解释纠偏框架:一种基于标准解释的归因分数生成方法

邢钟毓[1),2)] 梁嘉旋[1),2)] 余国先[1),2)] 王 峻[2)]
郭茂祖[3)] 崔立真[1),2)]

[1)](山东大学软件学院 济南 250101)
[2)](山东大学人工智能国际联合研究院 济南 250101)
[3)](北京建筑大学智能科学与技术学院 北京 102616)

摘 要 模型可解释性研究面临一个关键挑战:对于同一数据集,不同模型尽管能达到相似的预测性能,但受训练过程中随机因素等变量影响,其输入特征的重要性评分(归因分数解释)存在显著不一致,这降低了解释的可信度。针对此问题,本文首先从理论上探讨了解释不一致与模型不确定性因素之间的联系,证明了归因解释中的SHAP(SHapley Additive exPlanation)方法在相似预测模型中的不确定性上界。在此基础上,我们通过实验深入研究了模型集合中模型训练随机因素等变量对特征归因方法的影响,发现模型不确定导致的解释不确定性普遍存在,而SHAP方法由于其上界的影响不确定性较低。据此,我们提出了一种基于不同模型的标准解释生成稳定归因分数解释的纠偏框架ASGM(Attribution Score Generation Method),以减少归因分数解释的不一致,提升模型解释的稳定性和可信度。该框架通过检测少量抽样模型解释与大量模型生成标准解释之间的差异,利用校正偏差的深度学习模型,生成代表规格不足集或罗生门效应集的归因分数解释,并能预测规格不足集解释间的不确定性。实验结果表明,ASGM可以生成受模型(尤其是随机因素)影响较小的解释,生成解释的质量高于对模型集合解释排名的均值,接近标准解释。此外,与标准解释相比,ASGM在罗生门效应集上的计算时间减少了20%~30%,在规格不足集上减少了17%~48%,这些结果验证了ASGM可有效提升解释稳定性和可信度。

关键词 模型不确定性;可解释人工智能;规格不足集;罗生门效应集;SHAP方法
中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2025.00949

Explanation Rectification Framework: An Attribution Score Generation Method Based on Standard Explanations

XING Zhong-Yu[1),2)] LIANG Jia-Xuan[1),2)] YU Guo-Xian[1),2)] WANG Jun[2)]
GUO Mao-Zu[3)] CUI Li-Zhen[1),2)]

[1)](School of Software, Shandong University, Jinan 250101)
[2)](Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101)
[3)](School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing 102616)

Abstract Model interpretability has become increasingly critical in artificial intelligence research, particularly for high-stakes applications where transparency and trustworthiness are paramount. A

收稿日期:2024-07-18;在线发布日期:2025-02-20。本课题得到国家重点研发计划项目(2023YFF0725500)、国家自然科学基金重点项目(62031003)、国家自然科学基金面上项目(62072380)资助。**邢钟毓**,硕士研究生,主要研究领域为可解释人工智能和生物信息学。E-mail:zyxing@mail.sdu.edu.cn。 **梁嘉旋**,硕士,主要研究领域为因果学习和生物信息学。**余国先**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员。主要研究方向为数据挖掘和生物信息。E-mail:gxyu@sdu.edu.cn。 **王 峻**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员。主要研究方向为机器学习,生物信息学。**郭茂祖**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员。主要研究方向为生物信息学,机器学习和数据挖掘。**崔立真**,博士,教授,博士生导师,中国计算机学会(CCF)理事。主要研究方向为软件与数据工程、数据库与知识工程、人工智能。

fundamental challenge in this field emerges from an intriguing phenomenon: models that achieve comparable predictive performance often yield inconsistent feature importance scores (attribution scores interpretations) for identical data. This interpretability dilemma, manifest in both the Rashomon effect set (diverse models with different architectures achieving similar performance) and its subset, the Underspecification set (identical architectures varying due to training randomness), significantly diminishes the credibility of models′ explanations. To address this challenge, this paper firstly explores theoretically the relationship between explanation inconsistency and model indeterminacy factors, and then proves that due to the local accuracy property of SHAP (SHapley Additive exPlanation), there exists an upper bound on the uncertainty of SHAP methods for models with similar predictions. On this basis, we thoroughly investigate experimentally the specific impacts of variables (i. e. , model training stochastic factors) on various feature attribution methods. It finds that explanation uncertainty arising from model indeterminacy is widespread, whereas SHAP methods exhibit lower uncertainty due to the impact of its upper bound. Based on these findings, we propose an explanation rectification framework called ASGM (Attribution Score Generation Method) to generate stable attribution score explanations using standard explanations obtained from diverse models, aiming to reduce the inconsistency of attribution score explanations and enhance the stability and credibility of model interpretations. ASGM identifies disparities between explanations from several sampled models and standard explanations generated from massive models on specified data. Through bias rectification deep network, ASGM efficiently generates attribution score explanations by calibrating the explanations of sampled models, thereby approximating the explanations representative of either the Underspecification set or the Rashomon effect set with few model sampling. ASGM can also predict the uncertainty between explanations of Underspecification set. Experimental results show that ASGM can generate explanations that are less impacted by the model, particularly by stochastic factors. Notably, the explanation evaluation frameworks differ between the two sets under investigation: For the Rashomon effect set, which encompasses diverse architectural designs and optimization strategies, we assess the explanation fidelity with respect to the underlying data distribution. In contrast, for the Underspecification set, where variations arise solely from training stochasticity within identical architectures, we evaluate the explanation fidelity relative to the model ensemble and effectively mitigate the influence of stochastic training processes. Our quantitative analysis reveals that ASGM consistently generates explanations with markedly superior fidelity across both sets compared to conventional methods, such as rank averaging of model set explanations, closely approximating the standard explanation. Furthermore, on the Underspecification set, ASGM demonstrates multiple advantages: it exhibits enhanced stability against data perturbations, maintaining consistent explanations when input data undergoes minor modifications; it demonstrates superior performance in predicting explanation uncertainty, with improvements in coefficient of determination (R-squared) varying from 4. 2% to as high as 40. 1% on different datasets when compared to traditional approaches. In addition, compared to the standard explanations, ASGM significantly reduces the computation time by 20% to 30% on the Rashomon effect set and by 17% to 48% on the Underspecification set. These results confirm that ASGM effectively enhances the stability and credibility of interpretations.

Keywords model indeterminacy; XAI; underspecification set; rashomon effect set; SHAP methods

1 引　言

模型的可解释性是当前人工智能研究的热点[1,2]。提升可解释性，即模型决策过程的透明度和可理解性，对提高模型可信度和用户信任度及降低风险等都具有重要意义。在高风险和监管环境，模型的可解释性（尤其是复杂模型）在公民的生命、财产和权利等领域的决策中极为重要。特征归因是可解释人工智能（Explainable Artificial Intelligence，XAI）中最重要的研究方向之一[3,4]，这类方法通过为输入特征分配不同的重要性分数（即归因分数）来解释输入特征对模型的重要性和影响程度，以帮助用户了解模型决策中的关键特征，从而提高模型的可解释性。与其它可解释性学习分支（如反事实解释[5]、规则提取方法[6]和自解释分类器[7]）相比，特征归因方法因其能够明确量化每个特征对单个预测结果的贡献，在医学影像诊断和金融风险评估等需要精确了解特定特征影响的应用场景中具有独特的重要性。因此，本文聚焦于可解释人工智能中的特征归因方法。

现有特征归因方法的研究主要集中在对单个模型的解释上，它们忽略了大量精度相近的模型对相同数据不同解释的具体含义，事实上，依赖单一模型进行解释意味着所有的解释都基于该模型的特定参数和结构。当存在模型不确定性，即同一任务下存在多个精度相近的模型时，这些模型可能会对同一输入产生不同的解释。这种解释的不确定性会削弱用户对解释的信任。例如，D′Amour等人[8]指出，不同的训练过程可能导致模型在部署时表现差异很大，原因是某些模型训练过程中可能进行了捷径学习（shortcut learning），即模型依赖伪特征在独立同分布的测试中表现良好，但在实际应用中不可靠或存在偏见。Brunet等人[9]发现即使一组模型在整体性能上非常相似，但在特定输入下的SHAP值解释可能存在显著差异，甚至矛盾。Shaikhina等人[10]通过实例表明，相似精度的模型集合中，单一模型的归因分数解释可能存在显著差异，各模型依赖于先验关键特征的不同部分，难以与领域先验知识对齐，需对模型集合的归因分数解释进行综合，从而实现与先验知识的一致，提高解释与模型的可信度。总体而言，模型的不确定性导致依赖单一模型的归因分数解释的可信度降低[9-11]，但针对解释不确定性的研究仍不充分。

罗生门效应集（Rashomon effect set）是指在相同的任务和数据集上，使用不同的模型架构、超参数和优化器并提供近似最佳精度的预测模型集合[9]。该集合保留了对应数据集的一系列有限真实模型，不会考虑假设空间中的所有模型。在罗生门效应集中，具有同样模型架构，训练时受不同的参数初始化、数据划分和Dropout（随机失活，防止过拟合）等随机因素影响且性能水平相当的模型集合被称为规格不足集（Underspecification set）[8-9]。罗生门效应集中的模型精度相似，但其内部对单个数据点的预测和特征重要性可能会有所不同，这反映了模型的不确定性，从而导致解释的不确定性，本文称这种现象为模型导致的解释不确定性。当用户发现模型间预测结果相近，无论是否调整训练过程中的超参数，模型间解释仍存在显著差异时，用户会对模型的归因分数解释的可靠性产生怀疑，从而降低模型的可信度。

模型不确定性对归因分数解释的影响主要体现在两方面[9]。一方面，当模型本身的固定变量（如模型架构、优化器等）改变时，例如将线性模型的架构调整为非线性以建模数据中的非线性信息。若是忠于模型的事后解释，当解释用于调试模型或数学说明时，此时归因分数解释的变化是有意义的；而如果是忠于数据的解释，则需要保证归因分数解释的稳定，减少变化。另一方面，重新训练模型会减少冗余特征对特定模型架构解释的影响[12]，即同样的信息在同一模型架构上也可以用另一个特征表示，此时解释是由于模型训练随机因素的影响发生变化，这是数据规模不足所引起的模型参数波动，导致解释难以复现。这对于一些对决策边界敏感的解释方法影响显著，所以归因分数解释的随机变化应稳定在均值附近，以提高模型解释的稳定性和决策可信度。但若为此对规格不足集的大量模型进行解释，则会耗费大量时间。因此，需要一个稳定、快速的解释方法以减少或者预测模型不确定性导致的归因分数解释不确定性，提高用户对模型归因分数解释的信任度。

关于如何评估归因分数解释，以及如何确保归因分数解释忠诚于模型，已有一些方法进行了探讨。部分方法所产生的标准归因分数解释并不忠于模型[13-14,15]；另有一些方法通过限制模型训练数据的方法以减少模型不确定性所带来的影响，但应用范围有限[16]；还有一些工作虽指定了生成稳定归因分数解释的策略，但效率较低[10-11,17]。为解决上述问题，本文提出了纠正解释偏差并生成稳定归因分数

解释的框架(Attribution Score Generation Method,ASGM),以减少模型不确定性对归因分数解释的影响,主要贡献如下:

(1)由于不同特征归因解释方法的计算方式各异,在面对不确定的模型集合时,各方法的解释稳定性也不同。本文从理论上证明当模型输出相近时,模型间SHAP方法不确定性的上界,并研究了模型集合中模型训练随机因素等变量对不同解释方法,尤其是SHAP方法的具体影响,研究结果表明,模型集合中的不同变量会影响归因分数解释结果的分歧程度,解释不确定性在多种解释方法中普遍存在,但当模型集合的预测结果较为一致时,由于SHAP方法的不确定性上界,SHAP方法的不确定性相对较低。

(2)提出了一种纠偏框架ASGM,该框架通过学习标准解释来减少模型不确定性对解释的影响,从而提高解释的可信度。该框架使用模型解释及其对应数据进行训练,输出的解释代表具有不同的训练随机因素但具有相同或不同模型架构和相似精度的模型集合。该框架还可用于预测模型导致的解释不确定性,并与具体的模型无关,适用于大多数受到不确定性影响的模型。

(3)本文在不同领域的数据集上使用纠偏框架进行了充分实验,并与解释的平均归因分数[10]和平均排名[11]进行比较,结果表明,即使仅使用少量模型平均解释,在面临随机因素和模型结构等变量影响时,本文的纠偏框架也能取得良好的性能。

本文第2节介绍模型不确定性的相关工作以及可解释方法。第3节介绍所需的背景知识,如针对单个模型的特征归因可解释技术和对归因分数解释的评估方法。第4节介绍SHAP方法的不确定性评估和生成稳定归因分数的解释纠偏框架ASGM。第5节通过实验证明所提理论与框架的有效性。第6节总结全文。

2 相关工作

除特征归因,还有多种解释模型决策的方法。反事实示例[5]与特征归因解释类似,都基于特征与模型输出关系的深入分析,但归因分数解释侧重于定量评估特征的贡献,而反事实解释侧重于特征变化如何影响模型输出,提供具体的修改建议,帮助用户理解决策边界,如在贷款审批中建议提升信用评分或增加收入以通过审批。规则提取方法[6]致力于从复杂模型中提取易懂的规则,简化模型决策过程,如将深度学习模型近似为决策树。自解释分类器[7]通过设计内置解释机制,使模型决策过程透明,如使用注意力机制或生成可理解的中间表示。这些方法各有其优点,但特征归因方法在提供细粒度解释方面表现出更大的优势,特别适用于需要明确特征贡献的应用场景。因此,本文将重点关注特征归因分数方法。

特征归因方法基于解释的对象可分为三类:只对数据进行解释,针对单个模型,针对性能相近的模型集合进行解释。在只针对数据的解释方法中,由于人工数据集的数据分布条件概率已知,有无需使用随机抽样等方法逼近去除的特征,因此可以准确地得到人工数据集上的真实解释,该真实解释可以作为现实数据集上解释的评估指标。目前已有一些工作采用人工合成数据集的方法评估解释的准确性。XAI-bench[13]通过调整多元高斯分布的参数来生成人工数据集以模拟自然数据集,以人工数据集的真实解释作为自然数据集的近似解释,但该方法的计算时间随着数据维度的升高呈指数性增长,只能用于特征维度小于15的数据。而且该方法无法保证合成数据一定代表真实数据以及解释忠实于模型。因为即使是达到了精度要求的模型也可能不依赖于数据真实解释认为的重要特征[18],而是依赖于其他特征,此时数据的真实解释反而无法准确地解释模型。OpenXAI[16]试图利用数据特征的相关性使得模型的解释接近数据的真实解释,但该方法要求训练数据能够分离出相互独立、无二义性的簇,并且对模型有所限制,只能应用于人工数据集。实际上,归因分数的不一致在面向长序列以及图像数据集的深度学习模型上更为明显,因为这类数据中冗余特征更多,产生的归因分数波动更大。图像数据集的真实解释计算通常基于遮盖图像中的目标物体[14-15],然而组成目标物体像素之间的归因分数应当不同,而不是统一遮盖。上述方法所得解释与模型无关,不受模型不确定性的影响,能够反映数据在现实中真正重要的特征,但并不忠实于模型,无法提高模型的可解释性。

与只针对数据的特征归因可解释方法不同,第二类归因方法着重于对单个模型的解释。比如基于梯度的显著图方法[19]、输入乘梯度[20]、逆卷积[21]、基于博弈论的SHAP方法[22]以及基于扰动的LIME解释[23],这些解释方法对应于一个固定的模型。为评估模型不确定性对这些解释方法的影响,Krishna等人[24]提出了sign agreement(SA)方法,Brunet等人[9]则提出了signed set disagreement(SSD)和

contradicting direction of contribution（CDC），这些解释不确定性的评估指标可以接受两个TOP-K解释集并输出与解释稳定性相关的数值，其中TOP-K解释集代表解释中前K个重要的特征。SA方法计算TOP-K解释集中共同且同号的比例，SSD代表在随机抽取多个解释子集时，两个TOP-K解释集中特征不同或出现矛盾的概率。CDC方法更加偏重是否存在互相矛盾的解释，即若一个特征在两个TOP-K解释集中共同出现但异号则判定解释矛盾。Brunet等人[9]通过上述评估方法发现，模型的不确定性会导致对模型解释不一致甚至相互矛盾，从而降低解释的可信度。

第三类归因方法探讨如何使归因分数代表整体模型集合，如选择性集合方法[17]探讨了模型不确定性问题，并关注模型预测中的不一致性。基于样本方差小于总体方差的事实，该方法通过计算大量模型对每个预测类别的归因分数平均值以降低解释的不确定性，但其计算耗时较长。Shaikhina等人[10]同样利用模型集合解释的平均值汇总集合的归因分数，并用方差衡量归因分数的不确定性，研究结果证明，平均解释代表了对模型集合解释的期望值。Schulz等人[11]通过平均不同模型归因分数的排名得到共识的解释，从而在模型不确定情况下提供一个稳定且尽量忠诚于所有模型的归因分数，但该方法忽略了解释中极端值对整体解释的影响。本文所提出的方法在纠偏框架训练过程中需要部分训练数据的平均解释，测试时仅对输入解释进行纠偏，具有更优的效率和灵活性。

本文在理论上分析了模型集合中使用SHAP方法[22]的理由，并通过实验进行了验证，实验过程中综合了对模型集合解释的不同评估方法以测量模型不确定性对解释方法的影响。本文还提出了使解释更加具有代表性的通用框架ASGM，不同于XAI-bench[13]等只针对数据集的真实解释，本文区分讨论了不同模型集合，并认为规格不足集上的平均解释可看作在该模型架构上对数据的忠实解释，在对模型的忠诚度方面也优于图像掩码[14-15]等忽略模型架构的真实解释。不同于Open-XAI[16]通过寻找最优模型以获取正确的解释，ASGM通过对精度相近的模型集合的平均解释作为参考进行纠偏，从而使解释与模型一致。此外，ASGM无需像Open-XAI对数据进行限制，应用范围更广。相比大量模型解释的均值[10]和对模型解释排名的平均[11]，ASGM通过学习已有标准解释，实现了更高的效率。与针对单个模型的解释相比，ASGM纠偏后的解释受模型不确定性的影响更小。多个数据集上的实验证明ASGM不仅使解释代表单个模型，还代表受随机因素影响的模型集合，适当训练后也可代表罗生门效应集，详细原理及流程将在4.3节描述。

3　背景知识

本节介绍特征归因的可解释方法及模型解释评估方法的相关概念。

3.1　解释方法

本小节主要介绍相关的特征归因可解释方法。包括基于梯度的显著图方法、输入乘梯度、逆卷积、LIME、特征消融、特征遮挡以及与SHAP相关的一系列方法，选择这些方法的原因在于它们在当前研究中被广泛采用，涵盖了基于模型的不同解释机制。其中，显著图方法、输入乘梯度以及逆卷积代表基于梯度的可解释方法，LIME代表基于代理模型的可解释方法，特征消融（Feature Ablation）、特征遮挡（Feature Occlusion）以及与SHAP相关的一系列方法则代表基于特征扰动或特征消除（removal-based）的可解释方法，这些代表性的特征归因方法也在不同的模型和任务中表现出良好的解释能力。

显著图方法[19]是一种解释深度学习模型的可视化方法，有助于理解决策过程中输入各部分的作用。本文主要介绍基于梯度的显著图方法（Gradient-based Saliency Map Method），该方法在计算机视觉中通过计算输入图像与输出结果的梯度确定每个像素对输出的贡献，揭示模型的关注点并提高可解释性。输入乘梯度[20]将显著图与输入数据相乘得到归因分数。逆卷积[21]对卷积层进行逆运算获得输入特征的归因分数。LIME[23]对每个预测进行扰动，在每个待解释的数据点附近生成一个特征可解释的线性局部代理模型。

基于特征消除或扰动的可解释方法通过系统性地移除或遮蔽输入特征，观察模型输出的变化，从而评估各特征的重要性。常见的方法包括特征消融、特征遮挡和SHAP方法。特征消融[25]是一种基于扰动的归因方法，通过将每个输入特征替换为预设的基线，并计算输出的差异来评估该特征的重要性。此方法不仅可以独立消融单个特征，还可以通过传递特征掩码将多个特征组合一起消融，例如在图像处理中替换整个区域或片段，从而衡量整个特征组合的影响。特征消融可以针对每个特征组合返

回一个标量。特征遮挡[26]是另一种基于扰动的归因方法，它通过替换输入中的连续矩形区域（通常称为超矩形）为基线，计算输出的差异以评估特征的重要性。该方法采用滑动窗口的策略，从输入的起始点开始，逐步替换窗口内的区域。遮挡法特别适用于具有空间连续性的输入数据（如图像），能够有效捕捉到区域级别的重要性。

SHAP方法[22]从博弈论的角度出发，计算单个个体对整体的贡献程度，即SHAP值。SHAP值可以理解为模型函数被固定的Shapley值。SHAP的核心思想是计算特征对模型输出的边际贡献，计算公式如下：

$$\phi_i(f,\boldsymbol{x})=\frac{1}{|M|}\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}\left[f_{\boldsymbol{x}}(\boldsymbol{x}_{S\cup\{i\}})-f_{\mathrm{x}}(\boldsymbol{x}_S)\right] \tag{1}$$

其中，i为待求取的特征，S为保留的特征，$\boldsymbol{x}$表示输入数据集中的任意一条输入数据，$\boldsymbol{x}_S$为保留S后对其他特征进行扰动的输入数据，M为总的特征集合，f为模型函数，$\phi_i(f,\boldsymbol{x})$为$\boldsymbol{x}_i$对特征i的归因分数，$f_{\boldsymbol{x}}(\boldsymbol{x}_S)=E\left[f(\boldsymbol{x})|\boldsymbol{x}_S\right]$，$E\left[f(\boldsymbol{x})|\boldsymbol{x}_S\right]$是输入特征的子集$S$的条件期望值。通过计算去除不同特征情况下$i$的边际贡献，进行加权平均，获得$i$特征对整体输出的贡献。SHAP值计算有不同的加速方法，包括基于深度学习模型的Deep SHAP，与模型无关的Kernel SHAP，基于树模型的Tree SHAP等。SHAP值具有以下几个特性：(1)局部精确要求，每个样本的简化解释模型的输出应等于原始模型（被解释模型）的输出。对于SHAP方法，这意味着解释输出的归因分数值之和等于原始模型输出减去一个基线值，基线值是模型输出分布的期望，可以被视为一个常数。因此，当样本间的SHAP值相同时，原始模型的输出必然相同；(2)缺失性，实例中不存在的特性其归因值为0；(3)连续性，当模型的变化导致简化输入的贡献增大，其归因值也应增加。本文选择使用SHAP方法进行实验，而不采用特征消融（Feature Ablation）和特征遮挡（Feature Occlusion），主要原因在于后者在评估特征重要性时存在计算成本高、难以捕捉特征间复杂交互和归因结果不稳定等不足。SHAP能够提供公平且一致的特征归因，适用于多种机器学习模型，并能更好地应对模型不确定性。

3.2 对模型解释的评估方法

相关工作中已经介绍了对模型集合解释的评估方法，而对单个模型的归因分数解释可以从忠诚性（fidelity）、针对数据扰动的稳定性（stability）、完整性检查（sanity check）、准确度（accuracy）和对比性（contrastivity）方面进行评估。

Dasgupta[27]等人认为对模型忠诚的解释应当具有以下属性，即相同解释的输入有相同的预测：

$$O^b(\text{sample})=\Pr_{\text{sample}'\in_u B_\pi}\left(Mo(\text{sample}')=Mo(\text{sample})\right) \tag{2}$$

其中，sample代表某个输入样本，Mo为模型函数，Pr为概率函数，B_π为sample的解释，u为sample的分布，$\text{sample}'\in_u\mathrm{B}_\pi$的含义为从分布$u$中抽取sample′，其解释为$\mathrm{B}_\pi$。该公式通过将解释相同的输入视为一个分布，计算在该分布内模型输出一致的概率，以衡量解释对模型的忠诚性。SHAP值由于局部精确性，当解释相同时其预测必定相同。但在现实情况下，解释完全相同的要求过于严格，因为特征太多导致解释完全相同非常困难。本文探讨解释接近与预测接近的相关性，并通过理论与实验证明：相比其他方法，SHAP方法更能反映解释与预测的关联，即使模型存在一定不确定性，预测的一致性仍可约束SHAP方法。

基于修改（feature removal）的方法也可以评估解释的忠诚度[28]，这类方法将归因分数解释降序排列，逐步遮挡解释认为重要的特征，观察模型准确率的下降程度，下降速度越快，解释质量越高。然而该方法存在一个缺陷，即在删除特征后样本分布发生了变化，这违背了训练数据与评估数据来自相同分布的基本假设。因此，在未进行重新训练的情况下，模型性能的下降无法明确归因于掩码分布变化还是被删除特征所导致的重要信息丢失[12]。如Remove and Retrain（ROAR）[12]，在数据遮挡特征后重新训练模型获得下降的准确率；Remove and Debias（ROAD）[29]通过对遮挡特征的掩码使用噪声线性插值的方式减少掩码分布对模型预测的影响，提高解释的评估效率。

针对数据扰动的解释稳定性评估方法认为，解释面对微小的数据扰动不应发生巨大变化。Yeh等人[30]通过对输入数据进行大量微小扰动，计算解释的最大偏离程度，即最大灵敏度，作为单个模型解释针对数据扰动稳定性的评估标准。本文中的稳定性指模型不确定性对模型集合解释的影响，针对数据扰动的稳定性则以完整的名称表示。

本文主要目的是解释多个模型，而不限于单个模

型，当对大量模型集合的参数和数据标签进行随机化，并观察代表模型集合的解释的变化时，可能在模型间获得截然不同的评估结果，这表明基于扰动模型参数和数据标签的完整性检查难以代表整个模型集合的行为和特性，因此对模型集合进行完整性检查的意义有限[31]。准确度[32]可以评估解释方法在识别真正重要特征方面的能力。它通过解释特征与先验重要特征的一致性来衡量。然而，非合成数据集上先验的重要特征普遍未知。对比性[33]考察解释方法能否在不同类别之间提供有区分力的解释，即不同预测结果的解释应具有明显差异。本文所研究的特征归因方法只能提供各特征在单一预测中的贡献，而对比性方法主要用于评估反事实解释，直接比较不同预测结果的归因分数无法捕捉到部分复杂的对比关系。当解释目标是模型集合而非单个模型时，这些模型对同一输入可能给出不同的解释，使得不同预测间解释结果的比较更复杂，降低了对比性评估方法的有效性。通过上述评估方法可以确定单个解释在特定模型上的质量，同时，这些评估方法也可迁移到模型集合中，一个受模型不确定性影响较小的解释应当对罗生门效应集或规格不足集中的所有模型保持较高的忠诚性，并具备针对数据扰动的稳定性。

4　减少解释不确定性的方法

4.1　近似精度模型集合中SHAP方法的不确定性评估

基于特征归因的解释方法可以为输入特征值赋予归因分数，但受模型不确定性的影响，不同的解释方法会产生不同程度的波动，如图1所示。即使是预测完全相同并且输出值之差在一定范围内的两个分类器，它们之间的梯度也可能完全不同。

SHAP值具有局部准确性，它可以将每个属性的归因值差缩小到一定范围内。Brunet等人[9]已简要总结了简单的线性回归模型下模型不确定性与SHAP值不确定性的关系，本文延伸到深度学习模型上，证明如下。

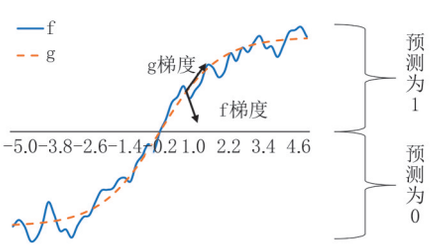

图1　两个预测完全相同的分类函数(f,g)以及它们对应的梯度解释

以不同分类器对原始数据的输出值之差以及对基线的输出值之差符合高斯分布$\mathcal{N}(0,\sigma^2_{\Delta OUT})$为前提，$f$,$g$是两个分类器的函数，它们的基线值分别是$f_0$,$g_0$，令$\boldsymbol{X}$为所有输入数据向量组成的矩阵，即输入数据集，$\boldsymbol{x}$为定义在输入数据集$\boldsymbol{X}$上的随机变量，表示某条输入数据。$i$,$j$为输入数据$\boldsymbol{x}$中的任意两个特征，其中$\boldsymbol{x}_i$表示向量$\boldsymbol{x}$的第$i$个特征，$\boldsymbol{X}_i$表示所有输入数据中第$i$个特征，且$\boldsymbol{x}_i\in\boldsymbol{X}_i$。基线的含义为模型输出分布的期望值[22]，即$E\left[f(\boldsymbol{x})\right]$以及$E\left[g(\boldsymbol{x})\right]$，而在模型线性近似的情况下，等价于$f(E[\boldsymbol{x}])$和$g(E[\boldsymbol{x}])$。

参考公式(1)，分类器f对特征i在$\boldsymbol{X}$上的SHAP归因值分布为

$$\phi_{i,f}=\frac{1}{|M|}\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}\left[f_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)-f_x\left(\boldsymbol{x}_S\right)\right] \tag{3}$$

其中，M为总的特征集合，大小为$|M|$，S为保留的特征，大小为$|S|$，$\boldsymbol{x}_S$为输入数据在特征集合S上的分布。$f_x(\boldsymbol{x}_S)=E\left[f(\boldsymbol{x})|\boldsymbol{x}_S\right]$，$E\left[f(\boldsymbol{x})|\boldsymbol{x}_S\right]$是输入特征的子集$S$的条件期望值。分类器$g$对特征$i$的SHAP归因值是

$$\phi_{i,g}=\frac{1}{|M|}\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}\left[g_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)-g_x\left(\boldsymbol{x}_S\right)\right] \tag{4}$$

两者之差为

$$\phi_{i,f}-\phi_{i,g}=\frac{1}{|M|}\left(\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}\left[f_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)-f_x\left(\boldsymbol{x}_S\right)-g_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)+g_x\left(\boldsymbol{x}_S\right)\right]\right) \tag{5}$$

由于本文不限制为线性模型，假设分类器输出时的非线性激活函数φ为RELU，输入特征之间相互独立，其中Var为方差函数，Cov为协方差函数，函数$w_2=\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x})\right)\right)\right)$，$k_1$,$k_2$为线性变换，$k_1(\boldsymbol{x}_S)$代表$\boldsymbol{x}_S$未扰动而$\boldsymbol{x}_i$扰动的情况。可观察到由于引入$\boldsymbol{x}_i$这一独立随机变量，虽降低了相关性，但协方差不变，即$COV\left(\varphi\left(k_1(\boldsymbol{x}_i,\boldsymbol{x}_S)\right),\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)=COV\left(\varphi\left(k_1(\boldsymbol{x}_S)\right),\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)$。

同时，RELU将负值设置为0，缩小了取值范围，只有当$k_1(\boldsymbol{x}_i)>0$，$k_1(\boldsymbol{x}_i,\boldsymbol{x}_S)>0$时，两者才存在依赖关系，所以RELU只会减小协方差。因此，

$$
\begin{aligned}
&COV\left(\varphi\left(k_1(\boldsymbol{x}_i,\boldsymbol{x}_S)\right),\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\\
&=COV\left(\varphi\left(k_1(\boldsymbol{x}_S)\right),\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\\
&=Var\left(\left(\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\right)\\
&<COV\left(k_1(\boldsymbol{x}_S),k_1(\boldsymbol{x}_S)\right)\\
&=Var\left(k_1(\boldsymbol{x}_S)\right)。
\end{aligned}
$$

同样地，由于RELU函数的截断作用，每层的RELU都会减小协方差，

$$
\begin{aligned}
&COV\left(\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_i,\boldsymbol{x}_S)\right)\right)\right),\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\right)\right)\\
&=COV\left(\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\right),\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\right)\right)\\
&=Var\left(w_2(\boldsymbol{x}_S)\right)\\
&<Var\left(k_2(k_1(\boldsymbol{x}_S))\right)。
\end{aligned}
$$

因此，可推广到多层的情况，在由多层RELU和线性变换构成的深度网络w_m中[34]，协方差最大值也不会超过$Var\left(w_m(\boldsymbol{x}_S)\right)$和$Var\left(k(\boldsymbol{x}_S)\right)$，其中$k$指网络中所有线性变换。并且因RELU为非负和非递减函数，$\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_i,\boldsymbol{x}_S)\right)\right)\right)$与$\varphi\left(k_2\left(\varphi\left(k_1(\boldsymbol{x}_S)\right)\right)\right)$均依赖$\boldsymbol{x}_S$，为正相关关系，即使是推广到多层RELU和线性变换，$Cov\left(w_m(\boldsymbol{x}_i,\boldsymbol{x}_S),w_m(\boldsymbol{x}_S)\right)$也同样依赖$\boldsymbol{x}_S$，不可能为负数。

(1)不同掩码数据集中，分类器输出具有不同的分布，设非线性网络中移除RELU后的方差最大值为t^2，由上文的RELU可知其协方差在0到$Var\left(f(S)\right)$之间，$\phi_{i,f}-\phi_{i,g}$的分布由方差公式可得

$$
\begin{aligned}
&Var\left(\phi_{i,f}-\phi_{i,g}\right)=\frac{1}{|M|^2}Var\sum_{S\subseteq M\setminus\{i\}}\left(\binom{|M|-1}{|S|}^{-1}\cdot\right.\\
&\left[f_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)-f_x(\boldsymbol{x}_S)-\right.\\
&\left.\left.g_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)+g_x(\boldsymbol{x}_S)\right]\right)\\
&=\frac{1}{|M|^2}\left(\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}\cdot\right.\\
&\left(Var\left(f_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)\right)+Var\left(f_x(\boldsymbol{x}_S)\right)\right.\\
&-2Cov\left(f_x\left(\boldsymbol{x}_{S\cup\{i\}}\right),f_x(\boldsymbol{x}_S)\right)\\
&+Var\left(g_x\left(\boldsymbol{x}_{S\cup\{i\}}\right)\right)+Var\left(g_x(\boldsymbol{x}_S)\right)\\
&\left.\left.-2Cov\left(g_x\left(\boldsymbol{x}_{S\cup\{i\}}\right),g_x(\boldsymbol{x}_S)\right)\right)\right)\\
&\leqslant\frac{4}{|M|^2}\left(\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-2}t^2\right)
\end{aligned}
\tag{6}
$$

由组合数的性质可得$\binom{|M|-1}{|S|}\geqslant 1$，所以

$$
\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-2}\leqslant\sum_{S\subseteq M\setminus\{i\}}\binom{|M|-1}{|S|}^{-1}=|M| \tag{7}
$$

$$
Var\left(\phi_{i,f}-\phi_{i,g}\right)\leqslant\frac{4t^2}{|M|} \tag{8}
$$

(2)同样地，不同掩码数据集中，分类器输出具有不同的分布，但设非线性网络中移除RELU后的分布方差均近似为t^2，由于SHAP计算需要特征组合对，数量较大，个别分布的方差不同不影响不等式成立。首先计算归因值之差的和的方差，由SHAP的局部准确性可知，SHAP值之和等于原始模型输出与基线输出的差，可得

$$
\sum_{i\subseteq M}\phi_{i,f}+f_0-\sum_{i\subseteq M}\phi_{i,g}-g_0=\mathcal{N}\left(0,\sigma^2_{\Delta OUT}\right) \tag{9}
$$

$$
\sum_{i\subseteq M}\left(\phi_{i,f}-\phi_{i,g}\right)=\mathcal{N}\left(0,\sigma^2_{\Delta OUT}\right)-f_0+g_0 \tag{10}
$$

由此可知归因值之差的和为$\mathcal{N}(0,2\sigma^2_{output})$分布。根据方差公式计算每个归因值之差的分布：

$$
\begin{aligned}
&Var\left(\sum_{i\in M}\left(\phi_{i,f}-\phi_{i,g}\right)\right)=\sum_{i\in M}\left(\right.Var\left(\phi_{i,f}-\phi_{i,g}\right)\\
&\left.+2\sum_{j\in M/\{i\}}Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)\right)
\end{aligned}
\tag{11}
$$

设H为$\{i,j\}$子集，归因值之差可定义为

$$
v(S\cup H)=f_{\mathrm{x}}(\mathbf{x}_{S\cup H})-g_{\mathrm{x}}(\mathbf{x}_{S\cup H}) \tag{12}
$$

根据SHAP计算公式可得，其中j可分为扰动或未扰动两种情况：

$$
\begin{aligned}
&\phi_{i,f}-\phi_{i,g}=\\
&\frac{1}{|M|}\sum_{S\subseteq M\setminus\{i,j\}}\left(\binom{|M|-1}{|S|+1}^{-1}\left(v(S,i,j)-v(S,j)\right)\right.\\
&\left.+\binom{|M|-1}{|S|}^{-1}\left(v(S,i)-v(S)\right)\right)
\end{aligned}
\tag{13}
$$

当不受RELU影响时，模型为线性，可得归因值之差的协方差矩阵

$$
\begin{aligned}
&Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)\\
&=\frac{1}{|M|^2}\left(\sum_{S\subseteq M\setminus\{i,j\}}\left(\binom{|M|-1}{|S|+1}^{-2}Var\left(v(S,i,j)\right)\right.\right.\\
&-\binom{|M|-1}{|S|+1}^{-2}\left(Var\left(v(S,j)\right)+Var\left(v(S,i)\right)\right)\\
&\left.\left.+\binom{|M|-1}{|S|+1}^{-2}Var\left(v(S)\right)\right)\right)
\end{aligned}
\tag{14}
$$

当函数中存在多层RELU，为非线性时，需要考虑S以及i,j输入到RELU时为负数的情况，$COV\left(v(S,i),v(S)\right)$最小值为0，最大值为$Var\left(v(S)\right)$。其他同理，因此协方差矩阵的范围如下：

$$\begin{aligned}&\frac{2}{|M|^2}\sum_{S\subseteq M\setminus\{i,j\}}-2t^2\left(\binom{|M|-1}{|S|+1}^{-2}\right)\\&<Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)\\&<\frac{2}{|M|^2}\sum_{S\subseteq M\setminus\{i,j\}}2t^2\left(\binom{|M|-1}{|S|+1}^{-2}\right)\end{aligned}\tag{15}$$

由组合数的性质可得$\binom{|M|-1}{|S|+1}\geqslant 1$，$\sum_{S\subseteq M\setminus\{i,j\}}\left(\binom{|M|-1}{|S|+1}^{-1}\right)=\frac{|M|}{2}$，即$\sum_{S\subseteq M\setminus\{i,j\}}\left(\binom{|M|-1}{|S|+1}^{-2}\right)\leqslant\frac{|M|}{2}$，所以

$$\frac{-2t^2}{|M|}<Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)<\frac{2t^2}{|M|}$$

$$\frac{-2(|M|-1)t^2}{|M|}<\sum_{j\in M/\{i\}}Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)<\frac{2(|M|-1)t^2}{|M|}$$

$$-2(|\mathcal{M}|-1)t^2<\sum_{i\in M}\sum_{j\in M/\{i\}}Cov\left(\phi_{i,f}-\phi_{i,g},\phi_{j,f}-\phi_{j,g}\right)<2(|M|-1)t^2\tag{16}$$

结合归因值之差的和的分布与公式(11)可得归因值之差的波动范围：

$$Var\left(\sum_{i\subseteq M}\left(\phi_{i,f}-\phi_{i,g}\right)\right)-4(|M|-1)t^2<\sum_{i\in M}Var\left(\phi_{i,f}-\phi_{i,g}\right)<Var\left(\sum_{i\in M}\left(\phi_{i,f}-\phi_{i,g}\right)\right)+4(|M|-1)t^2\tag{17}$$

因方差大于0，可得单个归因值之差的极值：

$$\begin{aligned}&2\sigma^2_{\Delta OUT}-4(|M|-1)t^2<Var\left(\phi_{i,f}-\phi_{i,g}\right)\\&<2\sigma^2_{\Delta OUT}+4(|M|-1)t^2\end{aligned}\tag{18}$$

所以不同分类器在单个归因值上差的最大方差为$2\sigma^2_{\Delta OUT}+4(|M|-1)t^2$。

与其他解释方法（如输入乘梯度、LIME等）相比，SHAP归因值由于局部准确性的特性，当模型间输出相近时，其不确定性上界会限制归因值的波动范围。因此，在对规格不足集和罗生门效应集进行解释时，使用SHAP值更为有效。为此，本文选择主要对SHAP值进行纠偏，减少了不确定性。但SHAP值也难以从理论上根据单个模型的归因分数准确地推测整体规格不足集归因值的分布，也造成了解释的多重性。针对这一问题，本文通过基于深度学习的纠偏模型来减少归因分数的偏差。

4.2　以深度学习生成稳定归因分数的证明

为减少归因分数的偏差，本文希望根据抽样模型得到罗生门效应集，给定数据集D，设罗生门效应集为C，抽样模型为c，概率为P，由此可得到纠偏模型的目标：

$$\arg\max_{C}\left(P(C|c,D)\right)\tag{19}$$

即通过给定数据集和抽样模型推测出概率最大的罗生门效应集。由于罗生门效应集中的模型间输出基本相同，难以从中提取不同模型的特征，为此本文通过对特定数据的归因分数解释来近似反映模型内部所建模的概念，并将其作为模型的特征。设$d\in D$，$\Phi(c,d)$代表模型c对d的解释，c'是包含在C中的除c以外的其他任意模型，由于当给定罗生门效应集C时，对特定数据解释的方差与均值是确定值：

$$P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|C\right)=1\tag{20}$$

当数据量足够大时，罗生门效应集中几乎不存在解释完全相同的模型：

$$P\left(c|\left(\Phi(c,d),D\right)\right)\approx 1\tag{21}$$

所以纠偏模型可通过以下方式近似：

$$\begin{aligned}&P(C|c,D)\approx P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|C\right)\\&\cdot P(C|c,D)\cdot P\left(c|\left(\Phi(c,d),D\right)\right)\end{aligned}\tag{22}$$

即先通过模型对特定数据的解释$\Phi(c,d)$和数据集D推断抽样模型c，再通过c和D推测罗生门效应集C，最终得到罗生门效应集上对数据d的解释分布，包括方差与均值。由公式(20)、公式(21)可得，右侧表达式概率与通过c和D推测罗生门效应集C的概率相同。

当罗生门效应集中只存在简单的线性模型时，仅需一条数据和对应的SHAP值就可以确定抽样模型，此时抽样模型的SHAP值就是模型的系数乘以输入。

深度学习模型归因非线性函数，则需更多数据和归因分数解释拟合。给定$\Phi(c,d)$和D时，由公式(21)可确定c，所以此时C和c是独立的，可通过以下

方式转换公式(22)近似得到纠偏模型：

$$\begin{aligned}&P(C|c,D)\approx P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|C\right)\\&\cdot P(C|c,D)\cdot P\left(c|\left(\Phi(c',d),D\right)\right)\\&\approx P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|C\right)\\&\cdot P(C|c,D,\Phi(c',d))\cdot P\left(c|\left(\Phi(c',d),D\right)\right)\\&\approx P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|C\right)\\&\cdot P(C|\Phi(c',d),D)\cdot 1\\&\approx P\left(\left(\mathbb{E}_{c'\in C}\Phi(c',d),\sigma^2_{c'\in C}\Phi(c',d)\right)|\Phi(c',d),D\right)\end{aligned}\tag{23}$$

为此本文以训练数据以及抽样模型在训练数据上的解释作为输入特征，以模型集合在训练数据上的平均解释和解释方差作为训练标签，构建了一个纠偏模型ASGM。ASGM的具体流程将在4.3节详述。

4.3 基于标准解释的归因分数生成方法-ASGM

基于4.1节的近似精度模型集合中SHAP方法的不确定性评估和4.2节的以深度学习生成稳定归因分数的证明，本文提出了基于标准解释的归因分数生成方法ASGM(基本工作流程如图2所示)。该框架首先基于给定数据训练模型集合，其中模型集合可为罗生门效应集或规格不足集，模型架构可包括人工神经网络(Multi-Layer Perceptron，MLP)、随机森林(Random Forest，RF)、卷积神经网络(Convolutional Neural Network，CNN)等，为减少模型训练随机因素或模型架构等对解释集合的影响，ASGM将模型集合在抽样训练数据上解释的均值(即标准解释)作为纠偏网络的标准标签，以抽样模型解释及其对应数据作为输入，训练纠偏网络以学习抽样模型与模型集合上解释的关系。因SHAP解释受模型不确定性的影响较小，对模型集合中具体单个模型的解释在后续实现过程中均采用SHAP方法。ASGM也可以使用其他针对单个模型的解释方法。训练完成后，仅需将测试数据以及少量抽样模型对测试数据的解释输入纠偏网络，即可获得近似代表整个集合对测试数据的输出解释，这减少了解释对单个模型的依赖，提高了解释的稳定性。经过适当的修改和扩展，如图7所示，ASGM还可用于预测规格不足集上解释的不确定性。

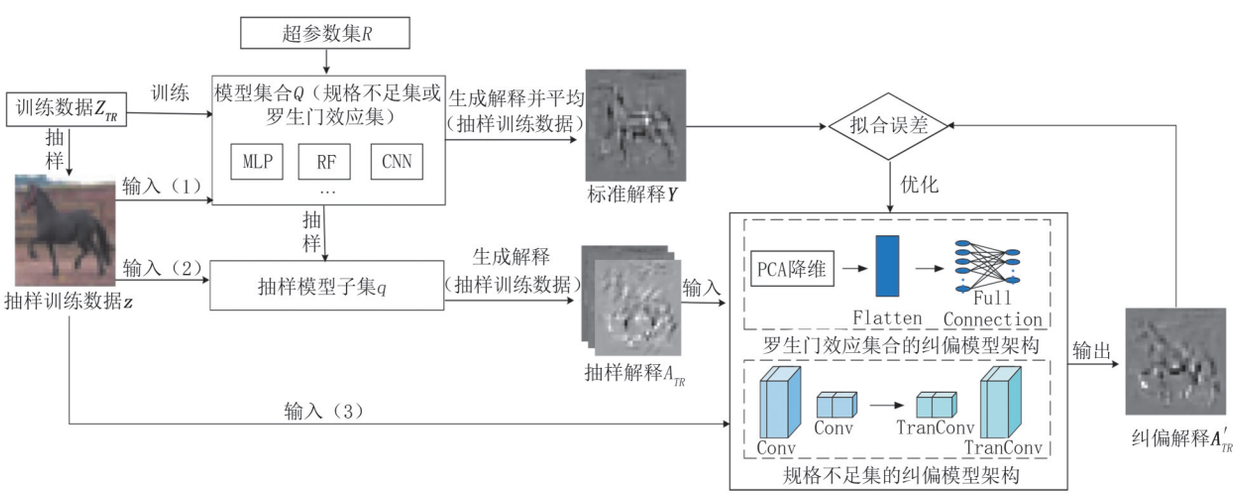

图2 解释纠偏框架ASGM(基于标准解释的归因分数生成方法)的工作流程

根据上述工作流程，ASGM算法实现如算法1所示。算法1描述了相同或不同模型架构的情况下，基于标准解释对解释集合进行纠偏的过程。首先生成超参数R，并基于R和训练数据筛选出符合精度要求的规格不足集或者罗生门效应集，即模型集合Q(第1～2行)。抽样部分训练数据z用于后续的解释，在Q上获得抽样训练数据对应的解释集合的均值Y，其中均值按模型维度平均(第3～4行)。其次，从Q中进行抽样得到模型子集q，并基于q对输入的z进行解释，获得抽样解释集合A_{TR}(第5～6行)。为消除单个随机模型的影响，将Y作为标准解释，并通过解释方差判断解释的不确定性[10]。在此基础上，将A_{TR}、z作为输入，Y作为标签以训练纠偏模型，纠偏模型用于减少模型输出的纠偏解释A'_{TR}与Y的拟合误差，此处选用欧式距离(L2 loss)，规格不足集以及罗生门效应集的纠偏模型架构如图2所示(第7行)。而在测试数据上无需标准解释，首先获得抽样子集q对测试数据的解释A_{TE}(第8行)，再通过纠偏神经网络对A_{TE}进行纠偏，使纠偏解释接近规格不足集或罗生门效应集对测试数据的平均解

释。另外，如果要预测规格不足集上数据点的解释方差以判断解释的不确定性，则需要对模型架构进行改造，具体见图7所示的模型架构（第9行）。不同于规格不足集，罗生门效应集因受模型架构等影响，解释间普遍方差较大，解释基本不一致，无需预测解释间的方差。

算法1. ASGM（基于标准解释的归因分数生成方法）

输入：训练数据Z_{TR}，测试数据Z_{TE}

输出：经过纠偏后的解释A'_{TE}（解释不确定程度见图7）

1. $R=generate()$//超参数集，当包含模型架构或者优化器时为罗生门效应集，仅包含随机因素时为规格不足集

2. $Q=Train(Z_{TR},R)$//根据超参数集训练模型，得到作为规格不足集或者罗生门效应集的模型集合

3. $z\in Z_{TR}$//抽样部分训练数据用于后续的解释

4. $Y=Explain(z,Q)$//获得模型集合对部分训练数据的归因分数解释，其均值为标准解释

5. $q\in Q$//从模型集合中进行抽样

6. $A_{TR}=Explain(z,q)$//根据抽样模型子集得到对部分训练数据的抽样解释集合

7. $A'_{TR}=Model(z,A_{TR}),\mathcal{L}=L2loss(Y,A'_{TR}),\mathcal{L}\xrightarrow{\text{反向传播}}$更新*Model*参数//根据抽样解释、抽样训练数据、标准解释训练纠偏模型

8. $A_{TE}=Explain(Z_{TE},q)$//根据抽样模型获得对测试数据的归因分数解释

9. $A'_{TE},\sigma^2_{A'_{TE}}=Model(Z_{TE},A_{TE})$//使用纠偏模型对测试数据在抽样模型上的解释进行纠偏，得到纠偏解释以及解释不确定程度（解释不确定程度如图7所示，将在5.4.3节详述）

与针对单个模型的解释相比，ASGM纠偏后的解释减少了模型不确定性所造成的误差。与其他针对模型集合的解释方法（如对归因分数的平均[10]或归因分数排名的平均[11]）相比，当这些方法只根据抽样模型的解释或其排名估算模型集合的解释时，则依赖于采样均值是总体均值的无偏估计，这些方法只使用了抽样模型，没有引入模型集合对抽样训练数据的平均解释，因此对模型集合的忠诚度较低。而当这些方法使用模型集合计算时，则会降低计算效率，本文后续会通过实验证明。

ASGM以模型集合对抽样训练数据的归因分数的均值作为标准解释，当模型集合为仅改变训练随机因素的规格不足集时，本文认为平均解释与数据、模型架构以及优化器相关，几乎不受训练过程影响，是模型集合解释的期望[10]。而当集合的架构、优化器和训练过程都变化时，模型集合为罗生门效应集，平均解释主要忠实数据，较少受模型影响。

综上所述，ASGM允许用户基于给定数据训练一个模型集合，该集合可以是罗生门效应集或规格不足集。ASGM采用抽样训练数据上模型集合解释的平均值作为纠偏网络的标准标签，同时将抽样模型的解释以及对应数据作为输入来训练纠偏网络。值得注意的是，纠偏网络训练完成后，只需固定少量抽样模型对测试数据进行解释，随后将这些抽样模型的解释以及测试数据输入纠偏网络，即可获得近似代表整个规格不足集或罗生门效应集对测试数据的输出解释。若训练过程中加入规格不足集上对训练数据归因分数解释的方差，如图7所示，ASGM也可用于预测模型导致的解释不确定性，具体流程将在5.4.3节详述。不同于大量模型解释值的平均[10]和对模型解释排名的平均[11]，ASGM在减少模型不确定性对解释影响的同时显著提高了计算效率。

5　实验分析

5.1　实验背景

为验证本文理论分析和ASGM的有效性，采用不同领域的多个数据集进行了实验。Swiss-Prot是一个高质量、人工注释的大型蛋白质序列数据库[35]。它当前包含超过50万条蛋白质序列，每个序列都经过详细的人工注释，包括序列信息、功能、结构、分类学信息和参考文献等，是研究蛋白质结构和功能的重要资源之一。DeepTFactor[36]是近期提出的基于深度学习判断蛋白质是否为转录因子的工具，它结合了多个卷积子网络，具有模块化设计和良好的正则化方法，通过多层次的特征提取和整合，适用于复杂的分类和特征提取任务。它不依赖同源性，而是使用序列特征。在本文实验中，DeepTFactor原模型固定为预训练模型，其训练数据为2020年4月前的Swiss-Prot数据，共计562 083条。本文从Swiss-Prot提取2020年4月至2023年2月的新数据6661条，随机分割为训练集和测试集，再根据优化器等超参数训练模型集合并存储。通过评估这些模型在旧数据和新数据上的表现，筛选出模型的罗生门效应集，以进行解释纠偏处理，生成稳定的归因分数。

MNIST[37]、Fashion MNIST[38]和CIFAR10[39]是三个常用的基准数据集，分别用于手写数字识别、服装识别和彩色图像分类。MNIST和Fashion MNIST各自包含10个类别，训练集60 000个实例，测试集10 000个实例，图像大小为28×28像素。

CIFAR10包含10个类别，是60 000张32×32像素的RGB彩色图像。

视觉变换器(Visual Transformer，ViT)是一种将Transformer架构应用于计算机视觉任务的模型[40]。它通过将图像划分为固定大小的块(patches)，并将这些图块作为序列输入到Transformer中，从而利用自注意力机制捕捉图像的全局特征。相比传统的CNN，Visual Transformer在图像分类、目标检测等任务中展示了优秀的性能和更强的建模能力。

实验中主要使用Pytorch实现框架和模型集合，仅对树模型如XGBoost使用Sklearn。实验环境：Python 3.10.12，Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz，NVIDIA GeForce RTX 3090，64 GB×8 RAM，Ubuntu 22.04.1 LTS。

5.2 规格不足集中的随机因素对不同特征归因解释方法的影响

本节验证模型训练的随机因素对不同特征归因解释方法的影响。

本文综合了相关工作中的SSD，CDS和SA评价方法，考虑TOP-K解释集中共同的特征以及同号的特征：设两个TOP-K解释集中各自有Num个特征，通过计算共同特征数$Attr_{CO}$，共同且同号的特征数$Attr_{SA}$，得到解释相似性分数(explanation similarity score，ESC)：

$$score=exp\left(\frac{1}{Num}\left(Attr_{SA}-\alpha\cdot\left(Attr_{CO}-Attr_{SA}\right)-\beta\cdot\left(Num-Attr_{CO}\right)-Num\right)\right) \tag{24}$$

其中，α，β分别代表出现矛盾的特征和不同特征的惩罚系数，exp为自然指数。当两个解释完全相同时，解释相似性分数为1。$Attr_{CO}-Attr_{SA}$代表在TOP-K解释集中相互矛盾的特征数，也就是共同特征的归因分数异号的数量，$Num-Attr_{CO}$代表不同的特征数量。解释相似性分数越高，说明在TOP-K解释集中相同的特征越多，该分数综合评估了矛盾特征和不同特征。相比较而言，CDC没有考虑不同的特征，而SA和SSD则简单计算了共同且同号的特征，没有权衡矛盾特征和不同特征的不同影响。

对MNIST和Fashion MNIST数据集，本文在根据3层全连接层(MLP)生成的规格不足集上使用不同解释方法计算3000张图像的归因分数，然后多次随机抽取图像，使用ESC评估同一图像数据的解释不确定性。在Swiss-Prot数据集中，随机选取500条序列，使用不同的解释方法在大小为100的根据DeepTFactor生成的规格不足集进行了解释。结果如表1所示，在SHAP、基于梯度的显著图方法、输入乘梯度、逆卷积方法以及LIME中，SHAP值间的相似性分数更高，说明SHAP值在规格不足集上更加相似。这与本文在4.1节中提出的模型间SHAP方法不确定性的上界推断相符，即当模型之间的预测接近时，SHAP在规格不足集上的波动保持在一个稳定的范围内。相比之下，其他解释方法更容易受到模型不确定性的影响，导致解释结果出现较大差异。尽管SHAP方法在稳定性方面具有显著优势，其解释结果仍存在一定的波动，因此需纠偏措施来提升解释的一致性和可靠性。总体而言，SHAP方法在减轻模型不确定性影响方面表现出更优越的性能，从而显著增强了模型解释的可信度和可靠性。上述实验结果证明了本文学习不同模型的

表1 规格不足集上不同解释方法所得解释的相似性分数(ESC)

数据集	α	β	SHAP	基于梯度的显著图	输入乘梯度	逆卷积	LIME
Swiss-Prot	0.8	0.8	**0.026±8.5e-04**	0.014±6.6e-04	0.010±3.5e-04	0.015±5.5e-04	0.003±6.6e-05
	0.8	0.2	**0.034±2.6e-04**	0.020±5.1e-04	0.014±5.6e-04	0.020±6.1e-04	0.003±1.1e-04
	0.2	0.8	**0.026±1.1e-03**	0.013±5.6e-04	0.010±5.2e-04	0.015±4.7e-04	0.002±7.3e-05
	0.2	0.2	**0.034±3.3e-04**	0.021±8.4e-04	0.014±6.1e-04	0.020±5.8e-04	0.003±1.2e-04
MNIST	0.8	0.8	**0.272±5.2e-05**	0.200±8.6e-05	0.267±6.6e-05	0.216±8.9e-05	0.168±1.1e-05
	0.8	0.2	**0.411±6.0e-05**	0.341±5.3e-05	0.392±1.8e-05	0.359±7.2e-05	0.285±4.8e-05
	0.2	0.8	0.277±9.06e-05	0.200±5.4e-05	**0.280±7.4e-05**	0.216±7.4e-05	0.182±4.0e-05
	0.2	0.2	**0.417±6.1e-05**	0.341±6.5e-05	0.411±2.2e-04	0.359±5.2e-05	0.305±2.2e-05
Fashion MNIST	0.8	0.8	**0.012±1.8e-04**	0.009±8.7e-05	0.009±1.9e-05	0.008±7.1e-05	0.006±5.0e-05
	0.8	0.2	**0.019±2.2e-04**	0.015±1.6e-04	0.016±1.1e-04	0.015±1.3e-04	0.006±6.5e-05
	0.2	0.8	**0.012±1.7e-04**	0.009±1.7e-04	0.009±1.5e-04	0.008±9.2e-05	0.005±4.6e-05
	0.2	0.2	**0.019±1.8e-04**	0.015±1.1e-04	0.016±1.2e-04	0.015±7.0e-05	0.006±6.4e-05

SHAP值的合理性，并为通过不同SHAP值学习标准解释的框架ASGM提供了依据。

5.3　规格不足集中预测对SHAP值不确定性的影响

本节通过实验证明在精度相近和模型架构相同的模型集合上，SHAP值会受到预测相似度的影响。

首先在实验数据集上基于不同的随机因素生成精度类似的树模型（RF），并计算模型之间对10 000张图像的预测相似度。随机选择初始模型Net_A，得到与Net_A预测最接近的模型Net_B以及用于对比的随机选择的模型Net_C，使用三个模型分别对500条数据进行解释，并比较Net_B与Net_A之间的SHAP值是否比Net_C与Net_A之间更加相似。三个模型均抽样了120次，设其中Net_A与Net_B的解释相似度为$\mathrm{sim}(Net_A, Net_B)$，$Net_A$和$Net_C$的相似度为$\mathrm{sim}(Net_A, Net_C)$，$\mathrm{sim}(Net_A, Net_B)$大于$\mathrm{sim}(Net_A, Net_C)$的概率即为解释与预测相似度呈现正相关，受到预测影响的占比。其中MNIST和Fashion MNIST解释使用结构相似度SSIM（structural similarity）评估相似度，SSIM主要用于衡量图像的相似性，取值范围在-1到1之间，完全相同的图像相似度为1。Swiss-Prot使用解释相似性分数（ESC）评估。结果如表2所示，不考虑跨数据集的情况下，在同一数据集的规格不足集内部，解释的确会受到预测的影响，预测接近的模型通常解释更加相似。因此当预测完全相同时，SHAP值的不确定性将进一步减小。预测相似度对SHAP值确定性的正向影响，进一步验证了SHAP值的局部准确性，能够将在规格不足集上的解释波动限制在一个稳定的区间内。这一发现与4.1节的推断相符，体现了相似的预测下SHAP值的不确定性上界可以提升SHAP值的一致性。

表2　不同规格不足集中预测对SHAP值的影响

数据集	模型数量	模型精度（%）	解释受到预测影响的占比
Fashion MNIST	60	79.5±1.8e-01	70%
MNIST	60	90.0±5.0e-01	67%
Swiss-Prot	100	97.0±3.0e-06	70%

5.4　ASGM在罗生门效应集以及规格不足集上的实验

本节将在不同数据集上验证解释集合经过纠偏后更加接近标准解释，根据获得标准解释的模型集合类型为罗生门效应集或规格不足集分为两类实验。如4.1节所述，SHAP方法对模型不确定性的影响较小，因此在后续ASGM针对具体单个模型的解释中，我们均采用SHAP方法。

5.4.1　ASGM在罗生门效应集上的稳定归因分数生成

在罗生门效应集上，进行了ASGM的论证实验。从MNIST数据集抽样1000张图像，Fashion MNIST抽样7000张图像使用SHAP方法进行解释，其中MNIST使用的罗生门效应集模型精度为$(98.5\pm5.0\mathrm{e}-01)\%$，包括CNN，RF等共4种模型架构，每种模型架构各训练了10个模型，根据这些模型获取标准解释，从5.3节MNIST的模型集合中任选3个模型作为抽样模型集合。Fashion MNIST的罗生门效应集模型精度为$(91.0\pm2.0\mathrm{e}-01)\%$，共106个，包括CNN，RF等4种模型架构，从5.3节Fashion MNIST的模型集合中随机选取了5个模型作为抽样模型集合。MNIST和Fashion MNIST的罗生门效应集内部的模型架构包括多种存在显著差异的机器学习模型和深度学习模型。Swiss-Prot的模型集合精度为$(97.0\pm3.0\mathrm{e}-06)\%$，共120个，架构基于DeepTFactor[36]，分为四种不同的优化器，训练了5个抽样模型。对500条序列进行SHAP解释。实验中使用模型集合对训练数据解释的均值作为纠偏网络的标准标签，抽样模型集合的解释及其对应数据作为纠偏网络的输入。训练完成后，将测试数据以及抽样模型在测试数据上的解释输入到如图2所示的纠偏网络，该网络包含PCA降维和多层全连接层，以进行罗生门效应集的纠偏。

对不确定模型集合的解释分布难以具体评估，Shaikhina等人[10]证明了通过大量模型解释的平均可以提高解释稳定性，因此，本文在实验验证中将罗生门效应集在测试数据上的平均解释作为稳定的标准解释，将纠偏前的抽样解释的均值以及纠偏后的解释与标准解释进行对比。评估过程中采用了多种度量方法：使用了L2 loss计算解释间的欧式距离，使用5.2节中的ESC分数评估解释的相似性，对MNIST图像数据集和Fashion MNIST图像数据集，引入了5.3节提到的图像结构相似度（SSIM）衡量图像解释

的结构相似性，具体实验结果如表3所示。ASGM纠偏后的解释与标准解释相比，SSIM相似度和ESC有显著提升，而L2 loss比纠偏前降低了50%以上，所以纠偏后的解释更接近罗生门效应集的标准解释。纠偏模型的具体效果如图3所示，抽样解释是ASGM的输入解释，纠偏后的抽样解释为输出，标准解释则是罗生门效应集中不同解释的平均。抽样模型解释的平均值容易受到抽样模型中特定模型的影响，未能准确反映图像的关键特征。标准解释则参考了不同模型架构的解释，因此测试数据上的标准解释均匀地体现出数据本身的特征，是一种理想状态。而我们的纠偏模型能够有效地调整解释，使纠偏后的解释更接近标准解释，降低单一模型对解释的影响，提高解释的稳定性，经纠偏的解释集合可以视为保留模型特点的同时又参考了其他模型的归因分数解释，更加接近罗生门效应集的平均解释（标准解释）以及数据本身，这也与表3的实验结果相符。

本文还测量了在本地服务器上解释的生成时间。具体而言，测量了生成1000条标准解释和1000条纠偏解释所需的时间，其中纠偏解释里包括了ASGM的训练时间，然后计算纠偏解释对于标准解释的时间占比，实验结果如表4所示。ASGM仅需标准解释70%～80%的时间。从计算复杂度来看，使用大量模型的标准解释通常具有线性时间复杂度，而ASGM通过引入深度学习模型校正偏差，在应用阶段仅需对部分抽样模型进行解释和纠偏处理，其纠偏时间和标准解释的计算时间与抽样模型数量相对于模型集合总量的比例近似，约为6%。即便考虑到训练过程，因通常仅需一次训练，而且使纠偏模型收敛的训练解释数量也少于标准解释，约为70%。所以ASGM整体计算时间在标准解释的70%～80%，因此，ASGM在处理大规模数据集和复杂模型时，展现出更优的计算效率，降低了计算代价。

表3 罗生门效应集上对随机抽样模型集合解释的纠偏

数据集	罗生门效应集内部主要区别	解释状态	SSIM(↑)	L2 loss(↓)	解释相似性分数(ESC)(↑)
Fashion MNIST	模型架构	纠偏前	0.37	209.18	0.54
		纠偏后	0.81	77.50	0.70
MNIST	模型架构	纠偏前	0.44	65.59	0.63
		纠偏后	0.76	25.21	0.78
Swiss-Prot	优化器	纠偏前	—	23.61	0.48
		纠偏后		2.11	0.59

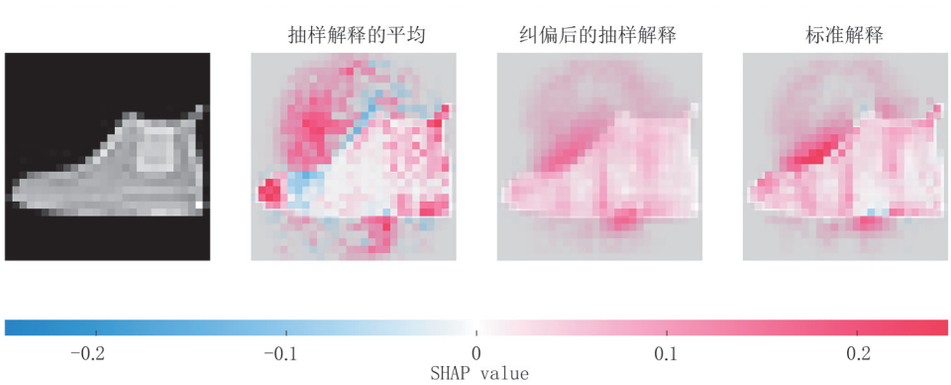

图3 Fashion MNIST数据集上的纠偏效果示例

表4 罗生门效应集上对测试数据解释生成时间的优化效果

数据集	纠偏解释相较于标准解释的所需时间占比
Fashion MNIST	79.2%
MNIST	74.9%
Swiss-Prot	71.3%

5.4.2 平均解释作为规格不足集标准解释的原因

本节通过ROAD方法评估数据的真实解释与规格不足集平均解释对特定模型的忠诚度。图像数据集维度过高，XAI-bench难以对其拟合得到真实解释，本文直接以对目标物体的掩码[14,15]作为数据的真实解释。罗生门效应集的平均解释同样减小了模型架构等因素的影响[10]，本文也通过5.4.1节中针对罗生门效应集的平均解释近似对数据的真实解释。实验中使用3层全连接层模型（MLP），在MNIST和

Fashion MNIST上通过随机重复训练分别得到50个准确率均超过80%的模型集合作为规格不足集，结果如图4所示。横坐标为基于归因分数降序排列的特征遮挡比例，纵坐标为全连接神经网络模型的准确率的均值，标题为数据集-模型架构，其中特征使用噪声线性插值的方式遮挡以减少遮挡分布对模型预测的影响，也就是Remove and Debias(ROAD)[29]方法。ROAD曲线的相对面积定义为曲线下的面积相对于其边界矩形面积的比例，特征遮挡后准确率下降越快，相对面积越小，说明解释对原模型的忠诚度越高。可以发现规格不足集解释的平均值在特定模型上的曲线相对面积更小，所以其针对规格不足集的忠诚度高于图像掩码等真实解释，因此规格不足集解释的平均值实际上代表了一种在规格不足集范围内稳定且具有代表性的解释。规格不足集解释的平均值受模型架构的影响，可以说明在该模型架构下所依赖的特征，所以规格不足集解释的均值能够用于对规格不足集解释的纠偏，而无论是罗生门效应集的平均解释还是数据掩码都几乎与模型无关，没有对给定的模型架构进行解释。

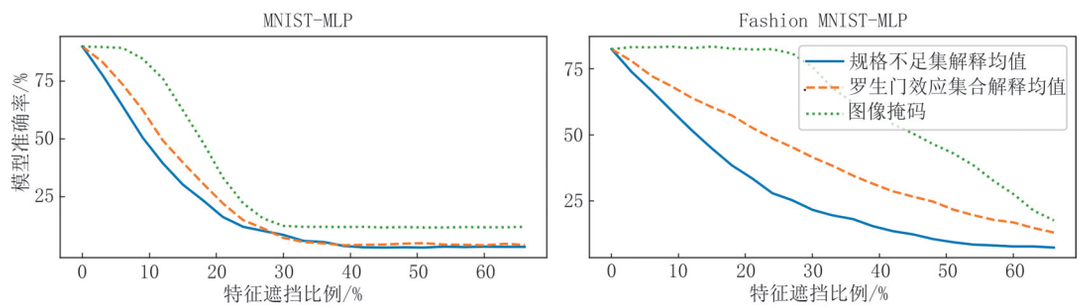

图4　数据真实解释与规格不足集平均解释的ROAD评估

5.4.3　ASGM在规格不足集上的稳定归因分数生成

为验证纠偏框架ASGM能够减少随机因素的影响，我们分别采用不同的深度学习模型，并根据数据集样本和规格不足集生成平均解释作为标准解释[10]。以MNIST、Fashion MNIST和CIFAR10数据集为例，针对MNIST、Fashion MNIST数据集使用了5.4.2中的规格不足集，在CIFAR10数据集上，我们采用了PyTorch官方提供的2层卷积和3层全连接层模型(CNN)以及Visual Transformer(ViT)[40]，这些模型均归类为非线性的深度学习模型。随后，针对每种模型架构，通过随机重复训练的方法，分别生成了50个精度相近且架构一致的模型集合作为规格不足集。然后利用DeepSHAP方法[22]得到不同数据集的各2100张图像的解释，这些解释和图像数据用于训练纠偏框架，训练完成后，对测试集3000张图像进行解释，并计算这些解释与标准解释之间的欧式距离，除欧式距离以外，还使用了ROAD等指标评估解释在模型集合上的表现，从而验证ASGM纠偏效果。本文在线性模型架构和如图2所示的针对规格不足集的卷积-逆卷积结构上进行实验，综合不同数据集，卷积-逆卷积模型最优超参数配置为5层卷积层，2层逆卷积层。同时根据不同的模型架构选择各自纠偏效果最好的抽样模型数量，即每条数据的输入解释数量，卷积-逆卷积结构在输入解释数量为1时效果较好，而线性模型架构的输入解释数量为3。结果如表5所示。

表5　在不同模型架构上训练，纠偏后解释与标准解释的欧式距离(L2 loss)

数据集	模型架构	解释状态	卷积与逆卷积模型	线性模型
CIFAR10	CNN	纠偏前	2.0e-04	1.0e-04
		纠偏后	6.0e-05	6.6e-05
Fashion MNIST	MLP	纠偏前	1.7e-03	1.2e-03
		纠偏后	9.0e-04	1.0e-03
MNIST	MLP	纠偏前	6.0e-04	4.0e-04
		纠偏后	2.0e-04	3.0e-04
CIFAR10	ViT	纠偏前	9.2e-04	3.1e-04
		纠偏后	4.6e-04	2.3e-04

经过以上两种模型架构纠偏后，归因分数解释都在一定程度上接近了平均解释。其中线性纠偏架构调整了不同解释间的权重，而图2中针对规格不足集使用的卷积-逆卷积架构则通过提取与还原图像特征进行纠偏。需指出，与罗生门效应集相比，规格不足集上的L2 loss明显更小，所以规格不足集上纠偏后解释与平均解释的欧式距离变小只能说明解释之间的数值接近，不足以证明模型学习到了标准解释在高维空间的分布和生成了较好的解释。为此，本文在实验中对解释在模型集合上的忠诚度和

灵敏度进一步评估。

为更深入地评估不同模型架构的纠偏效果，本文首先测试了不同解释在模型上的忠诚度，在规格不足集中随机抽取模型作为原模型，解释在原模型上的评估结果如图5所示，与5.4.2节类似，本文计算了ROAD曲线的相对面积。其中origin为通过原模型得到的解释，而average代表对规格不足集解释的平均，即标准解释，input代表规格不足集中除原模型以外随机选取的抽样模型解释的均值，即输入解释的平均[10]，不同模型架构的输入解释数量与表5的实验保持一致。output代表线性模型或卷积-逆卷积模型对输入解释纠偏后的解释，output-dropout代表在卷积-逆卷积模型中训练时启用Dropout的纠偏情况，mean_rank[11]代表对规格不足集内部所有模型的归因分数排名的均值，rand代表使用随机生成的数据作为解释，可以作为解释忠诚度的基线参考。

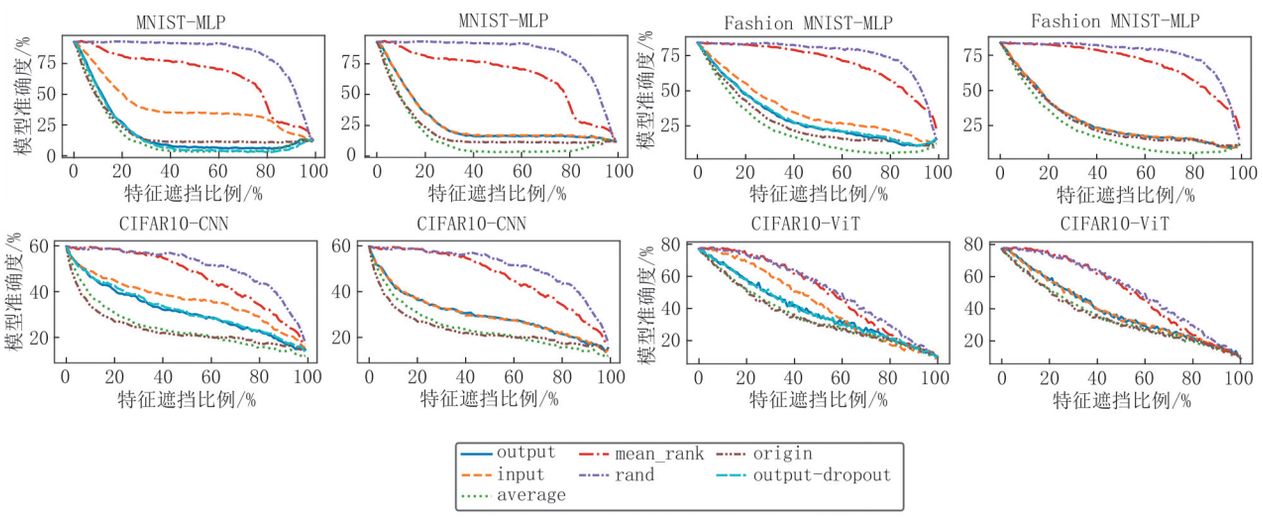

图5 在不同纠偏模型架构下纠偏解释的ROAD评估（第1、3列为卷积-逆卷积模型，第2、4列为线性模型）

从图5可以发现，其他所有解释都要强于随机解释。输入解释的忠诚度普遍低于原模型解释，这是由于所依赖的模型不同导致的。事实上，通过一个模型的归因分数去解释另一个模型会造成偏差，即使它们的模型架构与优化器完全相同。标准解释同时参考了大量模型解释，其忠诚度接近甚至高于依赖原模型的解释。由于线性模型仅调整不同抽样解释的权重，未能学习到标准解释所代表的信息，纠偏后的output解释的ROAD相对面积与输入解释相比未发生变化，没有提高解释的忠诚度。

ASGM中的卷积-逆卷积模型通过输入数据学习到特定模型解释的非线性偏差，对输入解释（input）进行了有效的纠偏。以纠偏效果最差的Fashion MNIST为例，与原本的输入解释相比，纠偏后的output的ROAD相对面积缩小了7%，在纠偏效果最好的MNIST数据集上其ROAD相对面积缩小了21%，这一结果有效证明了解释纠偏后其忠诚度更加接近标准解释。在应用Dropout后，output-dropout忠诚度与output基本相同。mean_rank的忠诚度远低于对归因分数的平均，这是因为mean_rank省略了具体的归因分数，只聚合不同模型间的归因分数排名，对原模型的解释较差。

本文还评估了图2中针对规格不足集的卷积-逆卷积模型生成的纠偏解释与其它解释的最大灵敏度以及所需解释数量的差异。随机解释（rand）与输入数据的扰动无关，无需评估灵敏度。平均排名（mean_rank）在输入数据的微小扰动下也基本不会变化。由之前的实验可知，随机解释和平均排名作为模型集合的解释忠诚度过低，其灵敏度和解释数量可以忽略。结果如表6，7所示。其中的抽样模型数量即为每条数据对应的输入解释数量，这里分别设置为1、3、5，标准解释则是包含50个模型的规格不足集解释的平均，本文通过数据扰动后解释的最大偏离程度量化最大灵敏度[30]，最大灵敏度与针对数据扰动的稳定性负相关。原模型解释与输入解释均为在规格不足集上随机选择的模型解释，多次实验时其最大灵敏度与所需解释数量是一致的。如表6所示，不同解释方法在不同数量抽样模型下的最大灵敏度及所需解释数量存在显著差异。特别是output和output-dropout方法所需解释数量远低于表7的标准解释方法。具体而言，在MNIST数据集上，output方法

在1、3、5个抽样模型下分别仅需3.0×10^3、9.0×10^3及1.5×10^4条解释，远少于标准解释所需的1.5×10^5条。这一差异在Fashion MNIST和CIFAR10数据集中同样存在。

表6　不同解释方法在不同数量抽样模型下最大灵敏度以及所需解释数量

数据集	模型架构	解释分类	不同数量抽样模型的最大灵敏度(↓)			3000条测试数据所需解释数量(↓)		
			1	3	5	1	3	5
MNIST	MLP	origin and input	5.4e-02±2.0e-03	4.4e-02±8.0e-04	4.2e-02±6.0e-04	3.0e+03	9.0e+03	1.5e+04
		output	3.9e-02±1.4e-03	3.4e-02±1.1e-03	3.1e-02±5.0e-04	3.0e+03	9.0e+03	1.5e+04
		output-dropout	3.7e-02±1.2e-03	3.2e-02±9.0e-04	3.0e-02±5.0e-04	3.0e+03	9.0e+03	1.5e+04
Fashion MNIST	MLP	origin and input	4.2e-02±1.9e-03	3.9e-02±1.8e-03	3.8e-02±1.8e-03	3.0e+03	9.0e+03	1.5e+04
		output	3.2e-02±1.6e-03	3.9e-02±1.3e-03	3.1e-02±1.5e-03	3.0e+03	9.0e+03	1.5e+04
		output-dropout	3.1e-02±1.6e-03	2.8e-02±1.0e-03	2.9e-02±1.4e-03	3.0e+03	9.0e+03	1.5e+04
CIFAR10	CNN	origin and input	2.8e-01±2.3e-02	2.1e-01±2.3e-02	1.8e-01±1.8e-02	3.0e+03	9.0e+03	1.5e+04
		output	2.0e-01±2.3e-02	1.6e-01±8.3e-03	1.5e-01±7.6e-03	3.0e+03	9.0e+03	1.5e+04
		output-dropout	1.9e-01±2.2e-02	1.6e-01±7.8e-03	1.5e-01±7.6e-03	3.0e+03	9.0e+03	1.5e+04
CIFAR10	ViT	origin and input	1.6e-01±6.5e-03	1.3e-01±6.4e-03	1.1e-01±6.4e-03	3.0e+03	9.0e+03	1.5e+04
		output	1.4e-01±.1.1e-02	1.2e-01±5.8e-03	1.0e-01±4.5e-03	3.0e+03	9.0e+03	1.5e+04
		output-dropout	1.3e-01±1.0e-02	1.1e-01±5.6e-03	1.0e-01±3.1e-03	3.0e+03	9.0e+03	1.5e+04

表7　标准解释的最大灵敏度以及所需解释数量(模型数量为50)

数据集	模型架构	解释分类	最大灵敏度(↓)	3000条测试数据所需解释数量(↓)
MNIST	MLP	average	3.5e-02±1.0e-03	1.5e+05
Fashion MNIST	MLP	average	3.7e-02±1.7e-03	1.5e+05
CIFAR10	CNN	average	8.2e-02±8.5e-03	1.5e+05
CIFAR10	ViT	average	4.0e-02±1.3e-03	1.5e+05

然而，output及output-dropout方法需要额外地解释训练纠偏模型，训练解释与测试数据所需解释数量之和通常是标准解释总量的50%～80%，由于纠偏模型的训练时间远小于解释的生成时间，因此纠偏解释的总体时间为标准解释的52%～83%左右。本文使用了约98 700条解释训练纠偏模型，所以其总体时间在标准解释与输入解释之间，测试数据被纠偏的解释越多，模型训练所需解释的占比就越小。

由于纠偏模型参考了未扰动前的标准解释，在不同的抽样模型数量下，纠偏解释(output以及output-dropout)的最大灵敏度均低于原模型解释(origin)和输入解释(input)，即纠偏解释针对数据扰动的稳定性更高。由于对解释的纠偏依赖于输入解释(input)，所以在一些简单的数据集上(MNIST和Fashion MNIST)，不同数量抽样模型的原模型解释和输入解释的最大灵敏度稍高于标准解释时，纠偏解释的灵敏度接近或低于所在数据集的标准解释。而在一些复杂的数据集上比如CIFAR10，采用CNN和ViT架构时，原模型解释和输入解释的最大灵敏度显著高于标准解释，导致纠偏解释对于数据扰动的最大灵敏度仍然高于标准解释。启用Dropout后，output-dropout的忠诚度略微下降，但相比output的最大灵敏度下降，减少了数据扰动的干扰，说明针对数据扰动的稳定性略有提升。

综合以上实验结果，图2中针对规格不足集的卷积-逆卷积模型纠偏后的解释在整个规格不足集上针对数据扰动的稳定性优于原模型解释(origin)和输入解释(input)，忠诚度接近标准解释，并且相较于标准解释效率更高，所需解释数量更少。

原模型解释作为依赖原模型的解释，其ROAD忠诚度在原模型上较高，但在规格不足集的其他模型上评估时，由于缺乏模型信息，其忠诚度与输入解释相当。这一点可以通过特征遮挡后重新训练模型的ROAR方法进行验证，为了减少遮挡分布的影响，遮挡方式选择了线性插值，实验结果如图6所示。评估方式使用了与5.4.2节类似的曲线相对面积，但标准为ROAR，在不同抽样模型数量(exp_num)，即每条数据对应的输入解释数量下的实验发现，由于特征遮挡后重新训练模型，可以去除原模型所依赖的

冗余特征影响，因此原模型解释(origin)与输入解释(input)的ROAR忠诚度接近。而模型集合平均解释(average)的忠诚度最高，这正是使用平均解释作为标准解释的原因所在。解释纠偏后，无论是否使用Dropout，纠偏解释(output以及output-dropout)的ROAR曲线相对面积与原本的输入解释相比均呈现出明显的减少趋势，这一趋势在不同的实验条件下表现各异。具体而言，在最差的情况下，即抽样模型数量为5，使用Fashion MNIST数据集时，相对面积收缩了2%；在最好的情况下，即抽样模型数量为1，使用MNIST数据集时，则收缩了31%。这说明纠偏解释的忠诚度均优于原模型解释和输入解释，即解释纠偏提高了解释质量，所生成的归因分数受模型不确定影响较小，稳定性较高。而模型归因分数的排名均值(mean_rank)的ROAR忠诚度仅高于随机解释，这是因为特征归因分数经常出现极端值(如极大值)，而平均排名的过程会减小极端值的影响，导致平均排名解释忠诚度较低。

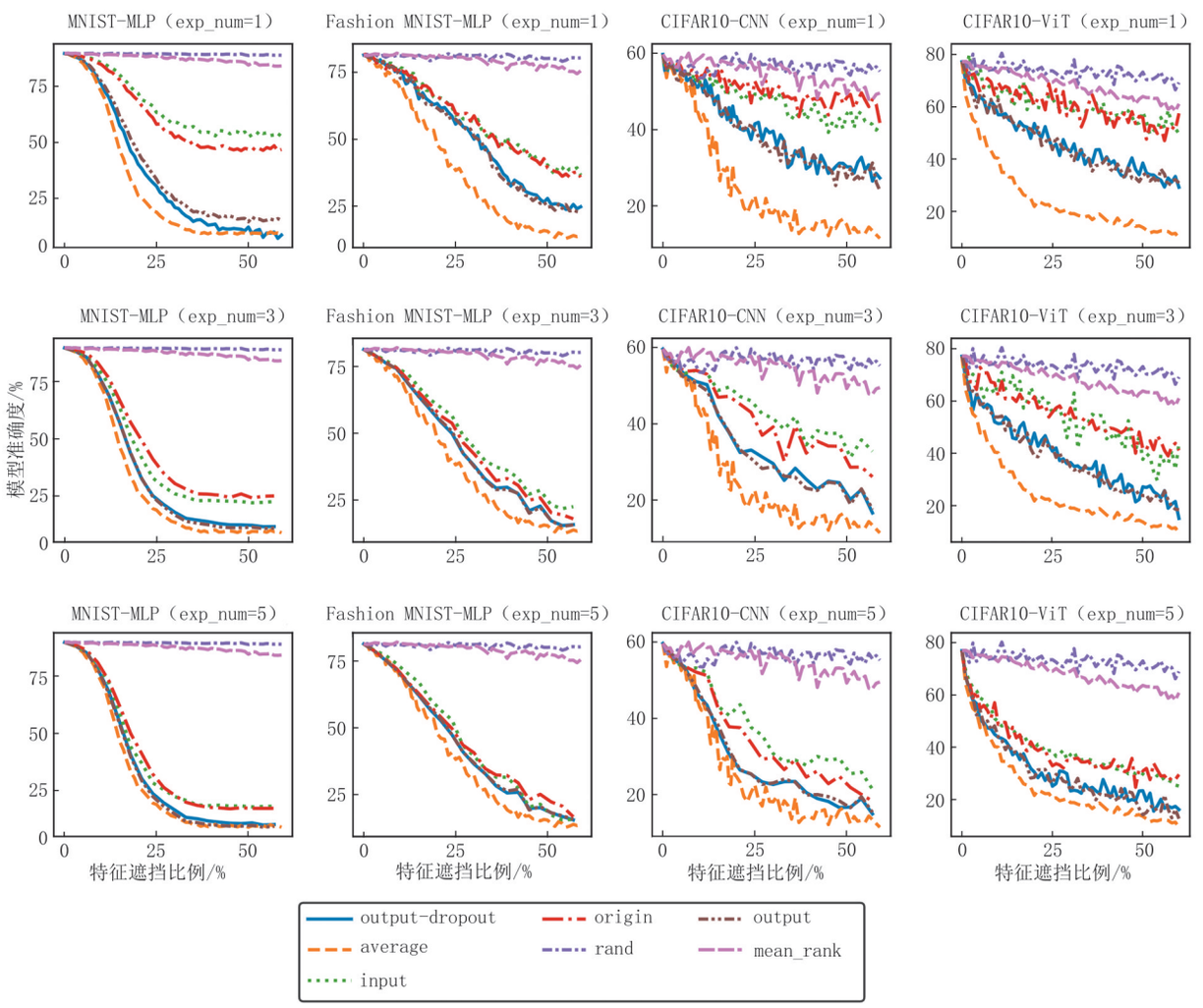

图6 在不同数据集、模型架构和抽样模型数量(exp_num)下，对图2中针对规格不足集的卷积-逆卷积模型解释的ROAR评估

本文的纠偏框架ASGM不仅可以使输入解释接近平均解释，还可以预测归因分数解释间的不确定性。具体地，模型通过预测数据中规格不足集解释的方差[10]以表示规格不足集上解释的不确定性。如图7所示，为预测模型导致的解释不确定性，在图2中针对规格不足集的卷积-逆卷积模型基础上进行了改进，除标准解释外，还通过规格不足集上对训练数据归因分数解释的方差训练纠偏模型。当抽样模型数量，即每条数据的输入解释数量为1时，ASGM可以预测解释不确定程度，但对于抽样模型的解释，即输入解释则无法计算方差。为方便对比，设置抽样模型数量为3，本文在各数据集上使用2100张图像进行训练，900张图像进行测试，使用r2_score(决定系数，一种衡量回归模型拟合优度的指标，数值越高表

示模型性能越好）评估模型的性能。在最优超参数下的实验结果如表8所示。从表8可以看出，因学习了规格不足集对训练数据的解释，与输入解释的方差[10]相比，ASGM的模型输出能够有效预测规格不足集对测试数据解释的方差，从而判断规格不足集对具体数据的解释是否出现了不一致的情况。

D′Amour等人[8]指出规格不足集是在同样的模型架构下对预测问题等价的一类解决方案，在经过纠偏后，输入的解释集合会受到规格不足集的影响，偏向于规格不足集中共同特征，即更加接近于标准解释。经过纠偏后，解释受到训练随机因素的影响更小。综合上述所有实验结果可以看出，本文提出的纠偏框架ASGM使输入解释更加接近标准解释，减少了模型不确定性对归因分数解释的影响。

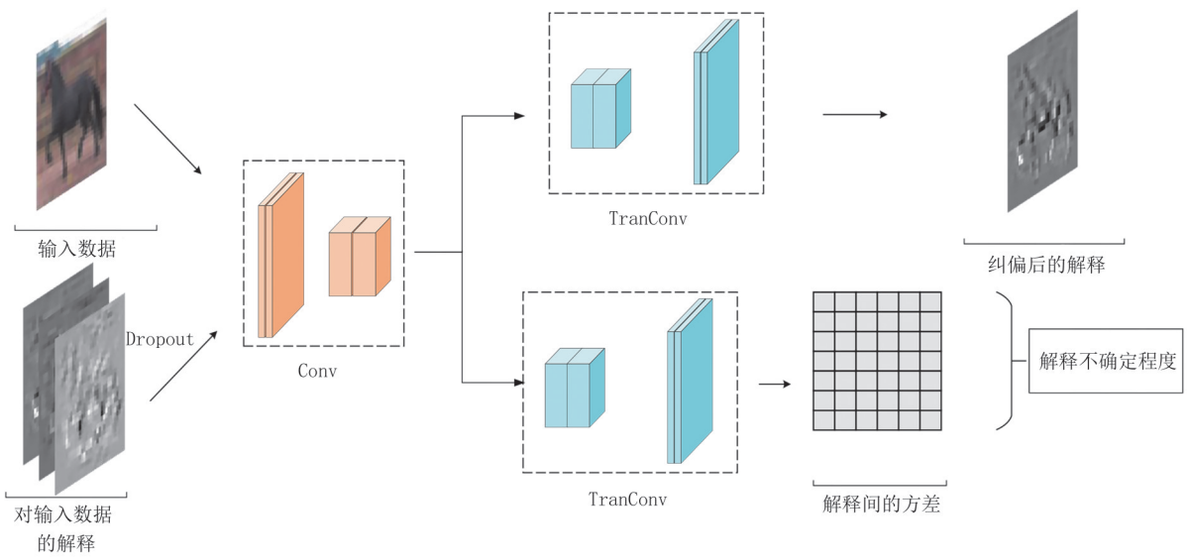

图7　预测解释不确定性的模型架构

表8　输入解释与解释不确定性预测模型的输出对规格不足集上解释间方差的拟合优度

数据集	模型架构	输入解释(%)	解释不确定性预测模型的输出(%)
MNIST	MLP	85.0±2.0e-00	94.4±8.5e-01
Fashion MNIST	MLP	89.6±4.2e-00	93.8±3.4e-00
CIFAR10	CNN	58.4±5.6e-01	82.0±1.2e+01
CIFAR10	ViT	51.0±3.2e-01	90.1±1.3e-00

6　总　结

本文首先分析了解释与导致模型不确定性因素之间的联系，从理论上证明了当模型输出相近时模型间SHAP方法不确定性的上界。还通过实验研究了模型集合中模型训练随机因素等变量对不同的解释方法，尤其是SHAP方法的具体作用。这一研究体现了解释不确定性的普遍性以及SHAP方法不确定性的上界对解释不确定性的影响。实验结果证明由于不确定性上界，SHAP方法受不确定性影响较小，为在ASGM中应用SHAP方法奠定了基础。虽然ASGM可以选择多种解释方法，但基于SHAP方法的优越性，ASGM以SHAP方法为主要工具。当解释与模型预测的关系越密切，对于预测相近的模型，解释就越相似，同时受到模型不确定性的影响越小。本文提出了一个通用的解释纠偏框架ASGM以生成稳定的归因分数，并对规格不足集和罗生门效应集分别讨论。在多类数据集上的实验结果证明，无论是在罗生门效应集还是规格不足集上，ASGM均高效率地降低了解释与标准解释的L2 loss，减少了模型不确定性的影响，输出的解释不再依赖于单个随机的具体模型，而是代表整个模型集合，提高了解释的稳定性和可信度。此外，ASGM还可用于推测规格不足集上解释间方差，从而预测模型导致的解释不确定性。与使用模型解释平均或者其排名平均相比，本文所提出的框架能更快地近似标准解释。进一步的研究工作包括：(1)改进获取标准解释的方式，从而在获得更少模型解释的情况下实现纠偏；(2)在罗生门效应集中，如何进一步减少模型不确定性的影响，提高归因分数解释的稳定性，并保证解释仍忠于抽样模型，需进一步研究。

参 考 文 献

[1] Dwivedi R, Dave D, Naik H, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Computing Surveys, 2023, 55(9): 1-33

[2] Ju T., Liu G., Zhang Z. et al. A review of probe interpretable methods in natural language processing. Chinese Journal of Computers, 2024, 47(4): 733-758 (in Chinese)
(鞠天杰,刘功申,张倬胜,等.自然语言处理中的探针可解释方法综述.计算机学报,2024,47(4):733-758)

[3] Chen H, Covert I C, Lundberg S M, et al. Algorithms to estimate Shapley value feature attributions. Nature Machine Intelligence, 2023, 5(6): 590-601

[4] Srinivas S, Fleuret F. Rethinking the role of gradient-based attribution methods for model interpretability//International Conference on Learning Representations. Virtual, 2021: 1-15

[5] Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery, 2024, 38(5): 2770-2824

[6] Zilke J R, Loza Mencía E, Janssen F. Deepred-rule extraction from deep neural networks//International Conference on Discovery Science. Bari, Italy, 2016: 457-473

[7] Gautam S, Höhne M M C, Hansen S, et al. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognition, 2023, 136: 109172-109185

[8] D'Amour A, Heller K, Moldovan D, et al. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 2022, 23(1): 10237-10297

[9] Brunet M E, Anderson A, Zemel R. Implications of model indeterminacy for explanations of automated decisions// Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 7810-7823

[10] Shaikhina T, Bhatt U, Zhang R, et al. Effects of uncertainty on the quality of feature importance explanations//AAAI Workshop on Explainable Agency in Artificial Intelligence. Virtual, 2021: 1-18

[11] Schulz J, Santos-Rodriguez R, Poyiadzi R. Uncertainty quantification of surrogate explanations: an ordinal consensus approach//Proceedings of the Northern Lights Deep Learning Workshop. Tromso, Norway, 2022:1-8

[12] Hooker S, Erhan D, Kindermans P J, et al. A benchmark for interpretability methods in deep neural networks//Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 9737-9748

[13] Yang L, Sujay K, Colin W, et al. Synthetic benchmarks for scientific research in explainable machine learning//NeurIPS Datasets and Benchmarks Track. Virtual, 2021:1-25

[14] Arras L, Osman A, Samek W. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion, 2022, 81: 14-40

[15] Oramas J, Wang K, Tuytelaars T. Visual explanation by interpretation: improving visual feedback capabilities of deep neural networks//International Conference on Learning Representations. New Orleans, USA, 2019: 1-29

[16] Agarwal C, Krishna S, Saxena E, et al. Openxai: towards a transparent evaluation of model explanations//Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 15784-15799

[17] Black E, Leino K, Fredrikson M. Selective ensembles for consistent predictions//International Conference on Learning Representations. Virtual, 2021:1-25

[18] Faber L, Moghaddam AK, Wattenhofer R. When comparing to ground truth is wrong: on evaluating gnn explanation methods// ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Virtual, 2021: 332-341

[19] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps//International Conference on Learning Representations. Banff National Park, Canada, 2014: 1-8

[20] Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks//International Conference on Learning Representations. Vancouver, Canada, 2018: 1-16

[21] Springenberg J, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net//International Conference on Learning Representations. San Diego, USA, 2015: 1-14

[22] Lundberg S M, Lee S I. A unified approach to interpreting model predictions//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 4768-4777

[23] Ribeiro M T, Singh S, Guestrin C. " Why should I trust you?" explaining the predictions of any classifier//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[24] Krishna S, Han T, Gu A, et al. The disagreement problem in explainable machine learning: a practitioner's perspective. Transactions on Machine Learning Research, 2024, 01: 1-34

[25] Merrick L. Randomized ablation feature importance. arXiv preprint arXiv:1910.00174, 2019

[26] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks//13th European Conference Computer Vision.ta S, Frost N, Moshkovitz M. Framework for evaluating faithfulness of local explanations//International Conference on Machine Learning. Virtual, 2022: 4794-4815

[28] Samek W, Binder A, Montavon G, et al. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(11): 2660-2673

[29] Rong Y, Leemann T, Borisov V, et al. A consistent and efficient evaluation strategy for attribution methods// International Conference on Machine Learning. Baltimore, USA, 2022: 18770-18795

[30] Yeh C K, Hsieh C Y, Suggala A S, et al. On the (in) fidelity and sensitivity of explanations//Advances in Neural Information

Processing Systems. Vancouver, Canada, 2019: 10967-10978
[31] Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps//Neural Information Processing Systems. Montréal, Canada, 2018: 1-11
[32] Guidotti R. Evaluating local explanation methods on ground truth. Artificial Intelligence, 2021, 291: 103428-103444
[33] van der Waa J, Nieuwburg E, Cremers A, et al. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence, 2021, 291: 103404-103423
[34] Montufar G F, Pascanu R, Cho K, et al. On the number of linear regions of deep neural networks//Neural Information Processing Systems. Montreal, Canada, 2014: 27-36
[35] ConsortiumUniProt. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 2019, 47(D1): D506-D515
[36] Kim G B, Gao Y, Palsson B O, et al. DeepTFactor: a deep learning-based tool for the prediction of transcription factors. Proceedings of the National Academy of Sciences, 2021, 118(2): 1-5
[37] Deng L. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142
[38] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017
[39] Krizhevsky A. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, Toronto, 2009
[40] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale//International Conference on Learning Representations. Virtual, 2021: 1-22

XING Zhong-Yu M. S. candidate. His research interests include explainable artificial intelligence and bioinformatics

LIANG Jia-Xuan, Master. His research interests include causal learning and bioinformatics.

YU Guo-Xian, Ph. D., professor, Ph. D. supervisor. His research interests include data mining and bioinformatics.

WANG Jun, Ph. D, professor, Ph. D. supervisor. Her research interests include machine learning and bioinformatics.

GUO Mao-Zu, Ph. D, professor, Ph. D. supervisor. His research interests include bioinformatics, machine learning and data mining.

CUI Li-Zhen, Ph. D, professor, Ph. D. supervisor. His research interests include software and data engineering, database and knowledge engineering, artificial intelligence.

Background

Model interpretability is a prominent topic in current deep learning research, particularly the feature attribution explanation of deep learning models. However, most studies on interpretable methods concentrate on the explanation of a single model and do not consider the specific implications of numerous models with similar accuracy for different explanations of the same data. In fact, research on how models influence attribute feature attribution explanations has demonstrated that the indeterminacy of deep learning models can negatively impact the credibility of feature attribution explanations. For the same tasks and datasets, there often exist deep learning models with similar performance but drastically different explanations.

Reducing explanation uncertainty faces the following challenges. Firstly, selecting appropriate attribute feature attribution explanation methods to minimize explanation uncertainty is essential. Secondly, balancing the impact of data and models on feature attribution explanation is crucial. Lastly, efficiently obtaining faithful and stable explanations using excellent model sets is a significant challenge.

To address these issues, this paper first proves that the fluctuations of SHAP explanations have an upper bound under data constraints based on the characteristic of SHAP explanation. Furthermore, experiments demonstrate that SHAP explanations are less impacted by model indeterminacy compared to other interpretable methods.

In addition, this article argues that the impact of data and models on explanation should be analyzed specifically based on changes in the model and the purpose of explanation. Explanations faithful to the data should ensure the stability of the explanation as much as possible, while post hoc explanations faithful to the model need to consider changes in the model. For the Rashomon effect set with changes in model fixed variables or the Underspecification set with different stochastic factors, this paper proposes a framework called ASGM (Attribution Score Generation Method) for inferring the explanation distribution of the overall model set based on the feature attribution explanations of a small number of

models. This framework conducts case-by-case discussions based on different model sets. When the model set is a Rashomon effect set or an Underspecification set, the explanations generated by the framework vary accordingly, ensuring that the explanation is faithful to the data or model set. Experiments on multiple datasets demonstrate that ASGM is more efficient than existing methods for obtaining feature attribution explanations on model sets. Moreover, the framework can be used to predict explanation consistency and is model-independent, making it applicable to most models subject to indeterminacy.

This work was supported by National Key Research and Development Program of China (2023YFF0725500), and National Natural Science Foundation of China (NSFC) (62031003, 62072380).