深度信息技术（精品）专辑
第六期

- 开源协同与数字主权
- 欧拉（openEuler）操作系统
- 龙蜥（OpenAnolis）操作系统
- 开放原子开源基金会受捐孵化情况
- 木兰开源社区和木兰许可证
- 国产微处理器和图形处理器
- 京东方商业模式创新
- 工业互联网
- 云原生
- 机器学习可解释性案例

中国开源软件推进联盟
China OSS Promotion Union

目录

[1] 本刊编者：陈伟、鞠东颖

■ **机器学习可解释性案例**

开源协同与数字主权

开源协同有力支持互联网数字主权

COPU 梁志辉 鞠东颖

2021.12.8

应 IGF（互联网治理论坛）的邀请，中国开源软件推进联盟（COPU）、印度政府、谷歌、哈佛商学院、GitHub 参加 2021 年 12 月 8 日在波兰召开的互联网治理座谈会。这是 COPU 代表的发言。

开源、共享、协同是开源的基本特征。《2021 中国开源发展蓝皮书》指出：“当今开源已成为全球的一种创新和协同模式。”

中国开源运动发展的一个重要体验是：开源协同有力支持互联网数字主权。

举例来说，百度研发自动驾驶与无人驾驶，建立 Apollo 平台，自 2013 年至今，已发表 Apollo 10 个版本（Apollo 6.0 是第 10 个版本），使 Apollo 成为全球最活跃的自动驾驶与无人驾驶平台之一。

在 Apollo 的 10 个版本中，百度拥抱开源，汇聚全球 97 个国家、6.5 万名志愿开发者，开发了 60 万行开源代码，并协同全球 210 家合作伙伴（包括奔驰、宝马在内的企业、大

学、研究机构等)，共建自动驾驶与无人驾驶生态和供应链，通过互联网支持分布在各地的数字主权。

开源协同可以帮助解决各地由于地区性利益对数字主权的分割控制，可以降低政府在基础设施上的成本，并提高各数字主权协同体之间的信任。

近来，由华为等开发的鸿蒙 (OpenHarmony)、欧拉（openEuler）操作系统和生态以及供应链，由阿里云等开发的龙蜥（Anolis OS）操作系统和生态以及供应链，也采取与百度相似的开源协同建设的模式，在支持分布于各地互联网数字主权方面取得了良好的效果。

欧拉（openEuler）操作系统

谈欧拉（openEuler)

陆首群

2021.11.10

华为开发两款操作系统，一款是鸿蒙（Hongmeng)，另一款是欧拉（Euler)。我曾受邀对鸿蒙做过点评，现在受邀请点评欧拉。

欧拉操作系统原创于华为，2019 年宣布开源后，命名为 openEuler。

一、openEuler 是一款 Linux 操作系统，这在 Linux 发展中是一件好事。

Linux 自诞生以来发展很快,但重点一直放在建设 Linux 内核（Kernel）上面，作为桌面的 Linux 操作系统，主要代表是 RedHat 和 SuSE 的操作系统；不少人认为智能手机的安卓（Android）操作系统不算 Linux 操作系统（只是采用了 Linux 内核)，但 Linux 创始人 Linus Torvalds 认为安卓也是 Linux 操作系统；Intel、Nokia（以及后来的三星）绑架 Linux 基金会，它们开发的 Meego（及后来改造为 Tizen)，命名为 Linux 操作系统，失败了！所以 Linux 在 PC 桌面系统的全球市场占有率一直只有 2%左右很低水平。Linux 开发者不满足这种状况，他们在采用 Linux 内核的基础上开发

了一批Linux操作系统衍生版，如2019 Linux系统Top100排行榜：

1、MX Linux

2、Manjaro

3、Linux Mint

4、Debian

5、Ubuntu

6、elementary

7、Solus

8、Fedora

9、Zorin

10、Deepin

11、antiX

12、CentOS

13、KDE neon

14、PCLinuxOS

15、ArcoLinux

16、openSUSE

17、Pop!_OS

18、Arch

19、Kali

20、Puppy

21、FreeBSD

22、Lite

23、ReactOS

24、Peppermint

25、EasyOS

26、EndeavourOS

27、SparkyLinux

28、Lubuntu

29、Slackware

30、Tails

……

其中,受到国内借鉴的主要有:Deepin,Ubuntu,Fedora,CentOS,Open SuSE,Lite,Lubuntu,Slackware。

二. Linux 操作系统发行版可分为产品版、企业版、网络版三个等级

随着发行版等级的提高,要求发行版满足越来越严格的工作负载的需求,即要求性能、可靠性、稳定性、适应长期运行的安全性越来越高。

CentOS 是红帽研制的 Linux 企业级服务器操作系统版本 RHEL 的克隆,一些民间人士组成社区以 RHEL 倒逻辑方式开发 CentOS(CentOS 社区不是 Fedora)。

openEuler 与 CentOS 是不同的企业级 Linux 服务器的社

区版，但 openEuler 完全不同于 CentOS，openEuler 沿袭红帽的 RPM 技术路线，使之能提供企业级 Linux 发行版。openEuler 还可向用户提供网络版（或电信版），以满足长期可靠性和稳定性的要求。

三、今年 11 月 9 日华为将 openEuler 捐给开放原子开源基金会进行孵化，决不是有人说的：他们不能赚钱而甩锅给基金会，也不是为图虚名（而且要指出，基金会就是他们出钱组建的）。

将原创技术开源后申请到基金会孵化，这是近几年来国际上出现（中国也参与）的新事物。这次中国走进了全球首批改革创举的前列，十分有远见！

把原创技术申请到基金会孵化，有利于集结开发者力量，共建生态和供应链，避免过度分散，可加快创新！

以 OpenEuler 开源社区为例，已汇集数百厂商、3 万个开发者，捐献给基金会后，将能汇集更大力量；目前 openEuler（社区创新版）已支持麒麟、统信、SuSE、中科创达等主要合作伙伴开发其商业发行版，相信在这次会议后将有更多企业前来拜廟！

这次 openEuler 的捐献活动将有力推动我国开源的大发展！

四. openEuler 有无技术优势

我想集中谈谈内核中文件系统。

众所周知,Linux 内核中的文件系统是 Ext4,沿用至今，很早就想改革，5 年前，Linux 基金会邀请 IBM 的安全、文件资深专家曹予德，他设计了分区文件系统，并被 Linux 基金会聘用为 CTO,那时我也聘请他担任 COPU 的智囊团高级顾问，他还答应尽早访华。可是不久他被谷歌挖走。当时华为从三星引进两位美西方文件系统专家，他们与华为有专家工程师合作，自创开发了 F2FS（号称 EROFS 超级文件系统），我曾与两位专家交谈，曹先生也告诉我他们自创的 F2FS 比较先进，2018 年谷歌在搭载手机 Pixel3 进行 Fuchsia 操作系统（微内核、跨平台）试验时，就优选华为自创的 F2FS 文件系统，现在 OpenEuler 的 EulerFS 文件系统就是脱胎于 F2FS 的。

龙蜥（OpenAnolis）操作系统

关于龙蜥的问答

（COPU 问，龙蜥社区答）

阿里巴巴　陈绪　报导

今年 11 月 4 日，龙蜥开源社区宣布，由阿里云、统信软件等 14 家单位联合开源的龙蜥操作系统捐赠给开放原子开源基金会，已全票通过项目孵化评审。

龙蜥是基于 Linux 的新一代云原生服务器操作系统，支持 X86、ARM、龙芯(LoongArch)等多种芯片架构和计算场景，性能和稳定性经受住了历年双 11 的严苛考验，为云上典型场景带来 40%的综合性能提升，故障率降低 50%，兼容 CentOS 生态，支持一键迁移，并提供全栈国密能力，致力打造数字经济基础设施的新底座。

针对 COPU 的提问，龙蜥社区进行了整理，具体回复如下：

问：为什么要把龙蜥捐给开放原子开源基金会？

答：龙蜥社区 2020 年 9 月成立后，接连推出 8.2 和 8.4 两个社区版本，影响越来越大，基于龙蜥的发行版也越来越多。社区理事会逐渐认识到，需要一个更具公信力和中立性的平台来承载这个项目。

开放原子开源基金会（以下简称“开源基金会”）的宗旨十分符合龙蜥的发展方向，而且受工信部指导。为了龙蜥的长远发展，今年 10 月，经 14 家社区理事单位表决，一致同意将龙蜥（包括品牌和开源项目）捐给开源基金会。

捐赠后，龙蜥的运营、研发、成员招募等工作将在开放原子开源基金会的指导下由社区成员继续贡献，龙蜥社区也将加大投入，和开源基金会的其他成员单位、其他项目深度互动，共同推动技术的发展和产业的创新。

问：龙蜥社区捐给开放原子开源基金会的内容包括哪些？

答：主要包括四部分：

1、龙蜥操作系统（Anolis OS）和龙蜥品牌、商标及其附属（包括网站、公众号等）；

2、社区创新项目 KeenTune、sysAK、Express UDP、T-one 等；

3、社区构建服务、测试系统等基础设施；

4、域名：OpenAnolis.cn、OpenAnolis.org 及系列子域名。

捐赠和开源不同，捐赠之后，就相当于龙蜥的所有权归属于开源基金会了，龙蜥之后的发展方向也由基金会把握。

问：和国内其他 Linux OS 比，龙蜥有哪些特色和优势?

答：首先，龙蜥是社区成员共同打造的，来自包括阿里云、统信软件、龙芯等多家企业的代码贡献，所以也综合了各个企业的优势，包括功能全面、适配简单、稳定可靠、安全可控、创新领先等。

其中，来自阿里云的代码贡献，是经过大规模、长周期的云计算技术实践的沉淀，使龙蜥性能和稳定性更优、安全性更强，在 Linux 内核技术、云原生操作系统、混合部署等层面均具备自主维护和把控能力；统信软件将长久以来自主维护的、基于 CentOS 7 衍生的相关调优代码贡献给社区，维护社区版本采用的、统信软件自主研发的 DDE 桌面环境，并对其组件 SIG 组进行长期维护投入；同时，龙蜥有国内外知名芯片厂商如英特尔、ARM、龙芯等的积极贡献，他们主要在编译工具链、基础库、内核及驱动、虚拟化、系统加速库及系统架构等方面提供支持，协助龙蜥进行操作系统生态建设，使龙蜥具备了成为国际先进操作系统的技术基础。

具体来说，在系统安全方面，除网络安全领域的操作系统层面加固外，龙蜥提供了基于 OpenSSL 衍生的 BabaSSL，能在密码应用场景使用国密算法实现国产替代；它还内置了首个机密计算开源容器 Inclavare Containers，奠定了解决可信、可控问题的基础。

在系统性能方面，龙蜥搭载的 Alibaba Dragonwell（龙井）Java 编译器，在 SPECjbb2015 基准测试中获得了迄今为止业内最好的成绩。

在系统管理方面，龙蜥提供了完善的迁移工具套件以便用户无缝迁移，也提供了一个全栈覆盖内核与核心组件的跟踪和诊断工具，能够增强在系统和应用两个层面的可观测性和可靠性，让业务的监控和诊断更加简单易用。

在系统生态方面，由统信软件、红旗、万里红等国产操作系统厂商联合共建生态，基于其自有的软硬件适配中心，组建社区兼容 SIG 组，面向芯片、服务器整机、外设、数据库、中间件等基础软硬件进行适配工作，致力于为国内用户提供“开箱即用”的操作系统软件。

问：龙蜥为何有两个内核？是出于什么考虑？

答：龙蜥搭载了 RHCK 和 ANCK 两种不同版本的内核，这种模式可以让用户按需选择。其中，RHCK 与 CentOS 8 内核同源，主要解决兼容性问题，稳定性得到了充分验证；ANCK 融入了更多创新点，让用户在尝鲜的同时不需要承担过高的风险，该内核已在阿里云公共云上使用，稳定性也经过了规模化验证，还支持海光、飞腾、兆芯、龙芯、鲲鹏等多个平台。

问：龙蜥操作系统兼容 CentOS 生态，能看做是 CentOS 换壳么？

答：显然不是。兼容和换壳不能划等号。

龙蜥和 CentOS 都是基于开源 Linux 制作的。CentOS 停服必然会导致用户迁移过程中的兼容性问题和性能问题，这些都是用户最关心的。为了让用户放心将业务迁移到龙蜥上，社区做了大量工作，形成了一整套兼容性分析体系，比如对语言运行时的分析、对系统参数和行为的分析、对内核的接口分析等。

而且，龙蜥既兼容已有的 CentOS 8，也提供自研的云内核（ANCK），后者拥有更多功能特性，如，稳定性提升、高效计算、高性能网络/存储协议栈、混部场景的资源隔离增强等。

问：龙蜥和其他国产操作系统厂商的关系是什么？

答：我们和操作系统厂商是合作关系。龙蜥是开源的、开放的，我们希望构建以龙蜥为基础的操作系统生态，提高国家数字基础设施的产业创新能力，进而成为全球数字创新的基石。

目前，统信软件、中国移动云、中科方德等已经基于龙蜥发布了各自的发行版本，在政务、金融、电信、交通、电力等多行业进行试点和落地，覆盖云上和云下场景，既有物理机部署也有虚拟机和容器化部署。

问：龙蜥开源社区的运行模式，和其他开源社区有哪些不同？

答：一般开源社区是操作系统和芯片厂商的双螺旋模式，龙蜥不太一样，是铁三角模式。龙蜥开源社区里既有统信软件等操作系统厂商，也有英特尔、ARM、飞腾、兆芯、海光等国内外领先的芯片厂商，还有阿里云、联通云、移动云、天翼云等头部云计算厂商。这是一种全新的社区合作模式和新的操作系统开源生态。

主要的区别在于有云厂商的加入。云厂商既是操作系统最大的用户，也理解云计算用户的真实需求，并有动力长期投入提升操作系统的能力，还能将自己的创新体现在操作系统层面，同时通过大规模的快速迭代和试错来缩短国产芯片和硬件的成熟路径，这把现代操作系统与传统开源操作系统区分开来。

我们认为，这种全新的社区模式将构建一个更具生命力、灵活性和更可持续发展的生态。

问：社区成员单位接下来将会为龙蜥做哪些投入？

答：总体而言，我们既会在研发方面加码，也会在推动行业标准建设、创新孵化、人才培养等层面进行投入，以建设更完善的操作系统生态。具体而言：

在研发层面，统信软件等将联合其他操作系统厂商负责支持 X86、ARM、LoongArch 架构的内核开发、优化、加固等商业落地工作；社区里的芯片公司将为龙蜥代码中硬件代码

相关部分做主要贡献，并发布基于龙蜥和各自芯片的专用版本；阿里达摩院操作系统实验室已经成立，阿里云将招募更多操作系统研发工程技术人员和专家为龙蜥服务。

在推动行业标准建设层面，积极参与国家相关机构的评测和认证；举办操作系统行业主题峰会；帮助社区会员参与国际相关会议。

在创新孵化层面，推动基于龙蜥的商业版本快速落地；挖掘操作系统上下游生态中有潜力的开源软件创企，孵化出更多独角兽。

在人才培养层面，撰写龙蜥技术图书，并在社区和高校推广；联合高校，共同培养操作系统人才；举办行业大赛，奖金激励优秀的个人开发者、开发团队。

我们希望这些投入能帮助龙蜥社区健康成长，推动国家基础软件产业的发展。

开放原子开源基金会受捐孵化情况

开放原子开源基金会任务和孵化的项目

本刊编者[2]

一、开源基金会概述

开放原子开源基金会（以下简称“开源基金会”）是在民政部注册的致力于开源产业公益事业的非营利性独立法人机构。 开源基金会的服务范围包括开源软件、开源硬件、开源芯片及开源内容等，为各类开源项目提供中立的知识产权托管，保证项目的持续发展不受第三方影响，通过开放治理寻求更丰富的社区资源的支持与帮助，包括募集并管理资金，提供法律、财务等专业支持。

开源基金会是开源项目的孵化器、连接器和倍增器。通过对开源代码的开放治理以便于形成事实标准，连接产学研共建开源生态，为开源项目找到更多的应用场景。

二、开源基金会的任务

1、接受多家开源系统软件和基础软件捐赠，负责孵化；

2、继续开发社区版；

3、维护升级社区版（支持社区版迭代）；

4、支持在多种芯片架构下的开发社区版；

5、生态建设。

[2] 本刊编者：陈伟、鞠东颖

三、开源基金会孵化中的主要项目

1、2021 年 11 月 3 日前，来自阿里巴巴、百度、华为、浪潮、腾讯、360 和招商银行等公司和机构捐赠的开源项目：Alios Things、XuperChain、OpenHarmony：OpenHarmony、PIKA、TKEStack、TencentOS Tiny。

2、2021 年 11 月 4 日来自阿里云、统信软件等 14 家家单位捐赠的开源项目——龙蜥（OpenAnolis）操作系统

3、2021 年 11 月 9 日来自华为等公司捐赠的欧拉（openEuler）操作系统。

木兰开源社区和木兰许可证

木兰开源社区和木兰许可证

中国电子标准化研究院　杨丽蕴

“木兰开源社区”建立于2019年8月，是国家重点研发计划重点专项“云计算和大数据开源社区生态系统”的核心成果。推动国家科技及企业创新成果开源，加强“政产学研用”之间的沟通，推动开源成果转化落地，为各类开源项目提供中立托管，保证开源项目的持续发展。

目前木兰开源社区**在开源规则构建方面**，基于我国法律语境，现已形成分别面向构建开放生态，保持技术体系一致性，以及规范数据集使用等三类不同应用需求的木兰开源许可证族。其中，木兰宽松许可证是全球首个由中国主导研制发布的中英文双语国际通用许可证，已经国际开源协会 OSI（负责批准国际开源许可证申请）批准，可被全球开源组织和开源项目使用，为我国开源项目出海提供保障。同时依托中国电子标准化研究院优势，系统开展开源标准化工作，目前在研6项开源标准涵盖术语、许可证框架、元数据等，正在推进4项国标立项,为我国开源产业化健康发展提供支撑。

建设木兰开源社区，颁发木兰许可证的意义：

木兰许可证的颁发是在国内建设、完善开源支撑系统；自木兰许可证颁发以来,已有10万余个开源项目得到应用，

日本、韩国也表示了兴趣；木兰许可证的颁发意味着开源基于中国本土语义、技术、法律环境适用性的建设，服务于全球最大消费者市场的开源建设，意义重大！

在开源社区运营治理方面，木兰开源社区形成了 1 个开源社区+9 位技术委员会专家+N 位开源项目导师的“1+9+N”的自主开源孵化治理模式，帮助开源项目孵化成长、引入更多参与方的同时，为我国开源项目出海提供通道。木兰开源社区现已托管汇聚了 160 余项国家重点研发计划的开源项目成果，同时也接受企业项目捐赠，目前在的孵化项目有 SRS、PiFlow、PostMan、DADI、LinkWechat、OceanBase-Client、Kube-OVN、Skyline、zCore、Furion 等，涵盖音视频、云原生、网络、大数据、操作系统等方向。

在生态系统建设方面，木兰开源社区依托中国电子技术标准化研究院与国际组织交流合作的优势基础，与国内外开源基金会、技术组织等广泛建立合作关系，在开源标准化、开源项目联合孵化、测试认证等方面广泛开展合作，为我国开源项目出海提供通道，同时为构建国内国际协同的开源开放生态提供平台。结合中国云计算标准和应用大会、木兰峰会、中国开源黑客松、木兰技术开放日等活动，拉通产学研用各方，开展开源代码优化和跨项目联合开发，探索开源人才培养，构建协同创新的开放合作模式、丰富开源应用生态。

下面概述有关问题：

1、 颁发木兰许可证的意义何在（本国颁发的开源许可证在本国开源建设中的作用）？

“木兰宽松许可证”是国家重点研发计划“云计算和大数据开源社区生态系统”项目重要成果之一，MulanPSL v2 经过严格审批，于 2020 年正式通过 OSI 批准为国际类别开源许可证。OSI 表示“中文版的开源许可证可以鼓励广大中国社区积极参与开源，同时也是对已批准开源许可证列表的宝贵补充”。

此次通过认证意味着木兰宽松许可证（MulanPSL v2）正式具有国际通用性，可被任一国际开源基金会或开源社区支持采用，并为任一开源项目提供服务，是木兰开源社区的重要成果。同时，木兰宽松许可证是首个由中国开源产业界联合编制并通过 OSI 认证的开源软件许可证，也标志着我国开源界立足中国贡献全球方面取得突破性进展。

2、木兰许可证颁发以来有哪些用户申请采用？

已在国内 10 万余项开源项目中得到应用，其中宽松许可证典型应用如 OpenEuler、OpenGauss、方舟编译器、XiOUS、SRS、龙蜥等；公共许可证典型应用如 OceanBase、建木等。

3、木兰社区发展状况如何？建立木兰公共许可证和木兰宽

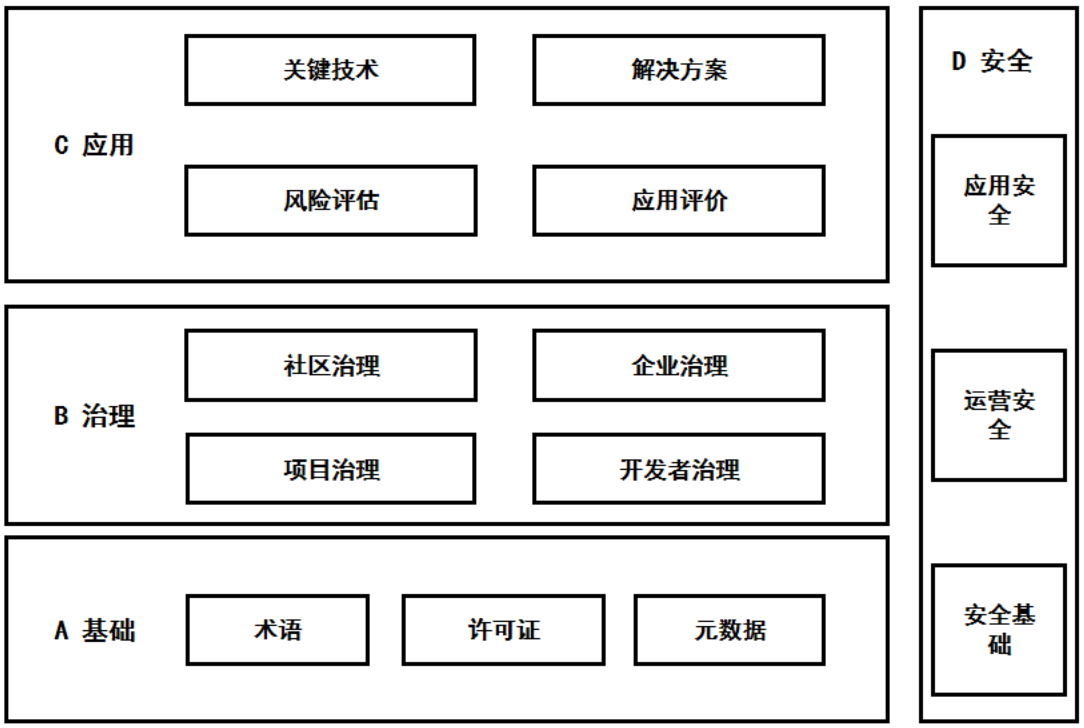
松许可证是应谁的请求？

社区发展情况详见第一页前面描述。

4、对开源标准化有哪些想法？

（1）国际开源标准化发展。国际三大开源基金会采取了不同方式在开源标准方面发力，标准化目标是社区治理、指导新项目建设以及推动实现互操作。

以 Linux 基金会为例，依托 CHAOSS 和 SPDX 开源社区，在国际 ISO/IEC 发布了首个开源流程和合规方面的国际标准 OpenChain:ISO 5230 和软件物料清单 SPDX:ISO5962,并被公认为国际开放标准安全性、许可证合规性和其他软件供应链的标准。



（2）国内开源标准化进展。目前初步形成标准体系草案，包括A基础、B治理、C应用、D安全等四个方面的开源标准体系；目前在研:开源术语、开源许可证框架、元数据、项目治理、社区治理、企业治理、开发者等七项开源标准。

京东方商业模式创新

京东方数据驱动商业模式创新

本刊编者[3]

2021 年 11 月央视新闻报导：投资 465 亿元，由中国企业自主设计、开发和建造的第 6 代柔性显示屏生产线在重庆京东方成功投产。该生产线将带动中国柔性显示产业上下游协同发展，加速柔性显示产品的应用普及。

近期，京东方向联盟汇报了工业互联网业务的进展，并介绍了集团的整体发展情况。

京东方凝聚 20 多年先进制造业数字化、智能化建设经验，通过全资子公司——北京中祥英科技有限公司拓展京东方工业互联网业务。围绕企业智能化生产、管理和协同需求，为客户提供高级排产、工业仿真、智能生产等业内领先的解决方案和平台产品，推动客户实现提质增效、降本减存、以数据为驱动的商业模式创新，专注于提供工业企业管理咨询、工业互联网解决方案及平台服务。依托人工智能、大数据、区块链等数字技术，已经为众多行业提供了丰富的解决方案，为客户持续创造价值。

附：京东方（BOE）发展简介

[3] 本刊编者：陈伟、鞠东颖

京东方科技集团股份有限公司（BOE）创立于 1993 年 4 月，是一家为信息交互和人类健康提供智慧端口产品和专业服务的物联网公司，形成了以半导体显示事业为核心，传感器及解决方案、MLED、智慧系统创新、智慧医工事业融合发展的“1+4+N”航母事业群。

作为全球半导体显示产业龙头企业，BOE（京东方）带领中国显示产业实现了从无到有、从有到大、从大到强。目前全球每四个智能终端就有一块显示屏来自 BOE（京东方），其超高清、柔性、微显示等解决方案已广泛应用于国内外知名品牌。全球市场调研机构 Omdia 数据显示，2020 年，BOE（京东方）在智能手机、平板电脑、笔记本电脑、显示器、电视等五大应用领域显示屏出货量均位列全球第一。

传感器及解决方案事业聚焦医疗影像、生物检测、智慧视窗、微波通信、指纹识别等领域，BOE（京东方）拥有从 12 英寸到 46 英寸的全尺寸 X-ray 平板探测器背板产品(FPXD)，广泛应用于欧美、日本、韩国等全球高端医疗器械公司；智慧视窗通过显示和传感技术创新，为交通、建筑等领域提供极具竞争力的传感器件及解决方案。

在 MLED 领域，BOE（京东方）以独有的主动式驱动架构、高速转印技术，为客户提供半导体工艺和先进微米级封装工艺的下一代 LED 显示系统及解决方案，目前已推出玻璃基 75 英寸 8K Mini LED、0.9mm 像素间距 Mini LED 显示产品等，

为人们带来全新的“视”界。

智慧系统创新事业通过人工智能、大数据、云计算技术，聚焦软硬融合的产品与服务，深耕智慧园区、智慧金融领域，聚焦智慧一体机、大尺寸智慧终端产品，为物联网细分领域提供整体解决方案。目前，BOE（京东方）智慧金融解决方案覆盖超过 1600 个网点，智慧园区解决方案在北京、天津、重庆等 20 余个城市落地应用，为人们生活带来更智慧的产品和服务。

智慧医工事业通过科技与医学融合创新，构建以物联网技术为支撑的智慧分级健康管理体系，形成智慧健康管理生态系统，构建以健康管理为核心、医工终端为工具、互联网医院及数字医院为支撑的全周期健康服务闭环，提升人们的健康生活品质。目前，BOE（京东方）已在北京、合肥、成都、苏州等地布局多家数字医院，提供以人为中心的全周期、全方位的健康管理服务。

截至 2020 年，京东方累计可使用专利超 7 万件，在年度新增专利申请中，发明专利超 90%，海外专利超过 35%，覆盖美国、欧洲、日本、韩国等多个国家和地区。美国专利服务机构 IFI Claims 发布 2020 年度美国专利授权量统计报告，京东方全球排名跃升至第 13 位，美国专利授权量达 2144 件，连续 3 年跻身全球 TOP20；京东方已连续多年在世界知识产权组织（WIPO）专利排名中位列全球前十。

BOE（京东方）在北京、合肥、成都、重庆、福州、绵阳、武汉、昆明、苏州、鄂尔多斯、固安等地拥有多个制造基地，子公司遍布美国、德国、英国、法国、瑞士、日本、韩国、新加坡、印度、俄罗斯、巴西、阿联酋等 19 个国家和地区，服务体系覆盖欧、美、亚、非等全球主要地区。

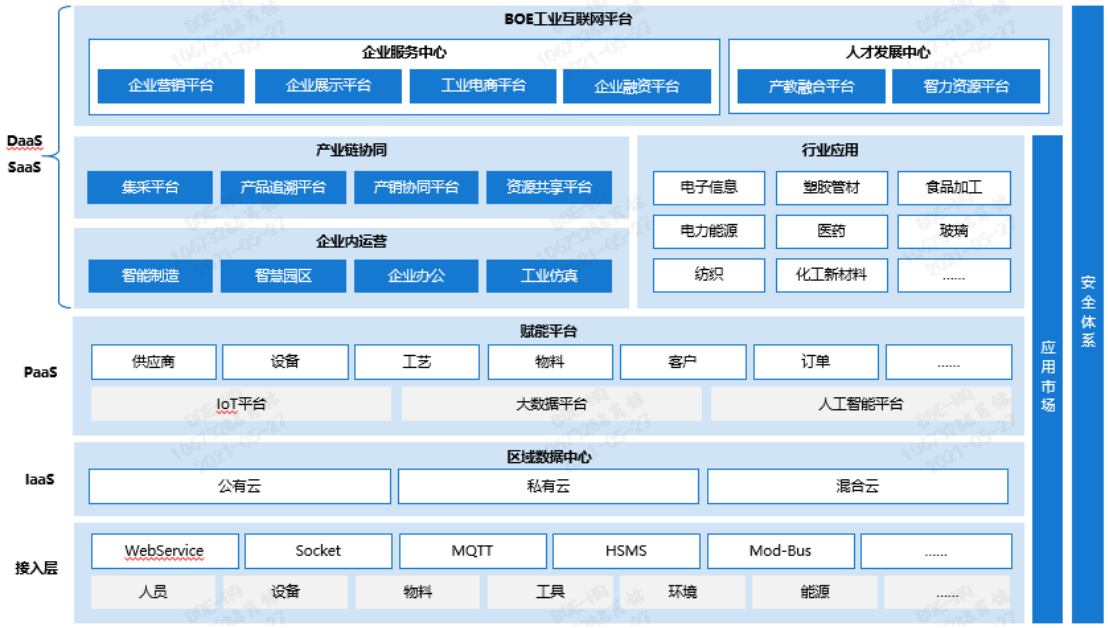
工业互联网

BOE 工业互联网平台

北京中祥英科技有限公司

一、平台简介

基于多年完备的制造体系，打造了泛半导体行业的全价值链工业互联网平台，打通了生产制造、园区管理到企业运营各环节，为客户提供支持多场景的平台和系统。



接入层：提供海量工业数据接入、转换、数据预处理和边缘分析应用等功能。一是工业数据接入，包括机器人、机床、高炉等工业设备数据接入能力，以及 ERP、MES、WMS 等信息系统数据接入能力，实现对各类工业数据的大范围、深层次采集和连接。二是协议解析与数据预处理，将采集连接

的各类多源异构数据进行格式统一和语义解析，并进行数据剔除、压缩、缓存等操作后传输至云端。三是边缘分析应用，重点是面向高实时应用场景，在边缘侧开展实时分析与反馈控制，并提供边缘应用开发所需的资源调度、运行维护、开发调试等各类功能。

IaaS 层：采用服务器、私有云和混合云多种形式进行部署实施。一是服务器部署，对于功能要求聚焦、资源容量不大的应用需求，可以将企业平台像普通应用软件一样安装部署在特定服务器之中进行操作访问，能够降低企业部署成本。但是由于服务器资源有限，未来平台能力拓展会受到一定限制。二是私有云部署，企业借助虚拟化、资源池化等技术支持，提供可灵活调度、弹性伸缩的存储和计算资源，支撑工业数据的管理和使用，确保所有核心数据停留在企业内部，避免敏感信息的泄露。三是混合云部署，企业在用私有云进行关键核心数据存储管理的同时，也使用公有云海量 IT 资源支撑，进行更为高效的业务处理，从而能够有效降低综合部署成本。

PaaS 层：提供资源管理、工业数据与模型管理、工业建模分析和工业应用创新等功能。一是 IT 资源管理，包括通过云计算 PaaS 等技术对系统资源进行调度和运维管理，并集成云边协同、大数据、人工智能、微服务等各类框架，为上层业务功能实现提供支撑。二是工业数据与模型管理，包

括面向海量工业数据提供数据治理、数据共享、数据可视化等服务，为上层建模分析提供高质量数据源，以及进行工业模型的分类、标识、检索等集成管理。三是工业建模分析，融合应用仿真分析、业务流程等工业机理建模方法和统计分析、大数据、人工智能等数据科学建模方法，实现工业数据价值的深度挖掘分析。四是工业应用创新，集成 ERP、SRM、MES、PMS 等生产管理、运营管理已有成熟工具，采用低代码开发、图形化编程等技术来降低开发门槛，支撑业务人员能够不依赖程序员而独立开展高效灵活的工业应用创新。此外，为了更好提升用户体验和实现平台间的互联互通，还需考虑人机交互支持、平台间集成框架等功能。

SaaS/DaaS 层：针对企业数字化发展短板，通过构建基础工业互联网平台，面向制造型的企业、产业、行业用户提供“产品智能化、产能数字化改造、不良品检测平台、数据资产化服务、产业大数据分析、应用生态集成等”服务，实现细分领域的工业互联网技术、数据和应用服务，充分解决跨行业跨领域的工业数字化转型问题。建设企业服务中心和人才发展中心，帮助企业扩充产品销售渠道和企业品牌营销平台，面向中小微企业现金流紧张的问题，可通过平台向金融机构申请快速贷款。建设人才发展中心，使生产与教育融合发展，提升管理人才和专业技术人才的能力，吸引高端企业或人才入驻。

二、实施案例

1、整体解决方案：

以泛半导体行业为例，当前企业面临了诸多挑战：其对自身工业园区管控更加精细化，要求实时掌握园区危险作业情况，及时响应紧急事件等；系统业务流程可执行性差，需对主线分段式工单管理，重工、抽检等业务流程进行改善；采集数据的自动化设备涉及多厂家、多 PLC 型号。

通过 BOE 工业互联网平台提供的整体解决方案，应用工业园区解决方案，实现园区内人员、车辆、设备、动环能耗多维度管控，异常情况快速识别告警；MES 产品的可便捷扩展性结合行业实践经验，协助客户完善系统业务流程，实现线上线下智能化管理；数据集成采用自有 EAS 产品，通过 Ethernet/IP 通讯，设备按工艺段集成管理，层次分明，且节约硬件和运维成本；成熟产品直接部署，试用

和适用优化同步进行，缩短项目交付时间。

实现对企业多维全局管控，数字化运营，提高管理效率；紧急情况准确定位，安全事件快速响应，保障工业园区安全；系统自动过账，节约作业人员成本，提高生产节拍，实现生产过程中账实一致；分享 BOE 管理经验，引领完善管理业务流程，打造标杆智造生产线。

2、SaaS 应用案例

某物流公司仓库业务量大，快进快出，异动频繁，防呆防错要求高；在全国有多个库房，采用多属地多仓库的运营模式，业务范围分布广，业务管理难度大；仓库内信息化程度低，无 IT 运维人员。

通过 BOE 工业互联网平台提供的标准化 WMS 应用，帮助物流企业完成仓储业务梳理及流程诊断的咨询服务，帮助客户完成智能云仓的整体规划设计；订阅式

SaaS WMS 产品，提供仓库管理涉及到的各类业务操作功能，实现仓储物流全生命周期管理；同时，多属地仓库集群数据联动，确保数据实时性和准确性；仓库智能化看板结合电子标牌产品实现仓库作业流程全程信息化、可视化管理。

为物流公司实现统一平台智能决策；仓库信息实时联动对接；信息共享，实现降本增效，成本降低 20%，效率提升 30%。

三、开源共创

工业互联网是链接工业全系统、全产业链、全价值链，支撑工业智能化发展的关键基础设施，是新一代信息技术与制造业深度融合所形成的新兴业态和应用模式，是互联网从消费领域向生产领域、从虚拟经济向实体经济拓展的核心载体。开源共创正是能够快速推广工业互联网大规模应用的核心。

工业软件是工业和信息产业的结合体，是将特定工业场景下的经验知识，以数字化模型或专业化软件工具的形式积累沉淀下来，是工业互联网的基石，也是工业互联网数据利用的关键，可以帮助工业互联网兑现价值。我国工业软件起步晚，CAD，CAE，PLM 等工业软件市场几乎被国外厂商垄断。近年来，我国提出“中国制造 2025”，多部门颁布智能制造发展政策，技术封锁加剧，加速了国产替代的进程，为我国

发展工业软件提供了发展舞台。拥抱开源，正是快速发展我国工业软件的捷径。中祥英与京东方的上下游企业，联合国内的高校，研究机构一起，积极拥抱开源，并在工业软件，AI 算法，云计算，边缘计算领域持续广泛合作，并期待有更多的志趣相投的伙伴拥抱开源，一起为工业软件的自主化贡献力量。

四、发展方向

中祥英开展工业互联网业务以来，秉承着成为中国制造“数智化”的使能者的企业愿景，积极围绕公司业务发展需求及京东方集团战略要求，将京东方智能制造行业的成功经验，通过工业互联网平台不断向其他细分市场拓展；以丰富的行业经验为核心，结合人工智能、工业大数据、工业视觉、数字孪生、边缘计算等高新技术赋能工业场景。

中祥英未来将持续打造具有市场竞争力和影响力的工业互联网平台；做透应用场景、打造极致客户体验；从服务于京东方及上下游产业链，将逐步拓展至显示、半导体、光伏、新能源等泛半导体行业，成为赋能企业数字化转型、智能制造的顶尖服务提供企业。不断强化技术能力创新，壮大产业生态；以当前的技术为基础，并紧盯新技术、新应用，发挥新一代信息技术优势，打造工业全要素、全产业链、全价值链互联互通的新型基础设施、新型应用模式和全新产业生态，激发数据要素作用，促进制造业数字化、网络化、智

能化升级。不断完善、拓展产品线，实现工厂、企业全流程覆盖，适配不同业务场景，为客户带来全新的制造过程数智化、工业园区数智化、企业运营数智化及产业链数智化，为客户持续创造价值。目标成为跨行业跨领域工业互联网平台的标杆，拓展合作伙伴，扩大“朋友圈”，带动全产业链互联互通，打造共生共赢、相互促进、互相迭代的生态体系，推动地区经济发展。

国产微处理器和图形处理器

自主研发国产微处理器和图形处理器

本刊编者[4]

一、国产微处理器（CPU）

CPU 作为 ICT 产业的核心基础元器件，是国家发展的一大“命门”。目前，在国际环境、产业政策、市场需求的联合驱动下，一大批国产 CPU 厂商在工艺、性能、生态建设等多个方面不断取得突破，为 CPU 的自主可控、安全可信做出了贡献，并在“好用”的市场化道路上逐渐迈向成熟。

在经历数十年的艰辛探索后，目前，国产 CPU 产业已初具规模，涌现出一批领军企业。根据 CPU 指令集体系进行分类：

1）复杂指令集（CISC）下，以 X86 架构为主，国内代表厂商包括海光、兆芯；

2）精简指令集（RISC）下，涉及 ARM 架构、Alpha 架构以及自研的 LoongArch 架构等,国内代表厂商包括鲲鹏(ARM)、飞腾（ARM)、龙芯（LoongArch)、申威（Alpha）等。

近两年发布的主要国产 CPU 介绍如下：

1、平头哥玄铁 910 处理器（开源 RISC-V 架构）

玄铁 910 是 2019 年 7 月阿里平头哥发布目前基于 RISC-

[4] 本刊编者：陈伟、鞠东颖

V 最强的开源处理器，采用 3 发射 8 执行的复杂乱序执行架构，单核性能达到 7.1Coremark / MHZ，主频达到 2.5GHZ，可用于设计制造高性能端上芯片，可应用于 5G、人工智能及自动驾驶等领域。

平头哥研发的玄铁 910，不仅性能较强，而且率先实现了适配安卓系统，这就意味着玄铁 910 不仅可以应用在工业领域，现在的手机等消费终端也可以使用。

2、龙芯处理器（LoongArch 架构）

2021 年 7 月，首款采用自主指令系统 LoongArch 的处理器芯片龙芯 3A5000 正式发布。

龙芯 3A5000 处理器主频 2.3GHz-2.5GHz，包含 4 个处理器核心，每个处理器核心采用 64 位超标量 LA464 自主微结构。集成了 2 个支持 ECC 校验的 64 位 DDR4-3200 控制器，4 个支持多处理器数据一致性的 Hyper Transport 3.0 控制器。支持主要模块时钟动态关闭，主要时钟域动态变频以及主要电压域动态调压等精细化功耗管理功能。龙芯 3A5000 处理器相较上一代龙芯 3A4000 处理器，性能提升 50%以上，功耗降低 30%以上。

3、华为鲲鹏 920 处理器（ARM 架构）

鲲鹏 920 处理器是华为在 2019 年 1 月发布的数据中心高性能处理器，由华为自主研发和设计，旨在满足数据中心多样性计算、绿色计算的需求。鲲鹏 920 处理器兼容 ARM 架

构，采用 7nm 工艺制造，可以支持 32/48/64 个内核，主频可达 2.6GHz，支持 8 通道 DDR4、PCIe 4.0 和 100G RoCE 网络。

4、飞腾 Phytium 通用计算处理器（ARM 架构）

FT-2000+/64 芯片集成 64 个飞腾自主研发的高能效处理器内核 FTC662，采用乱序四发射超标量流水线，芯片采用片上并行系统（PSoC）体系结构，集成高效处理器核心、基于数据亲和的大规模一致性存储结构、层次二维 Mesh 互联网络，优化存储访问延时，提供业界领先的计算性能、访存带宽和 IO 扩展能力。芯片兼容 64 位 ARMV8 指令集。 该产品适用于高性能、高吞吐率的服务器领域，如对处理能力和吞吐力要求很高的行业大型业务主机、高性能服务器系统和大型互联网数据中心等。

5、平头哥倚天 710 处理器（ARM 架构）

平头哥于 2021 年 10 月在“云栖大会”上发布倚天 710 通用服务器 CPU 芯片，采用业界最先进的 5nm 工艺，单芯片容纳高达 600 亿晶体管。

在芯片架构上，基于最新的 ARMv9 架构，内含 128 核 CPU，主频最高达到 3.2GHz，能同时兼顾性能和功耗。

在内存和接口方面，集成业界最领先的 DDR5、PCIe5.0 等技术，能有效提升芯片的传输速率，并且可适配云的不同应用场景。

为解决云计算高并发条件下的带宽瓶颈，倚天 710 针对

片上互联进行了特殊优化设计，通过全新的流控算法，有效缓解系统拥塞，从而提升了系统效率和扩展性。

在标准测试集SPECint2017上,倚天710以440分登顶，性能超出超过业界标杆20%，能效比提升50%以上。

二、国产图形处理器（GPU)

GPU,图形处理器又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。GPU按接入方式分成独立GPU与集成GPU,按应用端分为PC GPU、服务器GPU与移动GPU。PC和服务器GPU市场主要由Nvidia、AMD和Intel瓜分。

国产 GPU 的生产商包括两种：自主研发系以及引进系。自主研发系包括：中船重工709所、中船重工716、景嘉微、航锦科技、龙芯、上海兆芯等机构和公司；引进系则有凯桥资本收购的 Imagination。目前，国产 GPU 中，较为知名的为中船重工701所的凌久GP101、中船重工716所的JARIG12、景嘉微的JM7200。主要图形处理器（GPU）分别介绍如下：

1、凌久GP101

凌久GP101由中船重工709所研发,并在在2018年2月第一次流片成功。凌久GP101 GPU芯片支持HDMI、DVI、VGA等通用显示接口，支持2D、3D图形加速和OpenGL ES2.0，

支持 4K 分辨率、视频解码和硬件图层处理等功能，可以广泛应用于军民两用电子设备、工业控制、电子信息等领域。这标志着国内自主研发 GPU 芯片已经具备批量生产能力，为后续相关研发和国产化应用打下了坚实基础。

2、景嘉微 JM7200 系列

2018 年 9 月，景嘉微 JM7200 流片成功。JM7200 兼顾专用与民用市场，意味着国产 GPU 开始迈向民用市场。

JM7200/7201 采用 28nm CMOS 工艺，支持 4K 超高清显示，支持 4 路独立显示输出，支持多屏同时输出，提供多种丰富的外设接口，可高效完成 2D、3D 图形加速；支持 H.264、VC-1、VP8、MPEG2 和 MPEG4 等格式高清视频硬件解码，运行桌面系统时将 CPU 资源占用降至最低；提供符合 OpenGL 规范的驱动程序。

JM7200 芯片已完成与龙芯、飞腾、银河麒麟、中标麒麟、国心泰山、道、天脉等国内主要的 CPU 和操作系统厂商的适配工作，与中国长城、超越电子等十余家国内主要计算机整机厂商建立合作关系并进行产品测试，大力开展进一步适配与市场推广工作。

3、“风华一号”图形处理器（GPU）

2021 年 11 月 8 日 ，芯动科技发布最新科技成果：国产高性能图形处理器 GPU 芯片研制成功，定名为“风华一号”，即将在全国首发。

“风华一号”将用于5G数据中心、元宇宙、云桌面等高性能图形渲染领域。搭载的GDDR 6X是目前世界最先进的显存（英伟达的RTX3090和RTX3080搭载的显存就是美光独家供货的GDDR6X显存);Chiplet则是一种新型芯片封装技术。“风华一代”支持4K、HDMI 2.1等功能，以及安卓、Windows等各类操作系统。

人工智能

人工智能发展动态

陆首群

2021.11.23

当前人工智能研发的重点和趋势是从弱人工智能向强人工智能的转化。

1、弱人工智能的机器学习支持高效场景。

如人脸识别（商汤研究全球领先);

计算机视觉（旷视、商汤和阿里等数度在国际竞赛中获奖);

自动驾驶（L4 路况，Waymo 和百度处于前列);

国内研发出一批新抗生素（上世纪抗生素产生抗体);

基因医疗（谷歌用数千种蛋白质训练神经网络，并根据基因序列生命基本分子，探索基因医疗（Alphafold V1-2);

洪水预测系统（谷歌在印度、孟加拉等东南亚的研究);

六代战斗机（2019 年英国率先研制六代机原型“暴风雨”)。

2、研发可解释性机器学习，进入强入工智能（迄今 COPU 已汇集全球研究案例 70 多例）

3、异步脉冲神经网络和神经拟态计算系统

以英特尔研发的 Loihi-Pohoik Springs 大规模神经拟

态计算系统和曼彻斯特大学研发的 SpiNNaker 新颖大规模神经拟态计算+并行计算系统，最为突出。浙江大学基于达尔文-2 芯片及其集成支持的异步神经网络和神经拟态计算系统，正在跟踪研发中。

4、研发大规模语义网络（知识图谱）支持实现认知智能。这项研究由于未能直接理解知识、未能掌握知识推理的逻辑以及对常识、专业知识缺乏有效的知识表示和利用手段，以至离实现解决方案还很遥远。

5、下一代通用人工智能。目前一些国内外人工智能资深专家，提出了对下一代通用人工智能的研究课题。他们之中尚处于提出思路的阶段，距提出试点模型尚早，欲实现通用人工智能还是路漫漫。

6、脑机接口。在人脑中植入芯片，连接人脑皮层的神经元，对外连接计算机，由人脑思维指挥机器动作。目前中外已有几十例开展脑机接口工作，在治疗癫痫、瘫痪病人有奇效。近来已有采用脑外感应方式实现脑机接口（不用在脑中植入芯片）。

云原生

云原生为什么火?

本刊编者[5]

Kube 加入云原生，如虎添翼。

1、目前云原生已成为各行各业数字化转型的必修项，它对于软件的开发、部署和运维方式正在发生前所未有的改变；

2、云原生正在重塑企业数字化转型，构建现代企业应用技术基础架构的平台；

3、云原生技术是对企业应用开发方式的一次全方重构。利用容器、微服务等技术重写应用，利用 DevOps 重塑企业研发和运维流程，利用 GitOps、声明式架构重新定义企业的流水线和运维方式，利用可观测性和服务等级协议（SLA）升级原来的监控要求，利用云原生以身份为中心的安全体系，保障企业安全。

4、云原生技术成为主流或者趋势的原因：

云原生架构让开发可以快速持续发布，让服务按需快速伸缩，让系统更具弹性和可用性。

[5] 本刊编者：陈伟、鞠东颖

云原生国际峰会

本刊编者[6]

2021 年 12 月 9-10 日将在国内举行“云原生国际峰会”（KubeCon + CloudNativeCon + Open Source Summit China 2021，由 CNCF 主持）。受疫情影响以虚拟会议形式呈现。

作为 CNCF 的旗舰会议于 2018 年首登中国，大会每年都吸引全球 48 个国家的开源精英参会。今年将有 20 位来自全球各地（含中国）云原生资深专家线上作主题演讲（演讲内容都是开源的）。本次会议主题演讲的特点，除有关云原生容器部署、编排、管理等核心演讲内容，以及探讨云原生安全、多集群管理和云未来等前沿领域的内容外，一个特点是探讨金融云原生的内容（中国工商银行、浦发银行、华泰证券等均有主题报告）。

中国工商银行建设云原生基础设施平台，管理着各种异构硬件和大量资源，支持着成千上万业务，并为这些业务提供服务。作为一个多集群编排框架 Karmadar 的设计是针对 Kube Native API，这使生活变得更容易。工商银行大规模基础设施的关键挑战，基于 Kube 的多集群解决方案的评估和考虑，取得成绩、遇到问题和解决方法。

[6] 本刊编者：陈伟、鞠东颖

浦发银行报告“生态银行的数据云原生”的要点是：由生态银行的战略引发对金融科技理念、架构、管理的深层思考，围绕数字化时代最核心的数据要素，建设新一代的银行核心系统架构，在这一探索过程中，云原生技术成为浦发银行核心系统架构的基石性技术。

华泰证券的主题演讲是“证券数字化云原生演进”。

机器学习可解释性案例

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

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[7]，健壮性工具套件 ART[8] 和公平性工具套件 AIF360[9]。我们在第四集主要集中讨论了 AI 可解释性的技术背景及一个银行案例。

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

一、人工智能可解释性方法总结

AI 模型解释的背景：

[7] https://github.com/Trusted-AI/AIX360
[8] https://github.com/Trusted-AI/adversarial-robustness-toolbox
[9] https://github.com/Trusted-AI/AIF360

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 目标之前的映射关系,如果映射关系结构是清晰的,明确的, 就是可以解释的。

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

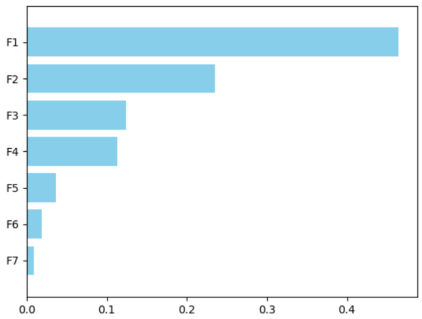
在特定场景,如果典型的,结构明确的算法模型(决策树或者线性回归)的预测或者识别的效果和复杂模型近似,就可以这些算法来解释复杂模型的算法。 比较成熟和应用广泛的就是对于一条实例的预测结果解释。 用复杂模型(比如深度学习,xgboost)对图像,或者文字,或者一个贷款记录作出了识别或者预测。 那么围绕着这条实例数据建立一个典型的线性回归模型(建立这个线性回归模型的数据由该1)条实例数据,2)该复杂模型和3)建立复杂模型数据共同生成, 以此来保证在这个实例上线性回归和复杂模型的等效性),用线性回归模型来解释这条记录的预测。0penScala 对于实例的解释采用该技术

▪从特征与目标之间的关系来理解和解释模型

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

特征重要性(Feature Importance)

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



（横轴是重要度的值，纵轴是各个特征，最上面的重要度最高的，往下依次降低）

2、选择特征

特征的选择基于既有的数据（客观存在）和 一些主观经验，可以使用一下方法

- **相关关系（correlation）法**

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

- **模型选择法**

将所有可特征作为预测变量，使用通用准确好的模型，比如 XGBoost。然后逐个从模型中去掉某特征变量，再建模，比较两次模型准确度变化，判断特征是否有用。

▪ **经验判断**

业务人员根据主观经验，预判哪些特征变量影响目标。

▪ **自动建模技术(Auto AI)**

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

3、依据特征和数据建模

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

4、根据模型求解算法

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

5、在计算基础上进行评估（人工或机器）

对于 AI 模型的预测结果评估有通用的评估方法，数据

一般会随机分为训练数据和测试数据两部分，训练数据主要用来训练模型，学习数据中的规律；测试数据对学习的结果评估，主要是从准确度角度，通过目标的观测值和预测值比较评估。区分回归模型（Regression）和分类模型（Classification）

- **回归模型（regression）评估指标**

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

- **分类模型评估**

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

混淆矩阵 ①

目标：DRUG

实测	预测					
	drugA	drugB	drugC	drugX	drugY	正确百分比
drugA	17	0	0	0	3	85.0%
drugB	0	13	0	0	6	68.4%
drugC	0	0	12	0	4	75.0%
drugX	0	0	0	47	6	88.7%
drugY	6	3	4	7	72	78.3%
正确百分比	73.9%	81.3%	75.0%	87.0%	79.1%	80.5%

不太正确 较为正确

6、进一步研究是否达到公平、公正、可信？！

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

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

二、人工智能可解释性案例分析

■ 案例分析一：AI 可解释性在银行贷款业务中的应用

1、背景

随着机器学习使用的不断普及，有时会被用来支持银行信用贷款审批流程，即针对用户贷款申请，通过机器学习模型来预测申请是被接受还是被拒绝。我们使用来自 FICO 可解释机器学习挑战赛的数据来讲述该场景，同时针对该场景中不同用户期望的解释来说明 AI Explainability 360

Toolkit （AIX360）的使用。此场景中涉及的三种类型的用户是：数据科学家，他在部署之前评估机器学习模型；信贷员，根据模型的输出做出最终决定；以及银行客户，他想了解申请结果的原因。

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

2、数据说明

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

下表列出了训练样本的主要特征，包括预测变量和目标变量。例如，NumSatisfactoryTrades 是一个预测变量，它计算过去与申请人签订的信用协议的数量，这些协议导致按时付款。要预测的目标变量是一个称为 RiskPerformance 的二元变量。 “差”值表示申请人在信用账户开立后的 24 个月内至少逾期 90 天或更糟一次。值“良好”表示他们已

付款，逾期未超过 90 天。预测变量和目标之间的关系为表中的最后一列。如果预测变量相对于坏的概率 = 1 单调递减，则意味着随着变量值的增加，贷款申请为“坏”的概率降低，即变得更“好”。例如，ExternalRiskEstimate 和 NumSatisfactoryTrades 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。ExternalRiskEstimate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，ExternalRiskEstimate 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，ProtodashExplainer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。

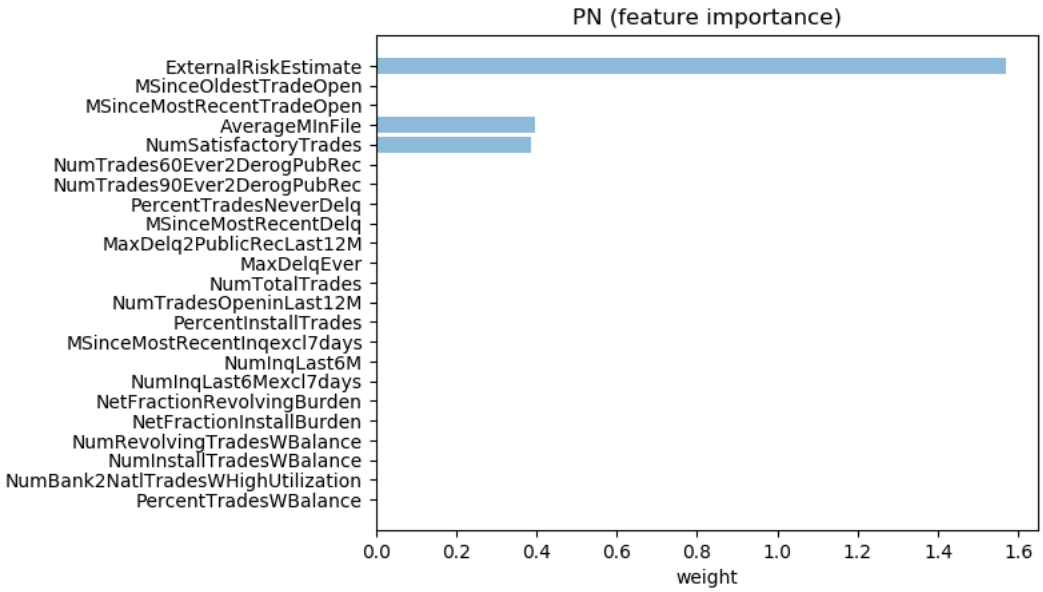
更进一步来说，对比解释算法输出由两部分组成：a) 相关否定 (pertinent negatives，PN) 和 b) 相关肯定 (pertinent positives，PP)。 PNs 识别一组最小的特征，如果改变这些特征将改变原始输入的分类。该方法实现这一点的方式是通过优化预测概率损失的变化，同时强制执行弹性范数约束，从而使特征及其值的变化最小。而 PP 标识了足以产生原始输入分类的最小特征集及其值，这里也有一个弹性范数项，因此所需的信息量最小。

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。

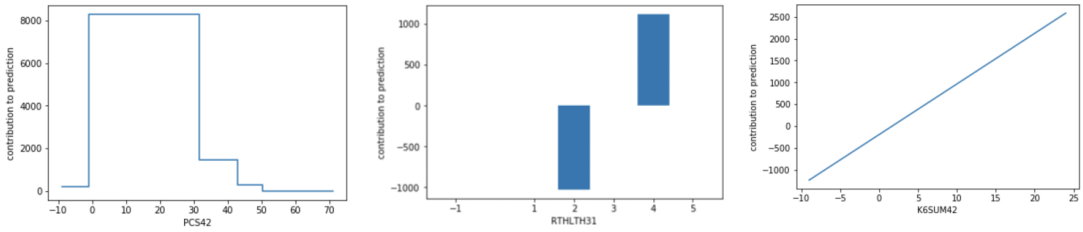


Figure 1　变量 PCS42，RTNLTH31，K6SUM42 与个人医疗支出的关系

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。

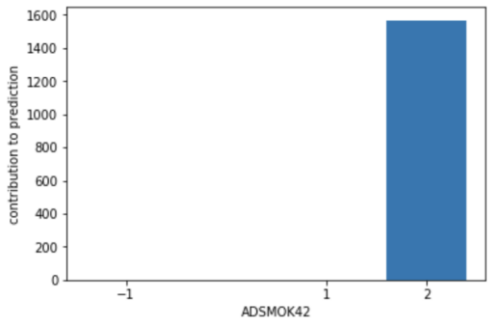


Figure 2 ADSMOKE 对个人医疗支出的影响

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。

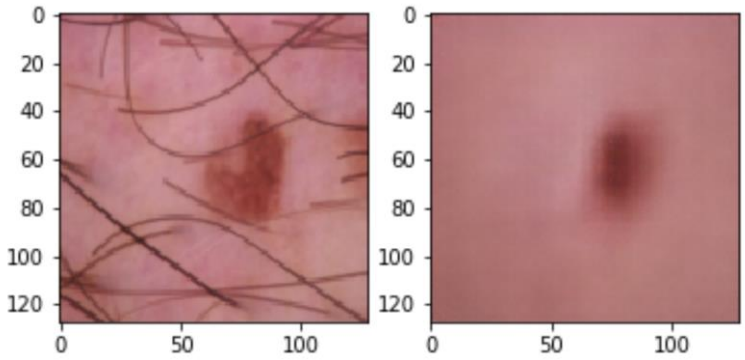


Figure 3 DIP-VAE 重建图与原始图对比

（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。

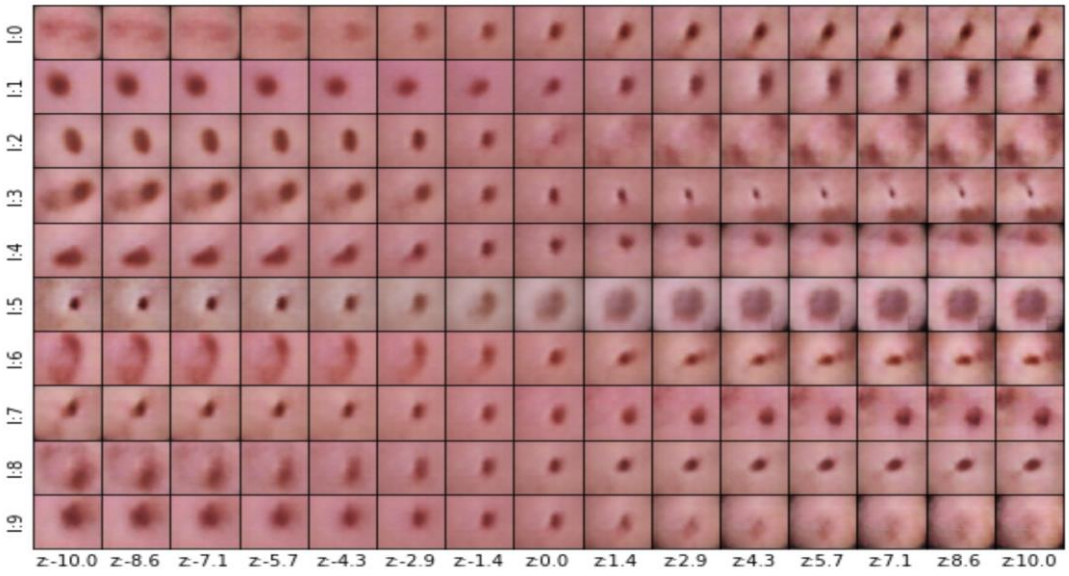


Figure 4 采用控制变量法得到可解释的高维隐藏特征

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。

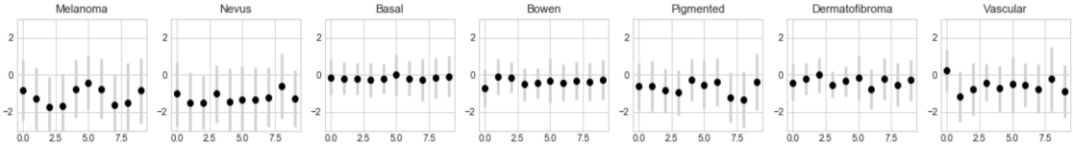


Figure 5 不同种类皮肤病下隐藏特征的分布

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

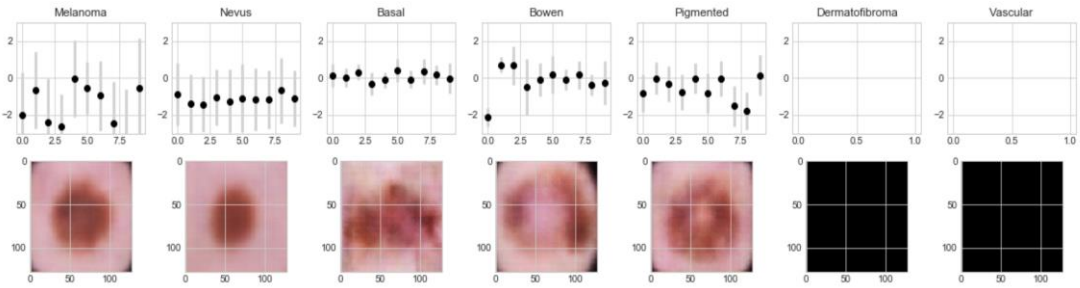


Figure 6 随机森林模型结果

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。

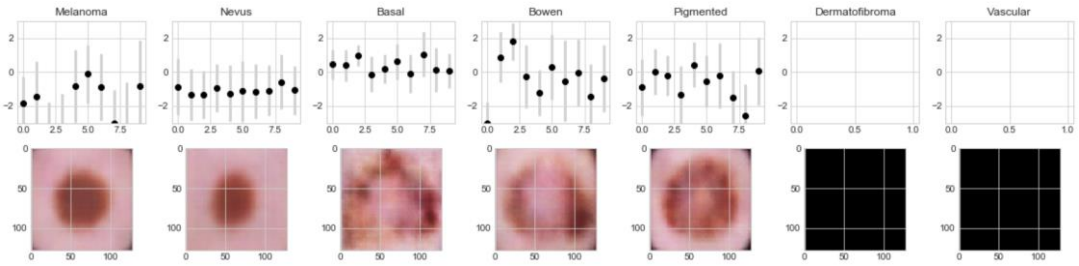


Figure 7 逻辑回归模型结果

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。

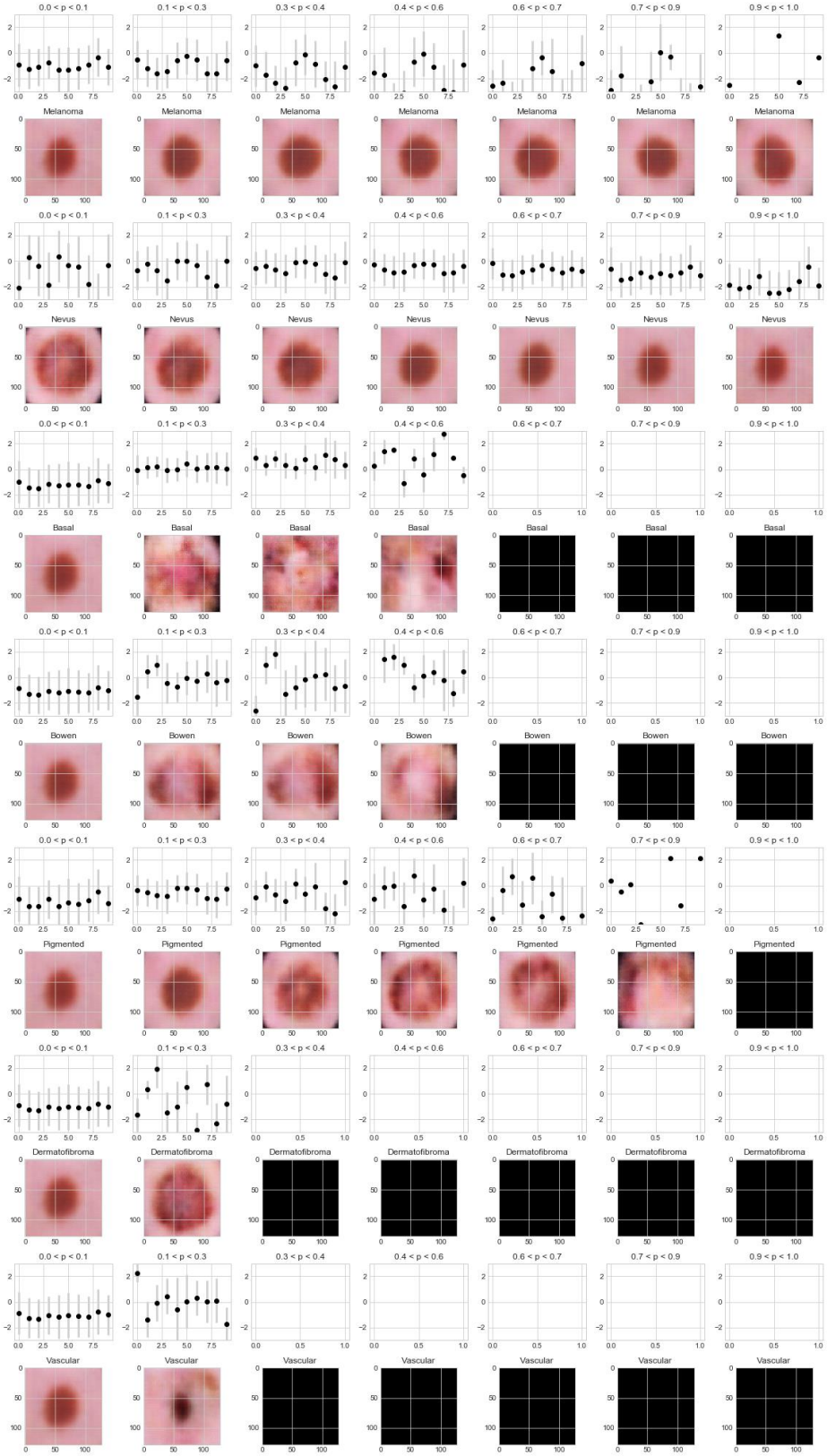

0.0 < p < 0.1
0.1 < p < 0.3
0.3 < p < 0.4
0.4 < p < 0.6
0.6 < p < 0.7
0.7 < p < 0.9
0.9 < p < 1.0
Melanoma
Nevus
Basal
Bowen
Pigmented
Dermatofibroma
Vascular

Figure 8

7、可解释性辅助模型评估

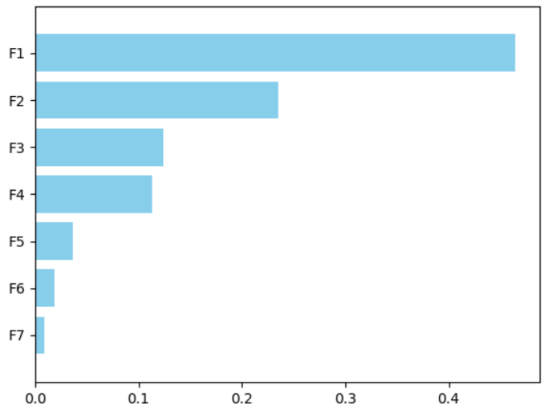
从本案例提到的两个简单机器学习模型来看，随机森林模型（精度 69%）和逻辑回归分类器（精度 65%），它们的精度略低于目前采用深度神经网络的最高精度（88%）。但这里有两点需要说明，首先这两个简单机器学习模型的输入，是由 DIP-VAE 提取的 10 维可解释的隐藏特征，由于存在手动设定的 10 维维度并不能完全表达所有重要的隐藏特征的可能性，所以如果增大设定的维数，有机会进一步提高机器学习模型的精度，或者采用其他的可解释机器学习模型，也可能获得更高精度。此外，即使按照现在的模型精度，虽然深度神经网络模型的精度更高，但是由于其推理过程的不可解释性，很难对其推理过程进行评估，在实际应用中受阻；而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[10]，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。

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

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。

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