使用 0 或平均值代替。

同时，BRCG 要求对数据做二值化处理，我们使用 9 个分位数阈值的默认值来二值化序数（包括连续值）特征，包含各个判断条件。以上表所示的

5 个申请样本中的特征 ExternalRiskEstimate 为例,样本 8960 的值为 64，条件“<=”下，59，63 为 0，其他大的值则为 1，条件“>”下，59，63 为 1，其他大的值则为 0,“==”NaN 为 0，否者为 1。

	<=									>									==	!=
value	59	63	66	69	72	75	78	82	86	59	63	66	69	72	75	78	82	86	NaN	NaN
8960	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
8403	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1949	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
4886	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
4998	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1

BRCG 算法旨在产生一个非常简单的 OR-of-ANDs 规则（更正式地称为析取范式，DNF）或一个 AND-of-ORs 规则（合取范式，CNF）来预测一个 申请人将按时偿还贷款（Y = 1）。 对于我们这里的二元分类问题，DNF 规则等效于规则集，其中 DNF 中的 AND 子句对应于规则集中的单个规则。 此外，可以证明 Y = 1 的 CNF 规则等效于 Y = 0 的 DNF 规则。

对于 HELOC 数据集，我们发现 Y = 1 的 CNF 规则（即 Y = 0 的 DNF，通过设置 CNF=True 启用）略好于 Y = 1 的 DNF 规则。训练，验证模型之后，可以输出该模型生成的规则。

```
Training accuracy: 0.719573146021883
Test accuracy: 0.696515397082658
Predict Y=0 if ANY of the following rules are satisfied, otherwise Y=1:
['ExternalRiskEstimate <= 75.00 AND NumSatisfactoryTrades <= 17.00',
'ExternalRiskEstimate <= 72.00 AND NumSatisfactoryTrades > 17.00']
```

如上所示，Y = 0 时返回的 DNF 规则确实非常简单，只有两个子句，每个子句都涉及相同的两个特征。 有趣的是，这样的规则已经可以达到 69.7%的准确率。 External Risk Estimate 是一些风险标记的合并版本（越高越好），而 NumSatisfactoryTrades 是合规信用账户的数量。 因此，对于拥

有超过 17 个合规账户的申请人来说，External Risk Estimate 对于预测好 (Y = 1) 和坏 (Y = 0) 的影响比具有较少合规账户的申请人略低（更宽松）。

4、信贷员

通过选取原型或类似用户申请，可以为银行员工（如信贷员）可能感兴趣的有问题的申请生成解释，这有助于信贷员了解与当前申请具备类似背景的训练样本是被接受或拒绝。

AIX360 提供的 ProtodashExplainer 可以用来选取原型。Protodash 算法将一个数据点（或一组数据点）作为输入，根据属于同一特征空间的训练集中的实例来解释该数据点。然后，该方法尝试最小化我们想要解释的数据点与它将选择的训练集中预先指定数量的实例之间的最大平均差异（MMD 度量）。换句话说，它将尝试选择与我们要解释的数据点具有相同分布的训练实例。该方法使用贪婪算法进行选择并具有质量保证，同时可得到选取的样本的权重，以此表明它们的相似程度。

该方法从训练数据集中选择在不同方面于要解释的贷款申请类似的申请。例如，一个用户的贷款申请可能因为合规账目数量与另一个用户申请一样低，或者因为债务与另一个用户申请一样高而被拒绝。任意一个原因单独来说都足够用来拒绝申请，并且该方法能够通过选定的原型来揭示各种此类原因。而使用欧氏距离、余弦相似度等指标的标准最近邻技术并非如此。因此，Protodash 能够提供更全面和全面的观点，说明为什么针对待解释的贷款申请的决定是合理的。

如下表所示，Protodash Explainer 在训练集中选取与申请 S0 最相似

的 5 个样本，并返回表示相似程度的权重。

	S0	S1	S2	S3	S4	S5
ExternalRiskEstimate	82	85	89	77	83	73
MSinceOldestTradeOpen	280	223	379	338	789	230
MSinceMostRecentTradeOpen	13	13	156	2	6	5
AverageMInFile	102	87	257	109	102	89
NumSatisfactoryTrades	22	23	3	16	41	61
NumTrades60Ever2DerogPubRec	0	0	0	2	0	0
NumTrades90Ever2DerogPubRec	0	0	0	2	0	0
PercentTradesNeverDelq	91	91	100	90	100	100
MSinceMostRecentDelq	26	26	0	65	0	0
MaxDelq2PublicRecLast12M	6	6	7	6	7	6
MaxDelqEver	6	6	8	2	8	7
NumTotalTrades	23	26	3	21	41	37
NumTradesOpeninLast12M	0	0	0	1	1	3
PercentInstallTrades	9	9	33	14	17	18
MSinceMostRecentInqexcl7days	0	1	0	0	0	0
NumInqLast6M	0	1	0	1	1	2
NumInqLast6Mexcl7days	0	1	0	1	0	2
NetFractionRevolvingBurden	3	4	0	2	1	59
NetFractionInstallBurden	0	0	0	0	0	72
NumRevolvingTradesWBalance	4	4	0	1	3	9
NumInstallTradesWBalance	1	1	0	1	0	1
NumBank2NatlTradesWHighUtilization	1	0	0	0	1	7
PercentTradesWBalance	42	50	0	22	23	53
RiskPerformance	Good	Good	Good	Good	Good	Good
Weight		0.7302	0.0690	0.0978	0.0498	0.0530

5、银行客户

通常，申请人想了解为什么他们没有资格获得信用额度，他们的申请中的哪些变化将使他们有资格获得贷款。另一方面，如果他们符合条件，他们可能想知道是哪些因素导致他们的申请获得批准。在这种情况下，对比解释（contrastive explanations）算法可以向申请人提供关于他们的申请资料的哪些最小变化会改变 AI 模型的决定的信息（pertinent negatives），从

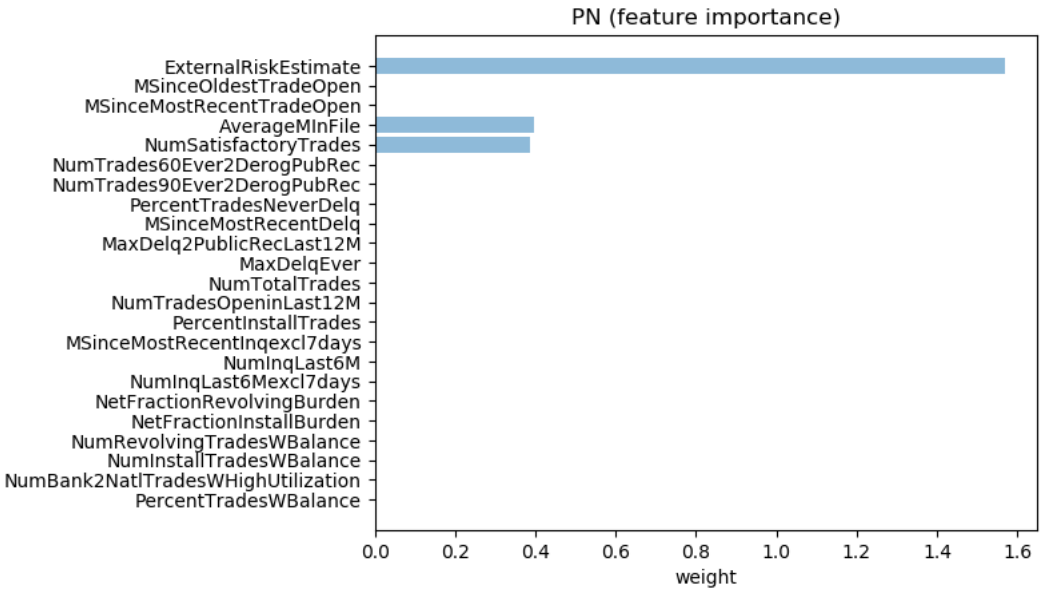
拒绝到接受或从接受到拒绝。例如，对于被拒绝的申请，保持其他不变，将合规账目数量增加到某个值可能会导致申请被接受。同时对比解释还可以从贷款申请信息中选出部分特征和取值以维持当前决策不变（pertinent positives)。例如，对于被接受的申请，即使将合规账目数量减少到较低的值，申请仍然可以通过。

更进一步来说，对比解释算法输出由两部分组成：a) 相关否定(pertinent negatives，PN) 和 b) 相关肯定 (pertinent positives，PP)。 PNs 识别一组最小的特征，如果改变这些特征将改变原始输入的分类。该方法实现这一点的方式是通过优化预测概率损失的变化，同时强制执行弹性范数约束，从而使特征及其值的变化最小。而 PP 标识了足以产生原始输入分类的最小特征集及其值，这里也有一个弹性范数项，因此所需的信息量最小。

以输出相关否定为例，AIX360 提供的 CEMExplainer 解释器计算与当前申请接近但结果不同的贷款申请样本，通过微量修改少量特征以改变模型的预测结果。这将帮助最初拒绝贷款申请的用户说，确定如何让贷款申请被接受。如下表中的被拒绝的贷款申请 X，通过 CEMExplainer 计算可以得到相关否定实例 X_PN。我们观察到，如果综合风险标记评分从 65 增加到 81，账目的平均时长大约 66 个月，合规账目数量增加到略高于 21，该申请则会被接受。

	X	X_PN	(X_PN - X)
ExternalRiskEstimate	65.000000	80.860000	15.860000
MSinceOldestTradeOpen	256.000000	256.000000	0.000000
MSinceMostRecentTradeOpen	15.000000	15.000000	0.000000
AverageMInFile	52.000000	65.620000	13.620000
NumSatisfactoryTrades	17.000000	21.400000	4.400000
NumTrades60Ever2DerogPubRec	0.000000	0.000000	0.000000
NumTrades90Ever2DerogPubRec	0.000000	0.000000	0.000000
PercentTradesNeverDelq	100.000000	100.000000	0.000000
MSinceMostRecentDelq	0.000000	0.000000	0.000000
MaxDelq2PublicRecLast12M	7.000000	7.000000	0.000000
MaxDelqEver	8.000000	8.000000	0.000000
NumTotalTrades	19.000000	19.000000	0.000000
NumTradesOpeninLast12M	0.000000	0.000000	0.000000
PercentInstallTrades	29.000000	29.000000	0.000000
MSinceMostRecentInqexcl7days	2.000000	2.000000	0.000000
NumInqLast6M	5.000000	5.000000	0.000000
NumInqLast6Mexcl7days	5.000000	5.000000	0.000000
NetFractionRevolvingBurden	57.000000	57.000000	0.000000
NetFractionInstallBurden	79.000000	79.000000	0.000000
NumRevolvingTradesWBalance	2.000000	2.000000	0.000000
NumInstallTradesWBalance	4.000000	4.000000	0.000000
NumBank2NatlTradesWHighUtilization	2.000000	2.000000	0.000000
PercentTradesWBalance	60.000000	60.000000	0.000000
RiskPerformance	Bad	Good	NIL

利用相关否定实例和原始申请的差距，可以进一步得出各个特征变化对最终预测结果的影响程度。



6、可解释性辅助模型评估

在上述的贷款审批流程中，辅助信贷员审批的 BRCG 模型，使用测试数据集验证准确率为 69.6%，符合基本上线的需求，但信贷员无法直接信任一个黑盒模型做出的预测，即使该模型在测试数据集上准确率为 100%，信贷员期望理解模型的预测，而开发模型的数据科学家和模型最终作用于的银行客户也都希望了解模型做出预测的策略，也就是说出了常见的可自动计算的指标（如准确率、召回率等）之外，评估模型对于相关人员是否具备可解释性也至关重要。

上述案例中使用 BRCG 算法训练得到的模型，其决策规则只有简单易懂的两条，并且数据科学家可以快速地通过历史数据和常识来演绎决策过程；而针对模型对于新的测试样本的推断，信贷员使用的 Protodash 解释方法可以进一步验证模型推断结果是否符合历史数据的规律，否则即使模型准确率再高也难以接受；而针对模型预测结果直接作用的银行客户使用的对比解释算法，可以验证模型的推断是否经得起提问和推敲，是否符合银行客户的认知。

由此也可见，可解释性即是AI系统需要满足的要求，也同时可以作为一种工具帮助相关人员从不同角度评估AI系统的工作原理和预测结果。可解释意味着模型决策过程的透明，透明意味着可控和可信，也只有如此，AI系统才能最终落地解决实际的问题。

7、总结

本案例基于银行的业务需求（利用机器学习辅助银行信用贷款审批流程）和业务对象（数据科学家、信贷员、银行客户）对于可解释的不同要求，利用AIX360工具集构建可直接解释模型，并为模型的使用者信贷员和银行客户提供不同角度的解释策略。

■ 案例分析二：可解释人工智能在个人医疗支出预测问题的应用

1、背景介绍

保险公司或者雇主想知道投保人或者员工未来一年的个人医疗支出，因为他们需要支付这些人的医疗费用。案例选取了AIX360中的两种全局可解释模型LinRR和BRCG来做预测。Linear Rule Regression（LinRR） 是一种广义线性规则模型，它产生一系列“AND” 规则并学习这些规则的权重得到线性组合。Boolean Rule Column Generation（BRCG）模型只产生简单的“OR of AND”分类规则。LinRR模型兼顾了准确性和模型的可解释性，在这个案例中用来做个人医疗支出的回归预测。有时回归预测无法准确预测“异常”样本，所以采用BRCG做二分类模型，专门识别医疗支出高的个体。

2、数据集介绍

案例数据来自于 MEPS。医疗支出小组调查(MEPS)是对美国各地的家庭和个人，及其医疗提供者和雇主进行的大规模调查，是关于医疗保健和医疗保险的成本和使用的最完整的数据来源。预测变量包括人口统计学特征（如性别，年龄)，社会经济学特征（如受教育程度，收入)，个人填写的健康状况等。

LinRR 和 BRCG 需要对非二元特征(即特征只有两种取值,如性别特征）进行二值化。每个连续特征都会先计算出它的 9 个分位数，再将分位数作为阈值做二值化。LinRR 使用原始特征和二值化特征做为输入，而 BRCG 模型只使用二值化特征。

预测个人医疗支出本质上是一个难题，特别是在美国医疗保健系统中。首先输入数据有限，例如对预测很有帮助的历史索赔数据就没有被纳入到特征当中。其次，预测变量的统计分布也增加了该问题的困难，该分布属于长尾分布，长尾由高支出的个体组成。具体来说，该分步的平均值是中位数的五倍，标准差是平均值的三倍，而支出最高的人则高达数十万美元。

3、使用 LinRR 模型预测个人医疗支出

为方便比较，先使用一个常见的机器学习模型梯度提升树 GBRT 建立基线模型，并且使用与 LinRR 相同的二值化特征作为 GBDT 的输入。LinRR 生成了一个基于规则的特征的线性回归模型。GBDT 在测试集上的 R 平方为 0.141，LinRR 的 R 平方 0.144 略高于 GBRT。更重要的是，LinRR 模型是可直接解释的。线性回归模型中包含的基于规则的和有序的特征及其系数如表 1 所示。作为线性模型，特征重要性自然由系数给出，因此列表按系数大小递减的顺序排序（注意系数可以为正或负)。

Table 1 LinRR 排名前 10 的系数：

序号	规则	系数
1	PCS42 <= -1.00	-8058
2	PCS42 <= 31.52	6827.75
3	RTHLTH31 != 5 AND PREGNT31 != 1	-6614.27
4	STRKDX == 1	4842.36
5	ADHDADDX != 1 AND PREGNT31 != 1 COGLIM31 != 1 AND DFSEE42 != -1	-3974.52
6	AGE31X	-3937.74
7	DIABDX == 1	3812.48
8	PREGNT31 != 1 AND ACTLIM31 != 1	-3778.59
9	CANCERDX == 1	3624.82
10	REGION != 1 AND DFSEE42 != -1	-2677.43

可以看到 LinRR 包含三种特征：

（1）未二值化的有序特征，例如表 1 中的第四个特征 STRKDX == 1；

（2）只含有一个条件的规则特征，例如表 1 中的第一个特征 PCS42 <= -1.00；

（3）含有两个或更多条件的规则特征，例如表 1 中的第三个特征 RTHLTH31 != 5 AND PREGNT31 != 1。

类别 1 和类别 2 中的特征一次只涉及一个原始非二值化特征（例如 AGE31X、PCS42），而原始特征之间的相互作用都属于类别 3。

为了便于解释，AIX360 提供的工具可以画出单个特征对因变量 y 的贡献。这些可以与领域专家的知识进行比较，以识别预期的行为以及可能令人惊讶的行为。

图 1 选取了三个典型的变量来说明单个特征对因变量 y 的影响。PCS42 代表 MEPS 调查日期前 4 周内的身体健康状况。它是根据 12 个回答计算得出的分数，它为诸如活动受限、疼痛干扰工作和爬楼梯困难等项目分配了

更高的权重。较低的值表示较差的健康状况，该图显示了医疗成本的相应增加，尤其是与 31 岁以下值相关的高成本。RTHLTH31 代表自我报告的健康状况， 1-5 对应于“优秀”、“非常好”、“良好”、“一般”和“差”。该算法仅对“非常好”和“一般”给出了非零系数，尽管人们可能认为“极好”健康的人应该看到至少与“非常好”健康的人一样大的成本降低。另一方面，由于状态是自我报告的，“优秀”不一定比“非常好”更好。健康状况不佳的非零系数的缺失可能是由于其在数据中的频率较低。K6SUM42 是一种用于测量调查前 30 天内的非特定心理困扰的分数。较高的值表示较高的压力，LinRR 算法发现 K6SUM42 与个人医疗支出呈正相关。

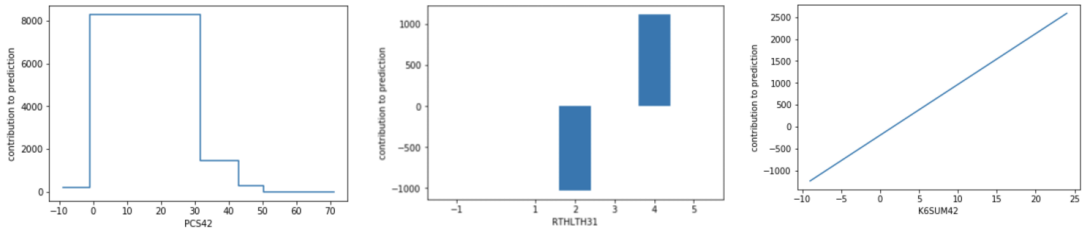


Figure 1 变量 PCS42，RTNLTH31，K6SUM42 与个人医疗支出的关系

当前吸烟状况变量 ADSMOK42)是需要进一步调查的反直觉发现的一个例子。2 表示不吸烟，但模型为其分配了对个人医疗支出的正贡献。这种关联的背后可能存在混淆。例如，吸烟者的平均年龄 (ADSMOK42 == 1) 为 44 岁，而非吸烟者的平均年龄为 49 岁，而老年人通常成本更高，因此模型认为不吸烟的群体（实际上年龄更大）医疗支出更大。

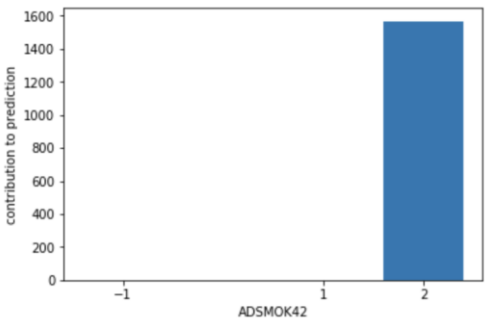


Figure 2 ADSMOKE 对个人医疗支出的影响

前面已经对模型中的线性项（特征类别 1）和一级规则（特征类别 2）进行了解释。现在考虑对类别 3 中的高阶规则进行解释。这些更高级别的规则自然更难以解释，并且需要更多的领域专业知识来做到这一点。表 2 只打印了 LinRR 模型的高阶特征系数，前三个规则可能是最简单的。当某些条件不存在时，它们会大大降低预测成本，共同因素是没有怀孕（PREGNT31 != 1）。如上所述，第一个规则 RTHLTH31 != 5 表示个人并未处于自我报告的“差”健康状态，则个人医疗成本低。第二个规则中，ADHDADDX 和 COGLIM31 分别指多动障碍和认知限制，这几个变量不等于某个值时，个人医疗支出低。在规则 4 中，REGION != 1 表示该个体不居住在东北人口普查区域，而 DFSEE42 != -1（也出现在规则 2 中）表示严重（甚至）戴眼镜看东西困难，规则 4 的系数为负值，这是一个与常识相悖的结论，无法做更进一步的解释。在规则 5 中，MARRY31X 的值 8 和 10 表明该个人在调查回合中丧偶或分居。同样不清楚为什么没有这些条件会导致更高的预测成本，但最后一个条件 PCS42 <= 50.22 确实有意义，因为它对应于身体健康状况不佳。在规则 6 和 8 中，INSCOV15 != 3 表示个人拥有健康保险，无论是公共的还是私人的。以此为条件，自我报告的健康状况低于“优秀”(RTHLTH31 != 1）和较差的身体健康状况（PCS42 <= 53.99）会导致更高的预测成本。最后在规则 7 中，PHQ242 是抑郁症的评分，数值越高，抑郁倾向越大。因此，PHQ242 != 5 表示没有高值，尽管不是最高值 6。 POVCAT15 != 5 表示个人收入不高（贫困线以上 400%），而 SOCLIM != -1 表示关于社会限制的问题是与常识一致的。

Table 2 LinRR 模型的高阶特征系数表

序号	规则	系数
1	RTHLTH31 != 5 AND PREGNT31 != 1	6614.27
2	ADHDADDX != 1 AND PREGNT31 != 1 AND COGLIM31 != 1 AND DFSEE42 != -1	3974.52
3	PREGNT31 != 1 AND ACTLIM31 != 1	3778.59
4	REGION != 1 AND DFSEE42 != -1	2677.43
5	AGE31X > 7.00 AND MARRY31X != 8 AND MARRY31X != 10 AND PCS42 <= 50.22	2211.48
6	RTHLTH31 != 1 AND INSCOV15 != 3	1640.62
7	SOCLIM31 != -1 AND PHQ242 != 5 AND POVCAT15 != 5	1565.76
8	PCS42 <= 53.99 AND INSCOV15 != 3	1277.63

4、使用 BRCG 分类模型识别高支出个体

为了演示布尔规则列生成 (BRCG) 算法，我们需要一个二分类任务，因为这是 BRCG 的设计目的。将医疗费用高支出定义为高于平均值（相对于中位数已经很高）的样本并相应地创建一个二值目标变量。输入特征与用于预测支出的特征相同。只有 21.5% 的人的成本高于平均值。

再次使用 GBDT 来建立基线模型，同时使用 BRCG 来执行相同的分类任务。 BRCG 生成了一组非常简单的规则（也称为 OR-of-ANDs 规则）来预测一个人的个人医疗支出是否高。 GBDT 在测试集上的准确率为 0.871，略高于 BRCG 的准确率 0.830。但 BRCG 模型的优势在于其简单性。BRCG 生成的模型为：若受教育程度为学士（EDRECODE == 15），并且受到工作、家务活学校活动的限制，则个人医疗成本高；若患有关节炎（ARTHTYPE != -1），身体机能受限（WLKLIM31 ！= 2），身体健康状况差（PCS42<=50.22），但有健康保险（INSCOV15 != 3）的个体个人医疗支出高；其他情况下个人医疗支出低。人们可能会推断出这两个群体之间的共同点是某种身体限制或健

康状况不佳，再加上收入（学士学位）或支付能力（保险范围）的代表。

5、LinRR 和 BRCG 可解释性对不同角色的意义

（1）数据科学家：LinRR 和 BRCG 这两种全局可解释的模型具有足够简单和透明的形式，可以从整体上理解模型的行为，而不仅仅是在特定实例中。它们不仅可以识别哪些特征最重要（如表 1 所示），还可以识别特征如何影响最终的结果（如图 1）。数据科学家可以将这些见解与医疗专家的领域知识进行比较，并以此决定是否调整模型。这种可解释性帮助数据科学家快速从业务的角度对模型进行改进，从而提升建模的效率。

（2）管理人员：作为保险公司或雇主方的管理人员，他们使用个人医疗支出预测模型进行审查。模型可解释性可以增加管理人员按预期执行的信心。此外，这些见解可以为干预措施提供信息，以降低成本，例如作为护理管理的一部分。

（3）个人：需要注意的是 LinRR 和 BRCG 并不适合作为针对个人诸如投保人或雇员的可解释工具。因为他们通常只关注自己的个人支出预测值为什么高以及应该怎样做出改变来改变自己的预测结果，而后者是 LinRR 和 BRCG 模型所不擅长的。

6、可解释性辅助模型评估

本案例展示了可直接解释的监督学习算法 LinRR 和 BRCG 能够生成准确且可解释的模型来预测医疗支出。LinRR 模型的精度高于无法解释的梯度提升树 GBDT，同时保留了线性模型的形式，并通过绘制各个特征与个人医疗支出的关系来增强模型的可解释性。BRCG 模型的准确性比 GBDT，但该模型仅包含两个易于理解的规则。我们相信，即便 BRCG 的准确率稍低，但如果

这种可直接解释的预测模型（不依赖于个别案例的事后解释）在与领域专家和下游决策者的人机协作中很有用，那么也会选择 BRCG 作为最终的模型。

■ 案例分析三：可解释人工智能在皮肤镜检查的应用

1、背景

皮肤镜检查是临床医学中的一个重要应用，具体过程为，医生使用皮肤镜获取的皮肤图像，来诊断包括皮肤癌在内的多种皮肤疾病。而深度神经网络的发展，使其能代替医生根据这些皮肤镜图像来判断皮肤疾病的种类。尽管某些深度神经网络模型的诊断能力甚至已经超过皮肤科专家，但这些模型却存在可解释性的问题。本案例使用 AIX360 中的 Disentangled Inferred Prior Variational Autoencoder (DIP-VAE)去捕获可解释的高维隐藏特征，进而帮助建立可信度高的机器学习模型。

2、数据集介绍

本案例识别 7 个种类的皮肤病，每个样本只属于以下某一种类别：

Table 3 皮肤病种类名称

英文种类名	中文种类名
Melanoma	黑素瘤
Melanocytic nevus	黑色素细胞痣
Basal cell carcinoma	基底细胞癌
Actinic keratosis / Bowen’s disease (intraepithelial carcinoma)	光化性角化病
Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis)	良性角化病
Dermatofibroma	皮肤纤维瘤
Vascular lesion	血管病变

3、利用 DIP-VAE 获取可解释的高维隐藏特征

利用DIP-VAE来进行解释性工作的基本流程为:将原始图片输入到DIP-VAE 编码器，经过编码可将原始图片转换为一组隐藏特征（也可叫隐藏表达 Latent Representation，比如一组 10 维的向量），然后再用 DIP-VAE 解码器将这组隐藏特征解码，解码可视为重建一张图片。

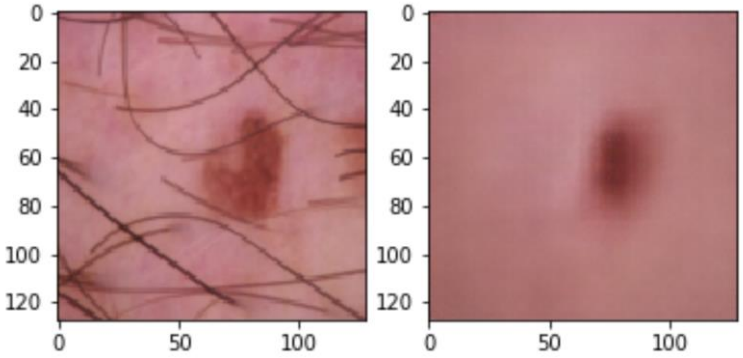


Figure 3 DIP-VAE 重建图与原始图对比

（左侧为原始图片，右侧为经过编解码后 DIP-VAE 重建的图片）

这里的关键在于，人们如何理解经过 DIP-VAE 编码得到的隐藏特征。显然，无法直接确定这里的隐藏特征（高维数组，而非图片，无法直接观察）

具体对应于图片的哪些实际特征（比如病患处的面积，是否对称，边界是否清晰等等）。

但是，可采用控制变量法，在一组隐藏特征中只改变其中一维，其他维固定，然后使用 DIP-VAE 解码器重建图片，观察图片的变化，理论上可根据该变化推理出被改变的维度对应的实际图像特征。如图 4 所示：每一行都只改变 10 维隐藏特征中的某一维特征，然后重建得到对应的图片。观察后不难发现，改变第 5 维（1:5）隐藏特征，会显著影响重建图片中病患处的直径，由此可推理出第 5 维隐藏特征对应于实际图片中病患处的直径大小信息。同理，还可观察出第 0 维，第 2 维和第 6 维隐藏特征对应的是实际图片中的边界信息，而第 1 维隐藏特征对应的是实际图片中的是否对称信息。由此，我们得到了可解释的高维隐藏特征。

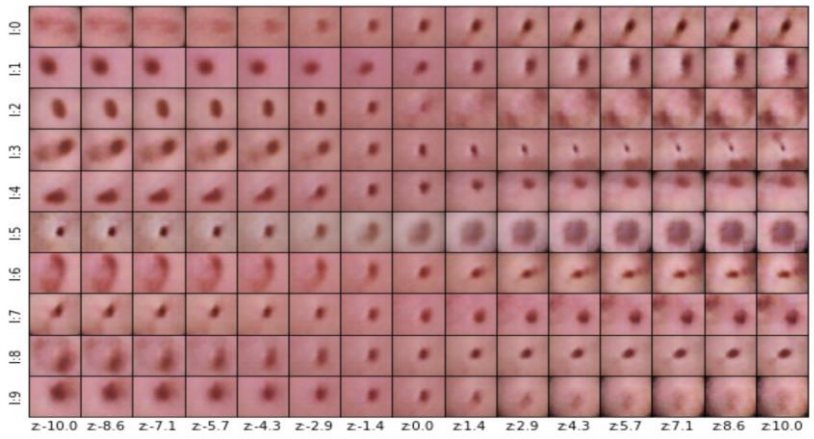


Figure 4 采用控制变量法得到可解释的高维隐藏特征

4、不同种类皮肤病下隐藏特征的分布

为了探索上述 10 维隐藏特征是否含有针对不同种类疾病的歧视性信息（这里的歧视性信息是指，某些特征是否对诊断某些疾病特别重要，而对其他种类影响不大），先使用 DIP-VAE 编码器得到所有原始图片（带有种类标签）的 10 维隐藏特征，然后按照种类，分别计算每个种类下所有样本每一

维隐藏特征的平均值和标准差等，如图 5 所示。

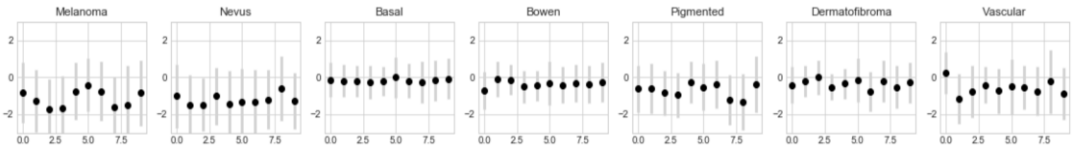


Figure 5 不同种类皮肤病下隐藏特征的分布

观察可得，不同种类的隐藏特征具有不同的模式。例如，从各个图中的第 0 维隐藏特征（即各图中的第一个黑点，从上文可知第 0 维隐藏特征对应的是边界信息），可以看出种类 Melanoma 和种类 Nevus 对第 0 维隐藏特征比较活跃，说明这两个种类对病患处的边界信息很敏感。然而，种类 Basel 和种类 Vascular 却对边界信息并不敏感（对应第 0 维黑点接近 0 值），即凭借边界信息很难对这两个种类进行诊断。

5、基于 DIP-VAE 得到的隐藏特征建立机器学习模型

将从原始图片中得到的 10 维隐藏特征（带种类标签）作为输入数据，分别使用两种机器学习模型进行预测皮肤病：

a)随机森林模型

预测的准确率为 69%，而在同一数据集上使用深度神经网络的最高准确率目前为 88%。使用 DIP-VAE 解码器重建图片进行观察，如图 6 所示。最后两个种类为空是因为模型没有预测出任何属于这两个种类的样本，从隐藏特征的分布上我们可以看到随机森林模型对各个种类下各隐藏特征的重要性反映

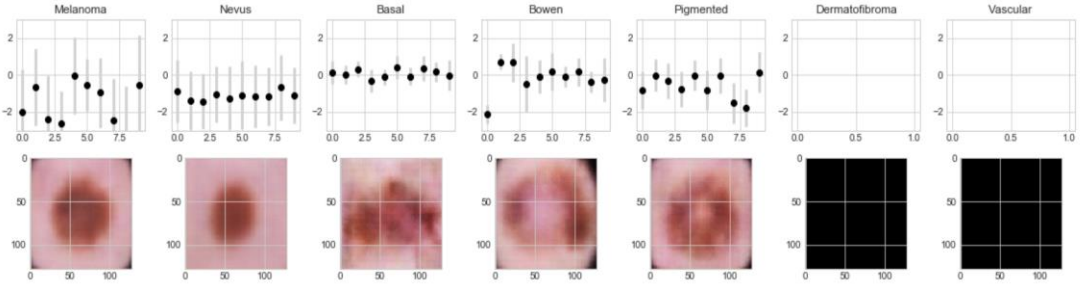


Figure 6 随机森林模型结果

b)逻辑回归分类器

预测的准确率为 65%,尽管比随机森林模型略低,依然属于不错的精度。类似地，重复 1 中的过程可得到图 7。

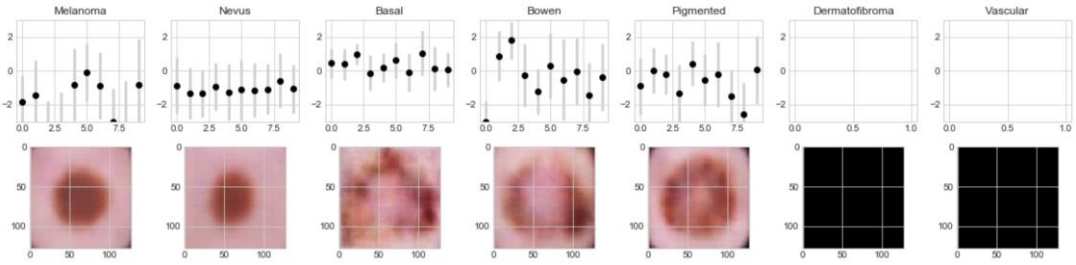


Figure 7 逻辑回归模型结果

此外，由于逻辑回归分类器还能输出它将一个样本判定为某个类别的可能性，因此我们按照预测结果的可能性大小，把被判定的各个种类再分成 7 个子集，最后使用 DIP-VAE 解码器可视化，如图 8 所示。这里我们关注那些可能性高的组，更容易看出不同类别的疾病对哪些隐藏特征更敏感。因为可能性越高的组，隐藏特征越稀疏，重要性更加明显。例如，随着可能性的增大，对黑素瘤和黑色素细胞痣来说，边界属性（对应第 0 维隐藏特征）变得愈发重要。

6、DIP-VAE 可解释性对不同角色的意义

在皮肤镜检查这一场景下（实际上其他根据病理图片诊断疾病的场景也类似），可解释性人工智能有着重要的应用意义。以建立在可解释隐藏特征（由 DIP-VAE 获得）的简单机器学习模型（简单机器学习模型本身的可解释性很高）为例，可解释性对该过程中的三个参与角色具有如下的意义：

（1）数据科学家：对于数据科学家来说，从整体上理解模型做出推理的过程是非常重要的。由于从 DIP-VAE 获得的模型输入是可解释的隐藏特征，而后采用的简单机器学习模型（如逻辑回归）的推理过程如果也可解释，

那么整个推理过程是透明的。

（2）专业医生：对于医生来说，了解到模型是根据病理图片的哪些特征做出的疾病诊断，可帮助医生依据其专业知识去评估模型的诊断结果是否合理与可信。例如，已知医学知识表明，疾病 A 的病理表现主要为病患处的面积和边界是否清晰，而模型做出疾病 A 的判断却主要根据病患处的颜色，那么该模型的推理逻辑明显不合理，其诊断结果不可信。因此，专业医生对可解释的人工智能模型可做出相对准确的可信度评估，这意味着通过评估的模型的诊断结果具有很高的可信度，可作为最终诊断结论的重要甚至主要参考。

（3）患者：对于患者来说，准确的诊断结果，是整个治疗过程的基础，而经过专业医生评估的可解释性机器学习模型，其诊断结果的可信度是有保证的，这对患者至关重要。

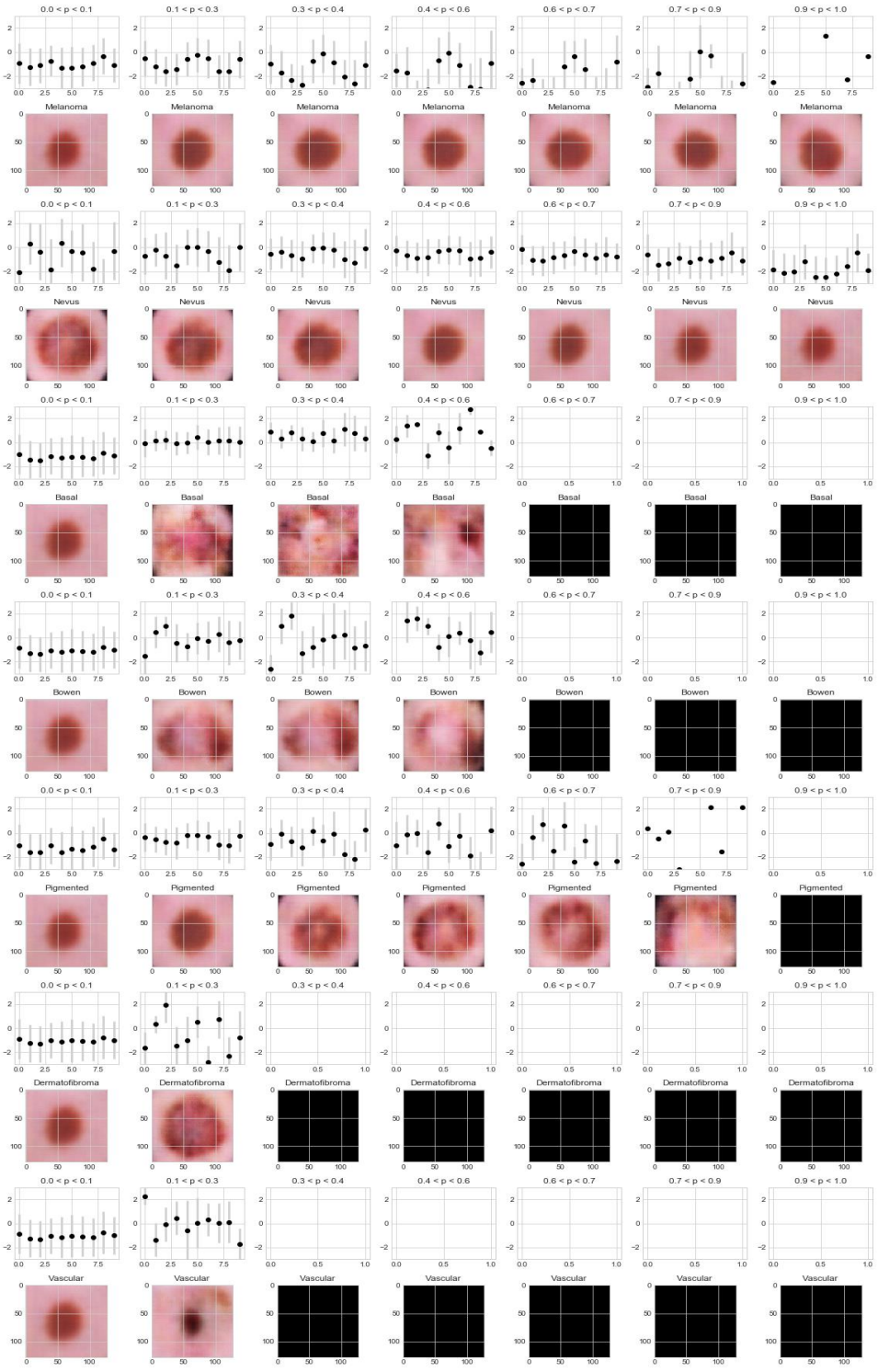


Figure 8

7、可解释性辅助模型评估

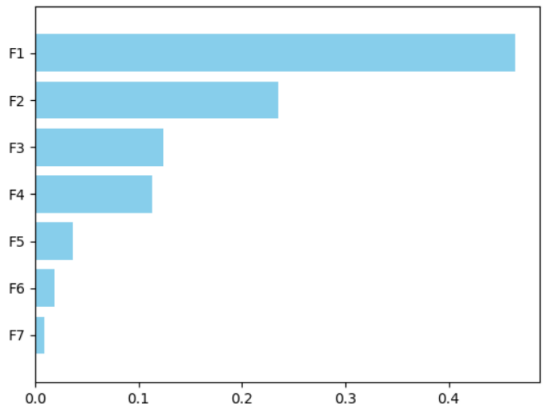
从本案例提到的两个简单机器学习模型来看，随机森林模型（精度 69%）和逻辑回归分类器（精度 65%），它们的精度略低于目前采用深度神经网络的最高精度（88%）。但这里有两点需要说明，首先这两个简单机器学习模型

的输入，是由 DIP-VAE 提取的 10 维可解释的隐藏特征，由于存在手动设定的 10 维维度并不能完全表达所有重要的隐藏特征的可能性，所以如果增大设定的维数，有机会进一步提高机器学习模型的精度，或者采用其他的可解释机器学习模型，也可能获得更高精度。此外，即使按照现在的模型精度，虽然深度神经网络模型的精度更高，但是由于其推理过程的不可解释性，很难对其推理过程进行评估，在实际应用中受阻；而可解释性的机器学习模型虽然精度仍有上升空间，但由于推理过程可解释，容易建立可信度，进而被接受乃至应用。

三、可信 AI 前景展望

1、其他解释性技术

IBM 除了 OpenScale，还有其他产品如 SPSS Modeler， SPSS Statistics 也涉及模型解释，例如，当模型的结构特别复杂，或者其结构很难解释时， 从特征与目标之间的关系来理解和解释模型，从宏观上看多个特征与目标之间的关系，有助于理性模型，对模型有宏观的，整体的认知。宏观的认知包括：**特征重要性**（Feature Importance）是在模型众多的特征中，计算出每一个模型的重要度值。 从这些值的排序中可以看到哪些些特征重要，哪些特征不太重要。典型的特征重要度如下所示



（横轴是重要度的值， 纵轴是各个特征，最上面的重要度最高的，往下依次降低）

还有一些特征相关的新功能正在研究开发中。

2、构建可信 AI 的原则及技术和产品

在第 4 集详细介绍的 AI 可解释性开源技术基础上，IBM Cloud Pak for Data 组件之一 OpenScale 作为一个在集成开源项目的商业产品，提供了更多，使用更方便，对于用户更友好的可信 AI 功能，包括：

（1）公平性检测与模型修正

AI 模型的结果是从训练数据中学习到的，当测试准确度达到要求的指标够，首先说明模型是准确的，完成了从数据中学习规律的任务， 是基于提供的数据是 “可信的”。 但训练数据可能并不完整，训练的模型可能出现以偏概全或偏向。

公正、公平是主观认知（基于法律和业务等），例如，根据法律或公司政策，不同性别的工资不应该有显著的差别，以保证公正、公平。因此，一个以工资为目标， 其他如年龄，学历，工龄，性别等为特征的模型，需要检测模型在性别方面是否有公平（Fairness）。

公平性检测就是要检测模型是否在某个特征上有明显的偏向。当检测

出模型有公平方面有问题后，提供修正模型的能力。

公正，公平的内容通常需要根据业务需求明确。

（2）监控模型使用与模型修正

监控 AI 模型的使用，通过了解有效数据和反馈数据，对已部署的模型采取行动以

确保业务应用程序中的模型持续有效运行；针对生产数据的模型使用的结果进行评估，提供 KPI 阈值和触发器来的智能得重新训练模型；监控模型在生产数据（而非训练数据）上的模型准确性和一致性的漂移。

当模型建成之后，应用于生产环境，一段时间后，往往发生"模型漂移"。所谓模型"模型漂移"是指在一段时间后，模型的预测精度与刚刚创建时相比， 发生了显著的下降，即变得不准确了。一般原因是：

- 用于预测的数据特征发生了改变，所以基于旧的数据创建的模型，不太适用于新的数据
- 目标内涵发生了改变

一旦发生了"模型漂移"，就需要检测，达到一定程度，就需要重新创建模型。持续监控模型，对于预测结果进行持续评估，变得十分重要，特别是当监控结果达到设定的条件是，自动触发后续动作，例如模型重建，是模型在生产环境必不可少的一部分。

IBM Open Scale 提供了模型监控，结果评估和后续动作触发等对应的功能，实现云环境，大量模型的模型自动监控功能。OpenScale 在模型创建好之后，把模型和创建模型的数据的特征导入云环境中，然后为模型目标配置 OpenScale 的订阅，开始监视模型运行。模型的每一次运行都会被记录，

并分析运行的结果和预定义的指标对比，以检测是否发生"模型漂移"并触发相应的动作。OpenScale 支持多个模型的同时监控。

在一个银行及金融行业监控模型准确性和数据一致性的案例里，客户使用 OpenScale 持续监控模型。模型中包含大量的特征，当模型漂移发生后，OpenScale 侦测了出导致模型偏移的特征，并区分出导致准确性偏移和数据一致性偏移的特征。这些特征的数据漂移（即该数据生产环境的特征与创建模型时的特征不同不一致）导致模型漂移。根据这些特征的取值，进一步找出来导致模型漂移的交易，并区分出导致准确率偏移的交易，导致数据一致性偏移的交易，以及导致准确性和数据一致性同时偏移的交易。为修正模型提供了准确的依据。

3、用简单的，结构清晰的模型来解释复杂模型

使用简单模型解释复杂模型的预测产生的结果，解释各个特征的产生的效果和贡献度， 并用图表展示。消除黑盒模型和允许业务用户以他们理解的方式理解 AI 结果。

4、生命周期管理

将 AI 模型度量指标集成到与业务和应用程序结果联系起来的通用报告工具中，实现 AI 模型生命周期编排框架化，实现 AI 和 IT 运营规模

在第 4 集我们详细介绍了 Linux 基金会 Data & AI 提出了构建可信任 AI 系统的 8 个原则[4]，除了本文详细研究的**可解释性**，以及上面提到的**公平**

[4]https://lfaidata.foundation/blog/2021/02/08/lf-ai-data-announces-principles-for-trusted-ai/

性外，还有**隐秘性，安全性，健壮性，可重现性，负责性**，和透**明性**。这些原则相互依赖和影响，共同作用以构建可信任的 AI 系统。

在 AI 系统构建和使用过程中，保证被训练数据，算法，推到数据这些重要数据和资产的**隐秘性**的**安全性**至关重要。IBM Security Cloud Pak for Security 提供完整的威胁管理能力，跨混合多云环境，获得威胁的可视性，主动分析异常行为，防止数据泄露，以及对核心资产（模型）的攻击和恶意访问行为。

基于零信任的思想，需要从访问控制、数据保护和威胁管理三个方面进行控制。无论是前台的数据科学家，还是后台的系统管理员，都需要通过严格的身份验证，处理提供用户名和口令以外，还需要引入 MFA 多因子认证，此外还需要对访问环境进行验证，如：非常规时间、异常地点、新的访问设备等，进行动态的策略验证。对于后台的特权用户，应用 Just in Time, Just Enough Privilege 等方式限制访问能力，不提供永久特权。访问全程需要记录操作过程，用于后期审计和调查取证。以上这些能力是 IBM Security Verify 提供的。

在数据测，最好的安全控制应该靠近被保护的数据资产，在数据全生命周期，提供不同的数据保护手段，包括但不限于：数据加密、数据活动监控、数据防泄漏等，利用安全策略和机器学习算法发现数据违规访问和操作活动异常。IBM Security Guardium 提供数据全生命周期保护能力。

四、结语

可信任是 AI 落地的基础。目前有很多可信任 AI 的学术文章和方法。IBM 也联合 Linux 基金会等开源社区在可信 AI 工具和原则做了一些卓有成

效的探索。但是目前实施的案例并不多。本文归纳总结了可解释 AI 的实用方法，并分析了可解释性在三个案例里面的应用，希望对国内的 AI 应用有所启发。

IBM 在 OpenScale，SPSS modeler 和 SPSS Statistic (Cloud Pak for Data 的组件)，以及 IBM Security Verify 和 IBM Security Guardium（Cloud Pak for Security 的组件）产品中扩产了开源技术的能力，这些产品的组合能为企业在混合云的环境下提供强大的可信 AI 能力。

但可信任是一个发展迅速的领域，需要持续的投入和努力，在此也欢迎更多有志于构建可信任的 AI 系统的人和组织能加入到这项工作中来。

806，面向自动驾驶应用的基于视觉的大规模 3D 语义映射

QingCheng 等，2022.3.2

美国加州初创公司（ArtisenseGmbH），慕尼黑工业大学，卡尔斯鲁厄应用技术大学

论文地址：https://arxiv.org/pdf/2203.01087v1.pdf

论文内容：想要实现自动驾驶中的全自动，其中一个巨大挑战就是三维感知，也就是如何使用车辆上传感器来认知三维世界。过去使用大量不同种类的传感器，现在则趋向使用更少、更便宜的传感器，轻量级和可伸缩的在线映射管道变得更受欢迎。

本文提出了一个完全基于立体相机系统的 3D 语义映射的管道，旨在构建可伸缩的和实时更新的地图。该管道包括一个直接稀疏视觉里程计算前端，和一个包括 GNSS 集成和语义 3D 点云标记并用于全局优化的后端。并提出了

一种简单而有效的时间投票方案，提高了三维点标签的质量和一致性。在KITTI-360数据集上对管道进行定性和定量评估。实验结果表明，本文提出的投票方案是有效的，并且该管道能够高效地进行大规模三维语义映射。通过展示一个覆盖8000公里道路的超大比例尺语义地图，进一步证明了管道的大比例尺映射能力，这种大规模语义地图可以作为完全矢量化高清地图的中间结果。此外，本文提出研究的下一步，是通过结合最先进的密集重建方法，在比例尺上构建语义3D体积图。

807，探索和指导解释性交互式机器学习的类型学

Felix Friedrich等，2022.3.10

德国汉森（Hessian）人工智能中心

论文地址：https://arxiv.org/pdf/2203.03668v1.pdf

内容：最近，越来越多的解释性交互式机器学习（XIL）方法被提出，其目的是通过集成人类用户对模型解释的监督来扩展模型的学习过程。这些方法通常是独立开发的，提供不同的动机，并源于不同的应用。值得注意的是，到目前为止，还没有对这些作品进行全面评估。通过确定一组通用的基本模块，并对这些模块进行深入讨论，我们的工作首次将各种方法统一为一种类型学。因此，这种类型学可用于根据已识别的模块对现有和未来的XIL方法进行分类。此外，我们的工作还调查了六种现有的XIL方法。除了对这些方法修改模型的总体能力进行基准测试外，我们还对错误原因修改、交互效率、反馈质量的稳健性以及修改严重损坏模型的能力进行了额外的基准测试。除了引入这些新的基准测试任务外，为了改进定量评估，我们还引入了一个

新的错误原因（ wrnospace）度量标准，用于测量模型解释中的平均错误原因激活，以补充定性检查。在我们的评估中，所有方法都证明能够成功地修改模型。然而，我们发现各个基准任务的方法之间存在显著差异，揭示了有价值的应用相关方面，不仅有助于比较当前的方法，也有助于激发在未来的 XIL 方法开发中纳入这些基准的必要性。

808，代表真相解释的数据（用于评估 XAI 方法）

Shideh Amiri 等，国际先进人工智能协会，2020.11.18

论文地址：https://www.aminer.cn/pub/5fb79c2e91e01122f29d69c9?conf=aaai2021

论文内容：目前，可解释性人工智能（XAI）方法的评估方法主要来自可解释性机器学习（IML）研究，该方法侧重于理解模型，例如与现有归因方法的比较，敏感性分析，特征的金集，公理或通过图像演示。这些方法存在一些问题，例如它们没有指出当前的 XAI 方法无法指导研究走向该领域的持续进展。它们无法衡量支持可靠决策的准确性，而且几乎不可能确定一种 XAI 方法是否优于另一种方法或现有模型的缺点，从而使研究人员无法就哪些研究问题将推动该领域发展提供指导。其他领域通常利用真实数据并创建基准。XAI 或 IML 中通常不使用表示真实解释的数据。一个原因是，在满足一个用户的解释可能不满足另一个用户的意义上，解释是主观的。为了克服这些问题，我们建议用标准方程式表示解释，这些方程式可用于评估 XAI 方法的准确性。本文的贡献包括创建代表真实解释的综合数据的方法，三个数据集，使用这些数据集对 LIME 的评估以及使用这些数据评估现有 XAI 方法所面临的挑战和潜在收益的初步分析。基于以人为本的研究的评估方法

不在本文讨论范围之内。

809，用于可解释的少机会学习的元决策树

Baoquan，哈尔滨工业大学,2022.3.7

论文地址：http://arxiv.org/pdf/2203.01482v1

内容：在本文中，作者们通过提出一种新颖的基于决策树的元学习框架，即 MetaDT，旨在向可解释的 FSL 迈出了一步，使用元学习的可解释决策树替换现有表示学习方法的最后一个黑盒 FSL 分类器。遇到的关键挑战是如何有效地学习决策树（即树结构和每个参数

节点）在 FSL 设置中。为了应对这一挑战，引入了一个树状的类层次结构作为先验：1）层次结构直接用作树结构； 2）通过重新将类层次结构视为无向图，设计了一个基于图卷积的决策树推理网络作为元学习器来学习推断每个节点的参数。最后，在框架中加入了一个双循环优化机制用几个例子快速适应决策树。文章中，为了展示 MetaDT 的决策可解释性，作者们在 miniImagenet 上进行了两个 5-way 1-shot 案例研究，包括正确和错误的决策案例，即从测试集中随机选择一个 5-way 1-shot 任务。然后只使用一个标记样本来构建和学习一个四层决策树。之后，随机选择一个预测正确的图像和一个错误预测的图像。性能比较和可解释性分析的大量实验表明了 MetaDT 的有效性和优越性。

810，基于 STDP 的尖峰神经网络监督学习算法

Zhanhao Hu 等，清华大学，2022.3.7

论文地址：https://arxiv.org/abs/2203.03379

内容：与基于速率的人工神经网络相比，SNN 提供了更具生物学合理性的模型为了大脑。但他们如何进行监督学习仍然是个谜。作者提出了一种基于尖峰时间依赖可塑性（STDP）的有监督学习算法，用于分层 SNN 由漏泄整合和激发（LIF）神经元组成。A.时间窗是为突触前神经元设计的，只有棘波在此窗口中，参与 STDP 更新过程。模型是在 MNIST 数据集上训练。分类精度的方法是具有类似结构的多层感知器（MLP）的标准的反向传播算法。作者描述了一种基于 STDP 的 SNN 监督学习算法，并得到了相应的结果在 MNIST 分类任务中取得了良好的结果。准确度接近具有类似体系结构的 MLP，这表明了该方法的有效性算法。与现有的 SNN 训练算法相比，本文提出了一种新的 SNN 训练算法算法已经取得了相互竞争的结果。该算法表明，生物神经元可能不会改变它们的突触一直在 STDP 规则下。STDP 仅在监管机构批准时生效信号被应用。此外，该算法表明，并不是所有的峰值突触前神经元参与突触的 STDP 学习过程。可能存在一个时间窗口，只有道琼斯指数在这场胜利中的峰值才会被计算在内。但需要生化证据来验证这些预言。

811，类神经拟态计算的替代梯度

BenjaminCramer 等，2022.1.14

德国海海德堡大学，瑞士弗雷德里克.米歇尔生物医学研究所

论文地址： https://www.pnas.org/doi/pdf/10.1073/pnas.2109194119

内容：为了以低代谢成本快速处理时间信息，生物神经元以模拟和的形式整合为输入，但及时与脉冲、二进制事件通信。类神经拟态硬件使用相同的原

理来模拟具有出色能效的脉冲神经网络。然而，由于设备不匹配和缺乏有效的训练算法，在此类硬件上实例化高性能脉冲网络仍然是一项重大挑战。替代梯度学习已成为一种有前景的脉冲网络训练策略，但尚未证明其对类神经拟态系统的适用性。在这里，我们使用循环方法演示了 BrainScaleS-2 类神经拟态系统的替代梯度学习。我们证明，学习可以自我纠正设备不匹配，导致在视觉和语音基准上具有竞争力的脉冲网络性能。我们的网络显示平均每个隐藏神经元和输入不到一个脉冲的稀疏脉冲活动，以高达每秒 85,000 帧的速率执行推理，消耗不到 200 mW。总之，我们的工作为类神经拟态硬件上的低能量脉冲网络处理设定了几个基准，并为未来的片上学习算法铺平了道路。

812，可池化分层图表示学习（语义网络，知识图谱）

Zhitao Ying 等，2022.3.10

美国斯坦福大学，加拿大麦吉尔大学，南加州大学

论文地址：

https://proceedings.neurips.cc/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7- Paper.

内容： 图神经网络 (GNN) 通过有效学习节点嵌入，彻底改变了图表示学习领域，并在节点分类和 链接预测等任务中取得了最先进的成果。然而，当前的 GNN 方法本质上是扁平的，并且不 学习图的层次表示——这一限制对于图分类任务尤其成问题，其目标是预测与整个图关联的 标签。本文提出了 DiffPool，这是一个可微的图池化模块，它可以生成图的层次表示，并且 可以以端到端的方式与各种图神经网络架构相结合。 DiffPool 为深度 GNN 的每一层的节 点学习可微分软集群分配，将节点映射到一组集群，

然后形成下一个 GNN 层的粗化输入。 我们的实验结果表明，与所有现有的池化方法相比，将现有的 GNN 方法与 DiffPool 相结 合，在图分类基准上的准确率平均提高了 5-10%，在五分之四的基准数据集上实现了新的最新技术。

813，基于动态卷积网络知识图谱补全模型

Haoliangpeng 等，2022.3.3,上海大学计算机科学和工程学院

论文地址：https://www.mdpi.com/2078-2489/13/3/133/htm

内容：知识图嵌入可以学习知识图实体和关系的低秩向量表示，一直是知识图补全的主要研究课题。最近的几项工作表明，基于卷积神经网络 (CNN) 的模型可以捕获头部和关系嵌入之间的交互，因此在知识图完成方面表现良好。然而，之前的卷积网络模型忽略了不同交互特征对实验结果的不同贡献。在本文中，我们提出了一种新的嵌入模型，名为 DyConvNE，用于知识库补全。我们的模型 DyConvNE 使用动态卷积核，因为动态卷积核可以为交互特征分配不同重要性的权重。我们还提出了一种新的负采样方法，它挖掘硬负样本作为额外的负样本进行训练。我们对数据集 WN18RR 和 FB15k-237 进行了实验，结果表明我们的方法在知识图谱补全方面优于其他几种基准算法。此外，我们在预测 WN18RR 和 FB15k-237 的 Hits@1 值时使用了一种新的测试方法，称为特定关系测试。与不使用此方法的模型相比，此方法在 Hits@1 方面的相对改进约为 2%。

814，通用人工智能技术（AGI）的认识过程论探析

陈凡、吴怡，2021.11，东北大学科学技术哲学研究中心

摘要：人工智能技术的发展被人们寄以从“专用”到“通用”的期盼，然而其实现的瓶颈在于循规蹈矩的程序恐难以真正理解符号背后的意义，由此表现为无法实现迁移学习的能力。马克思主义认为，人的认识是由感性上升至理性的过程。当前智能机器所做出的智能决策虽与人类由理性做出的决策结果相吻合，但因其缺少感性过渡与升华的阶段，故而与人类智能两异。从马克思主义对于人类认识发展过程的科学阐释来看，人工智能技术从专用转向通用，其间的感性要素是必要不充分条件。

往常传统的人工智能路径总是急于将人类的功能结果作为机器模拟的直接来源，抑或是将人类的神经元结构做以静态的拷贝。它们都是出于功能结果和片面结构的模拟，而无一种方式是完全按照人类认识的发展过程的规律（即按照感性认发展至理性认识这一顺序）进行模拟的，皆忽视了感性认识在智能机器设计中的基础性地位。马克思主义对于人类认识发展规律的科学解释，可从哲学的视角为通用人工智能技术提供一种全新的思路。虽然对于机器而言，这是个极为复杂繁琐的过程，但是我们没有捷径，应该遵循事物发展的客观规律。人类如此，机器亦如此。

815，关于人工智能的刑法思考（现实／未来）

赵秉志等，2019.1，北师大刑法科研院

摘要：新一代人工智能技术的发展及其应用将为人类世界提供新的发展机遇，在此过程中需要高度重视伴随而来的安全挑战。立足于人工智能的发展现状与未来趋势，刑法应当树立正确、理性的态度应对其蕴藏的刑事风险。

专用人工智能与通用人工智能存在本质差别，刑法在现阶段更应注重对围绕专用人工智能产生的刑事风险进行防控，对未来可能出现的通用人工智能保持克制态度。专用人工智能最具变革性的影响，在于推进网络空间与现实空间实现全面而深入的交融。专用人工智能的内部风险极有可能是围绕海量数据与智能算法而产生的，其外部风险则表现为对人工智能技术的滥用，以及由内部风险转化为现实危害的安全威胁。对此，刑法应坚守罪刑法定的总体底线和谦抑的内在品质，加强对数据安全的保护力度，实现对专用人工智能从设计到使用的全方位调控。此外，从刑法的哲学依据、价值观念、具体认定、适用效果等方面来看，将通用智能机器人作为与自然人并列的刑事责任主体存在较大难度，刑法对此应坚持谨慎克制的态度。

从刑法学角度的五个方面的质疑：通用智能机器人不具有相对的自由意志、通用智能机器人实施的“犯罪”难以与刑法学中的行为理论相兼容、承认通用智能机器人的主体地位会面临一系列判断难题、技术水平的突破并不等于价值观念的逾越、对通用智能机器人适用刑罚的合理性和必要性尚存疑问。

816，基于自适应脑电通道选择和变换的高效脑解码

JiaxingWang 等，2022.3.8，中科院自动化研究所复杂系统管理与控制国家重点实验室

论文地址：https://ieeexplore.ieee.org/document/9730833

内容：基于脑电图 (EEG) 的脑机接口 (BCI) 在神经康复和运动辅助方面具有广泛的应用。然而，从大量脑电图通道中获取的大脑活动与大脑解码任

务高度相关或无关，从而降低了解码效率和准确性。如何根据不同的试验自适应地选择最佳通道数仍然是一个很大的挑战。为了解决这个问题，本研究提出了一种高效的端到端大脑解码模型 AdaEEGNet。它可以通过自适应控制输入通道的数量来降低计算成本，并通过减少过拟合来提高分类精度。具体来说，我们设计了一个轻量级策略模块对当前需要那些通道解码脑电图进行分析。由于通道选择过程是不可微分的，我们建议使用 Gumbel-Estimator 反向传播梯度来训练整个框架。此外，我们还设计了在大脑解码准确性和效率之间进行权衡的权重系数。为了验证提议的 AdaEEGNet 在提高解码效率和准确性方面的可行性，在 BCI 竞赛 IV 数据集上进行了广泛的实验。结果表明，与基线方法相比，我们的方法可以将解码精度提高 2%，而计算成本仅为 65%。为了验证提议的 AdaEEGNet 在提高解码效率和准确性方面的可行性，在 BCI 竞赛 IV 数据集上进行了广泛的实验。结果表明，与基线方法相比，我们的方法可以将解码精度提高 2%，而其计算成本仅为 65%。

817，一个开放存取的基于脑电图的脑机接口数据集，用于内部语音识别

Nicolas Nieto 等，2022.2.14,信号、系统和计算智能研究所，滨海应用数学研究所，恩特雷里奥斯大学控制论实验室，布宜诺斯艾利斯大学应用人工实验室

论文地址：https://www.nature.com/articles/s41597-022-01147-2

内容：表面脑电图是测量脑电活动的一种标准且非侵入性式方法。人工智能的最新进展显著改进了大脑模式的自动检测，允许越来越快的、越来越可靠

的和可访问的脑机接口。不同的范式已被用于实现人机交互，并且在过去几年中，人们对解释和表征“内心声音”现象的兴趣有了广泛的显著增长。这种被称为内部语音的范式提高了通过思考执行命令的可能性，允许以“自然”的方式控制外部设备。不幸的是，缺乏公开可用的脑电图数据集，限制了内部语音识别新技术的发展。本文展示了和其他两个相关范例下获得的十人数据集，记录了 136 个通道的采集系统。这项工作的主要目的是为科学界提供一个可用于更好地理解相关大脑机制的内部语音命令的开放式多类脑电图数据库。

818，无人机多目标搜索的不确定性

Morstsinay，2022.3.18，Bar-llan 大学 Kraus 计算机科学部，以色列

论文地址：https://arxiv.org/pdf/2203.09476v1.pdf

内容：本文研究了一组无人机在不确定性条件下搜索目标的复杂问题。无人机团队的目标是在到达选定目标之前尽快找到所有移动目标。考虑的不确定性有三个方面：首先，无人机不知道目标的位置和目的地。其次，无人机的感知能力并不完美。第三，目标的运动模型未知。我们提出了一个无人机实时算法框架，结合熵和随机时间信念，旨在优化快速成功检测所有目标的概率。我们对该算法框架进行了实证评估，并与其他解决方案相比，显示了其效率和显著的性能改进。此外，我们还使用对等设计代理（PDA）评估了我们的框架，这些代理是模拟目标的计算机代理，并表明我们的算法框架在这种情况下优于其他解决方案。

819，通过高效的B样条路径构造加速基于深度神经网络的局部车辆调度规划

PiotrKicki 等，2020.3.14 波兹南工业大学

论文地址：https://arxiv.org/pdf/2203.06963v1.pdf

内容：尽管人们对自动驾驶汽车进行了深入的研究，但对这些车辆的运动规划的研究主要集中在管理交通场景和规则，而较少关注在拥挤的市中心停车、进入购物中心车库或避免与另一辆车发生意外碰撞所需的局部动作规划。当前规划算法难以在很短的时间（通常是几秒钟）内生成可行的路径，以避免在危险情况下发生碰撞，并满足类似汽车的约束条件。

本文提出一种使用 B 样条曲线来有效地表示规划路径方法，通过利用神经网络的诱导偏差构造过程,加速基于DNN的车辆运动规划器的推理和训练。作者在之前工作的基础上，利用 DNN 体系结构从过去的经验中学习车辆操控运动的规则，并引入了一种新的B样条路径构造方法，使得在几乎恒定的时间（约11 ms）内生成局部操纵动作成为可能，同时考虑了环境地图和车辆运动学所施加的一些约束。作者提出的神经网络架构和训练程序，允许运动规划器从自己的经验中学习。神经网络的训练采用深度强化学习框架，并使用基于梯度的策略搜索，以提高训练系统的性能。

在实验中，使用最新的Bench MR框架对新路径规划器进行了全面评估，该方法在测试集上实现了90.1%的精度，优于其他规划算法。最后作者也探讨了未来的一些工作，如通过预测当地地图的时间序列来直接考虑动态环境，以及使用路径生成器作为运动规划器的初始猜测，这有助于在可接受的时间内实现更高的精度。

820，基于特征的解释的努力：沙普利值与最小足够子集

Oana-MariaCamburu，2020.9.23 牛津大学

论文地址：https://www.aminer.cn/pub/5f6c6cf891e0119671e85880?conf=aaai2021

内容： 为了使神经模型赢得广泛的公众信任并确保公平，我们必须对其预测做出可理解的解释。近来，越来越多的作品致力于根据输入特征的相关性来解释神经模型的预测。在这项工作中，我们证明了基于特征的解释即使是解释琐碎的模型也带来了问题。我们表明，在某些情况下，至少存在两个基于事实的基于事实的解释，并且有时，它们都不足以提供对模型决策过程的完整了解。此外，我们显示出两种流行的解释器类别，分别是 Shapley 解释器和最少的足够子集的解释器，尽管显然隐含了以下假设：解释器应寻找一种基于特征的解释，但它们针对的是根本不同的地面真理解释。这些发现为开发和选择解释器带来了一个额外的考虑因素。

821，基于可解释神经网络的无监督关键词提取

Rishabh，2022.3.16 卡内基梅隆大学语言技术研究所，太平洋西北国家实验室，华盛顿大学

论文地址：https://arxiv.org/pdf/2203.07640v1.pdf

内容：关键词提取旨在自动提取文档中关键概念的“重要”短语列表。无监督关键短语提取的先前方法是通过嵌入相似性或图中心性来获得短语重要性的启发式概念，需要广泛的领域专业知识来开发它们。作者们的工作提出了另一种操作定义：基于预测文本最有用的短语是重要的关键短语的原因，为此，作者们建议 INSPECT——一个自我解释的神经框架，通过测量输入短

语对主题分类下游任务的预测影响来识别关键短语。文章表明，这种新颖的方法不仅减轻了对 ad-hoc 启发式的需求，而且在两个领域的四个不同数据集的无监督关键词提取方面取得了最先进的结果，并且实验也表明 INSPECT 是可推广的，可以使用现成的主题标签轻松适应新领域。最后，作者的研究提出了可解释神经网络作为 NLP 系统中固有组件的新用途，而不仅仅是作为向人类解释模型预测的工具。

822，自适应尖峰神经网络的精确平均场模型

LiangChen 等，2022.3.16，滑铁卢大学

论文地址：https://arxiv.org/pdf/2203.08341.pdf

内容：具有适应性的棘突神经元网络已被证明能够重现广泛的神经活动，包括突发群体爆发和棘突同步，这是大脑紊乱和正常功能的基础。从尖峰神经网络导出的精确平均场模型非常有价值，因为这样的模型可以用来确定单个神经元和网络参数如何相互作用，从而产生宏观网络行为。在本文中，作者推导并分析了具有尖峰频率自适应的神经网络的一组精确平均场方程。具体来说，作者的模型是一个伊兹克维奇神经元网络，其中每个神经元由一个二维系统建模，该系统由一个二次积分和火灾方程加上一个实现尖峰频率自适应的方程组成。之前的工作是为这种类型的网络推导平均场模型，依赖于适应变量足够慢的动力学假设。然而，这种近似并没有成功地在宏观描述和现实神经网络之间建立精确的对应关系，尤其是在适应时间常数不大的情况下。挑战在于如何通过包含自适应变量的平均场表达式来实现一组封闭的平均场方程。作者通过使用洛伦兹-安萨兹结合力矩闭合方法来解决

这一挑战，从而得到热力学极限下的平均场系统。由此产生的宏观描述能够定性和定量地描述神经网络的集体动力学，包括紧张性放电和爆发之间的转换。

823，节能随机游走计算的神经拟态缩放优势

J.Darby Smith 等，2022.2.14，美国桑迪国家实验室，神经探索与研究实验室

论文地址： https://www.nature.com/articles/s41928-021-00705-7.pdf

内容：神经拟态计算的目标是在合成硬件中复制大脑的计算结构和架构，它通常专注于人工智能应用。人们较少探讨的是，这种受大脑启发的硬件能否提供超越认知任务的价值。在这里，我们展示了脉冲神经拟态结构的高度并行性和可配置性，使其非常适合通过离散时间马尔可夫链实现随机游动。这些随机游动在蒙特卡罗方法中很有用，蒙特卡罗方法是解决广泛数值计算任务的基本计算工具。通过使用 IBM 的 TrueNorth 和 Intel 的 Loihi 神经形态计算平台，我们表明，与传统方法相比，用于生成扩散的随机游走近似的神经拟态计算算法在节能计算方面具有优势。我们还表明，我们的神经拟态计算算法可以扩展到更复杂的跳跃-扩散过程，这些过程在金融经济学、粒子物理和机器学习等一系列应用中都很有用。

824，用于异构 IIoT 数据集成的 ABGE 辅助制造知识图谱构建方法

LeiRen 等，北京航空航天大学，加拿大安蒂戈尼什圣弗朗斯泽维大学

论文地址：https://www.tandfonline.com/doi/full/10.1080/00207543.2022.2042416

内容：工业物联网 (IIoT) 为智能制造中新兴的数字服务化范式的发展奠定了基础。海量异构 IIOT 数据的深度融合，对实现制造业数字化服务具有关键作用。然而，不同制造领域之间存在知识鸿沟，这给工业大数据的高效整合和利用带来了挑战。为此，提出了制造知识图谱（FMKG）框架，用于从多源异构数据中提取行业知识三元组，以整合领域知识。此外，提出了一种基于注意力的图嵌入模型（ABGE）来发现和补充知识图谱中的隐式缺失关系，以获得完整的工业知识图谱。 ABGE 模型的有效性已在多个知识图谱数据集上得到验证。并以某航空航天企业生产过程为例，建立产品质量知识图谱，证明了该方法的可行性。

825，可池化分层图表示学习

ZhitaoYing 等，2022.3.10，美国斯坦福、加拿大麦吉尔、南加州大学

论文地址：

https://proceedings.neurips.cc/paper/2018/file/e77dbaf6759253c7c6d0efc5690369c7-Paper.pdf

内容：图神经网络 (GNN) 通过有效学习节点嵌入，彻底改变了图表示学习领域，并在节点分类和链接预测等任务中取得了最先进的成果。然而，当前的 GNN 方法本质上是扁平的，并且不学习图的层次表示——这一限制对于图分类任务尤其成问题，其目标是预测与整个图关联的标签。本文提出了 DiffPool，这是一个可微的图池化模块，它可以生成图的层次表示，并且可以以端到端的方式与各种图神经网络架构相结合。 DiffPool 为深度 GNN 的每一层的节点学习可微分软集群分配，将节点映射到一组集群，然后形成下一个 GNN 层的粗化输入。我们的实验结果表明，与所有现有的池化方法

相比，将现有的 GNN 方法与 DiffPool 相结合，在图分类基准上的准确率平均提高了 5-10%，在五分之四的基准数据集上实现了新的最新技术。

826，无开源不通用：通用人工智能机器生产工艺学批判

刘方喜，2022.3.15，中国社会科学院

摘要：技术工艺性应用的通用性,需要社会性应用的开源性与之匹配,在人工智能通用性不断提升的动态发展趋向中,构建与其匹配的开源性伦理规则,推动高效而无害的通用人工智能创造,具有重要意义。随着创造并使用物质和精神劳动工具活动的发展,人类智能的封闭性、非通用性被不断超越而社会性、开源性不断提升,现代科学和自动机器大大加速了这一进程。现代科学这种社会通用智能的工艺性应用,首先把物质劳动工具的使用技巧转移到能量自动化机器上而成为社会机械通用智能,超越了手工智能封闭于个体人身内的生物性的非开源性。当今人工智能正在使精神劳动工具的使用技巧也向机器转移,将进一步超越智能的个人生物性的非开源性,再进一步超越资本商业化社会性应用非生物性的非开源性,通用人工智能将成为高度自动化的社会机械通用智能。无开源不通用,构建公义创新动力机制,聚合非市场、非营利创新动力,人工智能将在通用性、自动性、开源性高度统一中充分发展并造福全人类。

827，通用人工智能需要在私人语言的层面上进行知识表征吗？——来自大森庄藏的启发

徐英瑾，2021.12.6，复旦大学哲学学院

摘要：通用人工智能语境中的私人语言,指的是这样一个意思:表征 A 在系统甲那里的知识表征方式与同一个表征在系统乙那里的表征方式必然会有所差异。因此,在预设推论主义语义学自身有效性的前提下,A 在甲中的意义集,总会有一个子集(无论这一子集有多小)仅仅为甲自身所拥有,而无法被任何一个别的系统所拥有。因此,任何一个与甲不同的别的系统,都无法彻底地理解甲对于 A 的意义把握方式。很显然,这样一种将机器表征与哲学史上的"第一人称哲学"传统相互结合的思路,是无法见容于后期维特根斯坦对于私人语言的著名反驳的。而为了与后期维特根斯坦论战,日本哲学家大森庄藏的思想资源便具有了很高的引用价值,因为他本身的哲学就可以被视为"维特根斯坦的话语方式与胡塞尔的思想内核"的日本式混合体。在对大森的哲学进行面向机器表征问题的重建的过程中,对于局域性原则与历史性原则的引入也是题中应有之义,以便为通用人工智能语境中建设私人语言的必要性提供辩护。而非公理性推理系统(纳思系统)所提供的技术手段,则会为这种想法的技术落地提供可能。

828，基于脑电图的脑机接口机器学习技术综述

Swati Aggarwal 等，2022.1.7，印度新德里内塔吉苏巴斯理大学

论文地址：https://link.springer.com/article/10.1007/s11831-021-09684-6

内容：脑机接口 (BCI) 框架使用计算机算法来检测心理活动模式并操纵外部设备。由于其简单性和非侵入性，最常用的成像技术之一是脑电图 (EEG)。用于评估基于 EEG 的 BCI 系统输出的评估方法是针对特定应用对 EEG 信号进行分类。人工智能技术的发展激发了研究人员使用机器学习 (ML)

技术和深度学习 (DL) 方法对基于 EEG 的 BCI 进行分类。机器学习技术使脑机接口能够在每个新会话中从受试者的大脑中学习，调整生成的规则以对思想进行分类，从而提高系统的效率。作者对基于 EEG 的 BCI 中各种 ML/DL 技术的使用进行了集中调查。使用了三种用于分类的 EEG 范例：运动想象、p300 和稳态诱发电位。此外，基于理想的信号处理方法、BCI 功能、性能评估和商业化，解决了最近基于 EEG 的 BCI 系统面临的挑战。作者希望收集到的信息有助于应用合适的机器学习技术，并为 BCI 研究人员增强未来的 BCI 系统提供基础。

829，使用卷积递归神经网络和跨数据集学习为帕金森病患者绘制基于 EEG 的情绪图表

Muhammad Najam Dar 等，2022.3.11， Islamabad 国家科技大学，Nanyang 技术大学等

论文地址：https://www.sciencedirect.com/science/article/pii/S0010482522001196

内容：基于脑电图（EEG）的情绪分类反映了真实和内在的情绪状态，可以帮助在娱乐消费行为、交互式脑机接口、患者心理健康监测等领域进行更可靠、更自然、更有意义的人机交互。实验环境和现实的不同和个体认知健康状况之间的差异给基于 EEG 的情绪识别的应用带来了极大的挑战。帕金森病(PD) 是第二常见的神经退行性疾病，其导致情绪识别和表达受损。 PD 患者情感表达的不足给医疗服务带来了挑战。本研究提出了 1D-CRNN-ELM 架构，它将一维卷积递归神经网络 (1D-CRNN) 与极限学习机 (ELM) 相结合，对 PD 患者的情绪检测具有较高的鲁棒性，也可用于各种跨数据集的实验环境。在所提出的框架中，经过 EEG 预处理后，经过训练的 CRNN 可以

用作特征提取器，ELM 作为分类器。这个训练好的 CRNN 可用于通过对其他数据集进行微调来学习新的情绪集。本文通过使用 PD 患者数据集进行训练并使用 AMIGOS 和 SEED-IV 的公开可用数据集进行微调来应用情绪的跨数据集学习。在 AMIGOS，PD，HC 数据集上进行六个基本情绪类别的分类准确度为 97.75 %，83.20%和 86.00%。在数据集之间进行交叉验证，AMIGOS 的平均准确率为 95.84%，PD 为 75.09%，HC 为 77.85%，SEED-IV 为 84.97%。仅来自 14 个通道的 1 秒 EEG 信号片段就足以检测情绪。我们所提出的方法在使用公开可用的数据集对基于 EEG 的情绪进行分类方面优于最先进的研究，能够跨数据集学习，并验证深度学习框架在帕金森病患者心理健康监测的实际应用中的鲁棒性。

830，训练还是不训练：说话人识别中的偏见缓解策略研究

Raghuveer Peri 等，2022.3.17，南加州大学

链接：https://arxiv.org/pdf/2203.09122.pdf

简介：说话人识别越来越多地用于一些日常应用，包括智能扬声器、客户服务中心和其他语音驱动分析。准确评估和缓解基于机器学习（ML）的语音技术（如说话人识别）中存在的偏见，以确保其广泛采用，至关重要。与人脸识别等其他以人为中心的应用相比，现代说话人识别系统中关于各种人口统计因素的 ML 公平性研究相对滞后。现有的说话人识别系统公平性研究大多局限于评估系统特定操作点的偏差，这可能导致对公平性的错误预期。此外，针对说话人识别系统开发的偏见缓解策略屈指可数。在本文中，我们系统地评估了说话人识别系统在一系列系统操作点中存在的性别偏见。我们

还提出了对抗式和多任务学习技术来提高这些系统的公平性。我们通过定量和定性评估表明，与使用数据平衡技术训练的基线方法相比，所提出的方法提高了 ASV 系统的公平性。我们还提出了一个公平-效用权衡分析，以联合检查公平性和整体系统性能。我们发现，尽管使用对抗性技术训练的系统可以提高公平性，但它们的效用往往会降低。另一方面，多任务方法可以在保持效用的同时提高公平性。这些发现可以为说话人识别领域的偏见缓解策略的选择提供参考。

831，带假设决策树的贪婪算法

Mohammad Azad 等，2022.3.16，朱夫大学，英特尔公司，香港大学，阿卜杜拉国王科技大学，卡托维兹广西里西亚大学

链接：https://arxiv.org/pdf/2203.08848.pdf

简介：我们研究了基于一个属性的传统查询和基于所有属性值假设的查询的 at 决策树。这种决策树类似于精确学习中研究的决策树，在精确学习中允许成员资格和等价查询。我们提出了基于不同不确定性度量的贪婪算法来构建上述决策树，并讨论了计算机对 UCI ML 存储库中各种数据集和随机生成的布尔函数的实验结果。我们还研究了由贪婪算法构造的决策树生成的决策规则的长度和覆盖率。

832，深度神经网络加速器硬件近似技术综述

Giorgos Armeniakos 等，2022.3.16,雅典国立技术大学，卡尔斯鲁厄理工学院

链接：https://arxiv.org/pdf/2203.08737.pdf

简介：深度神经网络（DNN）因其在机器学习（ML）中的各种认知任务中的高性能而非常流行。DNN 的最新进展已经在许多任务中超越了人类的准确性，但代价是计算复杂度很高。因此，为了有效地执行 DNN 推理，越来越多的研究工作利用了 DNN 固有的容错能力，并采用近似计算（AC）原理来解决 DNN 加速器不断增加的能量需求。本文对 DNN 加速器的硬件近似技术进行了全面的综述和分析。首先，我们分析了最新技术，通过确定近似族，我们根据近似类型对各自的作品进行了分类。接下来，我们分析所执行评估的复杂性（与数据集和 DNN 大小有关），以评估近似 DNN 加速器的效率、潜力和局限性。此外，还对更适合设计 DNN 加速器近似单元的误差指标以及针对 DNN 推理的精度恢复方法进行了广泛讨论。最后，我们将介绍 DNN 加速器的近似计算如何超越能源效率，并解决可靠性和安全性问题。

833，利用改进的高维神经网络为石墨烯基 2D-3D 界面开发势能面用于储能
Vidushi Sharma，Dibakar Datta，2022.3.12，新泽西理工学院
链接：https://arxiv.org/pdf/2203.08607.pdf
简介：设计新的异质结构电极有许多与界面工程相关的挑战。对模拟资源的需求和异质结构数据库的缺乏仍然是使用模拟来理解复杂界面的化学和力学的障碍。由二维（2D）和三维（3D）材料组成的混合三维异质结构由于其多变的性能，是无可争议的下一代工程器件材料。本文采用密度泛函理论（DFT）方法对二维石墨烯和三维锡（Sn）系统之间的界面进行了计算研究，并利用计算量大的模拟数据开发了基于机器学习（ML）的势能面（PES）。开发的 PES 可用于从 ML 模拟石墨烯-锡界面系统，具有接近 DFT 的精度。为

了开发 PES，高维神经网络（HDNN）依靠以原子为中心的对称函数来表示结构信息。对 HDNN 进行了修改，以训练界面系统的总能量，而不是原子能。在 5789 个石墨烯| Sn 界面结构上训练的改进 HDNN 的性能在同一材料对的不熟悉界面上进行测试，这些界面与训练数据集存在不同程度的结构偏差，包括新的 Sn 体相、缺陷和界面原子扩散。测试界面预测得到的最高 RMSE 为 0.458 eV/原子。结果表明，改进的 HDNN 方法在预测石墨烯和锡等复杂多晶界面的能量方面更为准确。基于 ML 的建模方法提高了精度，有望为异质结构储能系统中设计具有更高循环寿命和稳定性的接口提供经济高效的方法。

834，6G 无线网络机器学习算法综述

Anita Patil 等，2022.3.16，欧洲经委会，印度理工学院

链接：https://arxiv.org/pdf/2203.08429.pdf

简介：无线技术中人工智能/机器学习（AI/ML）集成的主要重点是减少资本支出，优化网络性能，并建立新的收入来源。用深度学习人工智能技术取代传统算法，极大地降低了功耗，提高了系统性能。此外，ML 算法的实现还使无线网络服务提供商能够（i）从适用于网络边缘的分布式 AI/ML 体系结构中提供高自动化水平，(ii)在接入网络上实现基于应用程序的流量控制，（iii）实现动态网络切片，以解决具有不同服务质量要求的不同场景，以及（iv）在各种 6G 通信平台上实现无处不在的连接。在本章中，我们回顾/综述了适用于 6G 无线网络的 ML 技术。并列出研究中需要及时解决的公开问题。

835，微型印刷电路机器学习分类的近似决策树

Konstantions Balaskas 等，2022.3.15,亚里士多德大学

链接：https://arxiv.org/pdf/2203.08011.pdf

简介：尽管印刷电子（PE）在传统的评估指标（如集成密度、面积和性能）上无法与硅基系统竞争，但 PE 提供了有吸引力的特性，如按需超低成本制造、灵活性和无毒性。因此，它针对的应用领域是基于光刻技术的硅电子产品无法触及的，因此还没有看到太多的计算扩散。然而，尽管 PE 具有吸引人的特性，但 PE 中较大的特征尺寸阻碍了复杂印刷电路的实现，例如机器学习（ML）分类器。在这项工作中，我们利用机器学习分类决策树的硬件友好性质，并利用近似设计的硬件效率，以生成适用于小型、超资源受限和电池供电的打印应用的近似 ML 分类器。

836，超越解释：基于 XAI 的模型改进的机遇和挑战

LeanderWeber 等，2022.3.15,弗劳恩霍夫.海因里希.赫兹研究所，新加坡理工学院，柏林学习和数控基础研究所，奥斯陆大学

链接：https://arxiv.org/pdf/2203.08008.pdf

简介：可解释人工智能（XAI）是一个新兴的研究领域，为高度复杂和不透明的机器学习（ML）模型带来了透明度。尽管近年来发展了多种方法来解释黑盒分类器的决策，但这些工具很少用于可视化以外的目的。直到最近，研究人员才开始在实践中使用解释来实际改进模型。本文全面概述了将 XAI 实际应用于改进 ML 模型各种属性的技术，并对这些方法进行了系统分类，比较了它们各自的优缺点。我们为这些方法提供了一个理论视角，并通过玩具

和现实环境的实验，以实证的方式展示了解释如何帮助改善模型泛化能力或推理等特性。我们进一步讨论了这些方法的潜在注意事项和缺点。我们得出结论，虽然基于 XAI 的模型改进即使在复杂且不容易量化的模型属性上也会产生显著的有益影响，但这些方法需要谨慎应用，因为它们的成功可能取决于多种因素，例如使用的模型和数据集，或使用的解释方法。

837，（脑机接口）联合神经成像检测嗜睡：综述和挑战

A-S-M-Sharifuzzaman 等，2022.2.27,世宗大学，孟加拉国美国国际大学，中国矿业大学

链接：https://arxiv.org/pdf/2202.13344.pdf

简介：脑-机接口（BCI）收集、分析大脑活动，并将其转换为指令，然后发送到检测系统。脑机接口在某些情况下，如基于注意力的任务中，在脑下活动中变得越来越流行。研究人员最近使用了 EEG+fNIRS 和 EEG+fMRI 等组合神经成像技术来解决许多现实问题。嗜睡检测或睡眠惯性是神经成像技术的核心研究领域之一。本文旨在研究基于脑机接口（BCI）的联合神经成像技术在嗜睡检测或睡眠惯性方面的最新应用。为此，这是基于神经成像的睡意检测系统的唯一综述。

敬请关注联盟微信公众号
COPU开源联盟

中国开源软件推进联盟秘书处
电话：+86 010-88558999
联盟公共邮箱：office@copu.org.cn
联盟官网：http://www.copu.org.cn
地址：北京市海淀区紫竹院路66号赛迪大厦18层