计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190456
57(9): 1971-1986, 2020

机器学习的可解释性

陈珂锐[1]　　孟小峰[2]
[1](河南财经政法大学计算机与信息工程学院　郑州　450002)
[2](中国人民大学信息学院　北京　100872)
(chenke0616@163.com)

Interpretation and Understanding in Machine Learning

Chen Kerui[1] and Meng Xiaofeng[2]
[1](*School of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou* 450002)
[2](*School of Information, Renmin University of China, Beijing* 100872)

Abstract　In recent years, machine learning has developed rapidly, especially in the deep learning, where remarkable achievements are obtained in image, voice, natural language processing and other fields. The expressive ability of machine learning algorithm has been greatly improved; however, with the increase of model complexity, the interpretability of computer learning algorithm has deteriorated. So far, the interpretability of machine learning remains as a challenge. The trained models via algorithms are regarded as black boxes, which seriously hamper the use of machine learning in certain fields, such as medicine, finance and so on. Presently, only a few works emphasis on the interpretability of machine learning. Therefore, this paper aims to classify, analyze and compare the existing interpretable methods; on the one hand, it expounds the definition and measurement of interpretability, while on the other hand, for the different interpretable objects, it summarizes and analyses various interpretable techniques of machine learning from three aspects: model understanding, prediction result interpretation and mimic model understanding. Moreover, the paper also discusses the challenges and opportunities faced by machine learning interpretable methods and the possible development direction in the future. The proposed interpretation methods should also be useful for putting many research open questions in perspective.

Key words　machine learning; interpretation; neural network; black box; mimic model

摘　要　近年来,机器学习发展迅速,尤其是深度学习在图像、声音、自然语言处理等领域取得卓越成效.机器学习算法的表示能力大幅度提高,但是伴随着模型复杂度的增加,机器学习算法的可解释性越差,至今,机器学习的可解释性依旧是个难题.通过算法训练出的模型被看作成黑盒子,严重阻碍了机器学习在某些特定领域的使用,譬如医学、金融等领域.目前针对机器学习的可解释性综述性的工作极少,因此,将现有的可解释方法进行归类描述和分析比较,一方面对可解释性的定义、度量进行阐述,另一方面针对可解释对象的不同,从模型的解释、预测结果的解释和模仿者模型的解释3个方面,总结和分析各种机器学习可解释技术,并讨论了机器学习可解释方法面临的挑战和机遇以及未来的可能发展方向.

收稿日期:2019-06-21;**修回日期**:2020-04-14
基金项目:国家自然科学基金项目(91646203,61941121,61532010,91846204,61532016,91746115);河南财经政法大学学术创新骨干支持计划项目
This work was supported by the National Natural Science Foundation of China (91646203, 61941121, 61532010, 91846204, 61532016, 91746115) and the Young Talents Fund of Henan University of Economics and Law.

关键词 机器学习;可解释性;神经网络;黑盒子;模仿者模型

中图法分类号 TP181

纵观机器学习的历史发展进程,其最初的目标是从一系列数据中寻找出可以解释的知识,因而在追求算法性能的同时,也很注重算法的可解释性.典型的代表譬如线性感知机、决策树、k 近邻算法等.进入 20 世纪 80 年代之后,伴随神经网络的复苏,机器学习算法在设计时开始放弃可解释性这一要求,强调提高算法泛化的性能.神经网络的激活函数的选择不再局限于线性函数,而采用非线性的譬如 Sigmoid,tanh,Softmax,Relu 等函数,一方面其表示能力大幅度提高,另一方面,随着其模型复杂度的增加,算法的可解释性就更差.

然而,机器学习解释技术具有巨大的潜在应用空间.譬如科学家在知识发现的过程中,可解释的机器学习系统可以帮助他们更好地理解输出的知识,并寻找各种因素之间的相关性;对于一些复杂任务的端到端系统,几乎无法完全测试,也无法创建系统可能失败的完整场景列表,人类无法枚举出所有可能出现的计算上或者逻辑上的不可行输出,系统的可解释性对于系统的理解则至关重要;需要防范可能产生某些歧视的场景,即使我们有意识将某些特定的受保护类编码到系统中,也仍然存在考虑欠缺的先验偏见,譬如种族歧视[1-3]、性别歧视等.

对机器学习的可解释性需求不仅仅来源于上述的需求,同时还来源于法律法规.欧盟于 2018 年 5 月生效的 GDPR(General Data Protection Regulation)中有条例明确规定,当机器针对某个个体作出决定时,该决定必须符合一定要求的可解释性.

NIPS2017 的工作组曾针对“可解释性在机器学习中是否必要”这一问题展开激烈的讨论[4].并非所有的机器学习系统都需要可解释性,譬如邮政编码分类、航空器防撞系统等都是在没有人类干预的情况下运行,不需要解释.但是在医疗保健、金融等行业而言,模型的可解释性不仅重要而且非常必要.譬如在医疗保健方面,护理人员、医生和临床专家都依赖于新的医疗技术来帮助他们监控和决策患者护理,一个良好的可解释性模型被证明可以提高临床工作人员的解决问题的能力,从而提高患者护理质量[5-7].通常对于系统出现不可接受的结果且无法造成重大后果的情况下,或者在实际应用中,人们已经充分地研究和验证出现的问题,即使系统表现不太完美,人们也愿意相信系统的决定.在类似的场景下,对可解释性是没有需求的.

近几年来针对机器学习的可解释性综述性的工作陆续出现,每个学者从不同的研究角度和侧重点进行概述说明.

Miller[8]从哲学、心理学和认知科学的角度对解释的定义、生成、选择、评估和呈现给予说明,展现人们在研究机器学习可解释过程中的某种认知偏见和社会期望.Zhou 等人[9]认为机器学习缺乏解释既是实际问题也是道德问题,根据解释的概念和黑盒子系统的类型不同,对目前的解释技术进行了分类总结.

Gilpin 等人[10]重点描述了可解释技术在人机交互领域(human computer interaction, HCI)、黑盒模型和其他领域的应用说明.Carvalho 等人[11]阐述可解释性问题的重要性,并粗粒度地给出 3 种体系的分类归纳:Pre-Model VS In-Model VS Post-Model、内在(intrinsic) VS Hoc 以及特异性模型(model-specific) VS 不可知模型(model-agnostic).Brian 等人[12]提出可解释地预测与可解释模型之间的区别,前者侧重于解释模型预测的结果,通常以个体特征的贡献角度来诠释,而后者从模型本身出发进行解释.还有部分的研究者关注特定研究领域的可解释性.譬如:Zhang 等人[13]聚焦卷积神经网络(convolutional neural networks, CNNs)的可解释研究工作.Tjoa 等人[14]则关注医疗领域的可解释性工作.纪守领等人[15]侧重可解释技术的应用和安全领域的研究工作.

本文立足于机器学习的可解释技术,借鉴和扩展 Brian[12]提出的分类框架,对可解释技术的最新研究进展进行综述.一方面对可解释性的定义、度量进行阐述,另一方面针对可解释对象的不同,从模型的解释、预测结果的解释和模仿者模型 3 个方面,总结和分析各种机器学习可解释技术.

1 基础知识

1.1 可解释定义

目前,关于机器学习的可解释性没有明确的定义,Liu 等人[16]给出定义为:“解释是指解释给人听的过程”.Doshi-Velez 等人[17]也提出类似的定义.解释

意味着提供可理解的术语来说明一些概念.这些定义隐含地假设,解释是由一些可理解的术语表达概念来构成,这些概念是自包含的,不需要进一步解释.

目前文献中用于描述可解释性的英文单词有解释(interpretation)、解释(explanation)和理解(understanding). Montavon 等人[18]给出了区别定义:Interpretation 表示将抽象概念(例如预测类)映射到人类可以理解的领域中;Explanation 是一个可解释域的特征集合,用于解释给定实例的决策(譬如分类、回归等)处理过程;Understanding 指对模型的功能性解释.

1.2 形式化描述

令 $D=\{x_1,x_2,\cdots,x_m\}$ 表示包含 m 个示例的数据集,(x_i,y_i) 表示第 i 个样例,$y_i\in y$ 是示例 x_i 的标记,y 表示输出空间.给定一个数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 和一个预测器 p.

1) 模型解释

模型解释的任务是从数据集 D 和预测器 p 中建立映射 $f:(x^m\rightarrow y)\times(x^{n\times m}\times y^n)\rightarrow(x^m\rightarrow y)$,解释函数 $f_E:(x^m\rightarrow y)\rightarrow\varepsilon$,$\varepsilon$ 表示人类能理解的逻辑值.

2) 预测结果解释

预测结果解释的任务是从数据集 D 和预测器 p 中建立映射 $f:(x^m\rightarrow y)\times(x^{n\times m}\times y^n)\rightarrow(x^m\rightarrow y)$,解释函数 $f_E:(x^m\rightarrow y)\times x^m\rightarrow\varepsilon$,解释过程中使用数据记录 x^m 的特征值.

3) 模仿者模型解释

模仿者模型解释的任务是从数据集 D 和预测器 p 中建立映射 $f:x^m\rightarrow y$,解释模型函数 $f_E:(x^m\rightarrow y)\rightarrow\varepsilon$,且 $\varepsilon\approx y$.

2 主要研究方向和可解释研究框架

2.1 主要研究方向

解释技术最早出现在基于上下文规则的专家系统中,早在 20 世纪 70 年代,Shortliffe 等人[19]就讨论了解释专家系统决策过程的必要性.

目前,可解释技术的研究方向主要由表 1 所示,包括解释理论和机器学习的可解释.解释理论的研究最早可以追溯到 20 世纪 90 年代,Chajewska 等人[20]在通用的概率系统中提出解释的正式定义.最近则是 2017 年,Doshi-Velez 等人[21]提出的分别以应用为基础、以人类为基础、以功能为基础的 3 种方法的分类,从而评估机器学习的人类可解释性.其理论的发展伴随应用场景的变化而发生改变.

Table 1 The Main Research Directions of Interpretation and Understanding Technology

表 1 解释技术的主要研究方向

Research Direction	Example
Interpretation and Understanding Theory	Chajewska, et al[20], Doshi-Velez, et al[21], Mohseni, et al[22], Johnson, et al[23], Yetim[24], Corfield[25]
Interpretation and Understanding in Machine Learning	Based in Statistic Machine learning: Bastani, et al[26], Andrews, et al[27], Bondarenko, et al[28] Deep Learning: Dumitru, et al[29], Simonyan, et al[30], Zeiler, et al[31], Yosinski, et al[32], Hou, et al[33], Zhou, et al[34-35], Bach, et al[36], Ribeiro, et al[37], Chu, et al[38], Hinton, et al[39], Frosst, et al[40], Balan, et al[41], Che, et al[42]

对于机器学习的可解释技术发展而言,早期主要关注传统统计机器学习的方法,譬如基于规则的浅层模型的解释、决策树等.现阶段大部分的研究聚焦于深度学习的可解释性,无论是学界还是业界,越来越多的学者注意到深度模型可解释的重要性和急迫性,未来在此方向将出现更多的研究思路和解决方案.

2.2 机器学习的可解释研究框架

人类认知科学中,人类会根据详细的逻辑推理来做决定,于是可以通过一步一步地展示推理过程来清楚地解释决策是如何做出.在这种情况下,决策模型是透明的.另外,人类也会先做出直觉决策,然后寻求对决策的解释,这属于事后解释法.依据这 2 种建模哲学构建机器学习的可解释技术研究框架如图 1 所示:机器学习处理流程通常将数据集划分为训练集和测试集,训练数据集经过训练模型得到

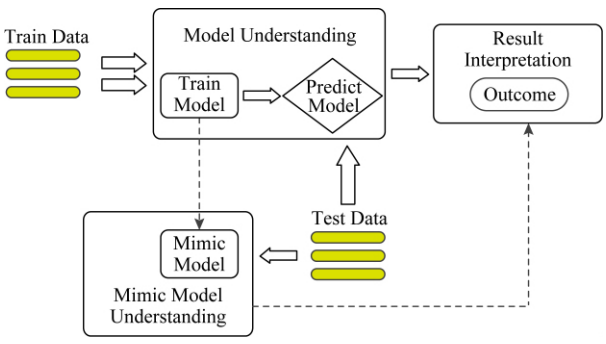

Fig. 1 The framework of interpretation and understanding in machine leaning

图 1 机器学习的可解释研究框架

预测模型,测试数据流入到预测模型,最终给出预测结果.围绕机器学习的处理流程,可解释工作主要围绕在模型和结果解释(result interpretation)两个环节上,对于模型的解释又分为模型解释(model understanding)和模仿者模型解释(mimic model understanding)两种方式,因此,本文将现存在的可解释技术按照上述的框架进行研究和总结分析.

3 机器学习模型的解释技术

3.1 基于规则的解释

基于规则的解释通常使用容易被人类理解的规则模型,譬如决策树和决策列表.Bastani 等人[26]提出一种学习决策树的模型提取算法,该算法对新输入的数据主动采样,并利用复杂模型对其进行标记,生成新的训练数据集,最后使用决策树作为全局解释.该学习决策树是非参数的,又高度的结构化,因此是可解释的.Andrews 等人[27]概括总结各种基于解释规则的方式,提供对复杂模型的理解.

除了树模型的规则解释之外,还有针对神经网络的规则提取.Bondarenko 等人[28]总结基于神经网络规则提取的分解法(decompositional rule extraction method),为网络中每一个隐藏单元都映射一条规则,最终形成复合规则库,并用于整个复杂网络的解释.

3.2 激活值最大化

激活值最大化思想主要是寻找能使一个给定的隐层单元的激活函数值最大的输入模式,即理解哪些输入会产生最大的模型响应.

Dumitru 等人[29]将激活值最大化技术应用于受限玻尔兹曼机(restricted Boltzmann machines, RBMs)进行叠加和自编码器去噪后所得到的网络中,通过研究网络中单个单元的响应,更好地深入理解该网络的体系结构和表示.

激活值最大化可看作一个优化问题,假设 θ 表示神经网络的参数(权重或者偏置),$h_{ij}(\theta,x)$ 是给定层 j 对给定单元 i 的激活函数,x 表示输入样本,ε 是用于解释的输入特征值,激活最大化的目标变为

$$\varepsilon=\arg\max h_{ij}(\theta,x). \tag{1}$$

式(1)问题通常是非凸优化问题,也就是该问题存在诸多个局部最大值.目前最简单易行的方法是通过梯度下降法(gradient descent)来寻找一个局部最大值.最终模型解释借助于一个或者多个最大值进行描述解释.

将上述的激活值最大化应用到深度置信网络(deep belief network, DBN)中,可转化为寻找 $\arg\max_x P(h_{ij}=1|x)$ 的问题.进而推广到深度神经网络(deep neural network, DNN)框架下,假定 DNN 分类器映射一系列数据 x 到一组类 ω_c 中,则转化为求解 $\max_x \log P(\omega_c|x)-\lambda\|x\|^2$.该问题在优化的过程中有诸多的优化策略,可以采取类似于 L2 范数正则化或者 Gaussian RBM 的专家策略,或者进行特定抽样,然后在 decoding 函数下映射到原始输入域.Simonyan 等人[30]将该方法推广到卷积神经网络上,构造了一个深度卷积网络 ConvNets,采取 L2 正则化进行优化.

激活值最大化方法相比于基于规则的解释,其解释结果更准确.但是该方法只适用于连续型数据,无法适用于自然语言处理模型.

3.3 隐层神经元分析

隐层神经元分析方法的主要思想是借助分析与可视化神经网络模型中隐层神经元的局部特征,从而解释深度神经网络的预测行为.该方法常见于图像处理领域.

对于隐层神经元的分析工作最初见于 AlexNet, Krizhevsky 直接可视化了神经网络的第 1 个卷积层的卷积核,其重构出的图像基本是关于边缘、条纹及颜色的信息,因此该方法局限于可视卷积核的 1 层卷积层[43].

Zeiler 等人[31]利用反卷积的方法对隐藏层的特征进行可视化,反卷积操作可看作卷积网络的逆操作.该方法以各隐藏层得到的特征图为输入,进行反卷积操作,最终得到的结果再去验证各个隐藏层提取到的特征图.实验结果表明经过 CNN 网络的学习,各个卷积层学习到的特征是有辨别性的.对于图像领域而言,隐藏层的最初几层主要学习到背景信息,而随着层数的提高,其学到的特征更加抽象.Zeiler 的工作真正可以可视化多层卷积层.

上述 2 种方法都属于无参数化的可视技术.提出的方法旨在可视化和理解具有最大池化和校正线性单元的卷积神经网络的特征,从而形成一个可视化的解释模式.

Yosinski 等人[32]在之前的可视化技术基础之上提出 2 种解释工具:第 1 种可视化实时卷积激活,可看出用户输入后如何实时地影响卷积层的工作;第 2 种利用图像空间中正则化优化技术,从而展示 DNN 每层的特征.Yosinski 在数据集 ImageNet 上进行训练,首先对所有的输入训练样本减去 ImageNet

中每个像素的均值,得到网络的输入数据 x 看作以0为中心的输入.然后构建一个目标函数:

$$\varepsilon=\arg\max_{x}(a_i(x)-R_\theta(x)), \tag{2}$$

其中,ε 是可视化的结果,$a_i(x)$ 是激活函数,而 $R_\theta(x)$ 是正则项.为了便于求解出结果,借助于公式

$$x\leftarrow r_\theta\left(x+\eta\frac{\partial a_i}{\partial x}\right) \tag{3}$$

进行更新.经过正则项来更新 x,Yosinski 等人[32]给出4种正则化方法:L2 衰变、高斯模糊、小范式裁剪像素(clipping pixels with small norm)和小贡献裁剪像素(clipping pixels with small contribution).

Yosinski 等人[32]提出的第2种工具属于参数化的可视工具,需要简单的配置安装,即可对 CNN 模型的隐层神经元进行可视化.

除此之外,隐层神经元分析解释的方法还可以借助重构图像的方法来实现,并取得较好的效果.

Dosovitskiy 等人[44]针对传统的计算机视觉图像特征 HOG(histograms of oriented gradient)[45],SIFT(scale invariant feature transform)[46],LBP(local binary patterns)[47]和 AlexNet 网络的每层特征2种类型进行图像重建.

类似的工作还有 Mahendran 等人[48],给定一个输入图片 $x\in\mathbb{R}^{C\times H\times W}$,其中 C 表示颜色通道,H 表示图片高度,W 表示图片的宽度,表征函数 $\Phi:\mathbb{R}^{C\times H\times W}\rightarrow\mathbb{R}^d$,特征值 $\Phi_0=\Phi(x_0)$,则重构图像 ε 即可表示为如下的目标函数:

$$\varepsilon=\arg\min_{x\in\mathbf{R}^{C\times H\times W}}\|\Phi(x)-\Phi_0\|^2+\lambda R(x), \tag{4}$$

其中,正则项优化主要采用 α 范式和总变差(total variation,TV),该方法主要优化了特征向量间的误差,并且需要借助人工设置的先验知识,比较而言,Dosovitskiy 等人[44]的工作更多地考虑图像重建误差,再者是根据隐型的方式学习了图像中的先验知识.

区别于前面几种图像领域的隐层神经元分析方法,侯博建和周志华[33]对递归神经网络(recursive neural network,RNN)进行解释,其基于如下的观察:如果 RNN 的每个隐藏层表示为一个向量或者一个点,向 RNN 中输入多个序列后,将出现积累大量的隐藏状态点,并且还倾向于构成集群的现象.这个观察假设在最小门控单元(minimal gated unit,MGU)、简版 RNN(simple RNN,SRN)、门控循环单元(gated recurrent unit,GRU)和长短期记忆网络(long short-term memory,LSTM)上通过实验都得以验证.于是,他们提出在训练数据集上训练 RNN,然后将验证数据集中所有隐藏层标注为一个点并执行聚类,最终学习到一个验证数据集上的有限状态机(finite state automaton,FSA),并用 FSA 对 RNN 做出解释,阐述 RNN 的性能如何受到门控数量的影响,以及数值隐藏状态转换背后的语义含义.该方法借用 FSA 对 RNN 内部机制进行透视.

虽然隐层神经元分析的方法提供了每个隐藏神经元的定性分析,然而该做法并不能对每个神经网络的整体机制提供太多可操作和定量的分析.

3.4 分离式表征

Zhou 等人[34-35]认为对于大型的深度神经网络而言,人类可理解的概念常常成为这些深度网络中的个体潜在变量,而这些网络可以自发学习分离式表征(disentangled representation),因而提出一种网路分割(network dissection)的方法来评估隐藏层和一系列语义概念之间的契合度,从而对深度网络做出解释.该方法处理如图2所示:

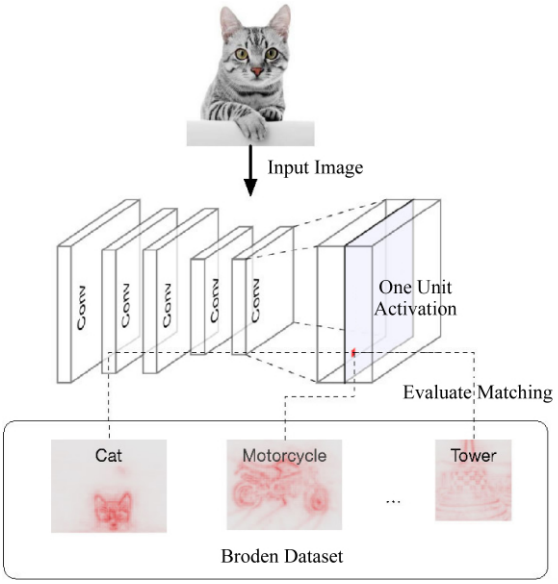

Fig. 2 The processing of disentangled representation

图2 分离式表征处理流程

分离式表征解释方法大致可以分成3步:

Step1. 人工创建一个视觉语义概念数据集 Broden,其中包含的每张图片都富含像素(pixel-wise)标签(颜色、纹理、场景、物体等),即对于每种语义概念都有一张标记映射(label map),图2中 Broden 数据集中每张图片标记成猫、自行车和塔等;

Step2. 对于一个训练好的模型 S,输入 Broden 所有的图片,收集神经网络中某个隐藏单元在 Broden 所有图片上的响应图,这些响应较大的区域即是该隐藏层的语义表征,将得到一个二值的 mask 值;

Step3. 利用 IoU 量化隐层的语义表征 mask 和概念对标记映射之间的匹配程度,从而利用标记映射(label map)解释神经网络的某隐藏层所表示的含义.

分离式表征的方法和 3.3 节中介绍的隐层神经元分析是一个相反的过程,前者是利用给隐藏单元计算匹配度并打标签的方式来正向解释隐藏层学习的特征,而后者是通过反向机制,重构各隐藏层的提取特征.分离式表征的解释方法效率较高,但是其准确度受限于语义概念数据集的大小以及其描述能力.

3.5 注意力机制

注意力机制(attention mechanism)[49]主要是在 Encoder + Decoder 模型框架下提出的,解决了该框架下输入数据中的各个部分具有相同权重的问题,为模型赋于区分辨别关键重要信息的能力.

目前广泛应用于图像处理[50-53]、自然语言处理[54]、语音识别[55]等领域,并取得较好的结果.在这些应用中依据对齐算法为每个部分赋予不同的权重,注意力机制可以很好地解释输入与输出之间的对齐关系,解释说明模型学到的内容,可以为我们打开机器学习模型的黑箱提供了一种可视方法.

Xu 等人[53]提出确定性软注意力(deterministic "soft" attention)和随机硬注意力(stochastic "hard" attention)两种机制.确定性软注意力是参数化的,可被嵌入到模型中直接训练.而随机硬注意力不会选择整个 Encoder 的输出为其输入,以概率采样的形式选择 Encoder 端输出的部分数据来进行计算,为了实现梯度的反向传播,通常需要采用蒙特卡洛采样的方法来估计模块的梯度.2 种注意力机制各有利弊,因为前者可直接求导,进行梯度反向传播,因此,目前更多的研究和应用倾向于使用确定性软注意力.

注意力模型中采用多种对齐函数[56]:

$$align(\boldsymbol{m}_t,m_s)=\frac{\exp(f(\boldsymbol{m}_t,m_s))}{\sum_{s'\in s}\exp(f(\boldsymbol{m}_t,m_{s'}))},\quad(5)$$

其中,

$$f(\boldsymbol{m}_t,m_s)=\begin{cases}\boldsymbol{m}_t^{\mathrm{T}}m_s, & \text{dot},\\ \boldsymbol{m}_t^{\mathrm{T}}\boldsymbol{W}_a m_s, & \text{general},\\ \boldsymbol{V}_a^{\mathrm{T}}\tanh(\boldsymbol{W}_a[\boldsymbol{m}_t;m_s]), & \text{concat}.\end{cases}\quad(6)$$

其中,$f(\boldsymbol{m}_t,m_s)$表示源端到目标端的对齐程度,常见有点乘(dot)、权值网络映射(general)和 concat 映射 3 种方式.

目前,注意力机制被用于解释各类任务的预测.Xu 等人[53],对于给定输入数据为图像,而输出数据为该图像的英文描述的任务,使用注意力机制来解释输出的英文描述中某个词语与图片中某个区域的高度依赖关系.

Chorowski 等人[55]采用基于混合注意力机制的新型端到端可训练语音识别方法,应用于基于注意力的递归序列生成器(attention-based recurrent sequence generator, ARSG)之上,借助内容和位置信息,选择输入系列中下一个位置用于解码,并很好地解释输入端的声音片段和输出序列的音素之间的对应关系.Bahdanau 等人[57]利用注意力机制表示输出序列中每个单词与输入序列中的某个特定单词的关联程度,从而解释法语到英语单词之间的对应关系.

Rocktäschel 等人[58]应用长短期记忆网络 LSTM 的神经模型,可 1 次读取 2 个句子来确定它们之间的蕴含关系,而非传统地将每个句子独立映射到一个语义空间方式.该模型利用逐词(word-by-word)的注意力机制解释了前提和假设中词和词之间的对应关系.

Rush 等人[59]设计了基于注意力机制的神经网络用于摘要抽取工作,注意力机制解释了输入句子和输出摘要之间的单词对应关系.

根据注意力的 2 种机制和对齐函数的分类标准,将各种神经网络的注意力机制整理成表 2 所示:

Table 2 The Summary of Attention Mechanism Explanation Methods

表 2 注意力机制解释方法总结

Method	Aligning Model Calculation	Type
Xu, et al[53]	concat	stochastic "hard" attention deterministic "soft" attention
Chorowski, et al[55]	concat	stochastic "hard" attention
Bahdanau, et al[57]	concat	soft attention
Rocktäschel, et al[58]	concat	soft attention
Rush, et al[59]	dot	soft attention
Choi, et al[60]	general	soft attention

注意力机制能否用于模型解释,目前仍存在一些争议.Jain 等人[61]认为基于梯度的机制下,注意力机制学习到的注意力权重值不总能够正确地解释特征的重要性,同时不同的注意力分布可能也会得到相同的预测结果,因此认为注意力机制不能作为模型解释的一种手段.部分学者认为其实验设计有诸多不妥,例如基准的注意力权重值是随意设置的,本应该由模型的其他图层参数共同决定;模型预测

结果的变化和注意力得分变化之间缺乏可比性等.本文认为注意力机制是可以被用来解释模型决策,但是该方法缺乏解释的一致性,相似的 2 个数据点,其解释的注意力分布和注意力权重值可能会有变化.

4 预测结果和解释技术

4.1 敏感度分析

敏感度分析[62]是研究如何将模型输出不确定地分配给不同的模型输入.该方法应用在预测结果的解释上,多数是建立在模型的局部梯度估计或者其他的一些局部变量测量的基础之上[63-65].该方法的理论基础来源于 Sundararajan 等人[66]认为深度学习模型具有 2 个基本公理:敏感性和实现不变性.

敏感度分析常使用如下的公式来定义相关性分数:

$$R_i(x)=\left(\frac{\partial f}{\partial x_i}\right)^2, \tag{7}$$

其梯度的值在数据点 x 处估计,最终输出那些最相关的输入特征,也即是最敏感的特征.该方法并不能解释函数 $f(x)$本身,仅能解释函数 $f(x)$的变化.

Cortez 等人[63-65]使用梯度和变量等因素来衡量敏感度的程度.另外,Baehrens 等人[67]引入解释向量来解释分类器分类的行为,其定义贝叶斯分类器为

$$g^*(x)=\arg\min_{c\in\{1,2,\cdots,C\}} p(Y\neq c\mid X=x). \tag{8}$$

而解释向量定义为

$$f_{\mathrm{E}}(x_0):=\frac{\partial}{\partial x}P(Y\neq g^*(x)\mid X=x)\mid_{x=x_0}, \tag{9}$$

其中,$f_{\mathrm{E}}(x_0)$和 x_0 维度相同,都是 d,分类器 $g^*(x)$将数据空间$\mathbb{R}^d$至多划分成 C 份,g^* 是常量.解 释向量 $f_{\mathrm{E}}(x_0)$在每个部分上都定义了 1 个向量场,该向量场表征是远离相应类的流向,从而具有最大值的 $f_{\mathrm{E}}(x_0)$中的实体突出显示了影响 x_0 的类标签决策特征,然后使用高亮技术可视化高度影响决策结果的那些特征,从而很好地解释决策结果.

为了更好地量化类似梯度、变量等因素的影响,Datta 等人[68]设计一套定量输入影响(quantitative input influence, QII)用于衡量模型的输入因素对预测输出结果的影响.

4.2 泰勒分解

采用泰勒分解的方法来解释预测结果,主要依靠分解函数值 $f(x)$为相关分数之和[69].简单的泰勒分解通过识别函数在某个根点 $\tilde{x}$ 处的一阶泰勒展开式的项,得到相关度的得分,该根点 $\tilde{x}$ 是满足 $f(\tilde{x})=0$ 的点,则一阶泰勒展开式为

$$f(x)=f(\tilde{x})+\sum_{i=1}^{d}R_i(x)+b=$$
$$0+\sum_{i=1}^{d}\frac{\partial f}{\partial x_i}\Big|_{x=\tilde{x}}(x_i-\tilde{x}_i)+b, \tag{10}$$

其中,$R_i(x)$为相关度分数,d 是输入数据的尺寸大小,b 表示二阶或者更高阶的多项式.对于多数的线性模型,譬如 ReLU 函数,其二阶或者更高阶的多项式趋向为 0,因此可以将式(10)简化为

$$f(x)=\sum_{i=1}^{d}R_i(x).$$

Li 等人[70]在泰勒展开式基础上,还利用表示绘图方法对自然语言处理(natural language process, NLP)领域中的文本进行解释.Montavon 等人[71]将其扩展为深度泰勒展开式,重新分配当前层和其下一层之间的相关度值.深度泰勒展开式为

$$f(x)=\left(\frac{\partial f}{\partial\{x_i\}}\mid_{\{\tilde{x}_i\}}\right)^{\mathrm{T}}\times(\{x_i\}-\{\tilde{x}_i\})+b=$$
$$\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\partial\frac{\partial R_j}{\partial x_i}\mid_{\{x_i\}}\times(x_i-\tilde{x}_i)+b, \tag{11}$$

其中,$\sum_{j\in\mathbf{N}}$ 表示当前层的所有神经元,$\sum_{i\in\mathbf{N}}$ 表示更低一层的神经元.通过将解释从输出层反向传播到输入层,有效地利用了网络结构.该方法借助空间响应图[72]来观察神经网络输出,同时在像素空间中滑动神经网络来构建热图.根据泰勒展开式的拟合特性,深度泰勒分解[71]准确度明显高于简单的泰勒分解[69],但前者比后者的计算量和复杂度更高.泰勒分解的方法适合神经网络下的各种简单或者复杂网络.

4.3 相关度传播

Bach 等人[36]提出的分层优化的相关度传播(layer-wise relevance propagation, LRP)从模型的输出开始,反向移动,直到到达模型输入为止,重新分配预测的分数或者相关度值,该方法常用于神经网络的预测结果解释.

1) 传播定义

假设 1 个 DNN 网络中具有 2 个神经元 j 和 k,j 和 k 所在的隐藏层是连续的,R_k 表示较高层的神经元 k 的相关度得分,$R_{j\leftarrow k}$表示神经元 k 到神经元 j 分享的相关度得分,则相关度的分配满足:

$$\sum_{j\in\mathbf{N}}R_{j\leftarrow k}=R_k. \tag{12}$$

具体传递流程如图 3 所示.w_{13}表示正向传播神经元节点 1 到神经元节点 3 的权重,$R_{1\leftarrow 3}^{\{1,2\}}$表示神经元节点 3 到神经元节点 1 在 1,2 层之间传播的相关得分.神经元之间的传递只能是连续层,不可跨层

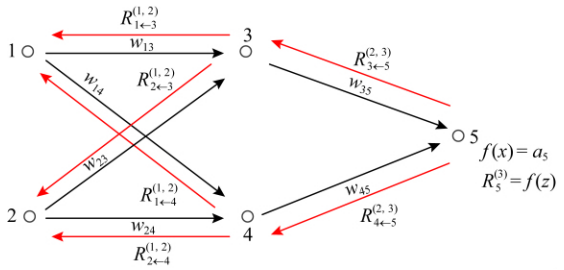

Fig. 3 The sample diagram of LRP propagation mechanism

图 3 LRP 传播机制示例图

传递,即不可能出现类似 $R_{1\leftarrow 5}^{(1,3)}$ 的情况.从传递机制可以看出,

$$R_{3\leftarrow 5}^{(2,3)}+R_{4\leftarrow 5}^{(2,3)}=R_5^{(3)}, \tag{13}$$

$$R_5^{(3)}\frac{a_3w_{35}}{\sum_{i=3,4}a_iw_{i5}}+R_5^{(3)}\frac{a_4w_{45}}{\sum_{i=3,4}a_iw_{i5}}=R_5^{(3)}, \tag{14}$$

$$R_3^{(2)}+R_4^{(2)}=R_5^{(3)}, \tag{15}$$

$$R_1^{(1)}+R_2^{(1)}=R_3^{(2)}+R_4^{(2)}. \tag{16}$$

2) 传播规则

针对 DNN 网络,使用 $\alpha\beta$ 原则实现相邻层之间的相关度传递.

假设 DNN 网络的神经元激活函数为

$$a_k=\sigma\left(\sum_{j\in\mathbf{N}}a_jw_{jk}+b_k\right), \tag{17}$$

其中,a_k 表示神经元 k 的激活值,j 表示神经元 k 所在隐藏层的前一层的所有神经元之一,w_{jk} 表示权重,b_k 为偏置项.

则 $\alpha\beta$ 原则定义为

$$R_j=\sum_{k\in\mathbf{N}}\left(\alpha\frac{a_jw_{jk}^{+}}{\sum_{j\in\mathbf{N}}a_jw_{jk}^{+}}-\beta\frac{a_jw_{jk}^{-}}{\sum_{j\in\mathbf{N}}a_jw_{jk}^{-}}\right)R_k, \tag{18}$$

其中,+表示正例,−表示负例,α 和 β 满足 $\alpha-\beta=1$,$\beta\geqslant 0$ 约束.从而不同的 $\alpha\beta$ 组合解释预测结果的不同行为.

不同的任务、不同的网络以及不同的数据上,各种 $\alpha\beta$ 原则组合表现出不同的效果.Montavon 等人[73]给出多种 $\alpha\beta$ 组合,譬如 $\alpha_2\beta_1$,$\alpha_1\beta_0$ 等,以及 $\alpha\beta$ 组合选取的原则,并且在实验中将敏感度分析、简单泰勒分解以及相关度传播的方法进行比较,明显看出其预测结果解释的准确度由大到小的排序为:相关度传播的方法、简单泰勒展开式、敏感度分析.

5 模仿者模型解释技术

模仿者模型解释方法的基本思想是通过训练一个可解释的模仿者模型 M 来解释复杂的原模型 S.相同的输入 $x_1,x_2,\cdots,x_N$,模仿者模型 M 和复杂的原模型 S 具有相似的输出,即

$$y_1\approx\tilde{y}_1,y_2\approx\tilde{y}_2,\cdots,y_N\approx\tilde{y}_N.$$

5.1 线性分类器拟合

局部解释法的主要思想是在一系列输入实例中采样获得一组近邻实例,然后训练出一个线性模型来拟合神经网络模型在该近邻实例上的决策输出,最后使用训练好的线性模型来解释复杂的神经网络模型.该方法典型的代表是 LIME[37],训练出的模型可用于本地局部解释预测结果,该方法适用于任何分类器的预测解释,作者还通过文本处理的随机森林模型和图像分类应用的神经网络数据集为例证明其方法的灵活性.

此类方法其优点在于模型设计训练过程简单,但由于近邻实例的抽样极具随机性,训练出的线性解释模型是不稳定的,极易造成对于相似的输入实例解释不一致的问题,以及对同一输入实例的多次解释不一致的问题,同时,近邻实例的选择也极大地影响解释结果的准确度.

Chu 等人[38]研究了激活函数为分段线性函数的分段线性神经网络(piecewise linear neural network, PLNN)的解释问题,提出 OpenBox 的解释模型.

以激活函数为 PReLU 的深度神经网络为例,其激活单元可分为 0 和 1 这 2 种情况,因为 PReLU 激活函数的线性性质,则可推导出无论神经元处于何种激活状态,其输入和输出始终保持线性关系.

解释模型 OpenBox 的处理流程如下所示:

给定一个输入实例 x,将隐藏层中所有神经元的激活状态按逐层顺序排列成一个向量 $conf(x)$,该向量的每一个元素为 0 或 1,也称为 PLNN 网中输入实例 x 的配置.

那么,对于单个输入实例的解释使用 PLNN 网络中输入实例 x 的配置 $conf(x)$.当 $conf(x)$ 的元素排列值不变时,PLNN 中所有隐藏层的计算等价于一个简单的线性运算 $Wx+b$,即可构造 $F(x)=Softmax(Wx+b)$ 的线性分类器.

为了解决解释一致性的问题,为 PLNN 的每个隐层神经元的输入 z 加上一组线性不等式约束 r,因为输入 x 和每个隐层神经元输入 z 是线性关系,则等价于对每个输入实例 x 加上一组线性不等式约束.因而,所有满足 r 中线性不等式约束的实例 x 都具有相同的 $conf(x)$,这些实例共享着相同的线性分类器.

对于总体的决策行为解释依靠一个线性分类器组来解释,不同的隐层神经元激活状态对应不同的 $conf(x)$,因此具有多个不同的线性分类器,这个分类器组可作为 PLNN 的解释模型.

该方法时间复杂度为线性的,具有较好的解释性能,但是局限性太强,仅能解释 PLNN 类的网络结构,对于其他复杂的网络显得无能为力.

5.2 模型压缩

采取模型压缩的方式模拟深度网络,训练出一个层数较少的浅层网络,这个新的浅层网络可以达到深度模型一样的效果,实验表明浅层神经网络能够学习与深度神经网络相同的功能[74-75].基于上述思想,研究出一系列的模仿者模型用于解释复杂的模型.

从原复杂的深度模型 S 到模仿者模型 M,多数是通过模型压缩途径获取.模型压缩技术的研究动机主要是为了引入机器学习到移动终端,但是设备处理能力有限,因而设计各种算法减少隐藏层的节点数量和模型层数.Lei 等人[76]通过减少隐藏层中节点的数量和输出层中多元音素(senone)方式压缩模型,最终在移动设备上安装 CD-DNN-HMM(context-dependent deep neural network hidden Markov model).Li 等人[77]利用最小化模型 S 和模型 M 输出分布之间的 KL(Kullback-Leibler)偏差进行层次压缩,使用对数模型和高斯模型之间的等价关系对多元音素进行压缩.多数学者利用压缩模型简单易解释特性用于复杂模型的可解释性工作.

5.3 知识蒸馏

知识蒸馏也称为模型蒸馏或模型模拟学习方法,属于模型压缩方法的一种.其基本思想是从预先训练好的大模型,蒸馏学习出性能较好的小模型.该方法有效地减小模型大小和计算资源.

Hinton 等人[39]提供一种高效的知识蒸馏的方法,蒸馏主要通过软性的 Softmax 概率来实现.对于 Softmax 的输入 z 而言,其对于每个子类的输出概率为

$$q_i=\frac{\exp(z_i/T)}{\sum_{j\in \mathbf{N}}\exp(z_j/T)}, \tag{19}$$

其中,当 $T=1$ 时,即为普通的 Softmax 变换,当 $T>1$ 时,即得到软化的 Softmax 的概率分布.通过式(19)生成软标签,然后结合硬标签同时用于新网络的学习.

最后用 KL 散度来约束模仿者模型 M 和原模型 S 的概率分布较为相似,即:

$$D_{\mathrm{KL}}(p^S,q)+\sum_{M\in A_k}D_{\mathrm{KL}}(p^M,q), \tag{20}$$

其中,p^M,p^S 分别表示模仿者模型 M 和原模型 S 的概率分布,A_k 表示一组模仿者模型,q 表示原模型 S 和模仿者模型 M 所包含所有类别的最小子集的概率分布.

Frosst 等人[40]在 Hinton 提出的知识蒸馏方法的基础之上,提出利用软决策树来模拟和解释复杂的原深度神经网络.

Balan 等人[41]利用蒙特卡洛算法实现从教师模型 S 中蒸馏出学生模型 M,并使 M 近似 S 的贝叶斯预测分布.该方法可简化问题的复杂性,但是大量的抽样将导致计算量较大.

Xu 等人[78]设计了 DarkSight 解释方法,利用蒸馏知识压缩黑盒分类器成简单可解释的低维分类器,并借助可视化技术对提取的暗知识进行呈现.

5.4 其他方法

Che 等人[42]利用梯度提升树(gradient boosting trees)来学习深度模型中的可解释特征,并构造出 GBTmimic model 对模型进行解释.其基本处理流程如图 4 所示:

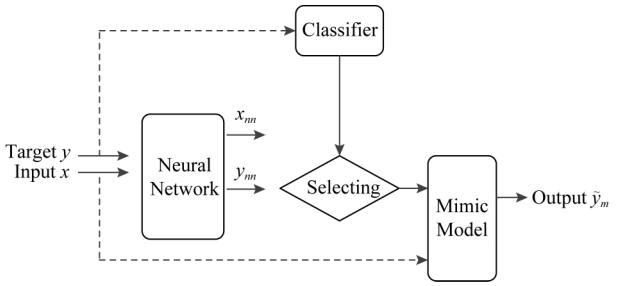

Fig. 4 The processing of GBTmimic model

图 4 GBTmimic 模型处理流程

给定输入特征 x 和目标 y,输入特征 x 进入原模型 S 后,输出 x_{nn} 和 y_{nn}.原模型 S 可能是多层降噪自动编码机(stacked denoising autoencoder, SDAE)或者 LSTM,都具有几个隐藏层和一个预测层,x_{nn} 是选择从最高隐藏层的激活函数中提出的特征,y_{nn} 是从预测层获得软预测分数.接下来,目标 y 和 x_{nn} 同时进入 Classifier,Classifier 选择 Logistics 回归,在相同的分类任务上,x_{nn} 进入分类器获得软预测分值 y_c.最后,选择 y_c 或 y_{nn} 以及特征 x 作为模仿者模型 M 的输入,通过最小均方差得到最终输出 $\tilde{y}_m$,模仿者模型即梯度提升回归树(gradient boosting regression trees).

Wu 等人[79]提出树规则化的方法,使用二分类

决策树模拟深度时间序列模型的预测,该模型比深度神经网络更容易解释,但是构造二分类决策树开销较大.

然而,由于模拟模型的模型复杂度降低,将无法保证具有超大维度的深神经网络可以被更简单的浅模型成功模拟,因此,模仿者模型不适合超深层的神经网路.同时,由于学习到一个更加简单的神经网络模型,解释模型的复杂度则达到某种程度的降低,而效果是以牺牲部分模型的准确度为代价而取得.

6 性能评估

对于可解释的评估因为任务的不同、解释模型的不同等诸多因素造成目前无法使用普适的方法.多数的方法都采用热力图[29,31-32]、错误率、准确率或者 AUC[42]等方法进行评估.为了考虑到可用性,Zhou 等人[9,34]引入人类评估作为基线.本文试图从可解释研究框架的角度给出如下的评估标准:

给定数据集 $D=\{X,Y\}$,对于任意的 $x\in X$,得到原预测模型 S 的值 $\hat{Y}=\bigcup_{x\in X}S(x)$,其解释模型 M 的预测值是 $\hat{Y}=\bigcup_{x\in X}M(x)$.

6.1 解释的一致性

一致性是指对于 2 个相似的数据点 x 和 x',其预测解释 $M(x)$和 $M(x')$也应该是接近相等的.解释的一致性可表示为:

$$\max_{x\neq x'}\frac{\|M(x)-M(x')\|_1}{\|x-x'\|_2}. \tag{21}$$

6.2 解释的选择性

Bach 等人[36]和 Samek 等人[80]提出解释的选择性可由相关度最高的特征移除时,激活函数 $f(x)$ 降低的速度来评价.该方法也称为像素翻转(pixel-flipping),不仅适合图像,同样适用于文本数据.

其执行过程循环的执行步骤为:

Step1. 计算 $f(x)$当前的值;

Step2. 找到最高相关度特征 $R_i(x)$;

Step3. 从特征集合中移除该特征 $x\leftarrow x-\{x_i\}$.

6.3 解释的准确性

准确性是指预测模型的准确度,可使用准确度值、$F1$ 指数等来衡量,构造一个可解释性模型,该模型自身的准确度依旧需要保持高精度,可解释模型的准确度为 $accuracy(\tilde{Y},Y)$.

6.4 解释的保真度

解释的保真度主要描述解释模型在何种程度上准确模仿原模型.针对黑盒子结果而言,利用其准确度、$F1$ 指数进行评价.保真度即是评估 $fidelity(\hat{Y},\tilde{Y})$.

此外,除上述几个与可解释性严格相关的指标外,机器学习的模型具有其他的重要因素,例如可靠性、鲁棒性、因果关系、可扩展性和通用性等,从而意味着模型能够独立于参数或者输入数据保持一定的性能(可靠性与鲁棒性),输入的变化影响模型行为的变化(因果关系),此外要求能够将模型扩展到更大的输入空间(可扩展性).最后在不同的应用场景中,人们会使用相同的模型和不同的数据,因此需要能够通用的解释模型,而非定制的受限的,这也将为性能评估方法提出巨大的挑战.

7 总结与展望

本文从机器学习模型的解释技术、预测结果的解释技术和模仿者模型技术 3 个方法总结了现有的关于机器学习的可解释技术,并总结其相关信息如表 3 所示.

7.1 可解释技术的内部问题

纵观当前机器学习的可解释技术,仍然面临着 3 个方面的挑战.

1) 准确性和解释性的均衡.伴随着模型的愈加复杂,提高最后预测的准确性,然后要求其预测的可解释性,必将意味着模型的复杂度受到一定程度的制约,预测模型需要牺牲部分准确度来满足可解释性,预测精度的损失是一个令人担忧的问题,因此,这一工作领域的中心重点是在保持可解释性的同时将精度损失最小化.

2) 解释一致性问题.输入一系列数据,经过预测模型,其解释机制给出一个解释.当下次再次输入相同或者类似的数据,解释机制是否能给出相同或者一致的解释是至关重要的,否则很难取得用户的信任,并将其真正地应用于实际项目中.

3) 评估问题.如何评估机器学习的解释质量的评价标准对于持续提升系统至关重要,因为只有这样才能更明确地有针对性地设计技术方案进行改进.机器学习的评估指标除了第 6 节中提到的之外,还有待于深入研究.

Doshi-Velez 等人[17]提出如何考虑评估机器学习模型的人类可解释性方法.他们提出 3 种类型的方法:以应用为基础,以帮助人类完成实际任务的程度来判断解释效果;以人类为基础,以人类偏好或根据解释对模型进行推理的能力来判断解释;以功能

为基础,以无人类输入的情况下,来判断代理模型的解释效果.对于这3种方法,皆假设结果数据的矩阵因子法有利于识别出解释性的常见潜在因素.

Mohseni等人[81]尝试将机器学习的可解释任务根据目标用户分为数据新手(data novices)、数据专家、机器学习专家3类,在每个类别下分别给出用户心智模型、用户-机器任务性能、用户解释的满意度、用户信任和信赖度和计算性能4个维度的评估.Mohseni认为机器学习的可解释性评估需要跨学科学者的共同努力,并充分考虑到人力和计算等要素.

7.2 安全和隐私性的问题

对模型研究得越透彻,意味着更大的风险,便于攻击者的攻击.无论是从数据上,还是从模型上,模型的可解释性和安全存在某种程度的相冲突.譬如模型训练阶段的数据投毒攻击[82]可造成模型的预测和解释失败;推理阶段根据解释技术中的激活值最大化和隐层神经元分析的方法,攻击者可以依据模型的解释机制而发起模型完整性攻击的模型推测攻击(model inversion attack)[83-85]和模型窃取攻击(model extraction attack)[86].

Table 3 The Summary of Interpretation and Understanding Methods

表3 可解释技术方法汇总表

Methods Type	Methods	Data Type	Task	Network Type	Test Dataset
Model Understanding	Ref [26]	No Limit	Classfication/ Reinforcement Learning	Random Forest/ Control Strategy	diabetes risk cart-pole
	Ref [29]	Image	Classfication	DBNs/SDAE	MNIST
	Ref [31]	Image	Classfication	CNNs	Caltech-101/Caltech-256/ PASCAL VOC 2012
	Ref [43]	Image/Video	Classfication	DNN	ImageNet
	Ref [48]	Image	Classfication	CNNs	ImageNet /ILSVRC 2012
	Ref [34-35]	Image	No Limit	CNNs	Broden/AlexNet/GoogLeNet/VGG-16/ ResNet-152/DenseNet-161
	Ref [53]	Image/Text	Machine Translation/ Object Identification	LSTM	Flickr9k/Flickr30k/MS COCO
	Ref [55]	Speech	Speech Recognition	ARSG	TIMIT corpus
	Ref [57]	Text	Machine Translation/	RNN	WMT '14
	Ref [58]	Text	Text Entailment	LSTM	Stanford Natural Language Inference
	Ref [59]	Text	No Limit	NNLM	DUC-2003/DUC-2004
	Ref [60]	Text	No Limit	RNN	EHR dataset
Result Interpretation	Ref [67]	Image	Classfication	Any Classfication	The Iris flower /USPS digits
	Ref [70]	Text	No Limit	LSTM/Bi-LSTM	Stanford Sentiment Treebank
	Ref [71]	Image	Classfication	DNN	MNIST
	Ref [73]	Image	Classfication	DNN	MNIST
Mimic Model Understanding	Ref [37]	Image/Text	Classfication	SVMs	religion dataset
	Ref [38]	Image	Classfication	PLNN	SYN/FMNIST-1 /FMNIST-2
	Ref [39]	Image/Video	Speech Recognition	Softmax Classfication	MNIST/JFT
	Ref [40]	No Limit	Classfication/Regression	DNN	ToyClass/MNIST /ToyReg/Boston Housing
	Ref [42]	Text	Classfication	SDAE/LSTM	VENT dataset

目前,越来越多的人开始关注深度学习的隐私保护问题,该问题的目标是保护训练数据集和训练参数.主流的做法是Shokri等人[87]提出的分布式训练方法,将噪声注入到参数的梯度中进行训练,以保护神经网络中的隐私.在该方法中,注入噪声的大小和隐私预算与共享参数的数量都将成比例地累积.因此,它可能消耗不必要的大部分隐私预算,多方之间的训练时期的参数数量和共享参数的数量通常很大,可用性较差.

机器学习的可解释方法是否可以提高深度学习的差分隐私的可用性?Phan等人[88]设计自适应拉普拉斯扰动机制,尝试将"更多噪声"注入到模型

输出“不太相关”的特征中.预测结果的解释技术中的分层优化的相关度传播 LRP 算法是不错的解决方案,实验表明在 MNIST 和 CIFAR-10 数据集上都取得不错的效果.Papernot 等人[89]提出的 PATE (private aggregation of teacher ensembles)模型借鉴了模仿者模型解释技术中的蒸馏知识技术,包含敏感信息的教师模型不能直接被访问,蒸馏出的学生模型可以被查询访问,从而有效地保护模型和数据的隐私.综上所述,将机器学习的可解释技术和深度学习的隐私保护技术相结合,同时有效地解决机器学习的隐私和可解释性 2 个问题成为一种可能.

7.3 研究视角的拓展

当前机器学习的可解释框架主要从模型和结果 2 个角度进行解释,具有一定的局限性.DeepMind 团队的 Rabinowitz 等人[90]试图以心智理论的视角来研究机器学习的可解释性问题,其研究目标是让观察者在有限的数据集的基础之上自动学习如何应对新的智能体建模,区别于以往的模仿算法,将学习如何像人理解人一样来预测另一个智能体的行为.其团队提出 ToMnet 模型改变以往尝试设计能表述内在状态的系统的做法,利用中转系统、人机接口,缩小原系统的行为空间大小,从而以人类更好理解的形式转述.同时,从训练数据分析的视角来解释机器学习的预测结果,也越来越被研究者所关注.譬如,Papernot 等人[91]提出的深度 k 近邻(deep k-nearest neighbors, DkNN)混合分类器将 k 近邻算法与 DNN 各层学习数据的表示形式相结合,有效地解决数据投毒攻击和模型可解释性 2 个问题.

目前,机器学习技术已渗入到数据库、图像识别、自然语言处理等多个研究领域,而机器学习的可解释技术必将影响着这些领域产品由实验室实验阶段走向工业实际应用的进程.

在数据库领域,将机器学习融入到数据库管理的索引结构[92]、缓冲区管理[93]和查询优化[94-95]等多个环节中,出现一种机器学习化的数据库系统趋势.一方面可提高数据库的处理速度;另一方面,数据库系统可智能自动调配数据库系统模块.然而,机器学习的其可解释性较差的缺点日趋凸显,再者机器学习化的数据库中重要的组成模块事务处理要求事务处理过程具有可追溯性和可解释性[96].因此,将可解释性引入到机器学习化的数据库中,可有效地帮助数据库设计者和使用者更快、更好地设计和使用数据库.

在自然语言理解领域,如何更好地利用知识和常识成为一个重要的研究课题.很多情况下,只有具备一定常识的情况下,才便于对机器做出更深入的解释和理解.在人机交互系统中需要相关领域知识,从而能更加准确地完成用户查询理解、对话管理和回复生成等任务,受益于类似人机交互系统通常需要相关的领域知识这一特点,提高了基于知识和常识的可解释性 NLP 的可能性.

多数学者将领域知识引入到机器学习中,主要出于处理小数据场景或者提高性能的考虑,极少考虑到领域知识也可看作解释技术的重要组成部分.Rueden 等人[97]首次提出知情机器学习(informed ML),对知识的类型、知识表示、知识转换以及知识与机器学习的方法的融合做出详细的分类说明.譬如知识类型可分为:自然科学、处理流程、世界知识和专家直觉.在该框架指导下,用户可以逐步选择合适的知识类型、知识表示和融合算法实现对机器学习模型的可解释和预测结果的可解释.

除此之外,知识图谱具有海量规模、结构良好、语义丰富等优点,使其成为机器学习理解语言的重要背景知识成为可能.肖仰华团队针对词袋[98]、概念[99]、实体集[100]和链接实体[101]做出一系列的解释工作,探索性地帮助机器理解和解释概念.然而,大规模的常识获取以及将符号化知识植入到数值化表示的神经网路都缺乏有效手段,这些问题将得到普遍的关注和研究.

再者不同的应用场景对于机器学习的可解释性的要求不同,如果仅是作为技术安全审查而用,专业的解释即可;如果当机器解释成为人机交互的一部分时,其解释必须通俗易懂.总之,机器学习的可解释性解决方案源于实用性的需求.

参 考 文 献

[1] Caliskan-Islam A, Bryson J J, Narayanan A. Semantics derived automatically from language corpora necessarily contain human biases [J]. arXiv preprint, arXiv: 1608.07187, 2016

[2] Angwin J, Larson J, Mattu S, et al. Machine bias [J/OL]. Propublica, 2016 [2018-11-01]. http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Courtland R. The bias detectives as machine learning infiltrates society, scientists grapple with how to make algorithms fair [J]. Nature, 2018, 558(7710): 357-360

[4] Samek W, Montavon G, Vedaldi A, et al. LNCS 11700: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning [G]. Berlin: Springer, 2019

[5] Kerr K F, Bansal A, Pepe M S. Further insight into the incremental value of new markers: The interpretation of performance measures and the importance of clinical context [J]. American Journal of Epidemiology, 2012, 176(6): 482-487

[6] Peleg M, Tu S, Bury J, et al. Comparing computer-interpretable guideline models: A case-study approach [J]. Journal of the American Medical Informatics Association, 2003, 10(1): 52-68

[7] Lowry S, Macpherson G. A blot on the profession [J]. British Medical Journal, 1988, 296(6623): 657-658

[8] Miller T. Explanation in artificial intelligence: Insights from the social sciences [J]. Artificial Intelligence, 2018, 267: 1-38

[9] Zhou Bolei, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene CNNs [C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2015 [2018-10-22]. https://people.csail.mit.edu/khosla/papers/iclr2015_zhou.pdf

[10] Gilpin H L, Bau D, Yuan Z B, et al. Explaining explanations: An overview of interpretability of machine learning [J]. arXiv preprint, arXiv:1806.00069v3, 2019

[11] Carvalho V D, Pereira M E, Cardoso S J. Machine learning interpretability: A survey on methods and metrics [J]. Electronics, 2019, 8(8): 832-866

[12] Biran O, Cotton C. Explanation and justification in machine learning: A survey [C/OL] //Proc of the IJCAI-17 Workshop on Explainable AI (XAI). San Francisco, CA: Morgan Kaufmann, 2017 [2018-12-01]. https://pdfs.semanticscholar.org/02e2/e79a77d8aabc1af1900ac80ceebac20abde4.pdf

[13] Zhang Quanshi, Zhu Song-Chun. Visual interpretability for deep learning: A survey [J]. arXiv preprint, arXiv: 1802.00614, 2018

[14] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Towards medical XAI [J]. arXiv preprint, arXiv:1907.07374, 2019

[15] Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096)

[16] Liu Yan, Sun Jimeng. Deep learning for health care applications: Challenges and solutions [C/OL] //Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017 [2018-10-11]. http://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf

[17] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning [J]. arXiv preprint, arXiv: 1702.08608, 2017

[18] Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks [J]. arXiv preprint, arXiv:1706.07979, 2017

[19] Shortliffe E H, Buchanan B G. A model of inexact reasoning in medicine [J]. Mathematical Biosciences, 1975, 23(3): 351-379

[20] Chajewska U, Halpern J Y. Defining explanation in probabilistic systems [C] //Proc of the 13th Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1997: 62-71

[21] Doshi-Velez F, Kim B. A roadmap for a rigorous science of interpretability [J]. arXiv preprint, arXiv:1702.08608, 2017

[22] Mohseni S, Zarei N. A multidisciplinary survey and framework for design and evaluation of explainable AI systems [J]. ACM Transactions Interact Intelligent System, 2018, 1(1): 1-37

[23] Johnson H, Johnson P. Explanation facilities and interactive systems [C] //Proc of the 13th Intelligent User Interfaces. New York: ACM, 1993: 159-166

[24] Yetim F. A framework for organizing justifications for strategic use in adaptive interaction contexts [C] //Proc of the 16th European Conf on Information Systems. New York: Elsevier, 2008: 815-825

[25] Corfield D. Varieties of justification in machine learning [J]. Minds and Machines, 2010, 20(2): 291-301

[26] Bastani O, Kim C, Bastani H. Interpreting blackbox models via model extraction [J]. arXiv preprint, arXiv:1705.08504, 2017

[27] Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks [J]. Knowledge-based Systems, 1995, 8(6): 373-389

[28] Bondarenko A, Zmanovska T, Borisov A. Decompositional rules extraction methods from neural networks [C] //Proc of the 16th Int Conf on Soft Computing. Berlin: Springer, 2010: 256-262

[29] Dumitru E, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network [J]. University of Montreal, 2009, 1341(3): 1-13

[30] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014 [2018-11-11]. https://arxiv.org/pdf/1312.6034.pdf

[31] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2014: 818-833

[32] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization [J]. arXiv preprint, arXiv:1506.06579, 2015

[33] Hou Bojian, Zhou Zhihua. Learning with interpretable structure from RNN [J]. arXiv preprint, arXiv:1810.10708, 2018

[34] Zhou Bolei, Bau D, Oliva A, et al. Interpreting deep visual representations via network dissection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2131-2145

[35] Bau D, Zhou Bolei, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3319-3326

[36] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PloS ONE, 2015, 10(7): 1-46

[37] Ribeiro T M, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144

[38] Chu Lingyang, Hu Xia, Hu Juhua, et al. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2018: 1244-1253

[39] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. arXiv preprint, arXiv: 1503.02531, 2015

[40] Frosst N, Hinton G. Distilling a neural network into a soft decision tree [J]. arXiv preprint, arXiv: 1711.09784, 2017

[41] Balan K A, Rathod V, Murphy P K, et al. Bayesian dark knowledge [C] //Proc of the 29th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 3438-3446

[42] Che Zhengping, Purushotham S, Khemani R, et al. Distilling knowledge from deep networks with applications to healthcare domain [J]. arXiv preprint, arXiv: 1512.03542, 2015

[43] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of the 26th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105

[44] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4829-4837

[45] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] //Proc of the IEEE Computer Society Conf on the Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 886-893

[46] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[47] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987

[48] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 5188-5196

[49] Cho K, Bart V M, Caglar G, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1724-1734

[50] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention [C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2015 [2018-11-02]. https://arxiv.org/pdf/1412.7755.pdf

[51] Gregor K, Danihelka I, Graves A, et al. Draw: A recurrent neural network for image generation [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 1462-1471

[52] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention [C] //Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2204-2212

[53] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2048-2057

[54] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend [C] //Proc of the 29th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 1684-1692

[55] Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition [C] //Proc of the 29th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 577-585

[56] Luong M, Hieu P, Christopher D M. Effective approaches to attention-based neural machine translation [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1412-1421

[57] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2015 [2018-11-13]. https://arxiv.org/pdf/1409.0473.pdf

[58] Rocktäschel T, Grefenstette E, Hermann M K, et al. Reasoning about entailment with neural attention [C/OL] // Proc of the 4th Int Conf on Learning Representations. 2016 [2018-10-13]. https://arxiv.org/pdf/1509.06664.pdf

[59] Rush M A, Chopra S, Weston J. A neural attention model for abstractive sentence summarization [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 379-389

[60] Choi E, Bahadori T M, Sun J, et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism [C] //Proc of the 30th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 3504-3512

[61] Jain S, Wallace C B. Attention is not explanation [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: NAACL, 2019: 3543-3556

[62] Saltelli A. Sensitivity analysis for importance assessment [J]. Risk Analysis, 2002, 22(3): 579-590

[63] Cortez P, Embrechts M J. Opening black box data mining models using sensitivity analysis [C] //Proc of IEEE Symp on Computational Intelligence and Data Mining (CIDM). Piscataway, NJ: IEEE, 2011: 341-348

[64] Cortez P, Embrechts M J. Using sensitivity analysis and visualization techniques to open black box data mining models [J]. Information Sciences, 2013, 225: 1-17

[65] Cortez P, Teixeira J, Cerdeira A, et al. Using data mining for wine quality assessment [G] //LNCS 5808: Proc of Int Conf on Discovery Science. Berlin: Springer, 2009: 66-79

[66] Sundararajan M, Taly A, Yan Qiqi. Axiomatic attribution for deep networks [J]. arXiv preprint, arXiv: 1703.01365, 2017

[67] Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions [J]. Journal of Machine Learning Research, 2010, 11: 1803-1831

[68] Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems [C] //Proc of IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2016: 598-617

[69] Bazen S, Joutard X. The Taylor decomposition: A unified generalization of the oaxaca method to nonlinear models, AMSE Working Papers 1332 [R]. Marseille: Aix-Marseille School of Economics, 2013

[70] Li Jiwei, Chen Xinlei, Hovy E, et al. Visualizing and understanding neural models in NLP [J]. arXiv preprint, arXiv: 1506.01066, 2015

[71] Montavon G, Lapuschkin S, Binder A, et al. Explaining nonlinear classification decisions with deep Taylor decomposition [J]. Pattern Recognition, 2017, 65: 211-222

[72] Hansen K, Baehrens D, Schroeter T, et al. Visual interpretation of kernel-based prediction models [J]. Molecular Informatics, 2011, 30(9): 817-826

[73] Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks [J]. arXiv preprint, arXiv: 1706.07979, 2017

[74] Bucila C, Caruana R, Niculescu-Mizil A. Model compression [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2006: 535-541

[75] Ba J, Caruana R. Do deep nets really need to be deep? [C] //Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2654-2662

[76] Lei Xin, Senior A, Gruenstein A, et al. Accurate and compact large vocabulary speech recognition on mobile devices [C] //Proc of the 14th Annual Conf of the Int Speech Communication Association. Piscataway, NJ: IEEE, 2013: 662-665

[77] Li Jinyu, Zhao Rui, Huang Jui-Ting, et al. Learning small-size DNN with output-distribution-based criteria [C] //Proc of the 15th Annual Conf of the Int Speech Communication Association. Piscataway, NJ: IEEE, 2014: 1910-1914

[78] Xu Kai, Park H D, Yi Chang, et al. Interpreting deep classifier by visual distillation of dark knowledge [J]. arXiv preprint, arXiv: 1803.04042, 2018

[79] Wu M, Hughes C M, Parbhoo S, et al. Beyond sparsity: Tree regularization of deep models for interpretability [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1670-1678

[80] Samek W, Binder A, Montavon G, et al. Evaluating the visualization of what a deep neural network has learned [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(11): 2660-2673

[81] Mohseni S, Zarei N, Ragan D E. A survey of evaluation methods and measures for interpretable machine learning [J]. arXiv preprint, arXiv: 1811.11839, 2019

[82] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines [C/OL] //Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2012 [2018-12-01]. https://arxiv.org/abs/1206.6389

[83] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014 [2018-11-20]. https://arxiv.org/pdf/1312.6199.pdf

[84] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[85] Wang Yue, Si Cheng, Wu Xintao. Regression model fitting under differential privacy and model inversion attack [C] //Proc of the 24th Int Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 1003-1009

[86] Qiao Linbo, Zhang Bofeng, Zhao Ruiyuan, et al. Online mining of attack models in IDS alerts from network backbone by a two-stage clustering method [C] //Proc of the 26th Int Parallel and Distributed Processing Symp Workshops & PhD Forum (IPDPSW). Piscataway, NJ: IEEE, 2012: 1263-1269

[87] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310-1321

[88] Phan N, Wu Xintao, Hu Han, et al. Adaptive Laplace mechanism: Differential privacy preservation in deep learning [C] //Proc of the 17th IEEE Int Conf on Data Mining (ICDM 2017). Piscataway, NJ: IEEE, 2017: 385-394

[89] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data [C/OL] //Proc of the 5th Int Conf on Learning Representations, 2017 [2018-10-11]. https://arxiv.org/pdf/1610.05755v1.pdf

[90] Rabinowitz C N, Perbet F, Song H F, et al. Machine theory of mind [J]. arXiv preprint, arXiv:1802.07740, 2018

[91] Papernot N, McDaniel P. Deep *k*-nearest neighbors: Towards confident, interpretable and robust deep learning [J]. arXiv preprint, arXiv:1803.04765, 2018

[92] Kraska T, Beutel A, Chi E H, et al. The case for learned index structures [C] //Proc of the 2018 Int Conf on Management of Data. New York: ACM, 2018: 489-504

[93] Margaritov A, Ustiugov D, Bugnionz E, et al. Virtual address translation via learned page table indexes [C/OL] //Proc of Workshop on ML for Systems at NeurIPS Co-located with 32nd Conf on NIPS. 2018 [2019-02-20]. http://www-users.cselabs.umn.edu/clases/Spring2019/csci8980/papers/virt_addrtrans.pdf

[94] Marcus R, Papaemmanouil O. Deepreinforcement learning for join order enumeration [C] //Proc of the 1st Int Workshop on Exploiting Artificial Intelligence Techniques for Data Management Co-located with SIGMOD. New York: ACM, 2018: 3:1-3:4

[95] Ortiz J, Balazinska M, Gehrke J, et al. Learning state representations for query optimization with deep reinforcement learning [C] //Proc of the 2nd Workshop on Data Management for End-To-End Machine Learning Co-located with SIGMOD. New York: ACM, 2018: 4:1-4:4

[96] Meng Xiaofeng, Ma Chaohong, Yang Chen. Survey on machine learning for database systems [J]. Journal of Computer Research and Development, 2019, 56(9): 1803-1820 (in Chinese)
(孟小峰, 马超红, 杨晨. 机器学习化数据库系统研究综述 [J]. 计算机研究与发展, 2019, 56(9): 1803-1820)

[97] Rueden V L, Mayer S, Beckh K, et al. Informed machine learning-towards a taxonomy of explicit integration of knowledge into machine learning [J]. arXiv preprint, arXiv: 1903.12394, 2019

[98] Sun Xiangyan, Xiao Yanghua, Wang Haixun, et al. On conceptual labeling of a bag of words [C] //Proc of the 24th Int Joint Conf on Artificial Intelligence (IJCAI 2015). Menlo Park, CA: AAAI, 2015: 1326-1332

[99] Xu Bo, Xie Chenhao, Zhang Yi, et al. Learning defining features for categories [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence (IJCAI 2016). Menlo Park, CA: AAAI, 2016: 3924-3930

[100] Zhang Yi, Xiao Yanghua, Hwang S, et al. Entity suggestion with conceptual expanation [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence (IJCAI 2017). Menlo Park, CA: AAAI, 2017: 4244-4250

[101] Xie Chenhao, Chen Lihan, Liang Jiaqing, et al. Automatic navbox generation by interpretable clustering over linked entities [C] //Proc of the 2017 ACM on Conf on Information and Knowledge Management (CIKM 2017). New York: ACM, 2017: 1857-1865

Chen Kerui, born in 1983. PhD, lecturer. Her main research interests include Web data management, machine learning and privacy protection.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Fellow of CCF. His main research interests include cloud data management, Web data management, native XML databases, and flash-based databases, privacy-preserving, and etc.