# 《可解释机器学习》

# 人工智能汇集

## 国内外人工智能跟帖留言

（共102条跟帖）

中国开源软件推进联盟
China OSS Promotion Union

# 人工智能点评：可解释性机器学习（共 102 条跟帖）

陆首群

2022.4.10

可解释性机器学习目前已成为全球由弱人工智能转向强人工智能的热点。

按大数据建立起来的机器学习／深度学习 AI 系统是一种强大的数据分析工具，它属于弱人工智能的范畴。但机器学习／深度学习是有缺陷的，它本质上是黑盒子（或称暗箱）技术，其模型是不可理解、不可解释的，只有打破黑盒子，实现可解释的机器学习，才能使之转化为强人工智能，解决更加复杂的应用场景中的问题。

2018 年，AI 大师 Yoshua Bengio、Yann Lecun、Geoffrey Hinton 指出，深度学习本质上是一项暗箱技术或盲模型，其训练过程不可解释、不可理解、不可控，缺乏类人的推理能力。2018 年人工智能大师 John-Hopcroft 满怀信心要在 5 年内打破深度学习这个黑盒子。

2020 年 6 月，COPU 主办的《第 15 届开源中国开源世界高峰论坛》邀请 IBM 副总裁 Todd Moore 与会作“可信任人工智能（反欺诈、可解释、公平性）”的报告（这是发表可解释性机器学习报告属于全球最早之一者）。

在国内，2020 年 12 月，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 年 1 月，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”。

COPU 今发表迄今为止收到的国内外关于可解释性机器学习的 102 条跟帖留言，如下：

## 美国

Judea Pearl

John Hopcroft

麻省理工学院

卡内基梅隆大学

加州大学伯克利分校

斯坦福大学

南加州大学

耶鲁大学

佐治亚理工学院

华盛顿大学

乔治理工学院

芝加哥大学

密歇根大学

德克萨斯大学

诺特丹大学

佛罗里达科技大学

罗格斯大学

Rowan 大学

谷歌

IBM

微软

Mozilla 基金会

## 中国

姚期智

沈向洋

张钹

清华大学

上海交大

哈尔滨工业大学

香港大学

香港浸会大学

COPU

阿里巴巴

## 加拿大

YoshuaBengio

GeoffreyHinton

魁北克 AI 研究所

蒙特利尔大学

多伦多大学

## 英国

阿兰图灵研究所

剑桥大学

牛津大学

Swansea 大学

## 法国

里昂大学

Bordeaux 大学

LaBRICrsdela 实验室

## 德国

柏林大学

康斯坦茨大学

汉森（Hessian）AI 中心

奥芬堡大学

国家 AI 研究中心

## 西班牙

格拉纳达大学

圣地亚哥-德孔波特斯拉大学

## 日本

京都大学

Hosei 大学

## 丹麦

丹麦科技大学

哥本哈根商学院

## 挪威

Agder 大学

奥斯陆大学

## 瑞典

斯德哥尔摩 KTH 皇家技术学院

## 葡萄牙

里斯本 NOVA 大学

## 奥地利

polten 大学

## 澳大利亚

澳洲国立大学

## 新加坡

南洋理工大学

## 俄罗斯

未来技术数学研究与教育中心

## 波兰

雅盖隆大学

## 巴基斯坦

国家 AI 中心

COPU 认为，由于全球机器学习可解释性技术(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在把握上可能有些差异，最后评估也只能依靠人工，所以 XAI 演算结果或算法可能有出入，致使可解释机器学习推广应用增加了难度，而且演算结果还有待精准，演算程序也有待简化，为此 COPU 要求 IBM 人工智能研究所对 XAI 列举具体案例，并对其演算程序进行解析和说明（COPU 向 IBM 提出 8 个问题，双方经过反复多次商讨，IBM 两度提出 XAI 具体案例的演算程序，可参阅第七集 575 条、第十集 765 条、第十一集 805 条）。

从各国研发可解释性人工智能看来，国外研发阵容强大，国内做这

方面研发工作的不多，COPU 在其中发挥了不小作用。本文提供大家参考。

# 国内外人工智能跟帖共 102 条

**1-74，**MIT 技术评论（Sep9.2016）：深度学习、人工智能的可解释性是由物理学家而不是数学家解释证明的。

**2-97，**深度学习算法是一种强大的数据分析工具，是实现人工智能的路径之一，但深度学习也是有缺陷的，它本质上是黑盒子技术，其训练过程难以解释、不可控，也未必能适应更加复杂的应用场景，深度学习过度依赖数据，而利用数据建模有时与真实生活之间也未必能直接划上等号，用深度学习模型训练有时也未必成功。当下深度学习可解决一些问题，但不少问题还不能靠它来解决，需靠大量基础理论研究创立新算法予以支持。

**3-137，**微软 AI 发展致力于研究“可解释的 AI”领域。微软亚洲研究院副院长张益肇针对 AI、AI+的应用 X：AI 技术不断发展重塑我们的生活和各行各业，并推动产业向数字化、智能化转变，通过与 AI 结合，金融、交通、医疗或为首批获利企业。

我们常说 ABC，A—强大的算法 Algoiithm，B—海量的数据 BigDate，C—大规模的计算 Compute。如何推动行业与 AI 结合？AI 赋能行业 ABCDE（五大关键因素），D—专业的领域 Domain，E—生态链 Ecosystem。ABC 三大因素是 AI 的基石，也是数字化转型的基础，产业数字化转型还需 D-专业领域和 E 生态链两大关键因素，D-专业领域指任何 AI 落地的场景，需行业专家一起参与（如在 A+医疗方向，微软很多应用是与辉瑞等医药企业一起共同完成的）。E-生态链，除落地的场景，AI 在行业的发展还需要一个完善的生态链。

4-163，可解释性和信任及AI伦理将受到更多关注（采访Element AI咨询和支持部门）2019年见证了AI道德规范和风险管理的早期原则的实施，2020年将是AI值得信赖的一年。可解释性的概念也越来越广为人知。

当然2019年人们对AI伦理的关注日益增强（2019年初欧委会公布了开发AI道德指南），10月由深度学习先驱者之一Yoshua Bengio与MoZilla基金会合作共同建立的Element AI，创建了数据信任关系并推行AI道德规范，微软和谷歌等大型科技公司采取相应措施，使他们的AI开发符合道德规范。

人们对AI道德规范的兴趣日益增长。

到2020年，企业都将关注AI信任，希望看到风投也关注，新的初创企业将为解决方案提供帮助。

5-198，为什么人工智能需要可解释性？

假如人工智能将来应用到实践中，那么人类必须知道计算机是如何想的？也就是说人类要知道机器的工作原理。

6-210，IBM Watson在探索医疗人工智能认知计算过程中就遇到了缺乏常识带来的困扰，在痛定思痛后，他们提出了具人（embodiment）概念，必须与现场临床医生沟通、取得共识、统一行动，以弥补研究人员常识之不足（将基础理论和常识结合起来），才能完成医疗防治的任务。

7-217，深度学习模型不可理解、不可解释

人工智能的发展要重点解决可解释性问题，人工智能研究的难点是对认知的解释

与建构，而认知研究的关键问题则是自主、创意和情感等所谓主客观现象的破解。利用大规模语义网络或 NLP 函数来训练、提高对任务的理解和可解释能力。

目前 NLP 的缺点是不理解常识，或计算机最大问题就是缺乏常识。做常识库难度太大！现在的问题是如何攻克这最近的 50 米！

8-237，建立可理解、可解释的人工智能模型

知识图谱的作业流程

1）从数据库和知识库中分别获取数据资源和知识资源，并传送至知识图谱平台。

2）在知识图谱平台上（在其上配置数据获取、知识表示、知识存储、本体构造、知识计算、测试评估等工具），对数据资源和知识资源进行测试、评估和分类，将分类的数据资源和知识资源分别传送至基于表格的存储体中和基于图结构的存储体中。

3）建立大规模的语义网络，提取在表/图存储结构体中的数据/和识资源，通过语义网络进行融合、转化（数据→知识），在资源融合、转化时要建立知识表示（确定基于符号/基于向量的知识表示）、知识推理和知识建模（不同意有些专家提出的分数据驱动、知识驱动两路并进形成多模态信息处理方式）。

4）将知识建模传送到行业应用（第三方应用知识库），通过大规模语义网络（NLP 的发展）建立知识图谱语义理解，并补充建立常识图谱和行业（领域）应用知识图谱。在建立知识图谱过程中，充分利用图挖掘和知识计算技术。

5）从而生成可理解、可解释的 AI 模型，把人工智能从数据驱动以深度学习技术为代表的感知智能阶段推向到知识驱动以知识图谱/语义网络技术为代丢的认知智能价段。

9-447，人工智能走向何方？

喜看三条出发路线，五位世界人工智能大师发评论。

人工智能与人脑越来越远还是近，也有大师发评论。

人工智能未来发展的第一目标是人类智能或接近人类智能。

目前有三条路，正在探索走向发展目标：

一、从深度神经网络（或机器学习/深度学习模型）出发

有人说，深度学习已近天花板，似乎很难往前发展了。他们说，深度学习是一个强大的数据分析工具，带动了当前人工智能的繁荣，但它本质上也是一项暗箱技术或盲模型，其训练过程不可解释、不可理解、不可控，缺乏类人的推理能力，与人类大脑的运作机制差距很大，难以逾越；也有人说，深度神经网络潜力很大，自监督学习（训练）可使深度学习达到或接近人类智力水平，这时出现了发展的转机。说这话的人还是世界人工智能大师，如 YoshuaBengio、YannLeCun、GeoffreyHinton 等，他们坦率地谈了未来深度学习人工智能的研究趋势，认为自我监督学习是一种机器学习/深度学习的“理想状态”，可使之产生类人的推动力，变不可解释、不可理解、不可控为可！去年人工智能算法大师 John-Hopcroft 更是信心满满要在 5 年内打破深度学习这个黑盒子，他说人类知道它在学习，但不知它怎么学习，我们会在 5 年内大体能读出深度学习的数学理论。

对于走这条路是否受限于天花板一直有争议，现在看来突破天花板发展有转机！

二、从异步脉冲神经网络出发

异步脉冲神经网络与人类大脑神经元网络在结构、特征、功能、机制等方面比较

相似(或力求相似),因此它在对人类大脑意识处理的探索上比其他路径有优势,但我们对异步脉冲神经网络的研究还处于初级阶段,欲达到人类智能或接近人类智能的目标,还有很长的路要走,还会遇到很多挑战:

在神经形态计算出现后我们必须把传统的冯-诺伊曼计算架构转移到神经形态计算(类脑计算)架构上来,把目前采用的人工智能加速技术(AI 芯片)转移到神经网络拟态技术(芯片)上来(神经拟态芯片模拟人脑运作机制,主要采用异步脉冲神经网络);我们应与神经科学联系,对异步脉冲神经网络很多未知的关键技术、运作机制和功能表现有待深入研究与工程实践:关键还要进一步深入理解人类大脑神经元的生物特性和运作机制以用于我们的研发;研发基于脉冲信号信息处理的稀疏和时间的动态特性、脉冲时序编码机制、突触转移高效函数、异步脉冲传输机制及各项功效指标等;异步脉冲神经网络向类脑方向发展也离不开自监督学习和训练。

总之,对于走这条路很多人工智能专家是向往的,但日前研究尚处于初级阶段,迄今国内外均未拿出亮眼的成果。

三、从知识表示、驱动、推理,建设大规模语义网络出发

业内人工智能专家欢呼:2019 年自然语言处理(NLP)取得重大突破!

这条知识工程之路从感知智能奔向认知智能。上世纪 80 年代中期启动了知识工程,本世纪初又更新为新知识工程。新知识工程的重点是建设大规模语义网络(以提升知识图谱)。语义网络的发展过程是从自然语言处理系统到自然语言理解系统,再到大规模语义网络。IBM 沃森主张在以知识表示、驱动、推理的路上,由大规模语义网络支持的认知智能目标得以实现。

早年间，IBM“WatsonHealth”搞医疗人工智能走的就是这条路。IBM 认为，对人工智能最重要的能力是知识而非数据。他们探索知识表示、驱动、推理，以期医疗人工智能从不可理解、不可解释的感知智能阶段推向可理解、可解释的认知智能阶段。但 IBM 走的这条路是失败的。

IBM 的失败，其中主要原因之一是大规模语义网络还不够完善，还没有能力支持认知智能的实现。这里我们引用图灵奖得主、人工智能大师 YoshuaBengio 对此评论中的一段话：“NLP 虽然取得较大进步，但与人类相差还甚远”。

必须指出，对于常识、专业知识、专家经验，机器是很难识别的。IBM 提出具人（embodiment）概念，强调人工智能专家必须与临床医生结合，在疾病诊断时要取得共识。还有达到人类智能的另一道难题是：背景知识，这在学习和训练时是不可或缺的。

所以对于走这条路，未来是非常有前途的，但当下还不成熟，路还很长。

**10-525，**深度学习是有缺陷的，它并非是实现人工智能的一条完美的路径

美国一位 AI 专家在《IEEE Spectrum》(2019.4）上撰文质疑 IBM Watson 研发 AI 医疗应用前景时提出：“深度学习也是有缺陷的”，“深度学习是实现人工智能的路径之一，而非一条完美的路径”。“深度学习本质上是一项黑盒子技术，其训练过程具有难以解释、不可控的特点”。

**11-526，**深度学习最大的问题是不可解释和不可理解。

我记得张钹老师去年 8 月 14 日的谈话，其中提到：现在深度学习本质上是基于概率统计的学习，而概率统计最大的问题是不可解释和不可理解。

12-527，当前机器学习理论有局限性完全以盲模型的方式运行。

图灵奖得主、贝叶斯网络之父 Judea pearl 在 2018 年的论文中谈到，当前机器学习理论有局限性，完全以统计学或盲模型（即黑盒子）的方式运行，所以不能成为强 AI 的基础。

13-528，深度学习算法目前并不完美有待继续加强理论研究。

国外一位 AI 大师论文中谈到，近年来以深度学习算法为代表的 AI 技术快速发展，在计算机视觉、语音识别、语义理解等领域实现了突破，但其算法并不完美，有待继续加强理论研究。

14-529，打破黑盒子使深度学习转变为可解释。

深度学习本质上是一项黑盒子技术，其训练过程是不可理解、不可解释的。

科学的发展，打破了黑盒子这把锁，将转变深度学习模型原来的不可理解、不可解释为可理解、可解释。

据我们收到一些跟贴反映，有关机构/人士正在潜心研究这项破解黑盒子的理论。

15-530，如何通过知识工程实现可解释的人工智能。

目前，通过知识工程尚难实现可解释的人工智能（或者说，机器还不能实现像人类那样的认知智能）。

最后一公里的短板在哪里？短板在知识图谱或大规模语义网络。

知识图谱即为一种大规模语义网络。

大规模语义网络是在大数据时代体现新知识工程的核心技术。

自然语言处理（NPL）取得很大进步，人工智能资深专家吴恩达说，2019 年是自然语言处理飞跃的一年。

我们来看一下自然语言研究的发展轨迹：自然语言处理（NLP）系统——>自然语言理解系统（具备一定的理解和解释的能力）或语义网络——>大规模语义网络（包括语言建模和训练模式）。

迄今大规模语义网络还不够完善。人工智能大师 YoshuaBeno 认为，NLP 虽然取得较大进步，但与人类的认知能力相差还甚远。让大规模语义网络来支持实现可解释的人工智能其能力尚嫌不足。

在跟贴 84 中，谈到如何完善大规模语义网络，如何提升其中的理解和解释两个核心能力（从现有一定的基础上提升）。

还有不少知识是大规模语义网络所不能概括的，如常识，常识是难以定义、表达、表征的，目前的大规模语义网络尚不包括常识。除常识外，还有背景知识、专业知识、专家经验、隐性知识等，也不能被大规模语义网络所概括。

跟贴 457 主要谈到华盛顿大学叶锦才研发团队关于常识推理攻关研究的进展。

IBM Watson 在人工智能医疗诊断中，提出具人（embodiment）的概念，要求医学科学家与临床医生沟通，取得共识，以此来克服缺乏常识的困难。

**16-533，**打破黑盒子克服机器学习模型不可解释缺陷创建可解释的机器学习新模型

机器学习/深度学习取得巨大进步，今天已成为引领人工智能大发展的引擎，但机器学习/深度学习是有缺陷的，它的黑盒子特性使其行为背后缺乏透明度，用户很难理解机器学习/深度学习模型是如何做出决策的，其不可理解、不可解释

的缺陷限制了它扩大、提升应用。

对待机器学习/深度学习模型的发展有两种态度，一种是“深度学习（技术）已达天花板”,“其致命的不可解释缺陷将使其停止不前”,“人工智能发展只能捨此另觅出路”；另一种是“机器学习/深度学习潜力很大”,“决不放弃改造可解释的机器学习模型”!

目前不少人工智能专家正致力于改造机器学习/深度学习模型，致力于研究可解释的机器学习/深度学习技术，可以预见将为机器学习/深度学习发展带来转机!

此处介绍一篇论文“可解释机器学习技术”,刊载于 Communications of theACM,“Techniques for interpret able machineLearning”, 2020, 63（1）: 68-77。

可解释机器学习使得机器学习模型能够以易于理解的方式向用户解释或呈现其行为，我们称这种特性为可解释性（interpretecbility）或者解释性（explainability)。

17-538，机器学习/深度学习是一个强大的数据分析工具，带动了今天人工智能的繁荣。但机器学习/深度学习是有缺陷的，它本质上是一项暗箱操作（黑盒子）技术或盲模型，其训练过程是不可解释、不可理解的，缺乏类似人脑的推理能力。

在 527 条跟帖中，图灵奖得主、贝叶斯网之父 Judea Pearl 谈到，当前机器学习理论有局限性，完全以统计学或盲模型（黑盒子）的方式运行，所以不能成为强 AI 的基础。

在 1~535 条跟帖汇集中我们也能查到：ACM 图灵获奖得者、算法大师约翰·霍普克罗夫在 2019 年谈到，对深度学习这个黑盒子，人们知道它在学习，但不知它怎么学习，人类可能会在 5 年后大体读出深度学习的数学理论。

当下，打破机器学习黑盒子研发可解释的人工智能已成为一股热潮：

IBM 人工智能研究所 Vijay Arya 研发团队（19 人），为打破机器学习黑盒子研发可解释的人工智能，针对信用、打假、反篡改、公平等 4 个案例（课题），率先研发了可解释工具包 AIX360，目前已研发了 10 种算法。

马格德堡大学人工智能实验室 Sebastian Stober 团队也在研发可解释的人工智能技术，其研究项目的编号为为 CogxAI。

他们用分析深度学习人工神经网络的方法打破黑盒子，在认知神经科学启发下研发可解释、可理解的人工智能，使其学习、训练过程变得更加透明和容易理解。

## 18-539. 可信任人工智能（反欺诈、可解释、公平性）

### 打破机器学习黑盒子研发可解释人工智能

### IBM 开放技术副总裁 Todd Moore

（在《第十五届开源中国开源世界高峰论坛》线上会议上的报告）

大家好。我是 IBM 开放技术副总裁 Todd Moore。今天，我想和大家谈谈对我来说很重要的事。此话题我之前在 Linux 基金会的一个开源活动中也谈到过。那就是《可信的人工智能》、《负责的人工智能》。

当今世界正处于新冠肺炎大流行病的高危之中，我们面临着一场全球性健康危机，而要有效地拯救生命，这取决于数据，取决于我们如何收集这些数据，取决于我们如何有效地利用这些数据。同时，我们正处于一场围绕种族和平权的全球运动之中。这些事件相互交叉、影响。它们影响到我们如何利用人工智能、如何利用我们的模型、我们的数据,以及我们将如何从这里走向未来。

IBM 的核心价值观根植于权利平等。这些 IBM 长期以来的公司政策，可以追溯到

1953 年，托马斯·约翰·沃森那时所写的（IBM 公司）政策令第 4 号，确保 IBM 作为雇主将提供平等机会,确保 IBM 不会容忍因种族、肤色或宗教而产生的偏见。今天，阿尔温德·克里希纳也已肩负这项使命。阿尔温德日前宣布 IBM 不再提供面容识别服务，因而切断了一条可能会因其使用而在社会中造成偏见的途径。他也将这一问题提升为一项公众对话，以探讨这项技术是否有意义，我们的社会应该如何应对。但是，人工智能的可信、透明不仅仅是面容识别，当我们收集的数据一旦投入使用，隐私问题就呈现出来。伦理问题关乎这些数据如何使用、关乎如何转化、关乎我们的社会如何应对：所有这些方面涉及的伦理问题都亟需解决。IBM 为此提供了一系列工具。我们也将其中许多项目开源了。我们能够提供一个由一系列 Python 库组成的《稳健性工具包》，我们能够提供一个能在模型和数据中发现偏见的《公平工具包》。我们能够提供一个《可解释性工具包》，它可以让你能够快速检查你的潜在 AI 模型的黑匣子，从而把这个黑盒子变成一个你能给识别其行为方式的东西。这些都是帮助我们消除偏见必不可少的工具。

同时，我们也在开始一项我们称之为 AI FactSheets 的工作，这个项目是关于我们如何追究模型的责任。想想看，就如同食物产品上贴的标签（标签体系）花费了很长的时间才开发出来，但一旦有了这样的标签，你会从中得到相当多的信息，比如手里这罐汤，到底含有些什么东西。这也是我们研发这个 AI FactSheets 项目希望达成的效果。

现在，这些项目已经取得了进展。我们参加了 Linux 基金会人工智能组织 （LF AI）。我们希望这些项目都置于其开放治理模式之下，（LF AI 的）技术顾问委员会（TAC）目前已经投票赞成孵化这些项目。大家将看到其中三个项目已经发布，并通过 LF AI 机构开源。我们非常感谢大家这样做。我们认为这是一个很好的场

所能让其他人可加入进来公开影响和推动这些工作的进展。

我们认为这对社会很重要。LF AI 这个组织的团队已经有不少非常重要的公司和大学参与其中，此外还成立的《可信人工智能委员会》，我们希望大家也都能参加进来。因为他们一直在制定（人工智能）的基本原则，而我们将充分准照这些基本原则，使用由此而研发的 AI 技术所产生的数据和模型。 这个工作组宣布了七项基本原则：即公平性、再现性、透明性、治理、隐私、安全和问责。这些是这个工作组正在研究的七项基本原则 ，与时俱进中。所以，现在就和我们一起来吧。

自从上次我们谈过这件事以来，就如我刚刚提到过的，AI FactSheets 已经提上议事日程，我们已经开始寻求其开源途径。AIF360 团队与 Scikit Learn 和 R Studio 展开合作，现在已经可以在（Learn 和 R Studio 中）使用了；Kubeflow MLOps 流水线也已经发布,我们可用这些流水线进行公平性和对抗性检测;Apache NiFi 也已经接纳 MLOps，现在我们也有了可用的（Apache NiFi）处理器。所有的项目也都取得了很大的进展，感谢一直以来为此做出贡献的人们，这些真的对我们的当今世界非常重要。作为小结，大家今天看到了不少东西，我们很高兴也能够提供非常有价值的（开源）社区提供支持，您如果有意愿，请加入这些项目或者加入 LF AI 组织的可信人工智能委员会，帮助我们确立发展方向。现在是至关重要的时刻，这个话题将帮助我们，作为世界性组织机构，作为开源贡献者，推动世界向前发展。 非常感谢大家！

**19-558，**ISO 的可解释 AI 标准项目

2020 年是国际标准组织 ISO/IEC JTC1 成立人工智能标准分委员会 SC42 的第三

年。在不久结束的第六次年会上，SC42 批准将新标准项目提案《机器学习模型和人工智能系统的可解释性之目标和方法》提交 SC42 的全体成员国投票表决。这一项目旨在本文档描述可用于实现 ML 模型和 AI 系统的行为、输出和结果方面的不同利益相关者的可解释性目标的方式方法。所谓利益相关涉及学术界、产业界、政策制定者和最终用户等。不出意外，此项目将最快明年春天开始，预计 2020 年中完成。

**20-570，**COPU 谈人工智能专题会议

——COPU 开源联盟秘书处

12 月 1 日（周二）陆主席召开 COPU 专题会议，讨论人工智能国内外 551 条跟帖评论：跟帖涵盖全球人工智能研发前沿，有 6 大部分，

1) 不可解释的机器学习/深度学习支持今天人工智能的繁荣，说它已无发展潜力、已近天花板是不妥的；
2) 打破机器学习的黑盒子研发可解释的人工智能，有所突破；
3) 基于异步脉冲神经网络的神经拟态计算系统，已有亮点；
4) 从知识工程出发，依托大规模语义网络（知识图谱）的支持，用以破解认知智能解决方案，还差最后一公里；
5) 脑机接口的理论和实践，目前国内外已有几十例试点；
6) 钟义信教授、张钹院士对不同学派提出的人工智能发展路径提出质疑，但钟、张所提的发展模式还是概念。此处还报导了清华、北大教授激辩：脑科学是否真能启发人工智能？！

现在看来，目前人工智能国内外跟帖评论热闹非常！朋友，如你有兴趣，不防参

加进来（先看后评）。

**21-574**，可解释的信用评分模型（为信贷建立可解释的信用评分模型）

Malta 大学人工智能部 Lara Marie Demajo，VinceVella，AlexieiDingli，2020.12.4

随着人工智能和金融技术的发展，信用评分模型已经引起了学术界广泛关注。信用评分可以帮助金融专家更好地决定是否接受信贷申请。最近一些法规，如《一般数据保护条例》(GDPR）和《平等信贷机会法》(ECOA)，都增加了对模型可解释性的要求，以保证算法决策的可理解性和一致性。论文作者提出了一个既准确又可解释的信用评分模型。该模型的流程是：首先对数据进行预处理，然后使用 XGBOOST 模型对数据实例进行分类，最后使用三种 XAI 方法对分类器进行扩展，提供一个全方位的解释框架。

数据预处理过程：数据清洗→特征生成/选择→数据集划分→标准化→交叉验证→平衡数据。

不同的人在不同情况下需要不同的解释，而单一的 XAI 方法不足以提供所有的解释。因此作者提出了一个可解释的信用评分模型，为各种角色产生一种解释。在信用评分模型中有三种不同的角色：

①信贷员，喜欢基于实例的局部解释。该解释提供了对单个实例预测的局部解释。信贷员更倾向于这样的解释，因为他们需要知道模型给出的预测结果是否合理。

②被拒绝的贷款申请人，喜欢基于特征的局部解释。这项解释是对某一特定预测结果的解释，说明模型是如何做出该预测的以及这样预测的原因。贷款申请人更倾向于这样的解释，因为他们最关心的是为什么他们的贷款申请被拒绝。

③监管者或数据科学家，喜欢模型全局解释。这项解释是对模型整体工作方式的

理解，解释了模型在做预测时背后使用的逻辑推理。监管机构和管理层通常更倾向于这样的解释，因为他们主要关注的是对信用评分模型的全局理解，而不是对每种情况的个别解释。

列出为各种角色提供解释的 XAI 方法如下：

| Explanation Type | XAI Method | Explanation Form |
|---|---|---|
| Global | SHAP＋GIRP | Decision Tree/IF-THENrules |
| Local feature-based | Anchors | DNF rule |
| Local instance-based | ProtoDash | Prototypicalinstances |

最后，作者进行实验，检验了模型的正确性、有效性、易理解性、细节充分性和可信度。

## 22-575. 可解释 AI 的上手实践

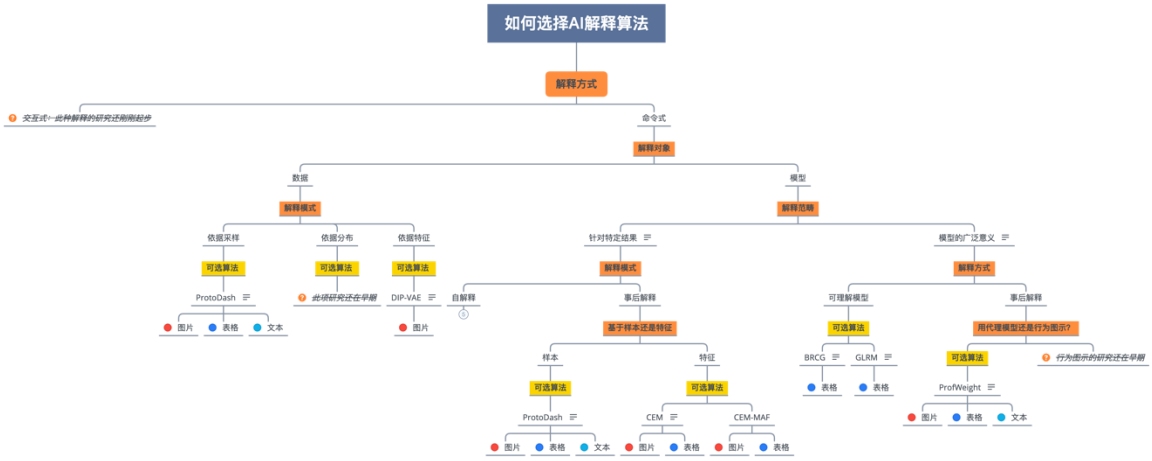
——IBM 田忠博士摘自 IBM AI 研究院

人工智能（AI）有意无意间已经成为我们生活的一部份。如何安心、放心、信心使用 AI 成为业界研究和实践的重点。学术界的研究相当活跃，每年都有若干专门的学术会议，如 WHI2020 （Workshop on Human Interpretability in Machine Learning)。2019 年，IBM 研究院多年关于可解释 AI 的研究成果开源，合称 AIX360，并捐给 Linux 人工智能基金会（LF AI)。 截止 2020 年初，AIX360 包含的算法，其中 8 个来自 IBM 研究院的科研成果，2 个是业界的流行算法，有关的代码、文档、演示可在 http://aix360.mybluemix.net 获得。和 AIX360 相辅相成，致力于可信赖 AI 的其它项目还有致力于公平性的 AIF 360、致力于健壮性的 ART 360、致力于真实性的 Factsheet 360。

可解释 AI 的意义、目的、方法因人而异。对于 AI 系统开发者、数据科学家、项

目经理而言，其目的多半是如何提高系统效率；对于 AI 系统使用者，如医生、律师、银行贷款经理、考官，则是需要对 AI 系统做出的推荐需要的是信心、放心、安心；对于主管当局，如欧盟委员会、纽约市政府、中国银保监会，他们主要关心的是如何确保 AI 的系统性的公平性；而对于最终受影响的用户，如病人、诉讼对象、贷款申请人、教师，他们需要的是能够理解影响结论的主次要因素，从而未来能够有所作为。

那么如何按需选择合适的解释算法呢？下面的树形结构可供参考。




我来看一个示例，一家银行使用了 AI 系统帮助基于公开可获得的 FICO HELOC Dataset 真实数据来辅助决策是否批准一项贷款申请。对于建设系统的数据科学家、银行的贷款经理以及贷款申请者对系统的解释性有不同的需求，因而选择了不同的算法获得洞察。

对于建设本 AI 系统的数据科学家来说，重中之重是向银行主管以容易理解的方式（比如一组简明规则）解释本系统的工作效果。为此，他需要系统执行一个命令来获得对决策模型的普适意义的解释模型。依照上面的选择路径，他因此选择了 BRCG 算法以生成一组布尔规则表，使用 GLRM 算法生成逻辑规则回归模型。

有了这个决定，他按如下步骤开展：加载整理数据、运行算法、显示结论。

加载整理数据

| | 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 |

运行算法

1）BRCG

```
# Instantiate BRCG with small complexity penalty and large beam search width
from aix360.algorithms.rbm import BooleanRuleCG
br = BooleanRuleCG(lambda0=1e-3, lambda1=1e-3, CNF=True)

# Train, print, and evaluate model
br.fit(dfTrain, yTrain)
from sklearn.metrics import accuracy_score
print('Training accuracy:', accuracy_score(yTrain, br.predict(dfTrain)))
print('Test accuracy:', accuracy_score(yTest, br.predict(dfTest)))
print('Predict Y=0 if ANY of the following rules are satisfied, otherwise Y=1:')
print(br.explain()['rules'])
```

```
Learning CNF rule with complexity parameters lambda0=0.001, lambda1=0.001
Initial LP solved
Iteration: 1, Objective: 0.2895
Iteration: 2, Objective: 0.2895
Iteration: 3, Objective: 0.2895
Iteration: 4, Objective: 0.2895
Iteration: 5, Objective: 0.2864
Iteration: 6, Objective: 0.2864
Iteration: 7, Objective: 0.2864
Training accuracy: 0.7195731460218 83
Test accuracy: 0.6965153970826 58
Predict Y=0 if ANY of the following rules are satisfied, otherwise Y=1:
```

2）LogRR

```
# Instantiate LRR with good complexity penalties and numerical features
from aix360.algorithms.rbm import LogisticRuleRegression
lrr = LogisticRuleRegression(lambda0=0.005, lambda1=0.001, useOrd=True)

# Train, print, and evaluate model
lrr.fit(dfTrain, yTrain, dfTrainStd)
print('Training accuracy:', accuracy_score(yTrain, lrr.predict(dfTrain, dfTrainStd)))
print('Test accuracy:', accuracy_score(yTest, lrr.predict(dfTest, dfTestStd)))
print('Probability of Y=1 is predicted as logistic(z) = 1 / (1 + exp(-z))')
print('where z is a linear combination of the following rules/numerical features:')
lrr.explain()
```

```
Training accuracy: 0.742536809401594
Test accuracy: 0.7260940032414911
Probability of Y=1 is predicted as logistic(z) = 1 / (1 + exp(-z))
where z is a linear combination of the following rules/numerical features:
```

| | rule/numerical feature | coefficient |
|---|---|---|
| 0 | (intercept) | -0.0686341 |
| 1 | MSinceMostRecentInqexcl7days > 0.00 | 0.680261 |
| 2 | ExternalRiskEstimate | 0.654248 |
| 3 | NetFractionRevolvingBurden | -0.553965 |
| 4 | NumSatisfactoryTrades | 0.551654 |
| 5 | NumInqLast6M | -0.463226 |
| 6 | NumBank2NatlTradesWHighUtilization | -0.448331 |
| 7 | AverageMInFile <= 52.00 | -0.43436 |
| 8 | NumRevolvingTradesWBalance <= 5.00 | 0.42154 |
| 9 | MaxDelq2PublicRecLast12M <= 5.00 | -0.418142 |
| 10 | PercentInstallTrades > 50.00 | -0.317566 |
| 11 | NumSatisfactoryTrades <= 12.00 | -0.312471 |
| 12 | MSinceMostRecentDelq <= 21.00 | -0.301566 |
| 13 | PercentTradesNeverDelq <= 95.00 | -0.273924 |
| 14 | ExternalRiskEstimate > 75.00 | 0.263437 |
| 15 | AverageMInFile <= 84.00 | -0.182118 |
| 16 | PercentTradesNeverDelq | 0.166518 |
| 17 | AverageMInFile | 0.15069 |
| 18 | PercentInstallTrades > 42.00 | -0.148802 |
| 19 | NumBank2NatlTradesWHighUtilization <= 0.00 | 0.135396 |
| 20 | MSinceOldestTradeOpen <= 122.00 | -0.132409 |
| 21 | PercentTradesNeverDelq <= 91.00 | -0.11771 |
| 22 | NumSatisfactoryTrades <= 17.00 | -0.11022 |
| 23 | ExternalRiskEstimate > 72.00 | 0.107613 |

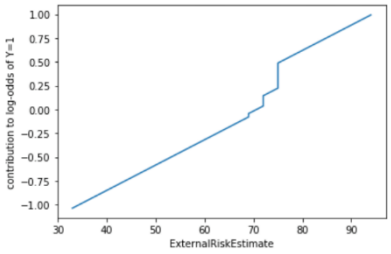
（图形）显示结论，如以 GAM 图示 LogRR 的结论

外部风险预估

**ExternalRiskEstimate**

As expected from the BRCG Boolean rule above, 'ExternalRiskEstimate' is an important feature positively correlated with good credit risk. The jumps in the plot indicate that applicants with above average 'ExternalRiskEstimate' (the mean is 72) get an additional boost.

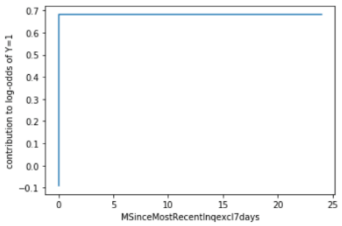
```
lrr.visualize(data, fb, ['ExternalRiskEstimate']);
```



**Credit inquiries**

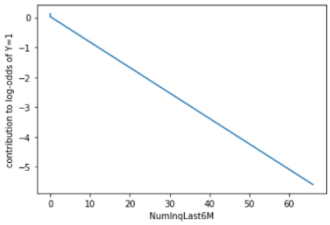
The next two plots illustrate the dependence on the applicant's credit inquiries. The first plot shows a significant penalty for having less than one month since the most recent inquiry ('MSinceMostRecentInqexcl7days' = 0).

```
lrr.visualize(data, fb, ['MSinceMostRecentInqexcl7days']);
```




The second shows that predicted risk increases with the number of inquiries in the last six months ('NumInqLast6M').

```
lrr.visualize(data, fb, ['NumInqLast6M']);
```

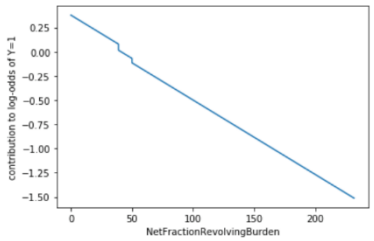



信用查询次数的影

**Debt level**

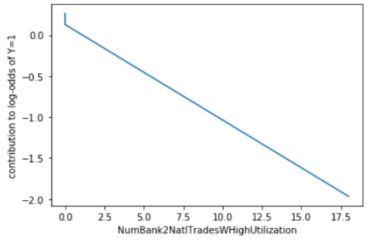
The following four plots relate to the applicant's debt level. 'NetFractionRevolvingBurden' is the ratio of revolving debt (e.g. credit card) balance to credit limit, expressed as a percentage, and has a large negative impact on the probability of good credit. A small fraction of applicants (less than 1%) actually have NetFractionRevolvingBurden greater than 100%, i.e. more revolving debt than their credit limit. This might be investigated further by the data scientist.

```
lrr.visualize(data, fb, ['NetFractionRevolvingBurden']);
```




The second 'NumBank2NatlTradesWHighUtilization' plot shows that the number of accounts ("trades") with high utilization (high balance relative to credit limit for each account) also has a large impact, with a drop as soon as one account has high utilization.

```
lrr.visualize(data, fb, ['NumBank2NatlTradesWHighUtilization']);
```




债务水平的影

对于使用这个AI系统的银行贷款经理而言,他的关心重点是贷款决定的一致性,是否存在系统性的歧视。同样他也只希望一个指令获得解释,以便增强对模型的

信心（即对模型普遍性的解释而不是单个案例的解释），而其依据是手边现有的案例为支撑（基于现有样本）对于特定结果的（事后）解释。依照上面的树形结构，我们自然能理解为啥银行贷款经理使用 ProtoDash 算法来寻求帮助。

有了这个决策，他按如下步骤开展：加载整理数据、运行算法、图示结论。

<table>
<tr><td>加载整理数据</td><td>

```
heloc = HELOCDataset()
df = heloc.dataframe()
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 24)
pd.set_option('display.width', 1000)
print("Size of HELOC dataset:", df.shape)
print("Number of \"Good\" applicants:", np.sum(df['RiskPerformance']=='Good'))
print("Number of \"Bad\" applicants:", np.sum(df['RiskPerformance']=='Bad'))
print("Sample Applicants:")
df.head(10).transpose()
```

```
Using Heloc dataset:  c:\users\ronnyluss\aix360\aix360\datasets\..\data\heloc_data\heloc_dataset.csv
Size of HELOC dataset: (10459, 24)
Number of "Good" applicants: 5000
Number of "Bad" applicants: 5459
Sample Applicants:
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| ExternalRiskEstimate | 55 | 61 | 67 | 66 | 81 | 59 | 54 | 68 | 59 | 61 |
| MSinceOldestTradeOpen | 144 | 58 | 66 | 169 | 333 | 137 | 88 | 148 | 324 | 79 |
| MSinceMostRecentTradeOpen | 4 | 15 | 5 | 1 | 27 | 11 | 7 | 7 | 2 | 4 |
| AverageMInFile | 84 | 41 | 24 | 73 | 132 | 78 | 37 | 65 | 138 | 36 |
| NumSatisfactoryTrades | 20 | 2 | 9 | 28 | 12 | 31 | 25 | 17 | 24 | 19 |
| NumTrades60Ever2DerogPubRec | 3 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| NumTrades90Ever2DerogPubRec | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| PercentTradesNeverDelq | 83 | 100 | 100 | 93 | 100 | 91 | 92 | 83 | 85 | 95 |
| MSinceMostRecentDelq | 2 | -7 | -7 | 76 | -7 | 1 | 9 | 31 | 5 | 5 |
| MaxDelq2PublicRecLast12M | 3 | 0 | 7 | 6 | 7 | 4 | 4 | 6 | 4 | 4 |
| MaxDelqEver | 5 | 8 | 8 | 6 | 8 | 6 | 6 | 6 | 6 | 6 |
| NumTotalTrades | 23 | 7 | 9 | 30 | 12 | 32 | 26 | 18 | 27 | 19 |
| NumTradesOpeninLast12M | 1 | 0 | 4 | 3 | 0 | 1 | 3 | 1 | 1 | 3 |
| PercentInstallTrades | 43 | 67 | 44 | 57 | 25 | 47 | 58 | 44 | 26 | 26 |
| MSinceMostRecentInqexcl7days | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NumInqLast6M | 0 | 0 | 4 | 5 | 1 | 0 | 4 | 0 | 1 | 6 |
| NumInqLast6Mexcl7days | 0 | 0 | 4 | 4 | 1 | 0 | 4 | 0 | 1 | 6 |
| NetFractionRevolvingBurden | 33 | 0 | 53 | 72 | 51 | 62 | 89 | 28 | 68 | 31 |
| NetFractionInstallBurden | -8 | -8 | 66 | 83 | 89 | 93 | 76 | 48 | -8 | 86 |
| NumRevolvingTradesWBalance | 8 | 0 | 4 | 6 | 3 | 12 | 7 | 2 | 7 | 5 |
| NumInstallTradesWBalance | 1 | -8 | 2 | 4 | 1 | 4 | 7 | 2 | 1 | 3 |
| NumBank2NatlTradesWHighUtilization | 1 | -8 | 1 | 3 | 0 | 3 | 2 | 2 | 3 | 1 |
| PercentTradesWBalance | 69 | 0 | 86 | 91 | 80 | 94 | 100 | 40 | 90 | 62 |
| RiskPerformance | Bad | Bad | Bad | Bad | Bad | Bad | Good | Good | Bad | Bad |

</td></tr>
<tr><td>运行算法</td><td></td></tr>
<tr><td>1）预处理训练数据</td><td>

```
# Clean data and split dataset into train/test
(Data, x_train, x_test, y_train_b, y_test_b) = heloc.split()
Z = np.vstack((x_train, x_test))
Zmax = np.max(Z, axis=0)
Zmin = np.min(Z, axis=0)


#normalize an array of samples to range [-0.5, 0.5]
def normalize(V):
```

</td></tr>
</table>

| | |
|---|---|
| | VN = (V - Zmin)/(Zmax - Zmin)<br>VN = VN - 0.5<br>return(VN)<br><br># rescale a sample to recover original values for normalized values.<br>def rescale(X):<br>return(np.multiply ( X + 0.5, (Zmax - Zmin) ) + Zmin)<br><br>N = normalize(Z)<br>xn_train = N[0:x_train.shape[0], :]<br>xn_test = N[x_train.shape[0]:, :] |
| 2）定义和训练模型 | # nn with no softmax<br>def nn_small():<br>model = Sequential()<br>model.add(Dense(10, input_dim=23, kernel_initializer='normal', activation='relu'))<br>model.add(Dense(2, kernel_initializer='normal'))<br><br>return model<br># Set random seeds for repeatability<br>np.random.seed(1)<br>tf.set_random_seed(2)<br><br>class_names = ['Bad', 'Good']<br><br># loss function<br>def fn(correct, predicted): |

```
    return tf.nn.softmax_cross_entropy_with_logits(labe
ls=correct, logits=predicted)

# compile and print model summary
nn = nn_small()
nn.compile(loss=fn, optimizer='adam', metrics=['accurac
y'])
nn.summary()

# train model or load a trained model
TRAIN_MODEL = False

if (TRAIN_MODEL):
    nn.fit(xn_train, y_train_b, batch_size=128, epochs=
500, verbose=1, shuffle=False)
    nn.save_weights("heloc_nnsmall.h5")
else:
    nn.load_weights("heloc_nnsmall.h5")

# evaluate model accuracy
score = nn.evaluate(xn_train, y_train_b, verbose=0) #Co
mpute training set accuracy
#print('Train loss:', score[0])
print('Train accuracy:', score[1])

score = nn.evaluate(xn_test, y_test_b, verbose=0) #Comp
ute test set accuracy
#print('Test loss:', score[0])
print('Test accuracy:', score[1])
```

| | |
|---|---|
| 3）验证现有数据集中类似案例将得到类似结论（即验证模型的一致性） | 比如，我们选择案例#8，其结论是 Good，可以放款，其特征及如右图所示。<br><br>那么，数据集中有同样结论的其它案例有哪些？他们是否也有同样的特征分布？如果有，则表明这个模型具有一致性。<br><br>我们先获得所有结论是 Good 的案例；<br>计算这些典型用户和#8 的相似度；<br>我们获得了如下的相似度对比表，第 0 列是我们的选定案例#8，其它四个是数据集中有同样 Good 结论的其它案例。显然，过半数的特征的是接近的。仔细研究对照表，贷款经理发现，能获得 Good 结论的客户都是没有负债的客户，这个发现让贷款经理对系统的结论更有信息了。 |

| | |
|---|---|
| ExternalRiskEstimate | 82 |
| MSinceOldestTradeOpen | 280 |
| MSinceMostRecentTradeOpen | 13 |
| AverageMInFile | 102 |
| NumSatisfactoryTrades | 22 |
| NumTrades60Ever2DerogPubRec | 0 |
| NumTrades90Ever2DerogPubRec | 0 |
| PercentTradesNeverDelq | 91 |
| MSinceMostRecentDelq | 26 |
| MaxDelq2PublicRecLast12M | 6 |
| MaxDelqEver | 6 |
| NumTotalTrades | 23 |
| NumTradesOpeninLast12M | 0 |
| PercentInstallTrades | 9 |
| MSinceMostRecentInqexcl7days | 0 |
| NumInqLast6M | 0 |
| NumInqLast6Mexcl7days | 0 |
| NetFractionRevolvingBurden | 3 |
| NetFractionInstallBurden | 0 |
| NumRevolvingTradesWBalance | 4 |
| NumInstallTradesWBalance | 1 |
| NumBank2NatlTradesWHighUtilization | 1 |
| PercentTradesWBalance | 42 |

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| ExternalRiskEstimate | 0.59 | 0.29 | 0.42 | 0.84 | 0.21 |
| MSinceOldestTradeOpen | 0.76 | 0.62 | 0.76 | 0.09 | 0.79 |
| MSinceMostRecentTradeOpen | 1.00 | 0.09 | 0.83 | 0.89 | 0.87 |
| AverageMInFile | 0.79 | 0.09 | 0.90 | 1.00 | 0.82 |
| NumSatisfactoryTrades | 0.95 | 0.39 | 0.74 | 0.39 | 0.15 |
| NumTrades60Ever2DerogPubRec | 1.00 | 1.00 | 0.08 | 1.00 | 1.00 |
| NumTrades90Ever2DerogPubRec | 1.00 | 1.00 | 0.08 | 1.00 | 1.00 |
| PercentTradesNeverDelq | 1.00 | 0.15 | 0.81 | 0.15 | 0.15 |
| MSinceMostRecentDelq | 1.00 | 0.36 | 0.22 | 0.36 | 0.36 |
| MaxDelq2PublicRecLast12M | 1.00 | 0.13 | 1.00 | 0.13 | 1.00 |
| MaxDelqEver | 1.00 | 0.41 | 0.17 | 0.41 | 0.64 |
| NumTotalTrades | 0.80 | 0.23 | 0.86 | 0.26 | 0.35 |
| NumTradesOpeninLast12M | 1.00 | 1.00 | 0.40 | 0.40 | 0.06 |
| PercentInstallTrades | 1.00 | 0.05 | 0.54 | 0.37 | 0.33 |
| MSinceMostRecentInqexcl7days | 0.08 | 1.00 | 1.00 | 1.00 | 1.00 |
| NumInqLast6M | 0.21 | 1.00 | 0.21 | 0.21 | 0.04 |
| NumInqLast6Mexcl7days | 0.26 | 1.00 | 0.26 | 1.00 | 0.07 |
| NetFractionRevolvingBurden | 0.96 | 0.88 | 0.96 | 0.92 | 0.09 |
| NetFractionInstallBurden | 1.00 | 1.00 | 1.00 | 1.00 | 0.08 |
| NumRevolvingTradesWBalance | 1.00 | 0.28 | 0.38 | 0.73 | 0.20 |
| NumInstallTradesWBalance | 1.00 | 0.13 | 1.00 | 0.13 | 1.00 |
| NumBank2NatlTradesWHighUtilization | 0.69 | 0.69 | 0.69 | 1.00 | 0.11 |
| PercentTradesWBalance | 0.67 | 0.12 | 0.36 | 0.38 | 0.57 |

| | |
|---|---|
| 类似地，我们也可以对结论是 Bad（贷款申请被拒绝）的客户用同样的步骤做同样的研究 | 结果也同样是对于指定的样本客户，同样获得 Bad 结论的客户，其特征指标中过半数非常接近。<br>贷款经理仔细研究这些接近指标发现，这些被判 Bad 的客户，大都有轻微犯罪前科。为此，贷款经理在处理这类客户时就要额外小心。 |

对于受这个 AI 系统影响的客户而言（即贷款申请人），尤其是申请被拒的人，需要了解哪些因素是关键，从而他们可以采取行动改善自己的财经状况，以便日后有机会成功申请。其步骤也是类似：加载整理数据、运行算法、显示结论。这个就不赘述了。

要亲自上手试验，请参考 AIX360 网站 http://aix360.mybluemix.net。

**23-577，**清华沈向洋教授结合创新谈人工智能和开源

清华大学双聘教授沈向洋最近在 CNCC2020 圆桌会议上谈到人工智能和开源的问题。

他说：我们的创新第一是人工智能，最重要的事情就是要做可解释的人工智能。还有，我们为什么需要拥抱开源？

他结合自己过去十几年在微软工作时思考的几个大方向，目前有三个方向：第一个是人工智能，最重要的事情就是要做可解释的人工智能，如今深度学习发展很快，但可解释性这边进展较为缓慢。第二个是量子计算，这个路途还很遥远，微软从很早以前就走一条所谓的拓扑量子计算这样的路，但真正的量子计算器还有相当长的路要走。第三就是微软一直在推的混合现实。大家都很期待苹果之后发行的手机，到底 AR 是怎么样的，5G 上来之后，会有什么样的变化。这些我都非

常期待。

我个人则希望做一些 AI 和神经科学之间的研究，神经科学研究实际上还处在早期阶段，数据不够，也做不了太多实验。我们能否解决一些真正的问题，例如阿尔兹海默症，中年忧郁症，儿童自闭症，这些都是大脑出现了问题。目前 AI 在学人脑，那么 AI 是否能反过来去帮助解决人脑的问题？也许有机会，清华的学生也好、其他人也好，大家能一起做一些事情。

我前面提到了，可解释性的人工智能有必要做。因为你如果不理解它，就很难去下判断，即这件事该不该做。

在谈到开源时，他说：创新就要做到极致，用开源的方式培养未来。他提出：为什么我们需要拥抱开源？

他说：过去 40 年，中国的科技发展非常大，因为出现了两个东西，一个是互联网，一个是开源。开源这件事情，对我们影响非常大。如果用开源的方式培养人才、提高水平，不要总提高 10 倍，只要提高 2 倍，对社会效率影响就会非常大。

24-578，论文：审查对可解释的人工智能的需求

（Reviewing the Need for Explainable Artificial Intelligence）

Julie Gerlings，Arisa Shollo，Ioanna Constantiou

哥本哈根商学院，2012

内容：人工智能应用程序在组织和社会中的普及推动了有关解释人工智能决策的研究。可解释的人工智能（XAI）领域正在以多种方式提取信息并可视化人工智能技术的输出（例如深度神经网络）而迅速扩展。但是，我们对 XAI 研究如何满足

可解释人工智能的需求了解有限。

本文作者对 XAI 文献进行了系统的回顾，重点关注与 XAI 有关的目的、定义和行为，并确定了关于 XAI 如何解决黑匣子问题的四个主题辩论。其次，作者从社会技术角度评估辩论，确定了两种未来的研究途径：a）利益相关者方法的必要性，并认识到不同利益相关者具有不同的可解释性需求；b）对可解释性需要整体看法和共同考虑社会问题以及 XAI 的技术、过程和结果方面，以及事实和讲故事方面。他们认为，要推进 XAI 的理论和实践，信息系统（IS）领域需要进行经验研究，以显示不同的 XAI 框架如何满足不同的利益相关者需求。

最后作者基于对 XAI 奖惩机制的这一批判性分析，作者将这些发现综合到未来的研究议程中，以进一步发展 XAI 知识体系。

其中 4 个主题介绍

在本节中，我们介绍了通过分析文章库而出现的四个主题辩论。

1） 激发对 xAI 的需求

探索有关 xAI 的最新文献和该技术的目的，我们观察到关于 xAI 定义的概念差异。

基本概念有各种解释，例如解释与解释及其相关概念。 一些研究人员可以互换使用这两个术语，而另一些研究人员则描述了两个概念块之间的差异。

Miller 阐明了如何将社会科学中的解释视为两步过程，包括：a）认知过程，描述事件原因的解释，其中选择了原因的子集作为解释（解释），以及 b）以交互方式在解释者和被解释者之间传递知识的社会过程。而 Brandão 等的立场是将“好的解释”描述为一种解释，其中解释者理解了解释对提出要求的人的意义，因为他们强调有必要调查其对开发商和其他研究人员的意义的社会意义。

正如布赖恩（Brian）和科顿（Cotton）指出的那样，可解释性和可解释性的术语（及其变体）相互交织，而且在其定义中仍然很混乱。“解释与可解释性的概念密切相关：如果系统的操作可以被人类理解，则系统可以解释，通过内省或通过详尽的解释。”

其他学者，则采取更为务实的观点，认为“解释”更接近于模型的发展，并且与“黑匣子”模型相反，在黑匣子模型中，人们寻求对机制的直接理解。模型的工作原理是可解释的机器学习的目标。

其他人则将解释定义为向人类解释或以可理解的术语呈现的手段，并以人类如何解释信息的方式指导研究。廖等人考虑到不同的用户需求，主张采用更加多样化的 xAI 方法。 他们将 xAI 描述为“……举一个例子，解释 ML 分类器做出的预测的最流行的方法之一，因为许多 XAI 算法都在努力做到这一点，它列出了对模型的预测有最大权重的特征”对开发人员而言，这可能具有很高的价值，但对于普通的外行而言却不然。

这些不同的定义表明，在 xAI 领域需要进一步的概念对齐。 以下各节介绍了 xAI 系统的主要驱动程序。

1.1） 产生信任，透明和理解。

产生信任是 xAI 的主要推动力，并且与透明度密切相关。DARPA 的 XAI 计划促进了对 xAI 的需求，因为我们需要进一步了解，信任和管理新兴的人工智能机器。沿着这些思路，进行了大量的工作和研究，重点是从模型中提取信息或构建更简单的模型，以期实现透明，理解并从而建立对模型的信任。Gilpin 等认为：“...能够概括神经网络行为原因，获得用户信任或产生有关其决策原因的见解的模型...”。与 DARPA 一起，机器学习性能和使用的普遍增长促使人们寻求对模型

的更好理解，以增加信任度，从而在业界增加机器学习的使用。此外，米勒认为，两种互补的方法将产生更加透明，可解释和可解释的系统，从而使我们更加有能力理解和信任模型：1）可解释性和可解释性被理解为人类对解释的理解程度在给定的上下文中；以及 2）对人（目标受众）的预测（决策）的解释。多数技术 xAI 方法旨在从模型（可能是神经网络或随机森林）中提取信息，例如特征重要性，相对重要性得分，敏感性分析，规则提取或其他方法以产生更大的透明度。这些 xAI 方法和框架主要是从透明性的感知出发，可以提高理解度，从而增加信任度 - 或相反，“黑匣子”模型不可信任。

很少有论文在所介绍的技术模型中包含社会技术方面的内容。然而，很少有人能解决利益相关者理解输出的障碍，其中包括将输出作为解释的社会技术方面的考虑，HCI 困境以及解决由开发者（庇护犯人）创建的为开发者提供解释的风险。例如，Zang 和 Zhu 提出了一种图形逻辑（或符号逻辑）来简化对卷积神经网络（CNN）的理解，而不仅仅是信息提取。而穆勒等。可视化用于确定狼的沙哑以通过 LIME 测试参与者的像素。通过这种方式，他们通过测试参与者对他们是否信任该算法的解释来测试对人类理解的需求。此外，文献强调，生成解释的 xAI 框架是由开发人员或技术人员构建的，专注于提取数据中的计算问题，这不一定能解决信任问题。

但是，许多概念性论文呼吁进行跨学科研究，并讨论了需要更多关注人类理解或可解释性的问题，而不仅仅是透明度。

1.2） 确保合规，遵守法规和 GDPR 法律。

对新法规和 GDPR 法律的众多反应之一就是要求 xAI 不仅向用户提供解释，而且向整个社会提供解释。这以及其他法规，使得从业者和行业迫切需要加大投资以

解释不透明的模型。GDPR 法规和“解释权”在研究和行业中引起了极大的轰动，将它们引向 xAI-作为合规性的可能解决方案。此外，一些研究者主张对 xAI 本身进行监管，或者为确保 xAI 的负责任使用而制定标准或质量措施的可能性，并避免建立有说服力的模型，而不是可以解释的模型。在 Gosiewska 和 Biecek 的实例中，很好地描述了构建有说服力的解释的谬误，其中示例是可加性模型如何导致对实例级别的解释产生误导性的指导，这一点得到了 Rudin 的支持，鲁丁反对最新的构建趋势（添加）可解释的事后“误导”解释。

1.3）为了履行社会责任，公平和规避风险。

特别是在医疗保健，临床和司法工作中，风险和责任是一个主要问题，因为它们潜在地影响着人们的生活，而不仅仅是成本效益分析。 将责任分配给各个专业人士可以避免风险。因此，为专家（例如临床）推理开发心理模型，以更好地理解深度神经网络和不透明模型背后的推理。此外，最近在不透明模型中出现的歧视和累犯事件引发了关于确保模型性能的公平性以及对模型构建方式的更深入了解的辩论。在 xAI 文献中，招聘过程中的少数群体案例，COMPAS 系统中的累犯和普遍公平都在增加。

1.4） 建立负责、可靠和合理的模型进行论证

对 xAI 产生巨大吸引力的一个主题是，通过审查模型或创建其合法性的证据来确保模型的公平性和公正性。Adadi 和 Berrada[21]生成了一种可证明的方法，用以捍卫算法决策的公平性和道德性。除此之外，Abdul 等人提出了一种更新颖的生成 xAI 的方法，即建立基于因果关系概念的反事实解释。Liao 和 Anderson 在形式论证的基础上提出了生成基于论证的理由和解释的方法，这些方法为模型的更好推理提供了依据。最后，Ananny 和 Crawford 提出了一个关于透明度不足以

管理和追究算法责任的讨论。他们声称，透明度不一定会建立信任，因为不同的利益相关者对系统的信任程度不同，这取决于他们对信息披露的时间和内容、以及信息的准确性和相关性的信任度。

1.5) 尽量减少模型性能和解释中的偏差和误解

模型的偏差和表现是 xAI 的一个重要驱动力，因为媒体经常报道模型的表现和性能不如人类，例如，在招聘过程中把合适的候选人筛选掉，或者未能识别出有色人种。特别是在用训练集数据训练神经网络模型时，有偏差的训练数据会成为影响模型输出有效性的重大问题。除了有偏差的训练数据、变量选择和表征之外，我们自身的认知偏差还会阻碍我们对模型可视化输出的解释，因为我们往往会过度简化信息。

1.6) 能够验证模型并且验证 xAI 生成的解释。

针对有偏差的模型和模型不达标的表现，研究人员提出了四种深度神经网络（DNN）的评估方法，分别是：（1）与原始模型相比的完整性；（2）替代任务的完整性；（3）检测模型偏差的能力；（4）人类评估。其他研究人员则提出了一种用于评估可解释性的完全分类法，其中成本最高的是基于应用程序的方法，该方法需要对已实现的解释进行测试，并最终让用户对其进行测试。此外，他们还提出了一种以人为本的评估，例如在时间限制下哪种类型的解释是最好的。

2)完整性与可解释性困境

从评估XAI的争论中，出现了关于是否能够做出正确解释的争论。研究人员认为，对可解释性的需求源于不完全性会产生不同的偏见，并认为用户专业知识的性质将影响解释可能包含的复杂程度。许多其他研究人员主张在完整性和解释性之间的权衡取舍。我们应该谨慎对待这种折衷，因为人类对简单描述有强烈的特定偏

见，这可能导致研究人员创建有说服力的系统而不是透明的系统。当健壮性较低时，人们会失去对解释的信任。

3)人的解释

关于如何解释和解释 AI 行为的许多研究是由构建 AI 的人而不是使用 AI 的人的需求所驱动的。用户可能具有的不同 AI 素养水平，而涉及利益相关者的多样性及其对 XAI 的不同需求的论文甚至更少。尽管有不同水平的 AI 素养和不同的学科领域，研究人员仍致力于开发以用户为中心的概念框架。只有少数几篇论文讨论了关于 XAI 生态系统的不同类型的角色和利益相关者，并认为一种解决方案可能不适合所有不同类型用户的目的，但我们需要包括利益相关者的背景，背景和知识，产生可以理解的解释。

4)技术产生了 XAI

近年来，在寻找打开臭名昭著的黑匣子的过程中，已经提出了许多不同的方法来构建更透明以及可解释的模型。可以分类如下：

本质透明：ML 模型具有更简单的特征，但不如其他更高级的模型（线性回归，逻辑回归，决策树）更精确

与模型无关的 XAI 框架：这些通常具有事后特征，这意味着它们旨在适应任何模型类型，并依赖简化模型的技术，显示特征相关性估计，可视化模型或生成输出的本地替代模型。 这些框架的共同点是它们产生某种视觉输出，以便于理解。

**25-579，**标题：在微博上下文中产生可解释模型的歧视性表达

(Discriminatory Expressions to Produce Interpretable Models in Microblogging Context)

作者：Manuel Francisco, Juan Luis Castro

机构：西班牙格拉纳达大学计算机科学与人工智能系（Department of Computer Science and Artificial Intelligence, University of Granada, Spain ）

内容：社交网站（SNS）是最重要的交流方式之一。特别是，由于微博站点的特殊性（及时性，简短文本等），它们被用作分析途径。有无数的研究以新颖的方式使用 SNS，但是机器学习（ML）主要集中在分类性能上，而不是可解释性和/或其他优势度量标准上。因此，最先进的模型是黑匣子，不应用于解决可能产生社会影响的问题。当问题需要透明时，有必要建立可解释的管道。可以说，管道中最具决定性的组成部分是分类器，但这并不是我们唯一需要考虑的事情。尽管分类器可能是可解释的，但生成的模型过于复杂以至于无法理解，因此人类无法理解实际的决策。本文的目的是提出一种功能选择机制（该流程的第一步），该机制可以通过使用较少但更有意义的功能来提高可理解性，同时在要求可解释性的微博环境中实现良好的性能。此外，本文提出了一种根据统计相关性和偏倚来评估特征的排名方法。为了评估模型的分类性能，泛化能力和实际可解释性，实验小组对五个不同的数据集进行了详尽的测试。实验的结果表明，就准确性，概括性和可理解性而言，本文的建议是更好的，并且到目前为止是最稳定的。

**26-583，**研发可解释的机器学习

——COPU 志愿者

打破机器学习的“黑盒子”研发可解释的人工智能，已经成为世界（尤其是美欧）当前的一大亮点。

所谓机器学习或深度学习技术一般是不可解释的，或用以作业是不透明的。不可解释的机器学习也是一种人工智能（初级阶段的人工智能），在用以作业（以提高智能）或对之训练（以提高算力）时，由于其带有“黑盒子”、盲操作的缺陷，将致使作业或训练成果较差（或较弱），这时的不可解释机器学习也称为弱人工智能（这是早期人工智能ノ。只有实现了可解释的机器学习（或给予机器学习模型以解释的能力），才有可能达致强人工智能。

为了打破机器学习中的“黑盒子”，导致使用机器学习作业及其训练透明化，变其不可解释为可解释，需要针对不同应用场景研发出各种不同、适用的可解释工具（包）或算法。使这些不同任务的解决方案或训练成果分别达到不同的目标：

①提高能效或绩效，提高性能或质量，

②提高判断、决策能力（或减低投资风险），

③开辟公平、信任或合规、遵法以及明责、打假等的新场

27-585，基于时间序列任务深度学习模型的可解释人工智能技术实证研究

（An empirical study explain able AI techniques on deep learning models for time series tasks）——UdoSchlegel，DanielaOelke，DanielA-keim，MennatallahEl-Assady

康斯坦茨大学，奥芬堡应用科技大学

2020.12.08 发表

机器学习黑盒子模型的决策解释通常是通过应用可解释 AI（XAI）技术生成的。但是许多建议的 XAI 方法的结果通常是未验证白输出。通常是通过人类手工对单

个图像或文本进行视觉解释来实现对模型解释的评估和验证。本文作者提出了一项经验研究理论和基准框架，用来将归因方法应用于为时间序列上的图像和文本数据开发的神经网络。这种方法可以使用扰动方法来自动确定时间序列的归因技术并识别其可靠的方法。

**28-586，**可解释人工智能：训练数据的子集如何影响预测

Explainable Arificial intelligence: How sunsets of the Training Data Aftect aprediction——AndreasBrandsaeter，IngridK.Glad

2020.12.7 发表

各种应用领域中，对机器学习模型和预测的解释与说明的兴趣及需求日益增长。本文考虑已经被开发、实施和训练的数据驱动模型，其目标是解释模型并解释和理解其预测。

由于数据驱动模型所做的预测严重依赖用于训练的数据，因此作者认为解释应传达有关训练数据如何影响预测的信息，为此提出一种称为 shapley 值，用于衡量训练数据子集的重要性。shapley 价值概念源自博弈理论，其发展目的是一组合作的参与者之间公平地分配支出。

作者描述和说明所提的方法是有用的，并在几个示例上证明其功能。从而展示了如何将提出的解释用以揭示模型和错误的训练数据中的偏差。

**29-590，**一个基于显性和隐性交互的广义相加模型的可解释性推荐系统

——YifengGuo，YuSu，ZebinYang，AijunZhang

香港大学保险学院统计系，深圳索信达控股公司

在过去的几十年里，预测用户喜好的推荐系统被广泛应用于多个领域，如电子商务、社交媒体、银行等。本文从统计建模的角度，基于显性交互和隐性交互的广义相加模型（Generalized Additive Model，GAM），提出一个可解释性推荐系统GAMMLI。

该系统可捕捉观察到特征的主要影响和显性交互，并发现未观测到特征的隐性交互，所有影响都可以用可视化的方式来解释。

与传统协同的过滤方法不同，GAMMLI 考虑了用户和项目的群体效应，这有利于提高模型的可解释性，也有利于冷启动推荐问题（Cold - start recommendation problem）。实验结果表明，GAMMLI 在预测性能和可解释性方面都具有优势。

（注：冷启动问题，如何在没有大量用户数据的情况下设计个性化推荐系统并让用户对推荐结果满意从而愿意使用推荐系统）

**30-591，**XAI-P-I：从实践到理论的可解释人工智能

（XAI-P-I：A Brief Review of Explainable Artificial Intelligence view of Explainable artificial intelligence from Practice to Theory）

——NazaninFouladgar，KaryFramling

2020.12.17 发表

随着新模型的出现，机器学习已经在不同应用中得到了巨大的应用，并随着时间的推移，机器学习也在不断地发展。然而这些模型的复杂行为阻碍了人们简洁地理解如何做出特定的决策。此限制要求机器通过解释来提供透明性。因此可以将“里盒子”分配给机器学习模型以做出决策，而将“白盒子”分配给这些模型的解释版本以“解释 AI”的主题。

问题表明，解释可以以因果关系和非因果关系的形式来说明。尽管因果关系已在AI 研究人员中解决算法决策问题变得更加普及，但非因果关系最近吸引了人机交互领域的学者。实际上，XAI 尚未成熟，并且有很多开放的空间可以提传实践和理论上的解释。尽管解决这些问题十分关键，从实践和理论上仍然缺乏对 XAI 领域目前的现状作简明的理解。

本文首先关注黑盒子解释的类别并给出一个实际的例子，然后讨论如何以多学科为基础讨论理论解释，最后提出未来工作的一些方向。

## 31-602，人工智能面对的一些挑战

——姚期智（图灵奖获得者、中科院院士、清华大学交叉信息研究院院长）在《2020浦江创新论坛》上的讲演

人工智能有三大技术瓶颈亟待突破，需要科学家“从 0 到 1”的原创研究。

脆弱性和不稳定性是人工智能面临的第一大技术瓶颈。

人眼识别十分稳定，一个图像如有微小改变，人仍能一眼看出它是什么，然而人工智能在图像识别方面有点“人工智障”，比如将一只小猪的照片加入一些图像“杂音”，一些机器视觉系统居然会把它识别为飞机。“小猪变飞机”这种漏洞给人工智能应用带来安全隐患，比如黑客可以攻击汽车自动驾驶系统，诱导它将马路上的“停止”标识当作“通行”，从而引发交通事故。

第二大技术瓶颈是机器学习算法缺乏可解释性，很多算法处于“黑盒子”状态。例如一个科研团队开发了一个房地产估价系统，这个系统通过一套算法学习了有关各地房地产价格的大数据，从而能自动评估房地产价格。然而，这套算法像黑盒子一样，很难给出估价的完整依据。这在商业应用上是一块很大的短板，房地

产商会怀疑：人工智能系统是否低估了价格？是不是有利益相关方对系统做了手脚，故意压价？因此，算法的可解释性问题亟待科研突破，否则会大幅限制人工智能的商业应用的进程。

第三大技术瓶颈是人工智能的对抗性较弱。

如今，一个无人机群可以轻松完成灯光秀、农林作业等任务，但要看到，这些任务都是在自然环境下完成的，如果是处在高对抗的人为环境中呢？比如在电子竞技和军事战斗中，无人机群的协同作战能力的强化学习、博弈论研究，让无人机群能够在高对抗环境中自主找到最优策略。

**32-603，**可解释的抽象训练数据集

（ExplainableAbstractTrainsDataset）

葡萄牙里斯本 NOVA 大学

ManueldesousaRibeiro，LudwigKrippahl

内容：可解释的抽象训练数据集是一个包含训练简化表示的图像数据集。它旨在为证明和解释提取算法的应用和研究提供一个平台。该数据集随附一个本体，该本体基于它们的视觉特征对所描绘的训练进行概念化和分类，从而可以精确地了解每个训练的标记方式。数据集中的每个图像都用描述训练特征的多个属性和训练元素的边框标注。

**33-608，**IBM 向 LFAI 捐赠 AIX360 以助力可解释 AI 实践

——IBM 田忠博士

早期的 AI 实践往往具有自解释的特点，因为那时使用的是规则库、决策树、抉

择表等这类比较直观的技术。目前机器学习/深度学习日益普及，但这类技术的模型对于用户而言往往是黑盒子，需要所谓“事后分析解释”（Post-hoc-interpretation）来帮助用户打破黑盒子，建立对于该系统决策的合理信心。

“事后分析解释”既可以用来理解所使用的数据，也可以用来理解所训练出的模型。对于前者，可以采用 DIP-VAE 算法以提取最有效特征，可以采用案例式推理算法 ProtoDash 建立典型案例。对于后者，可分为全局解释和局部解释。全局解释指的是向用户展示该系统的整体预期决策模型，以帮助用户理解系统决策的合理性，局部解释就特定案例进行分析，找出影响模型做出该结论的关键要素。

目前，可解释 AI 的发展能够就特定问题向用户呈现有意义的解释。

AIX360 是对于这些训练数据和模型建事后分析解释的工具包。研究人员给出了一个银行信贷决策系统的可解释实践。该 AI 系统基于全美公共真实数据（FICOHELOCDataset）来辅助决策是否批准一项贷款申请。项目建设团队需要在验收阶段向银行主管解释本旁统的决策效果，也就是要对所训练出的模型做出全局性的直观解释。因此，他们选择了 BRCG 和 GLRM 相互补充的规则生成算法，他们从当前的训练数据集上建立真值表，从而建立如下规则：

“没有至少 5 个户头或者拥有多过 5 个户头但负债超过 1000 美元的客户是高风险客户”

信贷经理关心的是这个系统给出的决策建议是否有系统谈歧视。他们用手边的正反案例来研究系统决策的合理性。ProtoDash 算法可以满足要求。经验算信贷经

理发现：能获得贷款的申请者过半数的特征值都是接近的，不能获得贷款的申请者其半数特征也是接近的。这给了他很大信心。同时他还发现未来能获得贷款的申请者大多有轻微违法犯罪记录，这可能有过于严苛并有失公允，要额外小心处理。

34-609，深度学习中的理论问题

（这些理论有助于人们对人工智能这个黑盒子的理解）

MIT 研究团队 TomasoPoggio，AndrzejBanburski，QianliLiao

虽然深度学习在许多应用程序中都是成功的，但理论上尚不十分了解。深度学习的理论表征应回答有关近似/优化动力学和样本外性能。

本文阐述在近似方面做出的工作，对深层/浅层架构近似能力的对比，训练神经网络的优化过程，尤其是梯度流对应动力学系统的属性，对复杂性控制方面也进行了理论表述，如对于特定类型的复合函数，卷积深度网络可以避免维数灾难。这些理论有助于人们对于人工智能这个黑盒子的玩理解。

35-610，语义和解释：为什么反事实的解释会在深度神经网络中产生对抗性示例

（Semantics and explanation: Why counterfactual explanations produce Adversarial examples in deep neural network s）

KieranBrowne，Ben Swift research school of Humanities &the Arts Australian National University

如果不首先解决语义的稀缺性，将无法解释深度神经网络。已经存在计算方法来产生模型不可知的解释。当将这些方法应用于DNN常见的模棱两可或低级表示时，

它们根本不能作为解释。这不仅仅是对现有解释方法的限制，而是没有语义就不可能有任何解释。由于深度学习通常对“原始数据”进行操作，几乎没有语义内容（例如像素和字符），因此这种实现有助于阐明可解释性的挑战，我们要么找到一种方法来提取假定存在于网络的隐藏层中的语义，要么承认失败。

最近在可解释人工智能（XAI）中的论文已经确定了该领域的理论基础存在的问题。该领域通常仅凭直觉来解释什么是解释，与人类解释和理解解释的方式不同，有人建议 XAI 根据心理学和社会科学研究的一系列原则采用“日常解释”。有的学者另外指定一种产生反事实解释的方法。认为反事实的解释、表演，与模型无关，可自动计算，并且对外行人员易于理解。作者认为，这些反事实的解释为向任何人解释复杂的算法系统提供了途径。但是，自 2014 年以来，等效计算已用于深度学习研究中，尽管没有给出解释。相反，在深度学习研究的背景下，反事实计算会产生“对抗性示例”，潜移默化地修改了输入，导致网络莫名其妙地分类错误。

这应该让我们思考：同一方法怎么可能一方面代表一种有前途的新方法向任何人解释深度神经网络的决策，另一方面又代表同一决策过程中令人困惑的脆弱性？我们称这种现象为解释性分歧。作者认为这种鸿沟揭示了 XAI 研究在语义上的盲点。

**36-633，**可解释机器学习/深度学习求解程序

一、审题：

首先区分有待求解的可解释人工智能案例是属于整体解释行为，还是局部（个别样本）解释行为？因为不同解释行为将采用不同求解方法。

谈到整体解决行为，属于宏观状态，如全国、行业、地区的规划性或整体性项目，而局部解释行为，属于微观状态，如企业或个体性质的项目。

在审题后，1，按整体或局部解释行为区分，采用不同的求解方法；2，按不同项目内容，开展可解释人工智能的分析、计算程序。

二，分析：

可解释人工智能多用于放贷风险评估项目、或其他决策项目，或反欺诈打假项目，及评定公平、正义或认信项目。

对审议中的不同项目内容，如属发放风险贷款而需进行风险评估的项目；又如属于反欺诈或打假（fake），要求识别假伪欺诈的项目。

分析步骤为：

1）取特征值或特征函数

如打假项目：

取技术特征识别率、环境干扰影响率、用户自身信用率等。

如放贷项目：

取回款、用户信用、担保等各项特征值。

2）采用算法

根据各项特征值及单项算法，计算综合特征风险率或综合特征欺诈率。

三，评估：

评估综合放贷风险率，确定风险率可承担红线：在红线以下，表示风险可承担，在红线以上，表示风险不可承担，或也可将风险是不可承担在红线以上分级处之。

评估综合欺诈率，确定风险可承担红线也是如此！

37-638，谷歌开发语言可解释性工具 LIT

——Jeff Dean

为了更好地理解语言模型的行为，谷歌 2020 年开发了语言可解释性工具(LIT)，这是一个可以更好地解释语言模型的工具包，使得交互式探索和分析语言模型的决策成为可能。

38-641，可解释人工智能：工程观点

FatimaHussain 等 SMIEEE，EkramHossainFIEEE

2021.1.10

深度学习的显著进步激发了人们在几乎每个领域使用人工智能技术的热情。但是，这些算法的不透明性，使其在安全关键型系统中的应用产生质疑。“可解释性”维度不仅对于解释黑盒子算法的内部工作至关重要，而且还增加了问责制和透明性维度，这对于监管机构、消费者和服务提供商至关重要。可解释人工智能（XAI）是将所谓的黑盒子算法转换为白盒子算法的一组技术和方法，这些算法获得的结果以及该算法采取的变量、参数和步骤达到所获得的结果，是透啊且可解释的。以自动驾驶汽车为例，讨论 XAI 在其不同组件（例如对象检测、感知、控制、动作决策等）中的应用。这项工作是一项探索性研究，旨在确定 XAI 领域新研究途径。尽管 XAI 在基线 AI 模型中添加了必要的功能，但评估与解释相关的其他功能的 AI 模型的性能也非常重要。必须考虑新功能可能会产生开销并可能影响准确性。

39-642，可解释人工智能和人类互动

美国陆军研究实验室和南加州大学创意技术研究所合作发布

如果人类无法理解同上下文进行有效交互，人工智能的优势无法体现，基于可解释人工智能（XAI）允许人工智能解释决策就可以缓解这种问题。

本文利用一些需求案例，探索对运营计划人员（高级分析师）被虚拟分析师代替，它们通过分析、搜索和呈现目标系统来实现需求。

通过将 XAI 集成到现实世界当中的双向工作流，能够创建计划并简要说明指令，一旦出现问题，就可以进行分析、实现理解。通过分析员共同努力，对威胁、脆弱性、事件等以应对未来可能出现的改击。

**40-644，**可解释性机器学习可信教程

上交大张拳石在 IJCAI2020 会上介绍可解释性机器学习可信教程（2021.1.11）

深度神经网络（DNN）无疑为计算机视觉、计算机语言学和人工智能的广泛应用带来了巨大的成功。但是，DNN 成功的可信赖性以及 DNN 对对抗攻击的适应力仍然很大。在可解释的人工智能范围内，对引人注目的但有争议的话题。相关问题包括①网络特征可信度的量化，②DNN 解释的客观性、鲁棒性，语义严格性，以及③可解释神经网络的可解释性的语义严格性等。

教程旨在将关注人工智能的可解释性、安全性和可靠性的研究人员、工程师以及工业从业人员。

本教程对当前可解释 AI 算法的优势和局限性的批判性讨论提供了新的前瞻性研究方向。预计本教程将对关键的工业应用产生深远的影响，例如医学诊断、金融和自动驾驶。

**41-645，**如何看待人工智能的发展方向？谷歌作为全球人工智能发展重镇之一，其看法和做法值得我们借鉴。现将谷歌高级副总裁 JeffDean 代表 Google Research 最近发文的总结(谷歌人工智能2020年的发展与成就,同时也展望2021年接下来的工作重点）简述如下。

他列出2020年谷歌人工智能在10大领域的发展与成就：

1) 新冠病毒和健康

基于机器学习算法，研究对新冠病毒的调研、检测、预防、诊治，以及研究新冠病毒流行对健康与经济的影响。

2) 天气、环境与气候变化

基于机器学习算法，研究天气、环境与气候变化。

3) 可访问性（Accessibility）

例如采用机器学习方法，帮助视力受损用户识别包装食品。

4) 机器学习在其他领域的应用

此处以脑机接口举例，研究机器学习与神经科学联系，研究果蝇的脑织织如何运作。

5) 负责任的人工智能

可解释的机器学习，谷歌开发语言可解释性工具（LIT)，这是更好地解释语言模型的工具包，使交互式探索和分析语言模型的决策成为可能。

6) 自然语言理解

7) 机器学习算法（深入研究）

8) 强化学习

9) AutoML

这是一个非常活跃的研究领域，将系统地重塑机器学习算法

10)更好地理解机器学习和模型

随着神经网络被做得更宽更深，训练得更好，泛化得更好的背景下

在上述 10 项工作中，机器学习是其重点，大多数项目都是基于机器学习的，对机器学习的算法与模型要加深理解、改进、创新。

2021 年谷歌人工智能的工作目标：一是以关键应用场景为导向，二是考虑 2020 年工作的延续，三是要对机器学习加深理解积极运用，四是根据人工智能国际形势发展有所创新。

**42-646，**IBM 向 LF AI 捐赠 AIX360 项目，助力可解释 AI 实践

人工智能的普遍应用需要达成知其然，也知其所以然。这是可解释 AI 的使命。早期的 AI 实践往往具有自解释的特点，因为那时使用的是规则库、决策树、抉择表等这类比较直观的技术。目前机器学习、深度学习日益普及，但这类技术的模型对于用户而言往往是黑盒子，需要通过“事后分析解释”(post-hoc interpretation)来帮助用户打破黑盒子、建立对于该系统决策的合理信心。

所谓“事后分析解释”既可以用来理解所使用数据，也可以用来理解使用特定数据所训练出的模型。对于前者,可以采用DIP-VAE算法以提取哪些是最有效特征，也可以采用案例式推理算法 ProtoDash 建立典型案例。对于后者，可分为全局解释和局部解释。全局解释指的是向用户展示该系统的整体预期决策模型，从而帮助用户理解系统决策的整体合理性。局部解释就特定案例进行分析，找出影响该模型做出该项结论的关键要素。

AIX360 是对于这些训练数据和模型建立事后分析解释的工具包。IBM 的研究人员

在http://aix360.mybluemix.net上给出了一个银行信贷决策系统的可解释实践。该AI系统基于美国的公开、真实的个人金融数据集(FICO HELOC Dataset)来辅助决策是否批准某项贷款申请。

本项目的建设团队需要在验收阶段向银行主管解释本系统的决策效果，也就是要对所训练出的模型做出全局性的直观解释。因此，他们选择了BRCG和GLRM相互补充的规则生成算法，他们从当前的训练数据集上建立真值表，从而建立如下的规则：

“拥有少于5个户头或者拥有多过5个户头且每个斧头负债超过1000美元的客户是高风险客户”银行的主管当局依据自己多年的工作经验，十分认同这样的规则，从而认可本系统所能给出的贷款决策建议。

银行信贷经理关心的是这个系统给出的决策建议是否有系统性歧视（比如种族、肤色、年龄、性别、价值观等)。为此，他要用手边的正反案例来研究系统决策的合理性。ProtoDash算法可以满足要求。经过演算分析，信贷经理发现：能获得贷款的申请者过半数的特征值都是接近的；不能获得贷款的申请者其半数特征也是接近的。这给了他很大的信心。通过分析这些共同特性，他发现未能获得贷款的申请者大多有轻微违法犯罪记录。这可能有过于严苛并有失公允，要额外小心处理。

要亲自上手试验，请参考AIX360网站http://aix360.mybluemix.net

**43-647，**可解释AI的算法选择和执行步骤

当前，采用机器学习、深度学习进行AI决策的系统往往有黑盒子特点。可解释AI通过“事后分析解释”(post-hoc interpretation)来帮助用户打破黑盒子、

建立对于该系统决策的合理信心。不同的人在不同的场合、不同的背景下，对于可解释 AI 有着不同的需求。

AI 系统的开发者，其目的多半是如何提高系统效率。AI 系统的使用者，则是需要建立对于这个系统所做决策的信心，能够放心、安心采纳其建议。而市场或者业务的主管乃至监管当局主要关心的是如果确保系统性的公平。最终受影响的用户则需要能够理解影响结论的主、次因素，从而未来能够有所作为。

下面的决策树概括了当前可解释 AI 的实践现状。

目前的可解释AI可以胜任事后分析解释,还不能完成交互型探究式的分步解释。不同的数据形式（图片、表格、文本）、不同的解释范畴（全局还是局部、特征分析还是案例分析）需要不同的解释算法。

要解释一个 AI 系统，从算法角度，大致分为三个步骤：

- 获取并加载、整理数据
- 依据需求选择、运行算法
- 整理并显示可理解结论

例如，一个银行贷款决策系统开发团队的项目经理要向甲方说明系统的功效，即无需经年累月的培训和项目经验积累，本系统能帮助贷款部的任何工作人员依照所积累的数据及时做出和资深经理依据多年经验同样的贷款决定。

项目经理选择了采用 BRCG 算法以生成一组布尔规则表，使用 GLRM 算法生成逻辑规则回归模型。为此，他按如下步骤展开工作：

获取和加载、整理数据

| | 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 |

选择和运行算法

1）BRCG

```
# Instantiate BRCG with small complexity penalty and large beam search width
from aix360.algorithms.rbm import BooleanRuleCG
br = BooleanRuleCG(lambda0=1e-3, lambda1=1e-3, CNF=True)

# Train, print, and evaluate model
br.fit(dfTrain, yTrain)
from sklearn.metrics import accuracy_score
print('Training accuracy:', accuracy_score(yTrain, br.predict(dfTrain)))
print('Test accuracy:', accuracy_score(yTest, br.predict(dfTest)))
print('Predict Y=0 if ANY of the following rules are satisfied, otherwise Y=1:')
print(br.explain()['rules'])
```

```
Learning CNF rule with complexity parameters lambda0=0.001, lambda1=0.001
Initial LP solved
Iteration: 1, Objective: 0.2895
Iteration: 2, Objective: 0.2895
Iteration: 3, Objective: 0.2895
Iteration: 4, Objective: 0.2895
Iteration: 5, Objective: 0.2864
Iteration: 6, Objective: 0.2864
Iteration: 7, Objective: 0.2864
Training accuracy: 0.719573146021883
Test accuracy: 0.696515397082658
Predict Y=0 if ANY of the following rules are satisfied, otherwise Y=1:
```

2）LogRR

```
# Instantiate LRR with good complexity penalties and numerical features
from aix360.algorithms.rbm import LogisticRuleRegression
lrr = LogisticRuleRegression(lambda0=0.005, lambda1=0.001, useOrd=True)

# Train, print, and evaluate model
lrr.fit(dfTrain, yTrain, dfTrainStd)
print('Training accuracy:', accuracy_score(yTrain, lrr.predict(dfTrain, dfTrainStd)))
print('Test accuracy:', accuracy_score(yTest, lrr.predict(dfTest, dfTestStd)))
print('Probability of Y=1 is predicted as logistic(z) = 1 / (1 + exp(-z))')
print('where z is a linear combination of the following rules/numerical features:')
lrr.explain()
```

```
Training accuracy: 0.7425368094015 94
Test accuracy: 0.7260940032414911
Probability of Y=1 is predicted as logistic(z) = 1 / (1 + exp(-z))
where z is a linear combination of the following rules/numerical features:
```

| | rule/numerical feature | coefficient |
|---|---|---|
| 0 | (intercept) | -0.0686341 |
| 1 | MSinceMostRecentInqexcl7days > 0.00 | 0.680261 |
| 2 | ExternalRiskEstimate | 0.654248 |
| 3 | NetFractionRevolvingBurden | -0.553965 |
| 4 | NumSatisfactoryTrades | 0.551654 |
| 5 | NumInqLast6M | -0.463226 |
| 6 | NumBank2NatlTradesWHighUtilization | -0.448331 |
| 7 | AverageMInFile <= 52.00 | -0.43436 |
| 8 | NumRevolvingTradesWBalance <= 5.00 | 0.42154 |
| 9 | MaxDelq2PublicRecLast12M <= 5.00 | -0.418142 |
| 10 | PercentInstallTrades > 50.00 | -0.317566 |
| 11 | NumSatisfactoryTrades <= 12.00 | -0.312471 |
| 12 | MSinceMostRecentDelq <= 21.00 | -0.301566 |
| 13 | PercentTradesNeverDelq <= 95.00 | -0.273924 |
| 14 | ExternalRiskEstimate > 75.00 | 0.263437 |
| 15 | AverageMInFile <= 84.00 | -0.182118 |
| 16 | PercentTradesNeverDelq | 0.166518 |
| 17 | AverageMInFile | 0.15069 |
| 18 | PercentInstallTrades > 42.00 | -0.148802 |
| 19 | NumBank2NatlTradesWHighUtilization <= 0.00 | 0.135396 |
| 20 | MSinceOldestTradeOpen <= 122.00 | -0.132409 |
| 21 | PercentTradesNeverDelq <= 91.00 | -0.11771 |
| 22 | NumSatisfactoryTrades <= 17.00 | -0.11022 |
| 23 | ExternalRiskEstimate > 72.00 | 0.107613 |

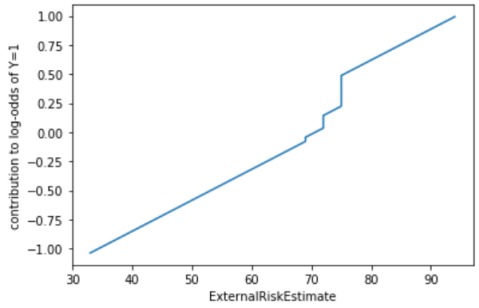
（图形）显示结论，如以 GAM 图示 LogRR 的结论

外部风险预估

**ExternalRiskEstimate**

As expected from the BRCG Boolean rule above, 'ExternalRiskEstimate' is an important feature positively correlated with good credit risk. The jumps in the plot indicate that applicants with above average 'ExternalRiskEstimate' (the mean is 72) get an additional boost.

```
lrr.visualize(data, fb, ['ExternalRiskEstimate']);
```

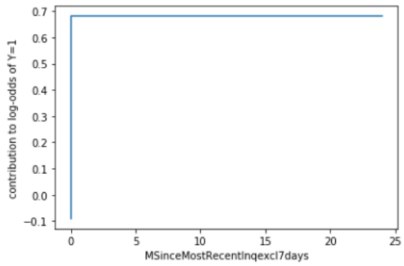



信用查询次数的影响

**Credit inquiries**

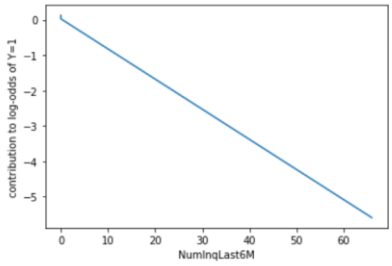
The next two plots illustrate the dependence on the applicant's credit inquiries. The first plot shows a significant penalty for having less than one month since the most recent inquiry ('MSinceMostRecentInqexcl7days' = 0).

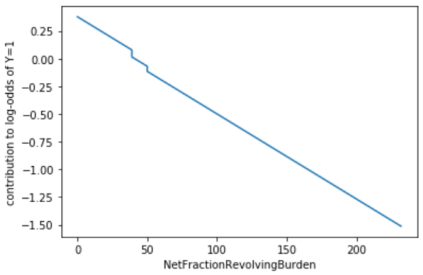
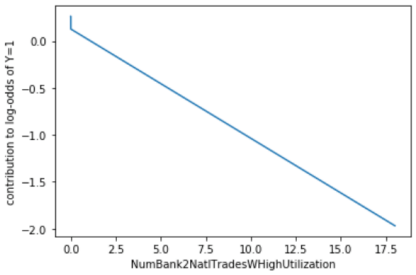
```
lrr.visualize(data, fb, ['MSinceMostRecentInqexcl7days']);
```




The second shows that predicted risk increases with the number of inquiries in the last six months ('NumInqLast6M').

```
lrr.visualize(data, fb, ['NumInqLast6M']);
```



| | |
|---|---|
| | 债务水平的影响<br><br>**Debt level**<br><br>The following four plots relate to the applicant's debt level. 'NetFractionRevolvingBurden' is the ratio of revolving debt (e.g. credit card) balance to credit limit, expressed as a percentage, and has a large negative impact on the probability of good credit. A small fraction of applicants (less than 1%) actually have NetFractionRevolvingBurden greater than 100%, i.e. more revolving debt than their credit limit. This might be investigated further by the data scientist.<br><br>`lrr.visualize(data, fb, ['NetFractionRevolvingBurden']);`<br><br>  <br><br><br>The second 'NumBank2NatlTradesWHighUtilization' plot shows that the number of accounts ("trades") with high utilization (high balance relative to credit limit for each account) also has a large impact, with a drop as soon as one account has high utilization.<br><br>`lrr.visualize(data, fb, ['NumBank2NatlTradesWHighUtilization']);`<br><br>  <br> |

以上图表的结论和资深经理的经验不谋而合，项目团队获得了甲方的认可。

相关算法的详情和代码、如何获取 FICO HELOC 数据集、亲自上手试验这些算法，请参考 AIX360 网站 http://aix360.mybluemix.net。

**44-648，**可解释人工智能（XAI）教程系列

AAAI 发布，2021.2.3，可参阅：

可解释人工智能只在通过象征性人工智能与传统机器学习优势结合来应对此类挑战。多年以来，各种不同的机器学习社区都以不同的定义、评价指标和动机、

结果等表现该主题。本教程主要是机器学习和 XAI 相关方法的介绍（分成五大部分）：

1) 从理论和应用的角度来描述和激发对可解释人工智能的需求（简介）

2) 可解释人工智能（不仅仅是机器学习）

3) 知识图在可解释人工智能中的应用

4) 可解释人工智能应用程序和课程

5) 可解释人工智能工具，编码实践和研究挑战

**45-661，**关于隐马尔可夫模型混合的可解释性

Towards interpretability of Mixtures of Hidden Markov Models

Negar Safinianaini,Henrik Boström

瑞典斯德哥尔摩 KTH 皇家技术学院  2021.3.23

隐性马尔可夫模型（MHMM）的混合通常用于顺序数据的聚类。与任何聚类方法一样,MHMM 的一个重要方面是它们的可解释性,从而可以从数据中获得新颖的见解。但是如果没有适当的衡量可解释性的方法，那么对新颖贡献的评估就很困难，并且实际上不可能设计出直接优化此特性的技术。在这项工作中，提出了一种用于 MHMM 的可解释性的信息理论测度（熵），并在此基础上找出了一种改进模型可解释性的新方法，即一种熵正则化期望最大化（EM）算法。新方法旨在减少 MHMM 中的马尔可夫链（涉及状态转移矩阵）的熵，即在聚类期间为常见状态转移分配更高的权重。有人认为，这种熵的减少通常会导致可解释性的提高，因为可以更容易地识别出群集中最有影响力和最重要的状态转换。一项实证研究表明，可以通过熵来改进 MHMM 的可解释性，而不必牺牲（但要提高）聚类性能和计算成本（分

别通过 v 度量和 EM 迭代次数来度量)。

**46-666，关于图神经网络的可解释性**

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji

华盛顿州立大学，2021.3.25

深度学习方法在许多人工智能任务中的表现越来越突出。深度模型的一个主要局限性是它们不易解释。这一限制可以通过研究解释技术来规避，从而产生了可解释性领域。近年来，图像和文本深层模型的可解释性取得了重大进展。在图形数据领域，图神经网络及其可解释性得到了迅速发展。本文提出对当前 GNN 解释方法的分类观点，阐明了现有方法的共性和差异，为进一步的方法发展奠定了基础。为了方便评估，本文生成了一组用于 GNN 解释性的基准图数据集。本文还总结了当前用于评估 GNN 解释性的数据集和度量。总之，本文为 GNN 解释性提供了一个统一的方法论处理和一个标准化的评估试验台。

对图神经网络可解释方法的分类如下附图所示，分为两大类：基于实例的方法和基于模型的方法。

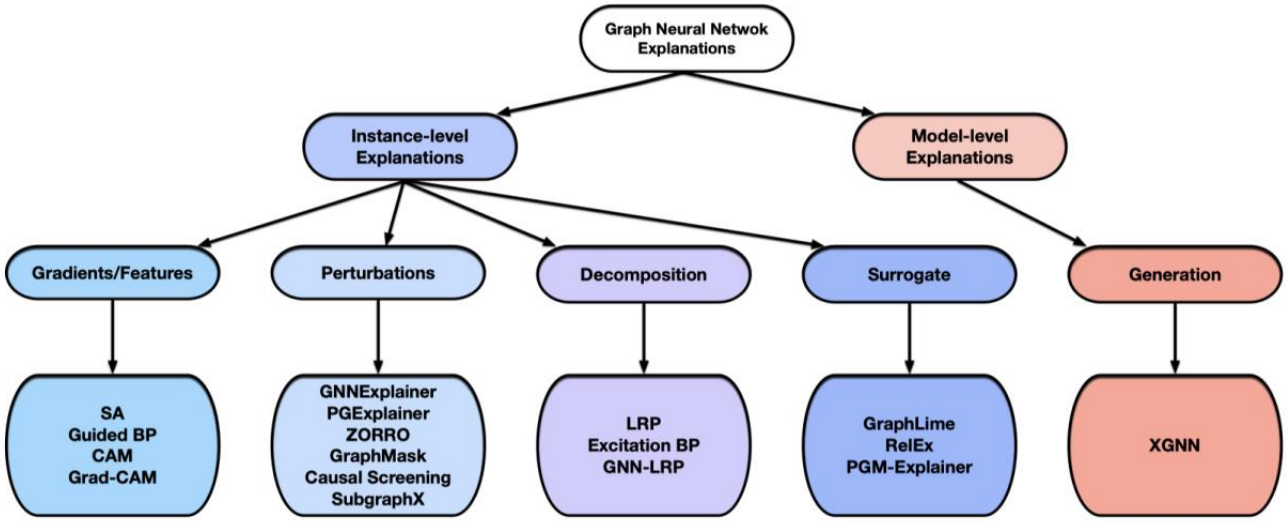
基于实例的方法为每个输入图提供依赖于输入的解释。给定一个输入图，这些方法通过识别用于预测的重要输入特征来解释深层模型。基于实例的方法又可分为四个不同的分支：基于梯度 / 特征的方法，基于扰动的方法，基于分解的方法和基于替代的方法。具体而言，基于梯度 / 特征的方法使用梯度或特征值来表示不同输入特性的重要性。基于扰动的方法可以观察预测结果在不同输入扰动下的变化，以研究输入重要性得分。基于分解的方法首先将预测分数（如预测概率）分解到最后一个隐藏层的神经元，然后将这些分数逐层反向传播到输入空间，并将

分解后的分数作为重要度分数。

基于模型的方法在解释图神经网络时，与输入无关的解释是高层次的，可以解释一般行为。唯一现有的基于模型的方法是基于图生成的 XGNN 方法。它生成图模式来最大化某个类的预测概率，并使用这些图模式来解释这个类。

总的来说，这两类方法从不同角度解释了图模型。基于实例的方法提供了对特定实例的解释，而基于模型的方法提供了对图模型如何工作的一般理解。对于实例的方法，为了验证和信任图模型，需要人对不同输入实例的解释进行验证。例如由于专家需要探索不同输入图的解释，因此需要更多的人为监督。对于模型级方法，由于解释是高层次的，因此较少涉及人为监督。此外，实例级方法的解释是基于实际输入实例的，因此它们很容易理解。然而，对于模型级方法的解释可能是人类无法理解的，因为获得的图模式甚至可能不存在于现实世界中。总的来说，这两种方法可以结合在一起，以便更好地理解图模型。

附图：



**47-668，**使用脉冲间隔对脉冲神经网络进行视觉解释

Youngeun Kim,Priyadarshini Panda

耶鲁大学，2021.3.26

脉冲神经网络（SNN）可以与异步二进制时间事件进行计算并进行通信，这可以通过使用神经形态硬件来大大节省能源，最近 SNN 相关算法工作已显示出在各种分类任务上的良好性能。然而，目前缺乏研究用于分析和解释这种深度 SNN 的内部脉冲行为的可视化工具。在本文中，我们提出了一种针对 SNN 的生物可视化的新概念，称为峰值激活图（SAM）。拟议中的 SAM 通过消除计算梯度以获得视觉解释的需要，规避了脉冲神经元的不可微特征，相反，SAM 通过在不同时间步长上正向传播输入脉冲来计算时间可视化图。SAM 通过突出显示具有短峰间间隔话动的神经元来产生与输入数据的每个时间步相对应的注意力图。有趣的是，在没有反向传播过程和类标签的情况下，SAM 会突出显示图像的区分区域，同时捕获细粒度的细节。借助 SAM，我们首次根据优化类型，泄漏行为以及何时遇到对抗性示例，对内部脉冲在各种 SNN 训练配置中的工作方式进行了全面分析。

**48-674，**解释人工智能所做决策的工作手册

David Leslie,Morgan Briggs，2021.3.20

英国阿兰图灵研究所

本手册由信息专员办公室和阿兰图灵研究所共同编制，概述了如何应用人工智能可解释性原则和实践。本手册介绍了解释人工智能决策的基础知识，提供了人工智能可解释性的四个原则、人工智能解释的类型，以及人工智能／多语言系统的解释性设计、开发和使用所涉及的任务。

人工智能可解释性的四个原则：

①透明。透明原则是 GDPR（合法性、公平性和透明度）中原则（a）的透明方向的延伸。在数据保护方面，透明度意味着对你是谁以及如何和为什么使用个人数据保持开发和诚实的态度。人工智能辅助决策的透明性建立在这些需求之上。它让你使用人工智能做决策的过程变得透明，并以一种有意义的方式向个人解释你所做的决定。

②负责。该原则源于 GDPR 中的责任制。在数据保护术语中，责任意味着承担遵守数据保护原则的责任，并能够证明遵守这些原则。负责解释人工智能辅助的决策将这些要求集中在设计和部署人工智能模型时执行的过程和操作上。

③考虑上下文。解释人工智能辅助决策不可以一刀切，需要关注几个不同但互相关联的元素，这些元素可以对解释人工智能辅助决策和管理整个过程产生影响。从概念到部署，以及向决策接受者介绍解释的各个阶段都需要考虑该原则。

④反思影响。在做出决策和执行任务之前需要有要负责的人去思考和推理，人工智能系统越来越多地充当人类决策的受托人。然而我们不能让这些系统直接对其结果和行为负责。在整个开发和实施阶段，你应该重新审视并反思 AI 项目初始阶段确定的影响。如果发现了任何新的影响，你至少应该记录这些影响，并思考如何减轻这些影响。这将帮助你向决策接受者解释你已确定的影响以及你如何尽可能减少任何潜在的有害影响。

解释人工智能所做决定的方式很多，现列出 6 种解释类型和说明，并对每种解释类型进行详细描述。

6 种解释类型和说明：

①基本原理解释

人工智能系统做决策的原因

②责任说明

谁参与了人工智能系统的开发、管理和实施，以及谁对决策进行人工审查

③数据说明

在特定决策中使用了哪些数据以及如何使用这些数据

④公平性解释

在人工智能系统的没计和实施过程中，为了确保其支持的决策是无偏见和公平的，所采取的步骤

⑤安全和性能说明

在人工智能系统的设计和实施过程中，为了最大限度地提高决策和行为的准确性、可靠性、安全性和稳健性，所采取的步骤

⑥影响解释

在人工智能系统的设计和实施过程中，为了考虑和监控人工智能系统的使用及其决策对个人和社会的影响，所采取的步骤

**49-675，**图像分类任务中卷积神经网络解释的白盒方法

Meghna P Ayyar,Jenny Benois-Pineau,Akka Zemmari

法国 Bordeaux 大学，LaBRI Crsdela 实验室，2021.4.6.

近年来，深度学习已成为解决来自多个领域的应用程序的普遍方法。卷积神经网络（CNN）特别展示了用于图像分类任务的最新技术性能，但是这些网络做出的决定并不透明，不能由人直接解释，已经提出了几种方法来解释以理解网络做出的预测背后的原因。在本文中，提出了一种基于这些方法的假设和实现对这些方法

进行分组的拓扑。作者主要关注白盒方法，这些方法利用网络内部体系结构的信息来解释其决策。给定图像分类和受过训练的CNN的任务，这项工作旨在提供一套全面而详细的方法概述，该方法可用于为特定图像创建解释图，这些解释图为图像的每个像素分配重要性得分基于其对网络决策的贡献。作者还建议根控其实现方式对白盒方法进行进一步分类，以实现更好的比较并帮助研究人员找到最适合不同情况的方法。

**50-679，**多跳推理真的可以解释吗？走向基准推理的可解释性

Xin Lv1,2 , Yixin Cao3 , Lei Hou1,2 , Juanzi Li1,2 Zhiyuan Liu1,2 , Yichi Zhang4 , Zelin Dai4

清华大学，南洋理工大学，阿里巴巴

近年来，多跳推理被广泛研究以获得更多可解释的链接预测。然而，我们在实验室中发现，这些模型给出的许多路径实际上是不合理的，而对它们的可解释性评估却做得很少。本文提出了一个统一的框架来定量评估多跳推理模型的可解释性，并设计了一个近似策略来使用规则的可解释性得分来计算它们。此外，我们手动注释了所有可能的规则，并建立一个基准来检测多跳推理的可解释性(BIMR)。在实验中，我们在基准上运行了9个基准。实验结果表明，当前多跳推理模型的可解释性不太令人满意，并且仍然远低于我们的基准所给出的上限。此外，基于规则的模型在性能和可解释性方面优于多跳推理模型，这为未来的研究指明了方向，即我们应该研究如何更好地将规则信号纳入多跳推理模型。

**51-683，基于通用模式理论的卷积神经网络可解释性**

Erico Tjoa，Guan Cuntai

南洋理工大学，阿里巴巴

2021.2.5

已有的工作和研究为深度神经网络（DNN）的可解释性提供了许多见解和贡献，但现有理论仍然无法完全理解和解释 DNN。提高 DNN 的可解释性具有很多好处，例如设计可靠性更高的方法，以及更好地对算法进行维护和改进。由于数据集结构的复杂性会加大解决由 DNN 机制引起的可解释性问题的难度，因此本文提出使用一种由 Ulf Grenander 提出的模式理论，其中数据可作为基本对象配置，从而使我们能够以组件方式研究 CNN 的可解释性。具体地，将扩展块附加到 Res Net 上来形成类似于 U-Net 的结构，从而使其可以在其 EB 输出广通道上执行与模式理论配置兼容的类似于语义分段的任务。通过这些 EB 模块来设计基于热图的可解释人工智能方法，以提取构成单个数据样本的单个生成点的解释，从而有可能减少数据集的复杂性对可解释问题的影响。包含上述模式理论元素的 MNIST 等效数据集可以让这种框架更加平滑，从而更加自然地通过图片生成的方式展示该理论。

**52-690，半监督文本分类中的虚拟对抗性训练增强注意机制的鲁棒性和可解释性**

Shunsuke Kitada，Hitoshi Iyatomi，2021.4.18 发表

日本 Graduate 科学与工程学校，Hosei 大学

本文提出了一种基于虚拟对抗训练（VAT）的注意力机制的新通用训练技术。VAT 可以在半监督的情况下从未标记的数据中计算出对抗性扰动，以用于先前研究中已报告的易受扰动的注意力机制。经验实验表明，本文的技术①与基于对抗训练

的传统技术以及基于 VAT 的技术在半监督环境下相比，提供了明显更好的预测性能；②证明了与单词重要性相关性更强，并且更好与人类提供的证据一致；③随着无标签数据量的增加，性能有所提高。

**53-691，**3D 脑肿瘤医学图像分割网络中的视觉可解释性

Hira Saleem、Ahmad Raza Shahid、Basit Raza

2021.4.26

巴基斯坦国家人工智能中心

医学图像分割是一项复杂而又重要的任务，它是医学诊断中最重要的环节之一。基于 3D 卷积神经网络（3DCNN）的模型在脑肿瘤图像分割方面取得了显著成果。然而由于神经网络的黑盒子性质，很难解释模型给出预测结果背后的基本原理，因此在医疗健康领域，集成此类模型以做出有关诊断和治疗决策的风险很高。因此在医学领域部署深度学习模型时，需要准确且透明的预测。在本文中，我们通过一种扩展的解释性技术来生成 3D 视觉解释，来提高 3D 脑肿瘤分割模型的可解释性。我们首先分析了无梯度可解释性方法相较于梯度依赖方法的优势。然后我们解释了分割模型对输入磁共振成像（MRI）图像的操作，并研究了该模型的预测策略。

我们还评估了其他多种针对医学图像分割任务的可解释性方法。为了证明我们的视觉不包含多余的噪声虚假信息，我们定量地对扩展方法进行了验证测试。该模型捕获的信息与人类专家的领域知识是一致的，从而使其风险性更低。最后我们使用 BraTS-2018 数据集训练 3D 恼肿瘤分割网络并通过可解释性实验以生成预测结果的视觉解释。

**54-701，**迈向可解释和可转移的语音情感识别：基于潜在表示的特征、方法和语料库分析

Sneha Das、Nicole Nadine Lønfeldt、Anne Katrine Pagsberg、Line H. Clemmensen

丹麦科技大学，2021.5.5

近年来，语音情感识别（SER）已经在从医疗保健到商业的多个领域得到广泛应用。除了信号处理方法外，SER 的方法现在还使用深度学习技术。但是对语言、语料库和记录条件进行概括仍然是该领域的挑战。此外，由于深度学习算法黑盒子性质，模型和决策过程缺乏解释性和透明性成为了新的挑战。当将 SER 系统部署在影响人类生活的应用程序中时，这一点至关重要。在这项工作中，我们通过对所提出的 SER 系统的决策过程进行深入分析来解决这一问题。为此我们提出了基于不完全和去噪自动编码器的很低复杂度 SER，对于四类情感分类该编译器的平均分类精度达到 55％以上。在此之上，我们调查了潜在空间中的情感聚类，以了解语料库对模型行为的影响并获得对潜在嵌入的物理解释。最后，我们探讨了每个输入功能对 SER 性能的作用。

**55-705，**在医学领域人类决策支持系统的可解释人工智能

瑞典皇家理工学院、芬兰阿尔托大学 Rohit Saluja，Samanta Knapic 等 2021.5.5

随着人工智能被应用于对人类有重大影响的更敏感的环境中，人工智能的可解释性就变得非常重要，人类必须能驹实时理解、再现和操纵机器决策过程。因此，人们越来越需要提高机器学习算法决策的可理解性，这些决策可以在实际应用中复制，特别是在医院领域。这就需要一个系统，允许直接地、可理解地和可解释

地做决策。可解释人工智能（XAI）有助于促进人工智能和机器学习在医学领域的应用，尤其有助于提高透明度和信任度。

本文提出了在医学图像分析领域对决策支持的人工智能方法进行实验，分别是两种事后可解释机器学习方法 LIME 和 SHAP，以及以语境价值和效用（ClU）为中心的解释方法 CIU。作者基于 LIME、SHAP 和 CIU 提供的解释进行了三次用户研究，来自不同非医学背景的用户在基于网络的调查环境中进行了一系列测试，并陈述了他们对给定解释的经验和理解。CIU 可解释方法在增加对人类决策的支持以及透明性和用户理解性方面比 LIME 和 SHAP 方法表现得更好。此外，CIU 比 LIME 和 SHAP 生成解释的速度更快。研究表明，在不同的解释支持环境 F，人类的决策存在显著差异。在此基础上，作者提出了三种可能的解释方法，这些方法可以在不同的医学数据集上推广，并为医学专家提供很好的决策支持。

**56-706，**用于欺诈检测的可解释机器学习

剑桥大学 Ismini Psvchoula 等 2021.5.13

应用机器学习来处理大型数据集在许多行业都有潜力，包括金融服务。然而，那些完全采用机器学习的实际问题，仍然集中在理解并能够解释复杂模型做出的决策和预测上。在本文中，我们通过研究在有监督和无监督模型上选择合适的背景数据集和运行时权衡，来探索实时欺诈检测领域中的可解释方法。

**57-708，**机器学习／深度学习可解释性算法尚未完全成熟有待完善

COPU，2021.5.19

打破机器学习中的黑盒子，实现可解释性人工智能（Explainable Artificial

Intelligence，XAI)，或从弱人工智能嬗变为强人工智能，已成为当今中外人工智能研发的亮点。

人工智能中外跟帖迄今已汇集发表 704 条，其中 40 条跟帖属于从理论上探讨机器学习／深度学习可解释性算法，以及从应用实践上体验可解释性算法的解决方案。可是，机器学习／深度学习可解释性算法尚未完全成熟，误差较大，作为最后评估有时结论分散。

下面谈一下机器学习／深度学习可解释性算法的演算步骤：

1）审题：根据不同项目的不同对象，选择不同的可解释性解决方案（或区分有待求解的可解释人工智能案例，不同的解释行为采用不同的求解方法)。在选择不同的 XAI 方案时可从如下排列组合中选择：分内在可解释性与事后可解释性两种情况，进一步还可分全局可解释性与局部可解释性两种情况。

2）确定／列出：可解释性算法分前提（条件）与任务的预期目标（结果）两方面，为了做好 XAI 求解，其前提条件最好能演译为（或确定）语义（网络)。有人说，没有语义就不能有任何解释，似乎有点夸大。

3）分析（演算步骤)：

①判断某项目求解的前提与结果是否为因果关系，

②收集数据、演译语义、捕捉特征、研究算法，

③纪计建模，

④建立 XAI 解决方案，显示结论，

⑤进行评估。

实际上机器学习／深度学习的 XAI 尚未完全成熟，有待完善。

1）在审题时，选择不同的 XAI 方案的背景中（4 种排列组合）是粗线条的，有相

当大的不确定性；

2）在演算步骤分析中，所谓没有语义就不能有任何解释（或不解决语义稀缺性将无法解释深度神经网络），似乎过于夸张；对于如何理解解释，有很多不同观点，如有人提出：仅据技术观点进行解释不够，要考虑心理学、社会科学等因素，有人提出用反事实解释来打破黑盒子，有人提出为建立可解释系统，决策者必须具有合理的信心等；对于最终的评估结论，其分散性和可信度均有待推敲！

58-709，XAI 手册：面向可解释人工智能的统一框架

德国人工智能研究中心（DFKI-G mbH）2021.5.14

可解释人工智能（XAI）领域已迅速成为一个蓬勃发展和多产的社区。然而这一领域有一个反复出现和公认的问题是对其术语缺乏共识。如“planation”和“interpretation”。本文提出了一个理论框架，不仅为这些术语提供了具体的定义，而且还概述了产生“explanation”和“interpretation”所需的所有步骤，本文证明了该框架在可解释性、可理解性和评价指标上符合要求。

本文对“explanation”和“interpretation”的定义如下：解释（explanation），指解释的任务，是描述（或解释）一个或多个事实的过程，而解释（interpretation），指解释的方法，是用以阐述解释（explanation）的含义，为了使解释（explanation）便于理解。

⊙大多数研发者侧重于解释方法的研究和使用，而忽略了解释任务的作用。

⊙解释通常没有明确定义域和共域(即 realm of the explanasand explanadum)。

在人工智能接管或协助人类专业工作者的领域（例如医疗应用），人们的兴趣通常从预测转到解释，以便人类专家学习并获得有关任务的新见解。

**59-710，**评价可解释人工智能分类算法的正确性

Swansea University 2021.5.20

近年来，可解释的 AI 通过特征归因算法吸引了许多研究者的关注，这些特征归因算法计算了预测中的“特征重要性”，因此变得越来越流行。但是对这些算法的有效分析很少。本文开发了一种通过创建具有已知解释基础事实的数据集来定量评估 XAI 算法正确性的方法，在实验中使用两种流行的特征归因解释器，即局部可解释模型不可知性解释（LIME）和 SHapley 可加性析构法（SHAP）。

关于解释性准确度，①分类准确度与解释性准确度呈正相关，②SHAP 比 LIME 提供更准确的解释，③解释准确性与数据集复杂度负相关。

**60-715，**多模态深层神经网络可解释性研究综述

Symbiosis Institute of Technology，2021.5.18

由深度神经网络提供支持的人工智能技术已在多个应用领域中取得了很大的成功，尤其是在计算机视觉应用程序和自然语言处理任务中。超越人类水平的表现推动了 AI 应用的研究，其中语言、视觉、感官、文本等不同模态在准确预测和识别中起着重要的作用。本文提出了几种采用深度学习模型的多模态融合方法。尽管它们具有出色的性能，但深层神经网络的复杂、不透明和黑盒性质限制了它们的社会认可度和可用性。这引发了对模型可解释性的探索，尤其是在涉及多模态人工智能方法的复杂任务中。本文对现有文献进行了广泛的回顾，以对多模态深度神经网络的可解释性，尤其是视觉和语言任务的可解释性进行全面的调查和评论。本文涵盖了多模态人工智能及其在通用领域的应用的几个主题，包括该领域的数据集、方法和技术的基本构件、挑战、应用和未来趋势。

61-718，用于可解释性推荐的个性化

Lei Li，Yongfeng Zhang，Li Chen，香港浸会大学，2021.5.25

自然语言生成的个性化在可解释性推荐、评论、摘要和对话系统等一系列任务中起至关重要的作用。在这些任务中，用户和项目 ID 是个性化的重要标识符。Transformer 具有很强烈语言建模能力，但由于 ID 标记与单词不在同一语义空间中 Transformer 没有个性化，不能充分利用用户和项目的 ID。针对这一问题，本文提出了一种个性化的可解释推荐变换器（PETER），并在此基础上设计了一个简单有效的学习目标，利用 ID 对目标解释中的词语进行预测，从而赋予 ID 语言意义，实现个性化 Transformer。除了生成解释，PETER 还可以进行推荐，这使得它成为整个推荐解释管道的统一模型。大量实验表明，我们的小规模无训练模型在生成任务上的有效性和效率都优于微调的 BERT,这突出了我们设计的重要性和良好的实用性。

62-719，扫视视觉如何有助于深层网络的可解释性

Research and Education Center，Mathematics of Future Technologies，Nizhing Novgorod，Russia

本文描述了现代深层网络的一些问题（可解释性，缺乏面向对象性）是如何通过采用一种生物学上合理的感知扫视机器来解决的，提出了这样一个扫视视觉模型的草图。实验结果证明了该方法的有效性。

63-721，可解释多跳科学问答的动态语义图构建与推理

WeiWen Xu，Huihui Zhang，DengCai，Wai Lam,香港中文大学

知识检索和推理是Web范围的多跳问答（QA）的两个关键阶段。当检索证据事实以填补知识空白时，现有方法的置信度不足并缺乏透明的推理过程。

在本文中，作者提出了一个新的框架，该框架可通过动态构建语义图并对其进行推理，从而在利用更多有效事实的同时获得多跳质量检查的可解释性。作者采用抽象含义表示（AMR）作为语义图表示。该框架包含三个新想法：①AMR-SG，一种基于AMR的语义图，由候选事实AMR构造以揭示问题，答案和多个事实之间的任何跃点关系。②一种新颖的基于路径的事实分析方法，利用 AMR-SG 从大型事实库中提取活动事实以回答问题。③利用图卷积网络（GCN）进行事实级别的关系建模以指导推理过程。在两个科学的多跳 QA 数据集上的结果表明，本文方法可以超越最近的方法,包括使用其他知识图谱的方法,AMR-SG 可以保持较高的解释性，并且可以成功地与强大的预训练模型相结合，从而实现对 OpenB ookQA 和 ARC-Challenge 的显著改进，而不是利用额外的 KG。

**64-723，**对话图：将可解释的策略图网络融入谈判对话

卡内基梅隆大学语言技术研究所，2021.6.2

有说服力的谈判策略的务实规划是必不可少的。虽然现代对话代理擅长生成流利的句子,但他们仍然缺乏语用基础,无法进行战略推理。本文提出 Dialo Graph，这是一个谈判系统，它使用图神经网络在谈判对话中结合实用策略。给定对话上下文，Dialo Graph 明确地结合策略序列之间的依赖关系，以实现对下一个最佳策略的改进和可解释的预测。本文基于图的方法在策略/对话行为预测的准确性和下游对话响应生成的质量方面都优于先前最先进的协商模型。本文定性展示了学习策略图在对话过程中提供有效谈判策略之间的明确关联方面的进一步好处，

从而导致可解释的战略对话。

**65-728，**Bounded Logitattention：学习解释图像分类器

Polten 大学，2021.5.31

可解释人工智能是试图通过称为“解释”的适当的辅助信息来阐明过于复杂而无法直接被人类认知访问的系统的工作原理。我们为卷积图像分类器提供了一个可训练的解释模块，称之为 BLA（Bounded Logitattention:）。BLA 克服了实例特征选择方法“学习解释”（L2X）的几个限制：BLA 可扩展到现实世界大小的图像分类问题；BLA 提供了学习可变大小解释的规范方法。由于其模块化，BLA 适合迁移学习设置，也可以用作训练分类器的事后附加组件。在用户调研中，我们发现 BLA 解释比流行的（Grad-）CAM 方法生成的解释更受欢迎。

**66-734，**人工解释的多样性和局限性

ChenhaoTan，芝加哥大学，2021.6.22

NLP 越来越多的努力旨在构建人工解释的数据集。然而，解释一词包含了广泛的概念，每个概念都有不同的属性和后果。本文的目标是提供一个不同类型的解释和人工局限性的概述，并讨论在 NLP 中收集和使用解释的意义。受心理学和认知科学先前工作的启发，将 NLP 中现有的人工解释分为三类：近似机制、证据和程序。这三类性质不同，并且对由此产生的解释有影响。例如，程序在心理学中不被视为解释，而是与从指令中学习的大量工作相联系。解释的多样性进一步体现在代理问题上，这些代理问题是标记者解释和回答开放式的为什么问题所需要的。最后解释可能需要与预测不同的，通常是更深层次的理解，这让人怀疑人工是否

能在某些任务中提供有用的解释。

**67-735，基于显著性的 XAI 方法众包评估**

Xiaotian Lu 等，京都大学，2021.6.27

理解深层神经网络预测背后的原因对于在许多重要应用中获得人们的信任至关重要，这反映在近年来对人工智能（XAI）可解释性的需求不断增加。基于显著性特征属性方法，特别是在计算机视觉领域，经常被用来作为一种 XAI 方法，该方法突出了图像中对分类器决策有重要贡献的部分。为了定量比较各种基于显著性的 XAI 方法，已经提出了几种自动评价方法；然而，不能保证这些自动化的评估指标能够正确地评估可解释性，自动评估方案的高评级并不一定意味着人类的高可解释性。在这项研究中，我们提出了一个新的基于人的评估方案，利用众包来评估 XAI 方法，而不是自动评估。我们的方法受人类计算游戏 Peek-a-boom 的启发，利用群体的力量对不同的 XAI 方法的显著性图进行评价。实验结果表明，基于人群的评价方案的评价结果不同于自动评价方案。此外，我们将基于人群的评估结果视为基本事实，并提供了一个定量的性能度量来比较不同的自动评估方案。我们还讨论了群体工作者对结果的影响，并表明群体工作者能力的变化对结果没有显著影响。

**68-736，多时间序列的可逆神经网络可解释非线性建模**

LuisMiguel 等，Agder 大学，2021.7.3

提出一种非线性拓扑识别方法，基于以下假设：时间序列的集合分两步生成：①潜在空间中的向量自回归过程，以及②非线性组件方式、单调递增的观测映射。

后面的映射被假定为可逆的，并被建模为浅层神经网络，因此可以对它们的逆进行数值评估，并且可以使用受深度学习启发的技术来学习它们的参数。由于函数反转，反向传播步骤并不简单，本文解释了应用隐微分计算梯度所需的步骤。虽然模型的可解释性与线性VAR过程相同，但初步数值测表明预测误差变小。

**69-740，**可信人工智能

HaoChen Liu等，密歇根大学，2021.7.12

在过去的几十年里，人工智能技术经历了飞速的发展，深刻地改变了人们的日常生活和人类社会的发展进程。开发人工智能的目的是减少人类劳动、为人类生活带来便利和促进社会进步。然而，最近的研究和应用表明，人工智能可能会对人类造成无意的伤害，例如在安全关键场景中做出不可靠的决定，或因无意中歧视某一群体而破坏公平性。因此，可信人工智能近年来受到了极大的关注，这就要求人们认真考虑，避免人工智能可能给人类带来的不利影响，使人类能够充分信任人工智能技术，并与之和谐相处。

近年来，人们对可信人工智能进行了大量的研究。本文作者从计算的角度对可信人工智能进行一个全面的调研，帮助读者了解最新的技术。可信人工智能是一个庞大且复杂的领域，涉及多个维度。本文作者关注六个最关键的维度:（i）安全性和稳健性,（ii）非歧视性和公平性,（iii）可解释性,（iv）隐私性,（v）责任性和可审计性，以及（vi）环境福利。对于每个维度，作者回顾了近期的相关技术，并总结了它们在实际系统中的应用。作者还讨论了不同维度之间的一致性和冲突性交互作用，并讨论了可信人工智能的未来研究方向。

人工智能系统性能的提高通常是通过增加模型复杂度来实现的。一个典型的例子

就是深度学习，它是大多数人工智能系统的核心。但是，它们被视为黑匣子，因为大多数深度模型过于复杂和不透明，人们无法理解。更重要的是，如果不解释模型背后的潜在机制，深度模型就不能完全可信，这就妨碍了它们在涉及道德、正义和安全的关键应用中的应用，如医疗保健、自动汽车等。因此，建立一个可信的人工智能系统需要了解特定决策是如何做出的，这导致了可解释人工智能领域的兴起。论文第六章对可解释人工智能的最新进展提供直观的理解和高层次的见解。首先，作者提供了人工智能中解释性的概念和分类。其次，根据前面提到的分类法回顾了人工智能系统中有代表性的可解释技术。随后，作者介绍了可解释人工智能技术的实际应用。最后，提供了一些综述和工具，并讨论了可解释人工智能的未来机遇。

在机器学习和人工智能文献中，explainability 和 interpretability 通常被研究者可互换地使用。最流行的可解释性定义之一是 Doshi Velez 和 Kim 的定义，他们将其定义为“以可理解的术语解释或呈现给人类的能力”。另一个流行的定义来自 Miller，他将可解释性定义为“人类能够理解决策原因的程度”。一般来说，人工智能系统的可解释性越高，人们就越容易理解某些决策或预测是如何做出的。同时，如果一个模型的决策比其他模型的决策更容易被人理解，那么它比其他模型更容易解释。虽然 explainable AI 和 interpretable AI 有着非常密切的联系，但有一些研究也讨论了它们之间的一些细微差别。

（1）如果模型本身能够被人类理解其是如何进行预测的，那么模型就是“interpretable”。 当查看模型参数或模型摘要时，人类可以准确地理解它如何做出某个预测/决策的过程，甚至给定输入数据或算法参数的变化，它是人类能够预测将要发生什么的程度。换句话说，这样的模型本质上是透明和可解释的，

而不是黑盒/不透明模型。interpretable models 的例子包括决策树和线性回归。

（2）explainable model 是采用了额外的（事后）解释技术来帮助人类理解为什么它做出了某个预测/决策，尽管该模型仍然是黑盒和不透明的。值得注意的是，这种解释往往是不可靠的，可能会产生误导。这类模型的例子是基于深度神经网络的模型，其中的模型通常过于复杂，任何人都无法理解。

人工智能的解释技术可以根据不同的标准进行分组。

根据模型用法的不同可分为 model-intrinsic 和 model-agnostic。如果可解释技术的应用仅限于人工智能模型的特定体系结构，那么这些可解释技术称为 model-intrinsic。相反，可以应用于任何模型的技术被称为 model-agnostic。

根据解释范围的不同可分为 local 和 global。如果该方法仅为特定实例提供解释，则它是局部解释，如果该方法可以解释整个模型，则它是全局解释。

根据解释方法的不同可分为 gradient-based 和 perturbation-based。如果这些技术利用输入实例的偏导数来生成属性，那么这些技术称为基于梯度的解释方法，如果这些技术侧重于输入数据的变化或修改，我们称之为基于扰动的解释方法。

还有一种技术通过其他方法进行解释，即 Counterfactual Explanations。该方法通常是包含因果关系的形式，例如：“如果 X 没有发生，Y 就不会发生”。一般来说，Counterfactual Explanations 方法与模型无关，可用于解释个别实例的预测（局部）。

每一类方法的代表模型如下表所示。

| Representative Models | Model Usage | Scope | Methodology |
|---|---|---|---|
| Linear model | Intrinsic | Global | - |
| LIME [267] | Agnostic | Both | Perturbation |
| CAM [369] | Agnostic | Local | Gradient |
| Grad-CAM [290] | Agnostic | Local | Gradient |
| SHAP [220] | Agnostic | Both | Perturbation |
| Saliency Map Visualization [300] | Agnostic | Local | Gradient |
| GNNExplainer [346] | Agnostic | Local | Gradient |
| Class Model Visualization [300] | Agnostic | Local | Gradient |
| **Surveys** | [27, 34, 103, 112, 113, 152, 175, 209, 238, 243, 314, 349, 361] | | |

可解释人工智能的应用有：（1）推荐系统。推荐系统（RecSys）在我们的日常生活中变得越来越重要，因为它们在缓解信息过载问题方面发挥着重要作用。这些系统提供个性化信息以帮助人类做出决策，并已广泛应用于各种面向用户的在线服务，如电子商务商品日常购物推荐（如亚马逊、淘宝）、就业市场就业推荐（如LinkedIn）等。近年来，基于深度学习的推荐模型在提高准确性和更广泛的应用场景方面取得了巨大的进步。因此，人们越来越关注理解基于深度学习的推荐系统为什么会推荐某些项目给最终用户，因为提供个性化推荐系统的良好解释可以充分激励用户与项目交互，帮助用户做出更好和/或更快的决策，增加用户对智能推荐系统的信任。（2）药物研究。在过去的几年中，可解释人工智能已被证明显著加速了计算机辅助药物发现的过程，例如分子设计、化学合成规划、蛋白质结构预测和大分子目标识别。（3）自然语言处理。作为人工智能应用最广泛的领域之一，自然语言处理（NLP）研究了如何使用计算机来处理或理解人类的语言。自然语言处理的应用无处不在，包括对话系统、文本摘要、机器翻译、问答、情感分析、信息检索等。最近，深度学习方法在许多不同的自然语言处理任务中取得了非常好的表现，但这是以模型变得不那么可解释为代价的。

由于人工智能的可解释性是一个相对较新的领域，一个发展中的领域，因此有许多问题需要考虑。（1）可解释人工智能的安全性。最近的研究表明，由于人工智

能模型的数据驱动特性，其解释容易受到恶意操作的影响。攻击者试图生成对抗性的示例，这不仅会误导目标分类器，还可能欺骗相应的解释器。这自然会在解释上引起潜在的安全问题。因此，如何防范翻译中的对抗性攻击将是今后的一个重要方向。(2) 评价方法。评价指标是研究解释方法的关键。然而，由于缺乏基本事实和人的主观理解，对某些预测的解释是否合理和正确的评价变得越来越困难。目前广泛使用的评价方法是基于人的评价，这种方法比较直观，同时也比较费时，并且存在偏见。(3) 从白盒到黑盒。大多数现有的解释技术要求对所解释的人工智能系统有充分的了解（表示为白盒)。然而，由于隐私和安全问题，在许多情况下，人工智能系统的相关知识通常是有限的。因此，一个重要的方向是理解如何在黑箱系统中生成决策的解释。

**70-741**，基于可解释 SincNet 的脑电信号情感识别深度学习

2Mila-Quebec 人工智能研究所，2021.9.23

机器学习方法，如深度学习，在医学领域显示出有希望的结果。然而，这些算法缺乏可解释性可能会阻碍它们在医疗决策支持系统中的适用性。本文研究了一种可解释的深度学习技术，称为 SincNet。SincNet 是一种卷积神经网络，它通过可训练的 sinc 函数有效地学习定制的带通滤波器。在这项研究中，本文使用 SincNet 来分析患有自闭症谱系障碍(ASD)的个体的神经活动,他们在神经振荡活动中经历了特征性差异。特别是,本文提出了一种新的基于SincNet的神经网络，用于使用 EEG 信号检测 ASD 患者的情绪。可以轻松检查学习到的过滤器，以检测 EEG 频谱的哪一部分用于预测情绪。本文发现本文的系统会自动学习 ASD 患者经常出现的高 α (9-13Hz)和 β (13-30Hz)频带抑制。这一结果与最近关于情绪识别

的神经科学研究一致，该研究发现这些频带抑制与在 ASD 个体中观察到的行为缺陷之间存在关联。在不牺牲情绪识别性能的情况下，SincNet 的可解释性得到了提高。

**71-744，**GLIME：一种用于可解释性模型不可知解释的新图形方法

可解释人工智能（XAI）是一个新兴的领域，在这个领域中，一系列的过程和工具使人们能够更好地理解由黑盒模型生成的决策。然而，大多数可用的 XAI 工具通常仅限于简单的解释，主要是量化各个特性对模型输出的影响。因此，人类用户无法理解特征之间的相互关系以进行预测，而训练模型的内部工作机制仍然是隐藏的。本文致力于开发一种新的图形化解释工具，该工具不仅能显示模型的重要特征，而且能揭示特征之间的条件关系和推理，捕捉特征对模型决策的直接和间接影响。提出的 XAI 方法称为 GLIME，它提供了全局（对于整个数据集）或局部（对于特定数据点）的图形模型不可知解释。它依赖于局部可解释模型不可知解释（LIME）与图形最小绝对收缩和选择算子（GLASSO）的结合，产生无向高斯图形模型。采用正则化方法将小的偏相关系数压缩到零，从而提供更稀疏、更易于解释的图形解释。选择两个著名的分类数据集（活检和 OAI）来证实 GLIME 在稳健性和一致性方面优于 LIME。具体来说，GLIME 在两个数据集上实现了特征重要性方面的稳定性提高（76%-96%，而使用 LIME 则为 52%-77%）。GLIME 展示了一种独特的潜力，通过提供信息丰富的图形化解释，可以打开黑匣子，从而扩展 XAI 当前最先进的功能。

**72-747，**稳健可解释性：深度神经网络基于梯度的属性方法教程

IanE.Nelsen 等，Rowan 大学，2021.7.3

随着深度神经网络的兴起，解释这些网络预测的挑战越来越受到人们的认可。虽然存在许多解释深度神经网络决策的方法，但目前还没有就如何评估它们达成共识。另一方面，鲁棒性是深度学习研究的热门话题；然而，直到最近才在可解释性方面谈论它。在本教程论文中，本文首先介绍基于梯度的可解释性方法。这些技术使用梯度信号来分配输入特征的决策负担。稍后，本文将讨论如何评估基于梯度的方法的鲁棒性以及对抗性鲁棒性在提供有意义的解释方面所起的作用。本文还讨论了基于梯度的方法的局限性。最后，本文介绍了在选择可解释性方法之前应该检查的最佳实践和属性。本文总结了该领域在稳健性和可解释性的融合方面的未来研究方向。

**73-748，**使用可解释的深度学习方法进行有效和健壮的模式识别

XiaoBai 等，德克萨斯大学奥斯丁分校，2021.7.23

深度学习最近在许多视觉识别任务中取得了巨大的成功。然而，深层神经网络（DNNs）通常被视为黑匣子，其决策过程和原理不易被人类理解，因此被禁止在关键安全应用中使用。本文介绍了 30 篇论文，这些论文都是关于 Explainable Deep Learning for Efficient and Robust Pattern Recognition 的特刊。它们主要分为三大类：可解释的深度学习方法、通过模型压缩和加速以实现高效的深度学习以及深度学习的鲁棒性和稳定性。本文对这三个专题的代表作和最新进展进行了综述，并简要介绍了各个专题已被接受的论文。这篇综述的整体结构如图 1 所示。

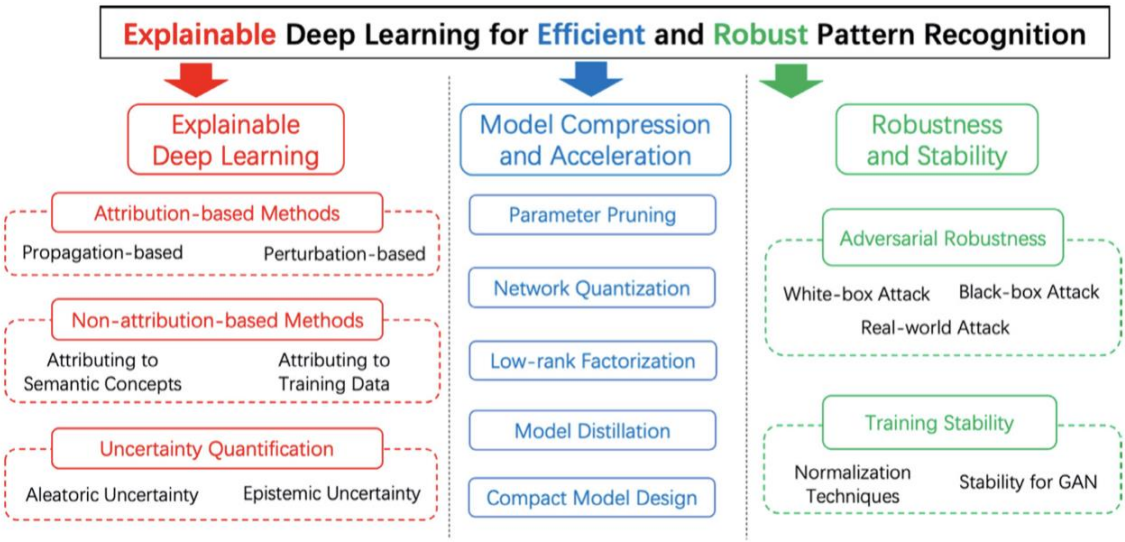



**Fig. 1.** The overall structure of this guest editorial.

许多可解释性方法旨在解释深层神经网络的工作机制。大多数的方法侧重于将 DNN 的预测归因于其输入特征。这种基于归因的方法涵盖了计算机视觉中的大多数可视化方法，它通过定位对决策贡献最大的区域，直接在输入图像的域中给出解释。除此之外，许多基于非归因的方法还从概念、训练数据、内在注意机制等方面进行了解释。从另一个角度看，不确定性暗示了网络决策的可靠性。这些信息是对网络解释的补充，在各种现实生活中的关键安全应用中是至关重要的。

深度神经网络在各种任务中以设置较多参数为代价获得了最高的精度，从而需要大量的计算资源和训练时间。因此，在部署到资源受限的设备和实时应用程序之前，业界对模型压缩和加速技术有着巨大的需求。近年来，越来越多的方法被提出来用于压缩和加速网络，同时对模型的精度做出最小的妥协。大多数方法可分为以下几类：参数剪枝、网络量化、低秩因子分解、模型提取和紧凑网络设计。

鲁棒性揭示了模型在数据噪声下提供可靠决策的能力。近年来，人们对与深度学习相关的鲁棒性的几个方面进行了研究。其中最热门的话题是对抗性健壮性，因为它与应用程序中的安全问题密切相关。稳定性是深度神经网络的另一个关键问题，它决定了网络能否成功收敛。一些技术有助于提高训练时的稳定性，如规范

化技术和网络优化的一些约束。

**74-750，**解决基于扰动的 XAI 技术所面临的分布外数据问题

LuyuQIU 等人，华为香港研究院、香港大学，2021.8.7

随着可解释人工智能（XAI）的迅速发展，基于扰动的 XAI 算法因其有效性和易实现性而变得非常流行。绝大多数基于扰动的 XAI 技术都面临着分布外(Out-of-Distribution，OoD)数据的挑战，即随机扰动数据的伪影与原始数据集不一致。OoD 数据导致模型预测中的过度置信问题，使得现有的 XAI 方法不可靠。但是，目前基于扰动的 XAI 算法存在的 OoD 数据问题尚未在文献中得到充分解决。本文，作者通过设计一个额外的模块来解决这个 OoD 数据问题，该模块量化了扰动数据和原始数据集分布之间的相关性，并将其集成到聚合过程中。作者的解决方案与最流行的基于扰动的 XAI 算法（如 RISE，OCCLUSION 和 LIME）兼容。实验证明，作者的方法在使用计算和认知度量的一般情况下表现出显著的改进。特别是在退化的情况下，与基线相比，作者提出的方法表现出了优异的性能。此外，作者的解决方案还解决了忠实度指标的一个基本问题，这是 XAI 算法的一个常用评估指标，似乎对 OoD 问题很敏感。

**75-751，**基于子组发现的黑盒子事件检测的可解释性总结

YoucefRemil 等，里昂大学，2021.8.6

随着监控系统和设备 软件用户报告的事件数量不断增加，预测性维护的需求也随之增加。在前线，待命工程师（OCE）必须快速评估事件的严重程度，并决定联系哪家服务机构采取纠正措施。为了使这些决策自动化，已经提出了几种预测模

型，但最有效的模型是不透明的（比如，黑盒），这大大限制了它们的采用。在本文中，提出了一个有效的黑盒模型，该模型基于过去 7 年中向本文公司报告的 170K 起事件，并强调在运行本文的产品（ERP）的数千台服务器上大量报告事件时自动分类的需要。可解释人工智能（XAI）的最新发展有助于为模型提供全局解释，但最重要的是，为每个模型预测 结果提供局部解释。不幸的是，在处理大量日常预测时，为人类提供每种结果的解释是不可想象的。为了解决这个问题，本文提出了一种源于子组发现的原始数据挖掘方法，这是一种模式挖掘技术，具有对共享其黑盒预测的类似解释的对象进行分组的自然能力，并为每个组提供描述。本文评估了这种方法，并给出了本文的初步结果，这给本文带来了采用有效 OCE 的良好希望。相信这种方法能为解决模型不可知结果解释问题提供了一种新的方法。

**76-754，**用于推荐知识图谱上的时间感知路径推理

Yuyue Zhao 等、中国科技大学、新加坡大学、中国风险感知与预防国家工程实验室，2021.8.5

由于知识图谱（KG）的推理能够提供明确的解释，因此已经研究了可解释性推荐的推理。然而，遗憾的是，当前基于 KG 的可解释推荐方法忽略了时间信息（如购买时间、推荐时间等），这可能会导致不合适的解释。在这项工作中，本文作者提出了一种新颖的时间感知路径推理推荐（简称 TPRec）方法，该方法利用时间信息的潜力以合理的解释提供更好的推荐。据作者所知，TPRec 方法是第一个将时间感知路径推理方法引入推荐系统的工作，并通过利用时间信息实现了显着的性能提升。首先，作者提出了一种有效的时间感知交互关系提取组件来构建具有

时间感知交互（简称 TCKG）的协作知识图谱，然后介绍一种新颖的时间感知路径推理方法进行推荐。作者对三个真实世界的数据集进行了广泛的实验。结果表明，所提出的 TPRec 可以成功地使用 TCKG 来获得可观的收益并提高可解释推荐的质量。

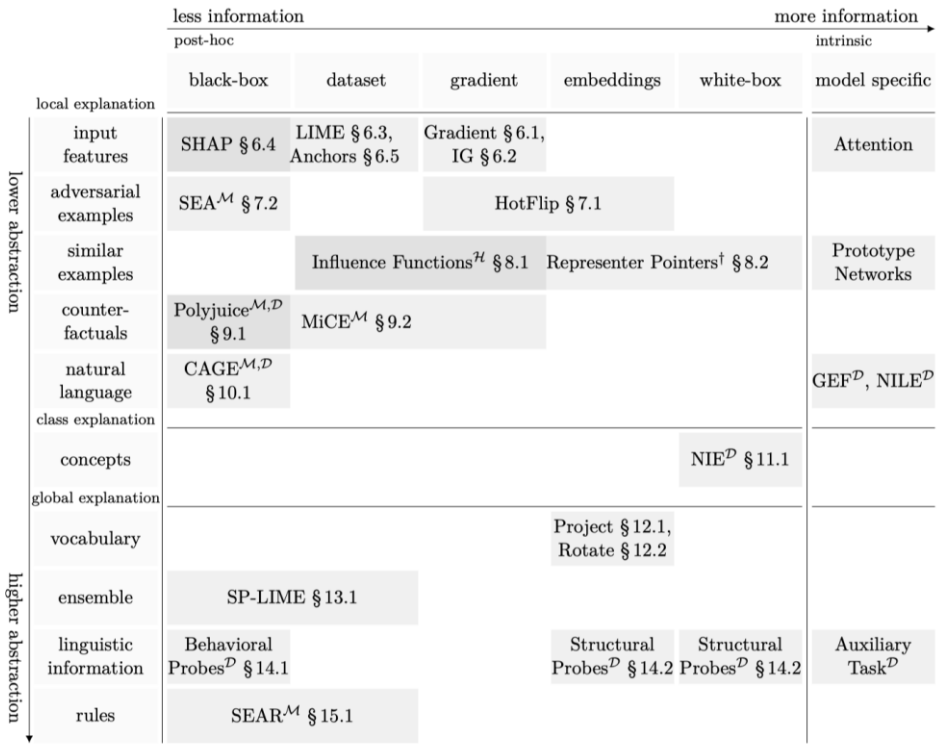
对于未来的工作，作者计划使用来自其他产品领域和其他主要在线供应商的更多数据集来评估 TPRec，并通过利用对抗性学习模型自动塑造奖励来扩展 TPRec 模型，以实现更准确的推荐结果。作者还计划将因果推断与我们的模型相结合，以实现更好的可解释性。

**77-756，**NLP 的事后可解释性研究综述

Mila-Quebec 人工智能研究所、脸书 AI 数据集、加拿大 AI 数据集，2021.8.10

自然语言处理（NLP）模型变得越来越复杂和广泛。随着神经网络的发展，人们越来越关注模型的可解释性。因此，本综述对可解释性方法进行了分类，并深入讨论这些方法的原理和特点。本综述侧重于调研事后（Post-hoc）可解释性方法，此类方法在模型训练完成后才提供解释，而模型通常是不可知的。

下表是事后可解释性方法的总结，其中 § 指出了该方法在文章的哪一部分中被讨论。行描述了如何进行解释，而列描述用于生成解释的信息。行和列分别按照抽象级别和信息量进行排序。




作者就 NLP 可解释性的未来研究方向进行了展望：

（1）衡量可解释性。目前可解释性的衡量方式各不相同。一般情况下，每篇论文都会介绍其衡量可解释性的方式。这降低了研究的相互可比较性。

（2）class explanations。关于方法本身，已经有很多研究。然而，在局部解释和全局解释之间，class explanations 仍然是一个未充分体现的中间地带。

（3）将事后方法与内在方法相结合。大多数内在方法（intrinsic methods）并不是纯粹的内在方法，它们通常有一个中间的表示，它在本质上是可以解释的。但是，生成此表示通常使用黑盒模型。因此，如果要理解整个模型，就需要事后解释方法。

**78-757，**基于结构匹配的可解释深度度量学习研究

中国 Tsinghua 大学，2021.8.13

神经网络如何区分两幅图像？理解深度模型的匹配机制对于开发可靠的智能系

统以用于许多危险的视觉应用（如监视和访问控制）至关重要。然而，现有的深度度量学习方法大多通过比较特征向量来匹配图像，忽略了图像的空间结构，因而缺乏可解释性。在本文中，提出了一种更透明的嵌入学习的深度可解释度量学习（DIML）方法。与传统的基于特征向量比较的度量学习方法不同，本文提出的结构匹配策略，通过计算两幅图像的特征映射之间的最佳匹配流来显式对齐空间嵌入。此方法使深度模型能够以更人性化的方式学习度量，其中两幅图像的相似性可以分解为若干部分相似性及其对整体相似性的贡献。此方法是模型不可知的，可以应用于现成的主干网和度量学习方法。本文在深度度量学习的三个主要基准（包括 CUB200-2011、Cars196 和斯坦福在线产品）上评估了该方法，并在更好的可解释性上实现了对流行度量学习方法的重大改进。

**79-761，**利用商业数据科学价值：确保解决方案的可解释性和公平性

本文介绍了人工智能中公平性和可解释性（XAI）的概念，旨在解决复杂的商业问题。为了公平，作者讨论了导致偏见的细节，以及相关的缓解方法，最后提出了一套在数据驱动的组织中引入公平的方法。此外，对于 XAI，作者审核了特定的算法和演示性的业务用例，讨论了大量的质量量化技术，并概述了未来的研究途径。

**80-765，**2020 年 6 月，COPU 主办《第 15 届开源中国开源世界 高峰论坛》，邀请 IBM 副总裁 Todd Moore 在会上作“可信 任人工智能(反欺诈、可解释、公平性)”的报告，从此 至今，COPU 已收到全球研发可解释性人工智能的跟帖 48 件。但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具有不 确定性，最后评估还只能靠人工，所以 XAI 演算结果或算 法可能有出入，致使可解释机器学习难以推广应用。为 此，COPU 要求 IBM Todd Moore 和人工智能研究所的 CTO Animesh 对 XAI 举出具体案例并进行解析和说明，对我们提 出的 8 个问题进行逐个解答:

① IBM 列出研发 XAI 的具体案例是什么?

② 选用下列哪种方法进入运算?

- 可直接解释(内在解释)
- 事后解释
- 全局(模型级)可解释性
- 局部(实例级)可解释性

③ 选择什么工具?

- 如:决策树、规划库、块择表等

④ 如何捕捉特征?

⑤ 如何建模?

⑥ 如何找到算法?

⑦ 如何进行评估?

⑧ 不但要导出本案例结果，还要使 XAI 在使用中确定是否能保持信任、公正、透明和可解释?!

**81-766，**工程设计的可解释人工智能：系统工程和基于组件的深度学习的统一方法

2021.8.29

由机器学习创建的数据驱动模型在设计和工程的所有领域都变得越来越重要。它们在帮助决策者创造具有更好性能和可持续性的新型人工制品方面具有很大潜

力。然而，这些模型的有限泛化和黑盒特性导致了有限的可解释性和可重用性。这些缺点严重阻碍了工程设计的采用。为了克服这种情况，作者提出了一种基于组件的方法，通过机器学习（ML）创建部分组件模型。这种基于组件的方法将深度学习与系统工程（SE）相结合。通过节能建筑设计的实例，作者首先通过准确预测不同于训练数据的具有随机结构的设计的性能，证明了基于组件的方法的泛化能力。其次，作者通过局部抽样、敏感性信息和来自低深度决策树的规则以及从工程设计角度评估这些信息来说明可解释性。可解释性的关键在于，组件之间接口处的激活是可解释的工程量。通过这种方式，分层组件系统形成了一个深度神经网络（DNN），该网络直接集成了工程可解释性信息。组成组件中的大量可能配置允许使用可理解的数据驱动模型检查新的看不见的设计案例。通过相似的概率分布匹配组件的参数范围，可以生成可重用、通用性好且可信的模型。该方法使模型结构适应系统工程和领域知识的工程方法。

**82-769，**基于梯度激活图（GAM）的可解释视觉相似性与分类

以色列开放大学和微软，巴伊兰大学，微软和特拉维夫大学，2021.9.3

本文提出了梯度激活图(GAM)——一种解释视觉相似性和分类模型预测的机制。通过从多个网络层收集局部梯度和激活信息，与现有替代方案相比，GAM 提供了改进的视觉解释。详细解释了 GAM 的算法优势，并通过经验进行了验证，其中表明 GAM 在各种任务和数据集上的表现优于其替代方案。

**83-770，**工程设计的可解释人工智能：一种结合系统工程和基于组件深度学习的统一方法

柏林大学，2021.8.29

由机器学习创建的数据驱动模型在设计和工程的所有领域都变得越来越重要。它们在帮助决策者创造具有更好性能和可持续性的新型人工制品方面具有很大潜力。然而，这些模型的有限泛化和黑盒特性导致了有限的可解释性和可重用性。这些缺点严重阻碍了工程设计的采用。为了克服这种情况，作者提出了一种基于组件的方法，通过机器学习（ML）创建部分组件模型。这种基于组件的方法将深度学习与系统工程（SE）相结合。通过节能建筑设计的实例，作者首先通过准确预测不同于训练数据的具有随机结构的设计的性能，证明了基于组件的方法的泛化能力。其次，作者通过局部抽样、敏感性信息和来自低深度决策树的规则以及从工程设计角度评估这些信息来说明可解释性。可解释性的关键在于，组件之间接口处的激活是可解释的工程量。通过这种方式，分层组件系统形成了一个深度神经网络（DNN），该网络直接集成了工程可解释性信息。组成组件中的大量可能配置允许使用可理解的数据驱动模型检查新的看不见的设计案例。通过相似的概率分布匹配组件的参数范围，可以生成可重用、通用性好且可信的模型。该方法使模型结构适应系统工程和领域知识的工程方法。

**84-775，**可解释人工智能的反事实评估

（美）罗格斯大学、阿里巴巴，2021.9.5

尽管近年来机器学习中出现了各种可解释的方法，但解释在多大程度上真正代表了模型预测背后的推理过程，即解释的可信度仍然是一个悬而未决的问题。衡量

可信度的一种常用方法是基于擦除的标准（erasure-based criteria）。基于擦除的标准虽然简单，但不可避免地会引入偏差和伪影。因此，作者提出了一种新的方法，从反事实推理的角度来评估解释的真实性。反事实评估过程是根据以下两个直观的观察结果精心设计的：（1）即使是最重要特征上的微小扰动也可能影响模型预测；（2）除非最不重要特征的扰动足够大，否则它们不会对模型预测产生太大影响。基于这两个观察结果，作者提出了一个新的框架，利用反事实的概念来评估解释的真实性。在反事实评估过程中，我们的目的是想要知道，如果我们改变了输入的特征，模型的预测会发生怎样的变化。作者提出了两种不同的算法，分别在处理离散或连续输入条件下找到适当的反事实，然后使用获得的反事实来衡量可信度。在多个数据集上的实证结果表明，与现有指标相比，作者提出的反事实评估方法能够实现与 ground truth 的最高相关性。

**85-785，**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

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。

**86-791，**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

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。

**87-794，**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

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。

88-795，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

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。

89-802，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：Victor Albiero 等，IEEE、圣母大学（诺特丹大学）、佛罗里达科技大学，2021.12.29

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。

90-805，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

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

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)。

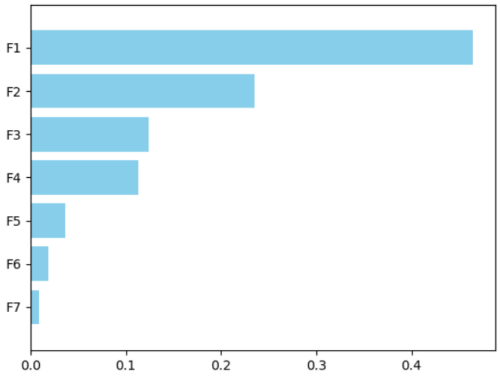
在特定场景，如果典型的，结构明确的算法模型（决策树或者线性回归）的预测或者识别的效果和复杂模型近似，就可以这些算法来解释复杂模型的算法。比较成熟和应用广泛的就是对于一条实例的预测结果解释。 用复杂模型（比如深度学习,xgboost)对图像,或者文字,或者一个贷款记录作出了识别或者预测。那么围绕着这条实例数据建立一个典型的线性回归模型(建立这个线性回归模型的数据由该 1）条实例数据，2)该复杂模型和 3）建立复杂模型数据共同生成，以此来保证在这个实例上线性回归和复杂模型的等效性)，用线性回归模型来解释这条记录的预测。OpenScala 对于实例的解释采用该技术

- 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）

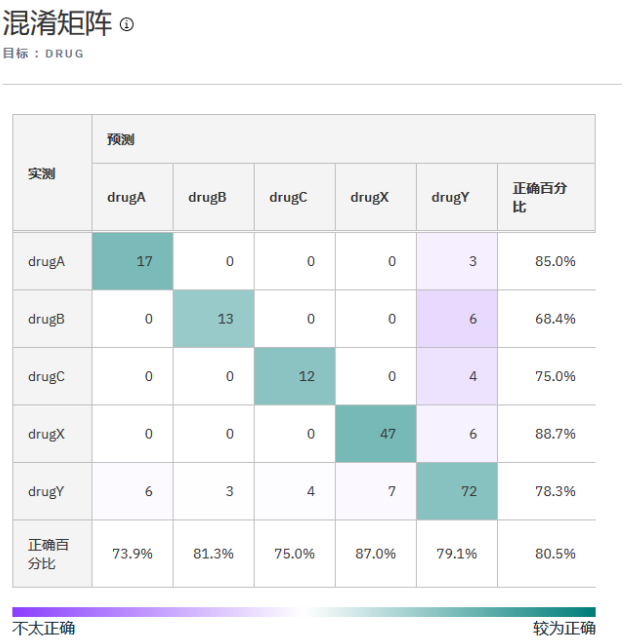
- **回归模型（regression）评估指标**

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，Num Satisfactory Trades 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 Risk Performance 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

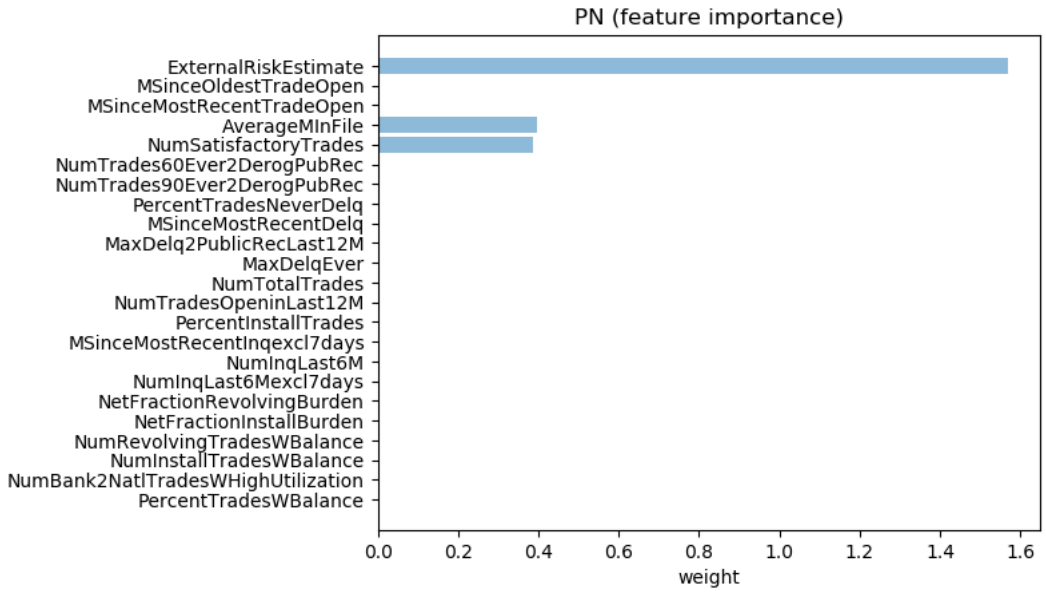
变 AI 模型的决定的信息（pertinent negatives），从拒绝到接受或从接受到拒绝。例如，对于被拒绝的申请，保持其他不变，将合规账目数量增加到某个值可能会导致申请被接受。同时对比解释还可以从贷款申请信息中选出部分特征和取值以维持当前决策不变（pertinent positives）。例如，对于被接受的申请，即使将合规账目数量减少到较低的值，申请仍然可以通过。

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。

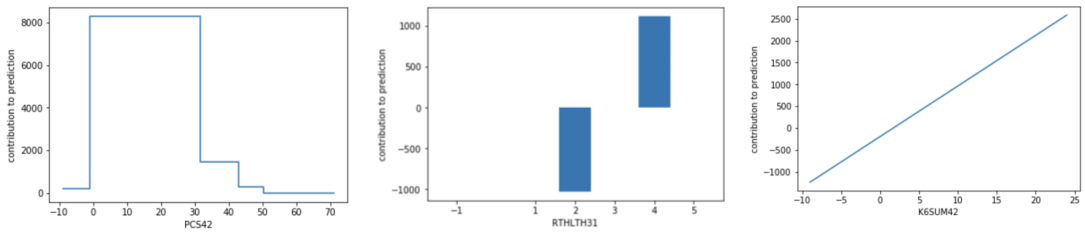



Figure 1　变量 PCS42，RTNLTH31，K6SUM42 与个人医疗支出的关系

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。

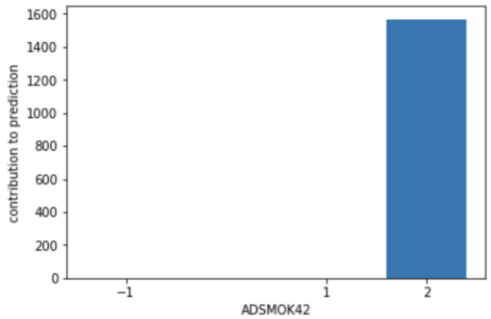



Figure 2 ADSMOKE 对个人医疗支出的影响

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。

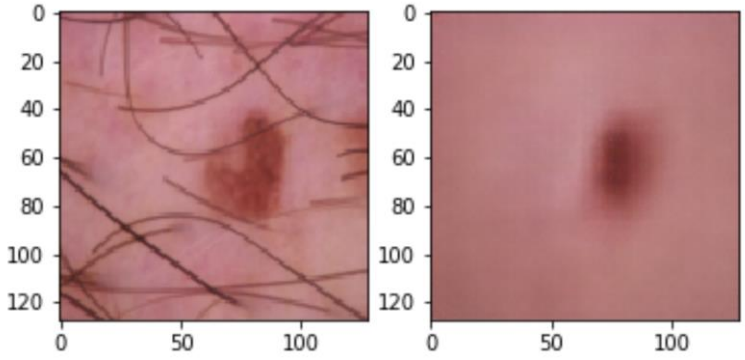



Figure 3 DIP-VAE 重建图与原始图对比

（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。

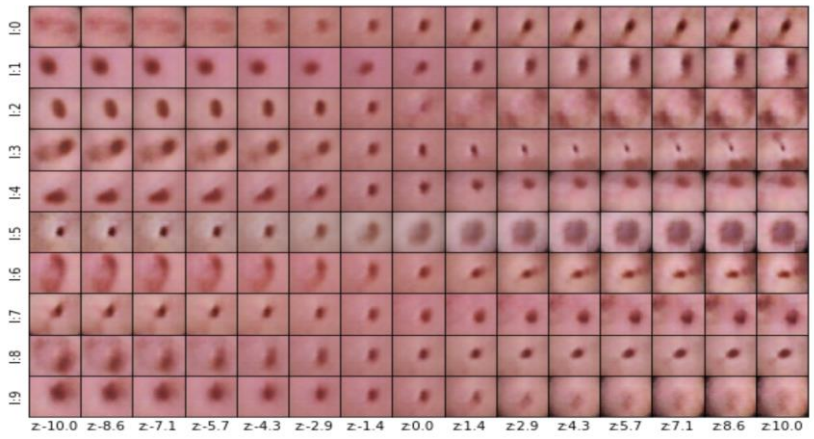



Figure 4 采用控制变量法得到可解释的高维隐藏特征

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。

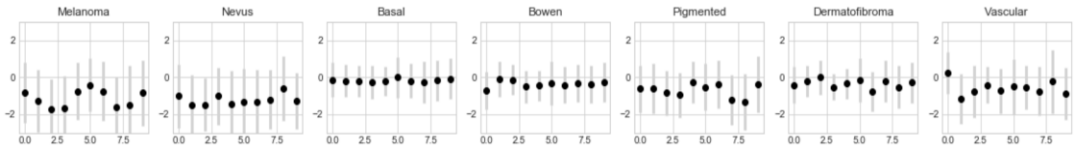



Figure 5 不同种类皮肤病下隐藏特征的分布

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

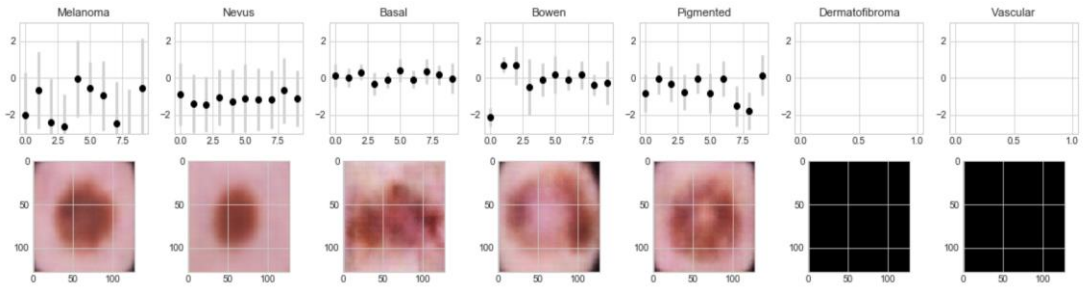



Figure 6 随机森林模型结果

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。

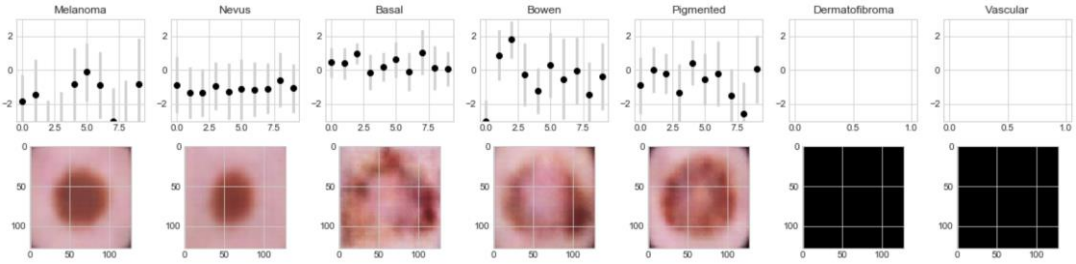



Figure 7 逻辑回归模型结果

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。

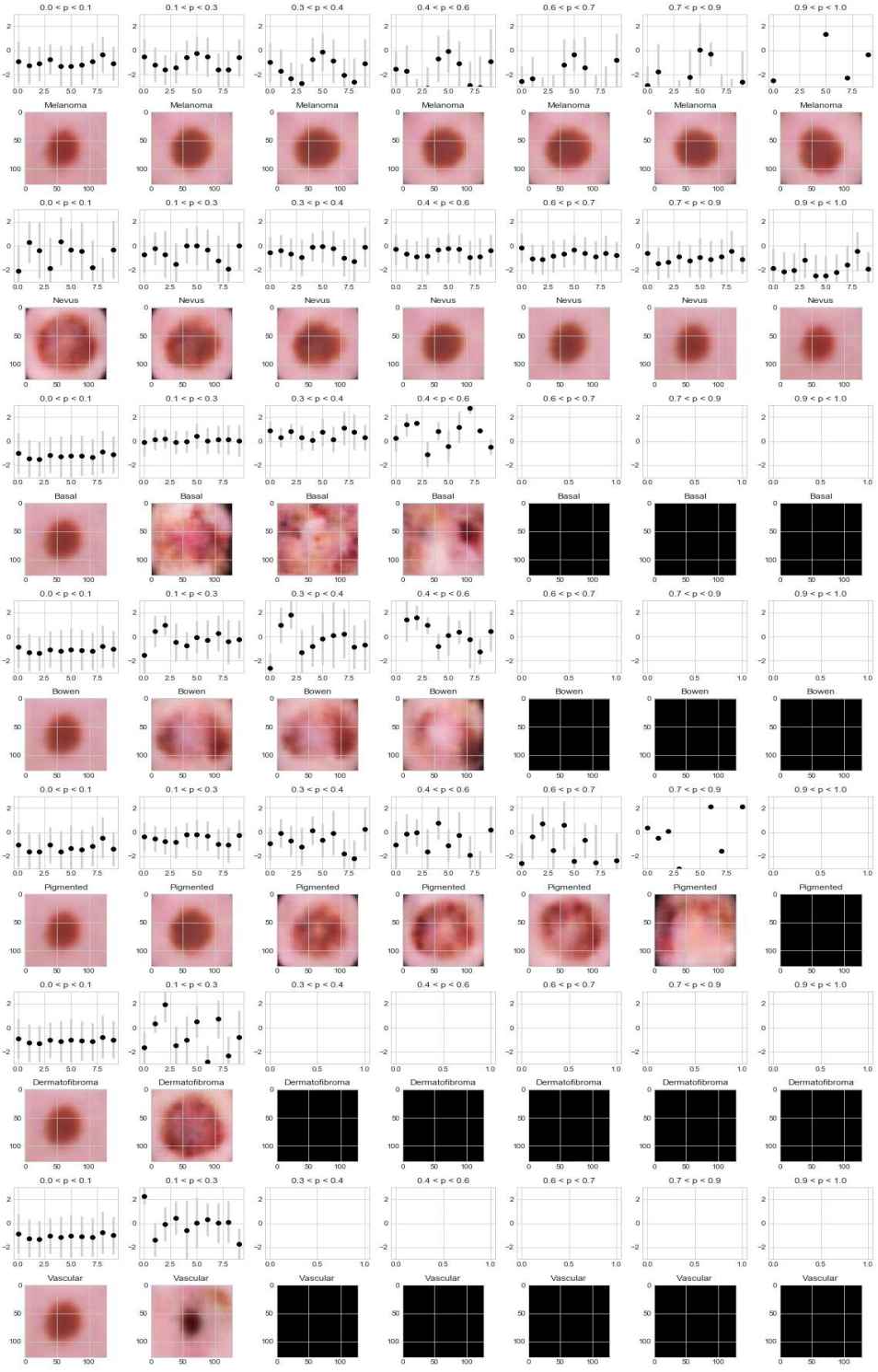



Figure 8

7、可解释性辅助模型评估

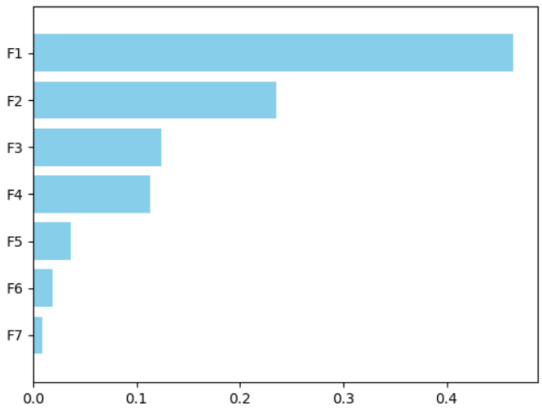
从本案例提到的两个简单机器学习模型来看，随机森林模型（精度 69%）和逻辑回归分类器（精度 65%），它们的精度略低于目前采用深度神经网络的最高精

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

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

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。

**91-807，**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

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。

92-808，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。

93-809，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

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。

**94-820，**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

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。

**95-821，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

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。

**96-836，超越解释：基于 XAI 的模型改进的机遇和挑战**

Leander Weber 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

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。

97-838，通过学习何时行动来实现可解释的强化学习

Alexis Jacq 等，2022.3.16，谷歌研究，法国国家信息与自动化研究所，里尔大学

简介：传统上，强化学习（RL）旨在决定如何对人工智能体进行优化。我们认为，决定何时采取行动同样重要。作为人类，当情况需要时，我们会从默认的、本能的或记忆中的行为转变为专注的、深思熟虑的行为。为了增强具有这种能力的 RL 代理，我们建议增强标准马尔可夫决策过程，并提供一种新的行动模式：懒惰，这将决策推迟到默认策略。此外，我们惩罚非懒惰行为，以鼓励最小的努力，并让代理只关注关键决策。我们将由此产生的形式主义命名为懒惰的 MDP。我们研究了惰性 MDP 的理论性质，表达了值函数并刻画了最优解。然后，我们根据经验证明，在惰性 MDP 中学习的策略通常具有某种形式的可解释性：通过构造，它们向我们展示了代理控制默认策略的状态。我们认为这些状态和相应的操作很重要，因为它们解释了默认策略和新的懒惰策略之间的性能差异。以次优策略作为默认

策略（预训练或随机），我们观察到，在 Atari 游戏中，代理能够获得竞争性性能，同时只在有限的状态子集中进行控制。

98-839，强化学习中的解释性：视角与立场

Agneza Krajna 等，2022.3.22，IEEE

简介：人工智能（AI）已经嵌入到人们日常生活的许多方面，让人工智能为人们做决策已经变得很正常。相对于其他机器学习范式，强化学习（RL）模型增加了可解问题的空间。一些最有趣的应用是在预期回报函数不可微的情况下，在未知或未定义的环境中运行，以及在算法发现方面，超过任何教师的表现，代理通过简单的反馈从实验经验中学习。应用范围及其社会影响非常广泛，仅举几个例子：基因组学、博弈（国际象棋、围棋等）、一般优化、金融投资、政府政策、自动驾驶汽车、推荐系统等。因此，通过解释来提高基于 RL 的系统的信任度和透明度至关重要。大多数关于人工智能中可解释性的文章都提供了与监督学习相关的方法，而在 RL 领域中很少有文章涉及这一点。造成这种情况的原因是信用分配问题、奖励延迟，以及无法假设数据是独立且相同分布的（i.i.d.）。本文试图系统地概述可解释 RL 领域的现有方法，并在现有方法的基础上提出一种新的统一分类法。立场部分描述了如何观察可解释性的语用方面。特别强调了接受解释和产生解释的各方之间的差距。为了缩小差距，实现解释的诚实和真实性，我们建立了三个支柱：主动性、风险态度和认识论约束。为此，我们对最短路径问题的简单变体进行了说明。

99-840，XAI 的信任和依赖——区分态度和行为测量

Nicolas Scharwski 等，2022.3.23，巴塞尔大学

简介：信任经常被认为是有效使用和实际部署人工智能的一个基本标准。研究人员认为，人工智能应该更加透明，以增加信任，使透明度成为 XAI 的主要目标之一。然而，关于透明度对信任的影响，关于这一主题的实证研究并不确定。对这种模糊性的一种解释可能是,信任在XAI 中的运作方式不同。在这份立场文件中，我们主张明确区分依赖的行为（客观）衡量标准和信任的态度（主观）衡量标准。然而，研究人员在试图获取信任时，有时似乎会使用行为测量，尽管态度测量更合适。基于过去的研究，我们强调，将信任和依赖分开是有充分的理论依据的。正确区分这两个概念可以更全面地理解透明度如何影响信任和依赖，从而有利于未来的 XAI 研究。

100-848，跨多个社交媒体平台的可解释错误信息检测

RaheeWalamhe 等，2022.3.23

印度浦那共生技术研究所，共生国际应用人工智能中心

内容：因为很大一部分人口依赖互联网获取信息，所以网络信息处理 (WIP) 对现代社会产生了巨大影响。社交媒体平台提供了传播信息的渠道，也为传播错误信息、在人群中制造混乱和恐惧提供了温床。检测错误信息的技术之一是基于机器学习的模型，然而，由于多个社交媒体平台的可用性，单独开发和训练基于人工智能的模型已成为一项繁琐的工作，因此，采用可解释的人工智能技术至关重要。作者在这篇文章中的贡献为：开发可跨多个社交媒体平台有效用于错误信息分类的分类器框架（当可用于训练的数据有限时，

这种方法会很有帮助)；上述架构的 DANN 实现优于 CoAID 数据集 [Cui, 2020] 上的最新结果；与 COVID-19 疫苗接种相关的新型错误信息数据集 (MiSoVac) 的开发，从新闻平台和社交媒体网站收集并展示了所提出的方法；用于解释目标标签的基于 XAI LIME 的方法在已开发的 DANN 模型中调用了更高的可信度，用于广义的可解释错误信息检测，值得关注的是，作者们的开发能够对相似领域的多个数据进行泛化的技术，在时间、处理和效率方面开发经济方法，无需在单个平台上训练模型，也考虑到社交媒体中的领域差异收敛到一般特征，并伴随着可解释的、值得信赖的采用范式来解决领域适应和可解释性。

**101-859，**领域知识驱动的可解释目标条件交互式弹道预测的伪标签

Lingfeng Sun 等，2022.3.30

加州大学伯克利分校，乔治亚理工学院，本田研究院

内容： 高度交互场景中的运动预测是自动驾驶中的一个具有挑战性的问题，在这种场景下，

需要准确预测交互代理的联合行为，以确保自动驾驶汽车的安全高效导航。在这个领域中，目标条件方法由于其性能优势和捕获轨迹分布中的多模态的能力而受到越来越多的关注。在本文章中，作者们研究了目标条件框架下的联合轨迹预测问题。特别是引入了基于条件变分自动编码器 (CVAE) 的模型，以将不同的交互模式显式编码到潜在空间中。然而，作者们发现香草模型遭受后塌陷，无法根据需要诱导信息丰富的潜在空间，为了解决这些问题，提出了一种新的方法来避免 KL 消失并用伪标签诱导可解释的交互式

潜在空间。伪标签允许使用者在交互中加入任意领域知识。在本文章中，作者们使用一个说明性玩具的示例来激发所提出的方法。此外，通过定量和定性评估在 Waymo Open Motion 数据集上验证了文章中提到的框架。

**102-860，**部署中的可解释机器学习

U-Bhatt 等，2020.6.10

卡内基梅隆大学

内容：可解释机器学习旨在通过特征重要性评分、反事实解释和有影响力的样本等技术，为各种利益相关者提供对模型行为的洞察。然而，在这一领域的最新进展中，没有对组织如何在实践中使用这些技术进行调查。本研究探讨了组织如何看待和使用利益相关者消费的可解释性。我们发现，大多数部署不是针对受模型影响的最终用户，而是针对机器学习工程师，他们使用可解释性来调试模型本身。实践中的可解释性与公共透明度的目标之间存在差距，因为解释主要服务于内部利益相关者，而不是外部利益相关者。我们的研究综合了当前可解释性技术的局限性，这些技术阻碍了终端用户使用它们。为了促进最终用户交互，我们开发了一个框架，用于建立明确的可解释性目标，包括关注规范性需求。

敬请关注联盟微信公众号
COPU开源联盟

扫描二维码
获取往期资料

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