文章编号：2095-2163(2020)05-0006-04 中图分类号：TP181 文献标志码：A

机器学习的可解释性综述

程国建，刘连宏
（西安石油大学 研究生院，西安 710065）

摘 要：机器学习系统日益普及，应用也在不断扩大，加速了向算法化的智能社会转变，这意味着基于算法的决策可能会产生重大的社会影响。新的法规以及高度规范的领域，已经强制要求对决策进行审计和验证，这增加了对机器学习结果的质疑、理解和可解释性的系统能力的需求。对于机器学习系统来说，算法的透明性是不可或缺的。本文对可解释性的研究现状进行综述，了解可解释性的发展以及面临的动机和挑战，包括可解释性及其要求和高风险决策。介绍了机器学习中可解释性的重要性，现有的可解释性评估方法，包括可解释性的评估指标和模型，及可解释性的发展趋势和展望。

关键词：人工智能；机器学习；可解释性；透明性

An overview of the interpretability of machine learning

CHENG Guojian, LIU Lianhong
(School of Postgraduate, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 The growing popularity of machine learning systems and their applications are accelerating the shift to algorithmic smart societies, meaning that algorithm-based decisions can have a significant social impact. New regulations and highly regulated areas have mandated the auditing and verifiability of decisions, increasing the need for systematic capabilities to question, understand, and interpret the results of machine learning. For machine learning systems, interpretability is essential. The purpose of this paper is to review the status quo of interpretability research, to understand the development of interpretability and the motivation and challenges it faces, including interpretability and its requirements and high-risk decision-making. Then, it introduces the existing methods of interpretability evaluation, including the evaluation indicators and models of interpretability and the importance of interpretability in machine learning, and finally obtains the development trend and prospect of interpretability.

【Key words】 artificial intelligence; machine learning; interpretability; transparency

0 引 言

自 20 世纪 70 年代以来，对智能系统的解释一直存在着零星的关注。首先是对专家系统的关注，接着是十年之后的神经网络，再到 2000 年之后的推荐系统[1~2]。但是，近年来关于机器学习的可解释性的研究尚未引起足够的重视，越来越多的关注放到了用机器学习实现预测能力和模型的研究，而解释决策过程的能力则退居其次。ML(机器学习)系统在许多领域都取得了瞩目的成就，使用越来越复杂和不透明的算法(如深度学习)，需要对上述系统的输出进行分析和解释。

人工智能可解释性也受到公众的关注，2016 年，白宫科技政策办公室公布的美国人工智能的报告，题目为“准备人工智能的未来”。这篇报告表示，人工智能系统应确保是透明的、开放的、可以理解的，这样就能清楚的知道人工智能系统背后的假设和决策模型。在 2017 年，美国计算机协会公共政策委员会(USACM)发布了一份“算法透明度声明和问责”，也要求算法透明，以达到可解释的目的。2018 年荷兰 AI 宣言的草案中，专注于可辩解的人工智能，要求人工智能的准确性与可解释性并重。同年 7 月，欧盟委员会发布了一份关于负责任的人工智能和国家人工智能的战略报告中，将不透明(黑盒风险)和可解释性风险列为人工智能的两个性能风险[3]。2019 年 4 月，欧盟委员会的人工智能高级别专家组(AI HLEG)发布了“值得信赖的人工智能”[4]，可解释性原则被列为人工智能系统中的伦理原则之一，透明性被作为可信任 AI 的七个关键要求之一[5]，并指出“一个系统要成为可信赖的，必须能够理解为什么它会以某种方式运行，为什么它会提供给定的解释”，强调了对可解释人工智能领域研究的重要性。

1 动机和挑战

1.1 可解释性

可解释没有确切的定义，Miller 给出了一个解

作者简介：程国建(1964-)，男，博士，教授，主要研究方向：计算智能、机器学习、模式识别等；刘连宏(1995-)，女，硕士研究生，主要研究方向：计算智能与机器学习。

收稿日期：2020-01-06

释是："可解释性是指一个人能够理解一个决定的原因的程度"[6]。也就是说，如果一个人能够容易的推理和追溯为什么模型能做出预测，那么模型的可解释性就较好。相比之下，第一个模型的决策比第二个模型的决策更容易让人们理解和接受，那么第一个模型就比第二个更具解释性。

1.2　可解释性的要求

如果一个机器学习模型表现的足够好，并且具有可接受的预测性能，为什么在信任它的同时还得明白它为什么做出某个决定？Doshi-Velez 和 Kim 认为，只采用单一的度量标准，如分类精度，是对大多数真实世界任务的不完整描述。值得注意的是并不是每个 ML 系统都需要解释能力，在某些具体的应用中，只需要提高预测性能，不需要解释决策。Doshi-Velez 和 Kim 认为有两种情况是不需要解释决策的(1)在没有重大影响或没有因结果不正确而导致严重后果的；(2)在实际的运用中该问题已经有足够多的研究和验证，可以完全信任系统的决策。前者指的是低风险的系统，如推荐系统、广告系统等，即使决策错误也不会造成严重甚至致命的后果[7]。后者指的是已经经过研究并且使用了一段时间的系统，如邮政编码的分拣和飞机碰撞系统[8]等。

1.3　高风险决策

在医疗保健和金融服务以及其他受到严格监管的领域，利用 ML 系统进行高风险决策的趋势越来越明显，这些决策对人类生活和社会都产生了深远的影响，进一步推动了对可解释的需求。ML 已经支持高风险决策的事实并不意味着它不容易出错，预测模型缺乏透明性和可靠性在不同领域中已经出现了严重的后果，如：人们被错误地拒绝假释[9]，错误的保释决定导致潜在的危险罪犯得到释放。决策系统不是百分百的完美，但并不是意味着 ML 模型原本就是坏的。模型取决于训练它的信息，将真实世界的数据提供给模型时，该模型将学习这些模式并返回针对于这些行为的预测。出现预测不准确的原因有两个：(1)在训练时如果提供的数据本身就存在不正确的因素；(2)将训练好的模型用于一个相似的领域的预测，也可能导致预测失败。有证据证明，不正确的建模假设至少对最近的抵押贷款危机负有部分责任[10]。

2　可解释性评估方法

2.1　评估指标

评价和比较机器学习的可解释性有两种指标：定性指标和定量指标。

定性解释能力指标：Doshi-Velez 和 Kim 提出了与解释相关的 5 个因素：(1)认知块形式，即这些解释的组成，如特征值、训练集的例子或是规则列表。在特定领域中，该形式也会有所不同，如在图像识别中认知块形式就是像素组。(2)解释所包含的认知块数量。考虑到一个示例包含的信息可能比一个特性多得多，是否可以在易于理解的程度上处理相同数量的信息，如果解释是由特性组成的，那么它是包含所有特性还是只包含少数特性(选择性)。(3)语义合成性，与认知块的组织结构有关。规则、层次结构和其他抽象可能会影响人类的处理能力。例如：一个解释可以定义一个新的单元(认知块)，它是原始单元的一个功能，并根据这个新单元提供一个解释。(4)单元之间的独立性和相互性，单元之间可以有线性、非线性或是独立的特性。(5)不确定性和随机性是指返回某种可理解的不确定性的解释。

定量可解释性指标：量化解释，为不同的解释提供一种直观的方法。Sundararajan 等人创建了一种基于公理的方法，把深度神经网络预测归因于原始输入特征[10]，发展了一种特定于模型的神经网络解释方法，称为综合梯度。Wilson 等人在他们的著作中对定量指标也进行了阐释。

一个好的解释应该做到：完整性，即解释的涵盖范围，包括解释所包含的实例数目。正确性，解释的结果是与事实相符的。简洁性，解释应该简洁，可通过决策规则中的条件数量和基于邻域的解释的特征维度来验证[12]。

2.2　可解释的模型

可解释的模型的算法包括线性回归、逻辑回归和决策树，它们的参数具有意义，可以从中提取有用的信息来解释预测结果。基于这些可解释模型所构成的神经网络也具有可解释性，神经网络的处理过程不再是暗箱操作，而是透明可解释的，具有明确的语义信息。Zhang 等人提出了基于可解释图的一种通过决策树来定量解释卷积神经网络的预测逻辑。该方法可在 CNN(卷积神经网络)的高层卷积层中学习物体部位的显示表示，同时在全连接层中挖掘潜在的决策模式。决策树通过一定的规则来对潜在的决策模式进行重组，从而达到定量解释 CNN 的预测逻辑。有文献提出了一种可解释的卷积神经网络，高对流层中的每个过滤器代表一个特定的对象部分。可解释 CNNs 使用与普通 CNNs 相同的训练数据，不需要任何对象部件或纹理的注释来进行监

督。在学习过程中,可解释的 CNN 自动地在一个高对流层中给每个过滤器分配一个对象部分。可解释的 CNN 中的显式知识表示可以帮助理解 CNN 内部的逻辑,从而达到可解释的目的。

2.3 机器学习中可解释的重要性

可解释性是满足人类好奇心和求知欲的一种手段,人们不需要对每件事都做解释,但是对于意外事件,就迫切的想要知道发生的原因。不透明的机器学习模型应用于科学研究时,如果该模型是只给出预测却不给出解释的一个黑盒子,不懂其中的原理就无法进行科学创新。基于 ML 模型的决策对人们的生活影响越来越大,这意味着解释机器的行为愈发重要。2016 年的著名论文《Why Should I Trust You?》的发表掀起了可解释性学习的热潮[13],随后 MIT 在 SIGKDD2016 上介绍了 LIME(局部可解释模型-不可知论解释)的概念[14]。目的是解释黑盒子分类器的预测,这意味着对于任何给定的预测和任何给定的分类器,它都能够确定原始数据中驱动预测结果的一小部分特征。

3 发展趋势与展望

3.1 发展趋势

近 5 年(2015~2019 年)发表在知网(CNKI)上的有关于可解释性的论文的数量也越来越多,统计了题目包含有:“interpretability”的相关论文,统计结果如表 1 所示。

表 1 2015~2019 年知网上关于题目包含可解释性的论文数量统计

Tab. 1 Statistics of the number of papers on topics including interpretability on cnki from 2015 to 2019

年份	中文文献	外文文献	总计
2015	0	-	0
2016	3	-	3
2017	4	1	5
2018	3	15	18
2019	13	42	55
总计	23	58	81

从统计结果来看,近 5 年来,共有 81 篇有关于可解释性的论文发表在知网上,统计调查后发现:

(1)总体来讲,可解释性已经得到了研究学者的关注。相比于国内学者,国外学者的关注度可能更高一些,也说明可解释性是当前的一个研究热点。关于可解释性的论文在 2015~2017 年之间较少,在 2018 年也只有 18 篇,其中中文的只有 3 篇,但在 2019 年却增长到了 55 篇,中文达到了 13 篇,外文达到了 42 篇。

(2)关于可解释性的研究呈现出快速增长的趋势,可预见,在之后的几年中,有关于可解释性的算法和论文数量会越来越多。

(3)在上述统计的论文中既包含了理论又包含了应用。体现了可解释性理论价值和应用价值。从侧面说明了可解释性的重要性。

3.2 展望

可解释性是一个很主观的概念,所以难以形式化[15],这是该问题尚未解决的一个重要原因。可解释性是一个特定领域的概念[16],无法给出一个通用定义,表明当涉及到 ML 可解释性时,需要考虑每个特定问题的应用程序域和用例。如今也有很多关注可解释性方面的工作,但只是关注现有的解释而不是创造新的解释。所以,应开发出一个模型无关的解释框架,这个框架能够在考虑问题域、用例和用户类型的同时,在可用的框架中推荐最佳的解释。

参考文献

[1] SWARTOUT W R. XPLAIN: A system for creating and explaining expert consulting programs[J]. Artificial intelligence, 1983, 21(3): 285-325.

[2] VANMELLE W, SHORTLIFFE E H, BUCHANAN B G. EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems[J]. Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, 1984: 302-313.

[3] SWARTOUT W, MOORE J D. Explanation in expert systems: A survey[J]. University of southern California, 1988.

[4] 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.

[5] CRAMER H, EVERS V, RAMLAL S, et al. The effects of transparency on trust in and acceptance of a content-based art recommender[J]. User Modeling and User-adapted interaction, 2008, 18(5): 455.

[6] HERLOCKER J L, KONSTAN J A, RIEDL J. Explaining collaborative filtering recommendations[C]//Proceedings of the 2000 ACM conference on Computer supported cooperative work. 2000: 241-250.

[7] Cédric Villani. AI for Humanity—French National Strategy for Artificial intelligence. 2018. Available online: https://www.aiforhumanity.fr/en/ (accessed on 22 January 2019).

[8] Rao, A. S. Responsible AI & National AI Strategies. 2018. Available online: https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/4% 20International% 20initiatives%20v3_0.pdf (accessed on 22 January 2019).

[9] High-Level Expert Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy Artificial Intelligence. 2019. Available online: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines (accessed on 3 May 2019).

[10] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning [C]//Advances in neural information processing systems. 2016: 3315-3323.

(下转第 13 页)