基于主动学习和二次有理核的模型无关局部解释方法

周晟昊, 袁伟伟, 关东海

引用本文

周晟昊, 袁伟伟, 关东海. 基于主动学习和二次有理核的模型无关局部解释方法[J]. 计算机科学, 2024, 51(2): 245-251.

ZHOU Shenghao, YUAN Weiwei, GUAN Donghai. Local Interpretable Model-agnostic Explanations Based on Active Learning and Rational Quadratic Kernel [J]. Computer Science, 2024, 51(2): 245-251.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300028

基于主动学习和二次有理核的模型无关局部解释方法

周晟昊　袁伟伟　关东海
南京航空航天大学计算机科学与技术学院　南京 211100
(cenhelm@nuaa.edu.cn)

摘　要　深度学习模型的广泛使用,在更大程度上使人们意识到模型的决策是亟需解决的问题,复杂难以解释的黑盒模型阻碍了算法在实际场景中部署。LIME作为最流行的局部解释方法,生成的扰动数据却具有不稳定性,导致最终的解释产生偏差。针对上述问题,提出了一种基于主动学习和二次有理核的模型无关局部解释方法ActiveLIME,使得局部解释模型更加忠于原始分类器。ActiveLIME生成扰动数据后,通过主动学习的查询策略对扰动数据进行采样,筛选不确定性高的扰动集训练,使用迭代过程中准确度最高的局部模型对感兴趣实例生成解释。并且,针对容易陷入局部过拟合的高维稀疏样本,在模型损失函数中引入了二次有理核来减少过拟合。实验结果表明,所提出的ActiveLIME方法引比传统局部解释方法具有更高的局部保真度和解释质量。

关键词:局部解释;扰动采样;主动学习查询策略;二次有理核
中图分类号　TP391

Local Interpretable Model-agnostic Explanations Based on Active Learning and Rational Quadratic Kernel

ZHOU Shenghao, YUAN Weiwei and GUAN Donghai
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China

Abstract　With the widespread use of deep learning models, people are more aware that the decision-making of model is a problem that needs to be solved urgently. Complex and difficult-to-interpret black-box models hinder the deployment of algorithms in actual scenarios. LIME is the most popular method of local interpretation, but the resulting perturbed data is unstable, leading to bias in the final explanation. To solve the above problems, local interpretable model-agnostic explanations based on active learning and rational quadratic kernel, ActiveLIME, is proposed, which makes the local interpretable model more faithful to the original classifier. After ActiveLIME generates the perturbed data, it samples the perturbation through the query strategy of active learning, selects the perturbation with high uncertainty for training, and uses the local model with the highest accuracy in the iteration to generate explanations for the instances of interest. And for high-dimensional sparse samples that are prone to local overfitting, a rational quadratic kernel is introduced into model's loss function to reduce overfitting. Experiments indicate that the proposed ActiveLIME has better local fidelity and quality of explanations than traditional local explanation algorithms.

Keywords　Local explanation, Perturbation sampling, Query strategy of active learning, Rational quadratic kernel

1　引言

如今复杂的实际场景如医疗、军事等领域对机器学习算法的性能以及人机互信的要求很高,仅仅达到模型精度的要求是不够的,而人类往往无法理解做出决策的深度学习模型[1]内部的决策逻辑,使得机器学习算法无法被使用者所信任,大大增加了部署难度。传统广义线性模型和树模型如线性回归、逻辑回归、决策树[2]等由于自身逻辑简单,因此具有较强的自解释性,但却无法满足实际场景中所需的性能要求。可解释技术的出现解决了高性能模型的自解释能力弱的问题,同时满足了各个领域对解释高性能黑盒模型的需求[3]。

可解释技术可以从3个角度进行划分[4-5]:(1)事前与事后;(2)模型无关与基于特定模型;(3)全局与局部[6]。LIME是一种流行的事后解释方式,需要黑盒模型训练完成后对感兴趣样本进行解释,并且LIME不需要特定模型,任意模型都可以解释,只需要提供黑盒模型的预测函数接口即可[7]。与全局解释[8]不同,LIME基于感兴趣实例,对单一样本生成解释。LIME首先会在感兴趣实例周围生成人工合成扰动数据集,使用欧氏距离对扰动样本加权后,将其作为训练集拟合一个局部稀疏线性模型,并将线性模型最终的特征权重系数

到稿日期:2023-03-03　返修日期:2023-06-25
基金项目:国防基础科研计划(JCKY2020204C009)
This work was supported by the National Defense Basic Scientific Research program of China(JCKY2020204C009).
通信作者:袁伟伟(yuanweiwei@nuaa.edu.cn)

作为解释。然而，由 LIME 生成的扰动数据具有不稳定性，无法保证生成数据的质量，如果不对扰动数据进行筛选，直接使用全部扰动数据会导致可解释模型的解释质量下降，并且会降低模型的局部保真性。

为了解决 LIME 生成的扰动数据的质量不稳定的问题，我们提出了 ActiveLIME 作为 LIME 方法的改进。与 LIME 使用全部随机扰动数据不同，ActiveLIME 引用主动学习查询策略[9]的思想采样不确定性高的扰动样本进行迭代式训练，直到满足指定迭代次数或者局部可解释模型达到精度要求，并且会选择迭代过程中性能表现最佳的模型作为最终的解释模型。ActiveLIME 主要使用基于版本空间缩减的采样策略[10]，其主张标记选择能够最大程度缩减版本空间的样本。传统局部解释方法以及 LIME-NN 和 DLIME 改进算法都使用线性模型作为局部解释模型，并通过线性模型生成解释。另外，ActiveLIME 是基于版本空间缩减中最著名的 Bagging-QBC 方法[11]，使用相同训练集训练多个线性基学习器，并投票选出争议样本，训练过程使用多个线性模型集成具有比单一弱学习器更强的拟合能力，最终解释也由每个基学习器的解释加权输出。同时，为了能更好地处理高维稀疏样本，我们放弃了原始局部解释方法中的指数核，转而在损失函数中引入二次有理核，获得了比传统解释方法更好的泛化能力和更少的计算成本。本文的贡献如下：

(1)借用主动学习中查询策略的思想，利用基于 Bagging-QBC 的查询策略采样不确定性高的扰动样本，并在局部融合多个线性及非线性基学习器。

(2)面对高维稀疏样本，为了避免局部过拟合，在损失函数中引入二次有理核并且对比了引入不同核函数的影响。

(3)在 4 个标准数据集上对比了 3 种局部解释方法，定量地证明了 ActiveLIME 相比传统局部解释方法具有更高的解释质量和局部保真度。

2 相关工作

局部解释技术的出现解决了复杂神经网络可解释性差、决策机制难以理解的问题。不同于从整体上训练近似黑盒模型预测的全局代理模型，局部代理模型旨在在局部拟合模型并对单条样本做出解释。

传统局部解释方法 LIME 通过在待解释实例周围生成扰动样本，然后在局部拟合稀疏线性模型，使用局部模型的特征权重作为解释表达。Zafar 等抛弃了在局部人工生成合成样本，而使用层次聚类对训练集进行划分，用 KNN 找到类后，将挑选出的训练集作为局部代理模型的输入[12]。使用黑盒模型的部分训练集作为局部解释模型的输入，虽然能够避免随机生成扰动样本的随机性带来的解释差异问题，但这时解释往往会带来错误，因为待解释样本来自于黑盒模型的测试集，而我们并不知道黑盒模型在测试集上会做出怎样的决策。Shankaranarayana 等提出去噪编码器对引入高斯噪声的训练数据复原作为扰动样本集，并且优化了局部线性模型的加权函数，提高了模型的保真度[6]。Ranjbar 等在使用去噪编码器的基础上，使用决策树代替局部线性模型，使非线性问题[13]得到了更好的处理。Laugel 等提出使用 KNN 寻找扰动样本生成时的约束边界，提高了局部模型的拟合能力[14]。Ribeiro 等提出了锚点的概念，解决了针对单条样本的局部解释可能无法适用于其他样本的问题，划分了解释边界，提高了解释的适用性[15-16]。Zhao 等提出利用先验知识和贝叶斯推理来提高单个预测重复解释的一致性和对内核设置的鲁棒性[17]。Adler 等通过观察在感兴趣实例的特征值周围进行扰动而对黑盒模型预测产生的影响，更加准确地概述了局部解释模型的决策行为[18]。Bramhall 等将 LIME 中的线性关系重新定义为二次关系，提高了局部模型处理非线性关系的能力以及解释的准确性[19]。

上述文献中提出的方法暴露了传统局部解释方法的两个问题。一方面，局部解释方法中生成扰动邻域是必要的步骤，邻域的生成很大程度影响了局部解释模型的预测精度和最终解释的质量，由于局部扰动样本是随机生成的，因此最终的解释具有不稳定性，无法保证最终解释的质量，虽然有方法使用黑盒模型的训练集来作为替代，但提取到的信息并不充分。另一方面，局部代理模型主要使用稀疏线性模型，对于很多非线性问题，其并不能很好地划分决策边界，导致产生最终解释的误差。因此，提高局部扰动样本生成的稳定性以及局部代理模型的拟合能力，对于模型的局部保真度以及最终解释质量是非常关键的。

3 ActiveLIME 模型无关局部解释方法

ActiveLIME 是对传统局部解释方法 LIME 做出改进的模型无关可解释方法，目的是在决策边界附近采样不确定性最高的样本从而优化扰动样本的生成，以及在损失函数中引入二次有理核使得局部解释模型更加忠于原始分类器，使最终得到的局部融合解释模型的精度更高，并且获得更高的局部保真度和解释质量。图 1 给出了 ActiveLIME 算法的逻辑流程图。

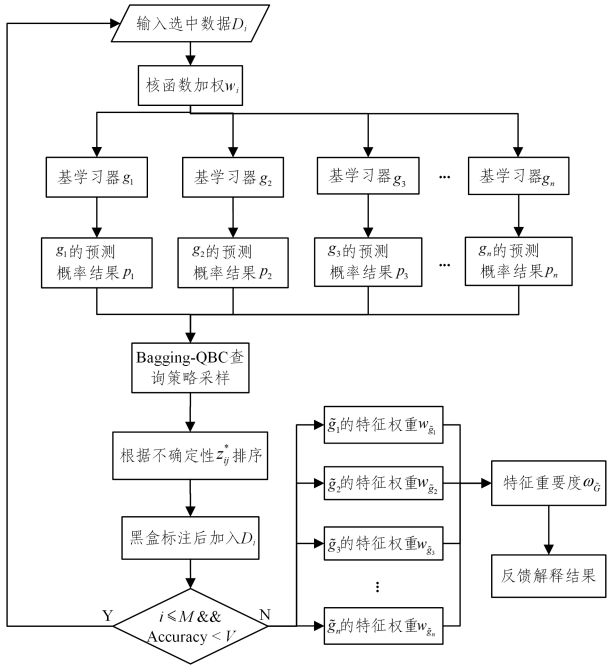


图 1 ActiveLIME 逻辑流程图

Fig. 1 ActiveLIME logic flowchart

设查询迭代次数为 M，扰动样本集为 D，核心流程为：

步骤 1　输入选中的数据 $D_i(D_i \subseteq D)$，$D_i=\{d_{ij} \mid d_{ij} \in D_i\}$，对第 i 次迭代中每条扰动样本 z_{ij} 进行加权得到每条样本权重 $w_{ij}(i=1,2,3,\cdots,M, j=1,2,3,\cdots,count(D_i))$。

步骤 2　训练多个不同的线性或者非线性基学习器 g_t $(g_t \subseteq G, t=1,2,3,\cdots,T)$，并对所有扰动数据进行预测。

步骤 3　根据预测概率结果 $p_t(p_t \subseteq P, t=1,2,3,\cdots,T)$ 使用主动学习中的 Bagging-QBC 查询策略采样不确定性 z_{ij}^* 最高的 k 个样本。

步骤 4　使用待解释的黑盒模型对步骤三中挑选出的 k 个样本进行预测，标注后加入 D_i 中。

步骤 5　当 i 达到迭代次数最大值 M 或者局部解释模型的精度达到要求值 V 后进入步骤 6，否则回到步骤 1。

步骤 6　将整个迭代轮次中准确度指标最高的一组基学习器集合 $\widetilde{G}$ 输出的特征权重 $w_{\widetilde{g}_t}$ 进行加权融合得到特征重要度 $\omega_{\widetilde{G}}$，最终反馈给用户解释结果。

ActiveLIME 局部解释方法的伪代码如算法 1 所示，迭代轮次为 M，在每一轮迭代中会训练 T 个基学习器，因此 ActiveLIME 算法的时间复杂度为 $O(MT)$。

算法 1　ActiveLIME

输入：扰动样本子集 D_i，待解释样本 x，扰动样本 z_{ij}，查询迭代轮次 M，基学习器 g_t，黑盒模型 f

输出：特征重要度 $\omega_{\widetilde{G}}$

1. Initialize $w_{ij} \leftarrow \{\}$, $p_t \leftarrow \{\}$, $\widetilde{G} \leftarrow \{\}$
2. while $i \leqslant M$ and $acc < V$ do
3. $\quad w_{ij} \leftarrow kernel(distance(x, z_{ij}))$
4. $\quad$ for $t=1,2,\cdots,T$ do
5. $\quad\quad g_t.fit(D_i, W_i)$
6. $\quad\quad p_t \leftarrow g_t.predict(D_i)', D_i \cup D_i' = D$
7. $\quad$ end
8. $\quad \widetilde{D}_i' \leftarrow BaggingQBC(D_i', z_{ij}^*, k) \widetilde{D}_i' \subseteq D_i'$
9. $\quad D_i \leftarrow D_i \cup \widetilde{D}_i'$
10. $\quad acc \leftarrow Accuracy\left(\frac{1}{T}\sum_{g_t \in G} g_t.predict(D), f.predict(D))\right)$
11. $\quad i=i+1$
12. end
13. $\widetilde{G} \leftarrow \{\widetilde{g}_1, \widetilde{g}_2, \widetilde{g}_3, \cdots, \widetilde{g}_n\}$
14. $\omega_{\widetilde{G}} \leftarrow \frac{1}{T}\sum_{\widetilde{g}_t \in \widetilde{G}} w_{\widetilde{G}_t}$
15. return $\omega_{\widetilde{G}}$

3.1　可解释表达以及目标函数

和传统局部解释方法类似，我们的解释使用人类可以理解的可解释表达，而非使用模型中的特征值作为解释。对于结构化数据，我们定义 x 为样本的原始表达，$x \in R^d$，x' 为样本的可解释表达的二进制形式，$x' \in \{0,1\}^{d'}$。

定义 g_t 为局部解释模型，$g_t \in G$，其中 G 是自解释较强的线性模型以及树模型等的集合，g_t 作用于样本的二进制形式的可解释表达 x'，由于不是所有 g_t 的自解释性都较强，因此定义 $\Omega(g_t)$ 衡量解释模型 g_t 的复杂度。对于决策树模型而言，其复杂度可以用树模型的深度来衡量；而对于线性模型，其复杂度可以用非零权重的数量来衡量。待解释的黑盒模型定义为 $f: R^d \rightarrow R$，在多目标分类问题中，$f(x)$ 为样本原始表达 x 属于某一类的概率。另外，为了比较扰动样本 z 和待解释样本 x 之间的位置关系，我们定义核函数 $\pi_x(z)$ 作为度量扰动样本 z 和待解释样本 x 的临近度。定义损失函数 $L(f, g_t, \pi_x)$ 衡量由 π_x 定义的局域中局部解释模型 g_t 对于黑盒模型 f 的近似程度，为了兼顾可解释性和局域性，必须最小化 $L(f, g_t, \pi_x)$ 并尽可能降低解释模型 g_t 的复杂度[7]。ActiveLIME 解释方法的目标函数公式如下：

$$\xi_t(x)=\arg\min_{g_t \in G} L(f, g_t, \pi_x)+\Omega(g_t) \tag{1}$$

3.2　基于 Bagging-QBC 查询策略的扰动采样

由于扰动样本生成具有随机性，因此产生了很多并不符合实际的数据。为了从扰动样本中挑选出最有价值且对模型决策帮助最大的样本，我们认为分布在决策边界附近的点为不确定性较高的点，即局部模型难以判断的样本。通过每次迭代输入查询策略挑选出不确定度最高的样本，可以有效避免生成的扰动样本的随机性影响模型决策。

考虑到原始 LIME 方法中局部解释模型采用岭回归的基础模型，对于非线性数据的拟合能力有限，而为了输出可解释表达，局部解释模型本身不能太复杂，自解释能力强的模型只有线性模型以及决策树模型。本文将局部解释场景延伸到多模型，通过多个线性模型或者决策树模型集成，融合成一个局部解释模型，使得该解释模型不仅能够很好地处理线性数据，也能更好地应对非线性数据，取得比单一线性模型更好的决策性能。通过多个模型投票的方式，选出局部解释模型相对难以做出区分的样本。

设局部解释模型集合 G，$G=\{g_1, g_2, g_3, \cdots, g_T\}$，局部解释模型必须为自解释性强的线性模型或者决策树模型，通过熵来衡量分类器区分不同样本数据的难度，计算式如下：

$$z^*=\arg\max_z -\sum_{c_i}\frac{V(c_i)}{S}\log\frac{V(c_i)}{S} \tag{2}$$

其中，基学习器类别集合 $C=\{c_1, c_2, c_3, \cdots, c_n\}$，$c_i$ 为类别编号，$V(c_i)$ 为所有基学习器中投票给类别编号为 c_i 的学习器的个数，S 为基学习器的个数。

在通过 Bagging-QBC 查询策略[11]采样完一批样本后，如果此时模型的精度未达到要求或者还未达到最大迭代次数，需要将本次迭代中挑选出的样本重新加入旧数据集中，然后对新数据集进行核函数加权，放入不同基学习器中学习，根据预测结果继续使用投票熵查询策略挑选样本，直到满足停止条件。在每次迭代中会调整最优局部模型，在满足停止条件后找到所有迭代轮次中精度最高的局部解释模型，然后通过对每个基学习器加权融合，输出融合之后的特征权重，作为可解释表达输出给用户。

3.3　基于二次有理核的稀疏融合解释

传统局部解释方法的指数核存在计算成本高和容易过拟合的问题。我们在平方损失函数中引入二次有理核函数[20]，降低了模型的计算成本，并且，在处理高维稀疏小样本的问题时，带有指数核的模型往往过于依赖训练集，导致模型的泛化性较差，在测试集上的表现不佳。引入二次有理核有效地避免了局部解释模型过拟合的问题，提高了局部代理模型的解释质量。二次有理核函数的定义如下：

$$\pi_x(\boldsymbol{z})=1-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\|\boldsymbol{x}-\boldsymbol{z}\|^2+\delta} \tag{3}$$

其中，$\boldsymbol{x}$ 为待解释样本的原始向量表达，$\boldsymbol{x}\in R^d$；$\boldsymbol{z}$ 为扰动样本的原始向量表示，$\boldsymbol{z}\in R^d$；δ 是可调参数，默认值为 $\delta=0.75\sqrt{d}$。

对于高维稀疏小样本，我们在局部拟合多个线性及非线性模型，并使用 bagging 思想进行加权融合，表达式如下：

$$g(z')=\frac{1}{T}\sum_{g_i\in G}\boldsymbol{w}_{g_i}^{\mathrm{T}}\cdot z' \tag{4}$$

其中 T 为基学习器的个数，z' 为扰动样本的二进制形式的可解释表达，$z'\in\{0,1\}^{d'}$，$\boldsymbol{w}_{g_i}$ 为局部基学习器的模型权重的向量表示。

对于不同局部代理线性模型 g_i，将引入二次有理核的加权平方损失 L_{i1} 或者对数似然损失 L_{i2} 作为 g_i 的损失函数[21]，表达式如下：

$$L_{i1}(f,g,\pi_x)=\sum_{z,z'\in Z}\pi_x(z)(f(z)-g_i(z'))^2 \tag{5}$$

$$L_{i2}(f,g,\pi_x)=-\sum_{z,z'\in Z}\pi_x(z)(f(z)\log(g_i(z'))+(1-f(z))\log(1-g_i(z'))) \tag{6}$$

其中，$\pi_x(\boldsymbol{z})$ 为二次有理核函数；f 为待解释黑盒模型，$f(z)$ 为黑盒模型 f 对扰动样本的原始表达对特定类的预测概率；g_i 为局部代理解释模型，$g_i(z')$ 为局部代理解释模型对扰动样本的可解释表达对特定类的预测概率。

4 实验

4.1 数据集

为了验证 ActiveLIME 方法的可行性，我们选用了来自 UCI 存储库和 Kaggle 上的 4 个数据集。Chess(King-Rook vs. King-Pawn)数据集每个实例都是国际象棋中残局的棋盘描述；ILPD(Indian Liver Patient Dataset)数据集是由印度收集的肝病患者与非肝病患者的记录；Statlog(German Credit Dataset)数据集中每个条目代表一个从银行获得信贷的人，根据属性集，每个条目都被分类为良好或不良的信用风险；PC1(NASA Defect Dataset)数据集由 NASA 缺陷数据组成。具体信息如表 1 所列。

表 1 实验数据集

Table 1 Experimental datasets

数据集	任务类型	属性数	实例数
Chess	多变量分类	36	3196
ILPD	多变量分类	10	583
Statlog	多变量分类	9	1000
PC1	多变量分类	37	759

4.2 实验结果

本文实验将 ActiveLIME 算法和原始局部解释算法 LIME 及其扰动改进算法 LIME-NN、DLIME 进行对比，对比方法的简单介绍如下。LIME 的核心思路包括：1)在感兴趣实例生成随机扰动样本；2)利用原始黑盒模型对新合成样本进行预测；3)对新的合成样本进行指数化加权；4)对特征进行筛选；5)利用新合成样本训练一个加权局部线性模型；6)生成解释返回[7]。由于 LIME 集中在感兴趣实例周围扰动采样，然而生成的邻域却不一定能够得到最佳决策边界，为了采样正确的邻域，基于 LIME 的最近邻域(Nearest Neighbour)方法 LIME-NN 被提出。该方法首先在训练集中找到与感兴趣实例 x 最近的一条实例 x_c，将该实例与感兴趣实例的距离作为扰动合成样本和感兴趣实例之间的最大距离约束[14]。LIME 中，因为扰动样本生成具有随机性，所以解释结果可能有差异。为了解决这个问题，DLIME 结合使用层次聚类 Agglomerative Hierarchical Clustering(AHC)对训练样本进行分类，并用 K 近邻(KNN)[22]挑选感兴趣实例周围的相关聚类，将被挑选出的训练样本固定为扰动集，解决了扰动样本随机性的问题。该方法首先使用层次聚类 AHC 算法[23]将训练集的数据分组，然后通过 KNN 算法找到距离感兴趣实例 x 最近的 N 的训练样本，选取 N 个样本中的主导类，将划分为主导类的训练集样本作为扰动集，然后和 LIME 一样在局部拟合线性模型，给出解释[12]。

我们在实验中使用的黑盒模型是随机森林模型，使用留出法划分训练集和测试集，将 70% 的数据作为训练集，将 30% 的数据作为测试集；每条实例周围生成的扰动数量统一为5000；ActiveLIME 中扰动样本选择的迭代次数为 30，每次挑选 top100 的样本，在训练局部代理模型时融合两个基学习器；DLIME 中生成扰动样本时层次聚类的簇数为 5，KNN 的最近邻居数为 100；LIME-NN 中使用均匀分布生成扰动样本，其他方法均为高斯分布生成。对于可解释评价指标，本文使用两个客观评价指标定量评价所提算法，分别是局部保真度以及解释质量。

4.2.1 局部保真度

局部保真度反映了局部解释模型对于待解释黑盒模型的拟合效果，和原始黑盒模型进行比较可得到分类性能[24]。实验中，我们使用待解释样本周围生成的扰动样本训练局部代理解释模型，将测试集的预测值与黑盒模型的预测值进行比较。通过计算精确率、召回率、准确率、F_1-Score 以及 AUC 作为局部保真度的度量。在多变量分类任务中，精确率描述了在所有预测为正的样本中实际为正的样本的概率；召回率描述了在实际为正的样本中被预测为正的样本的概率；准确率描述了预测正确的样本占总样本的比例；F_1-Score 是精确率和召回率的加权调和平均，F_1-Score 越大，模型性能越好；AUC 是 ROC 曲线下面积，ROC 曲线下方的面积越大说明模型性能越好。局部保真度的度量公式如下：

$$precision=\frac{TP}{TP+FP} \tag{7}$$

$$recall=\frac{TP}{TP+FN} \tag{8}$$

$$accuracy=\frac{TP+TN}{TP+FN+FP+FN} \tag{9}$$

$$F_1=\frac{2\times precision\times recall}{precision+recall} \tag{10}$$

其中，TP 是模型预测为正的正样本，FP 是模型预测为正的负样本，FN 是模型预测为负的正样本，TN 是模型预测为负的负样本。在局部保真度的概念中，样本的实际值等于黑盒模型的预测值，样本的预测值等于局部代理解释模型的预测值[25]。如图 2－图 5 所示，在 4 个数据集中，ActiveLIME 的局部保真度表现均优于传统方法 LIME 以及改进算法

LIME-NN 和 DLIME。可以看出,ActiveLIME 模型的拟合能力和泛化能力优于其他几种方法,尤其是相比原始局部解释方法 LIME 具有显著的提升。

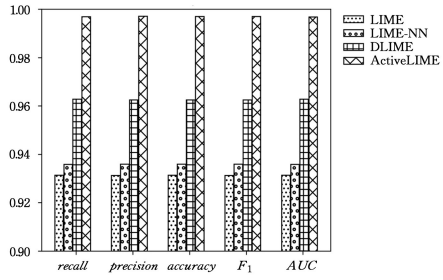


图 2 Chess 数据集上各模型的局部保真度

Fig. 2 Local fidelity of different models on Chess

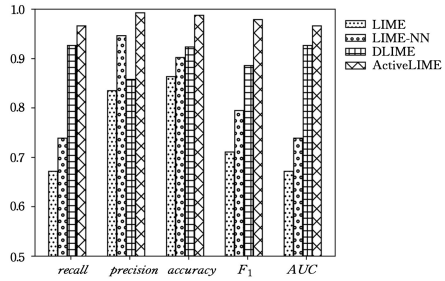


图 3 ILPD 数据集上各模型的局部保真度

Fig. 3 Local fidelity of different models on ILPD

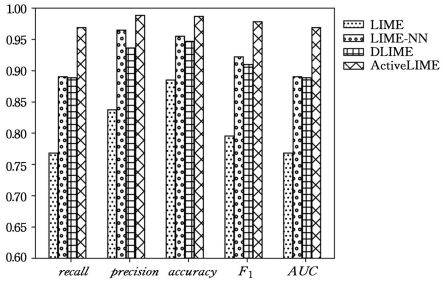


图 4 Statlog 数据集上各模型的局部保真度

Fig. 4 Local fidelity of different models on Statlog

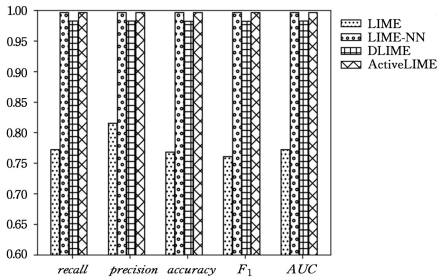


图 5 PC1 数据集上各模型的局部保真度

Fig. 5 Local fidelity of different models on PC1

4.2.2 解释质量

解释质量反映了局部代理可解释模型的可靠性,计算了局部代理模型的可解释表达与原始黑盒模型特征权重的相似度[26]。对于原始黑盒模型 f、局部解释模型 g,以及 d 维待解释实例 x,待解释样本源自测试集,测试集数量为 H,假定原始黑盒模型 f 的权重向量为 $\boldsymbol{e}=\{e_1,e_2,e_3,\cdots,e_d\}$,可解释模型 g 对于单条待解释实例 x_i 的解释向量为 $\boldsymbol{e}_i'=\{e_{i1}',e_{i2},e_{i3},\cdots,e_{id}\}$,实验中通过欧氏距离(Euclidean)、余弦相似度(Cosine)以及皮尔逊相关系数(Pearson)衡量 $\boldsymbol{e}$ 和 $\boldsymbol{e}'$ 之间的相似度,公式分别如下:

$$Q_1=\frac{1}{H}\sum_{i=1}^{H}\frac{1}{1+\|\boldsymbol{e}-\boldsymbol{e}_i'\|_2} \tag{11}$$

$$Q_2=\frac{1}{H}\sum_{i=1}^{H}\frac{\boldsymbol{e}\cdot\boldsymbol{e}_i'}{\|\boldsymbol{e}\|\boldsymbol{e}_i'\|} \tag{12}$$

$$Q_3=\frac{1}{H}\sum_{i=1}^{H}\frac{cov(\boldsymbol{e},\boldsymbol{e}_i')}{\sigma_e\sigma_{e_i'}} \tag{13}$$

实验中将 ActiveLIME 和其他 3 种方法生成的解释的质量进行比较,下表中的结果均为完整测试集计算的平均解释质量,ActiveLIME 和 LIME-NN、DLIME 算法以及传统局部解释算法比较的结果如表 2 所列。

表 2 不同模型的解释质量对比

Table 2 Explanation quality comparison of of different models

数据集	相似度	LIME	LIME-NN	DLIME	ActiveLIME
Chess	Euclidean	0.6406	0.6405	0.5225	**0.6613**
	Cosine	0.9468	0.9467	0.8621	**0.9505**
	Pearson	0.9321	0.9320	0.8218	**0.9375**
ILPD	Euclidean	0.4465	0.4413	0.4163	**0.4661**
	Cosine	0.7775	0.7223	0.6749	**0.7795**
	Pearson	0.3310	0.2371	0.2673	**0.3885**
Statlog	Euclidean	0.4289	0.3970	0.4143	**0.4366**
	Cosine	0.5905	0.4108	0.5252	**0.5952**
	Pearson	0.1778	0.2185	0.2202	**0.2205**
PC1	Euclidean	0.2933	0.3393	0.3463	**0.3637**
	Cosine	0.6648	0.5662	0.4672	**0.6895**
	Pearson	0.2135	0.1593	0.1838	**0.3208**

表 2 列出了 4 种模型在 4 个数据集上的比较结果。可以看出,4 种模型在提高局部解释模型复杂度的同时都会或多或少地损失解释质量,但 ActiveLIME 算法的解释质量和传统方法以及改进算法相比,均有不同程度的提升,其中在 PC1 数据集上的提升最为明显,Euclidean 相似度提升了 1.74%以上,Cosine 相似度提升了 2.47%以上,Pearson 相似度提升了 10.73%以上。这充分说明了 ActiveLIME 模型的解释质量更优。

4.2.3 核函数对比

ActiveLIME 方法除了在扰动样本生成后利用主动学习中的查询策略进行采样外,还在损失函数中引入二次有理核以缓解高维稀疏样本条件下传统局部解释方法中的指数核可能带来的过拟合问题。为了证明引入二次有理核的有效性,我们在保留使用主动学习的思想下,分别对比指数核、高斯核以及二次有理核[20,27]对模型的影响。其中指数核函数和高斯核函数的定义分别如下:

$$\pi_1(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{2\delta^2}\right) \tag{14}$$

$$\pi_2(\boldsymbol{x},\boldsymbol{z})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\delta^2}\right) \tag{15}$$

其中,$\pi_1(\boldsymbol{x},\boldsymbol{z})$是指数核函数的定义,$\pi_2(\boldsymbol{x},\boldsymbol{z})$是高斯核函数的定义,$\boldsymbol{x}$ 为待解释样本的原始向量表达,$\boldsymbol{x}\in R^d$,$\boldsymbol{z}$ 为扰动样本的原始向量表示,$z\in R^d$,δ 是核宽度。

我们分别在 4 个数据集上进行测试,选用以 F_1 度量的局部保真度指标和以余弦相似度度量的解释质量指标进行对比,实验结果如表 3 所列。

表 3 核函数对比结果

Table 3 Comparison results of kernel functions

数据集	指标	指数核	高斯核	二次有理核
Chess	F_1	0.9656	0.9781	**0.9969**
	Cosine	0.9442	0.9450	**0.9505**
ILPD	F_1	0.7188	0.7524	**0.9784**
	Cosine	0.7721	0.7730	**0.7795**
Statlog	F_1	0.8775	0.8886	**0.9782**
	Cosine	0.5801	0.5809	**0.5952**
PC1	F_1	0.9714	0.9821	**0.9964**
	Cosine	**0.6935**	0.6928	0.6895

表 3 列出了模型使用不同核函数时的评估指标结果，在前三个数据集中，无论是 F_1 指标还是余弦相似度指标，使用二次有理核函数的效果都好于使用指数核函数和高斯核函数。在 PC1 数据集中，二次有理核对 F_1 度量的局部保真度相较于指数核和高斯核分别提升 2.5%和 1.43%，但是对于余弦相似度度量的解释质量指标，二次有理核相较于指数核降低了 0.4%。上述实验说明，使用二次有理核比指数核和高斯核具有更好的泛化性以及更加准确的解释。

结束语 本文提出了一种基于主动学习和二次有理核的扰动优化的模型无关局部解释方法，主要解决了两方面的问题。首先，由于传统局部解释方法生成的扰动具有不稳定性，最终的解释产生偏差，因此 ActiveLIME 使用了主动学习中基于 Bagging-QBC 的查询策略思想进行迭代式采样，每次挑选出一批不确定性最高的样本加入模型重新训练直到满足迭代次数或者达到预期精度，并且在局部融合多个线性模型或非线性模型，提高了局部解释模型的分类性能。其次，在损失函数中引入二次有理核。传统局部解释方法使用指数核在处理高维稀疏小样本时容易陷入局部过拟合，并且计算成本过高，ActiveLIME 使用二次有理核减少了计算成本，并提高了模型的泛化能力。为了验证本文方法的有效性，我们在 4 个数据集上进行了测试，对比了传统的局部解释方法 LIME 以及两种对扰动集生成改进的算法 DLIME 和 LIME-NN，实验结果证明了本文方法具有更高的局部保真度和解释质量。由于可解释性缺乏一些公认性的指标，在未来的研究中，我们需要探究更多评价可解释算法的指标。本文虽然对解释这一目标提出了定量的相似度指标进行衡量，但是缺乏定性评估，且专家经验可能使解释更容易被人为理解。另外，局部解释技术的解释范围也是一种重要的研究方法，对单条样本的解释得到的经验也许并不适用于所有测试样本，如何划定解释范围也是未来研究的一个重要方向。

参 考 文 献

[1] BAI X, WANG X, LIU X, et al. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments[J]. Pattern Recognition, 2021, 120: 108102.

[2] BREIMAN L. Classification and regression trees [M]. Routledge, 2017.

[3] LINARDATOS P, PAPASTEFANOPOULOS V, KOTSIANTIS S. Explainable ai: A review of machine learning interpretability methods[J]. Entropy, 2020, 23(1): 18.

[4] DU M, LIU N, HU X. Techniques for interpretable machine learning[J]. Communications of the ACM, 2019, 63(1): 68-77.

[5] WANG F, KAUSHAL R, KHULLAR D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? [J]. Annals of Internal Medicine, 2020, 172(1): 59-60.

[6] SHANKARANARAYANA S M, RUNJE D. ALIME: Autoencoder based approach for local interpretability[C]// Intelligent Data Engineering and Automated Learning—IDEAL 2019: 20th International Conference, Manchester, UK, November 14-16, 2019, Proceedings, Part I 20. Springer International Publishing, 2019: 454-463.

[7] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" Explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1135-1144.

[8] LAKKARAJU H, BACH S H, LESKOVEC J. Interpretable decision sets: A joint framework for description and prediction [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1675-1684.

[9] SETTLES B. Active learning literature survey [R]. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[10] MUSLEA I, MINTON S, KNOBLOCK C A. Active learning with multiple views[J]. Journal of Artificial Intelligence Research, 2006, 27: 203-233.

[11] SEUNG H S, OPPER M, SOMPOLINSKY H. Query by committee[C]// Proceedings of the fifth Annual Workshop on Computational Learning Theory. 1992: 287-294.

[12] ZAFAR M R, KHAN N. Deterministic local interpretable model-agnostic explanations for stable explainability[J]. Machine Learning and Knowledge Extraction, 2021, 3(3): 525-541.

[13] RANJBAR N, SAFABAKHSH R. Using decision tree as local interpretable model in autoencoder-based lime[C]// 2022 27th International Computer Conference. Computer Society of Iran (CSICC), IEEE, 2022: 1-7.

[14] LAUGEL T, RENARD X, LESOT M J, et al. Defining locality for surrogates in post-hoc interpretablity [J]. arXiv: 1806.07498, 2018.

[15] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: High-precision model-agnostic explanations[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[16] RIBEIRO M T, SINGH S, GUESTRIN C. Nothing else matters: Model-agnostic explanations by identifying prediction invariance[J]. arXiv: 1611.05817, 2016.

[17] ZHAO X, HUANG W, HUANG X, et al. Baylime: Bayesian local interpretable model-agnostic explanations[C]// Uncertainty in Artificial Intelligence (PMLR). 2021: 887-896.

[18] ADLER P, FALK C, FRIEDLER S A, et al. Auditing black-box models for indirect influence[J]. Knowledge and Information Systems, 2018, 54: 95-122.

[19] BRAMHALL S, HORN H, TIEU M, et al. Qlime-a quadratic local interpretable model-agnostic explanation approach[J]. SMU Data Science Review, 2020, 3(1): 4.

[20] HOFMANN T, SCHÖLKOPF B, SMOLA A J. Kernel methods in machine learning[J]. The Annals of Statistics, 2008, 36(3): 1171.

[21] JANOCHA K, CZARNECKI W M. On loss functions for deep neural networks in classification[J]. arXiv: 1702.05659, 2017.

[22] GUO G, WANG H, BELL D, et al. KNN model-based approach in classification[C]// OTM Confederated International Conferences on The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. 2003: 986-996.

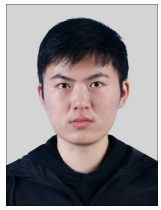
[23] NIELSEN F, NIELSEN F. Hierarchical clustering[J/OL]. Introduction to HPC with MPI for Data Science, 2016: 195-211. https://link.springer.com/chapter/10.1007/978-3-319-21903-5_8.

[24] VILONE G, LONGO L. Notions of explainability and evaluation approaches for explainable artificial intelligence[J]. Information Fusion, 2021, 76: 89-106.

[25] GRANDINI M, BAGLI E, VISANI G. Metrics for multi-class classification: an overview[J]. arXiv: 2008.05756, 2020.

[26] JIA Y, BAILEY J, RAMAMOHANARAO K, et al. Improving the quality of explanations with local embedding perturbations[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 875-884.

[27] KEERTHI S S, LIN C J. Asymptotic behaviors of support vector machines with Gaussian kernel[J]. Neural Computation, 2003, 15(7): 1667-1689.



ZHOU Shenghao, born in 1999, master. His main research interests include data mining and machine learning interpretability.

YUAN Weiwei, born in 1981, Ph.D, professor. Her main research interests include data mining and intelligence computing.

（责任编辑：何杨）

2023 年“CCF 最高科学技术奖”评选结果公告

“CCF 最高科学技术奖”授予在计算机科学、技术和工程领域取得重大突破，成就卓著、贡献巨大的资深中国计算机科技工作者。CCF 奖励委员会决定授予北京航空航天大学李未教授、北京信息科技大学苏东庄教授 2023 年“CCF 最高科学技术奖”，以表彰他们为中国计算机事业的发展做出的卓越贡献。

特此公告。

中国计算机学会

2024 年 1 月 12 日

李　未　北京航空航天大学教授，CCF 会士，中国科学院院士。

获奖理由：李未教授建立了并发语言的翻译与变换理论，给出对错误进行修正的 R-演算系统和版本序列理论，提出非结构化数据的四面体模型，建立了互联网群体智能的理论框架，对计算机学科建设和计算机教育质量提升做出了杰出贡献。

苏东庄　北京信息科技大学教授。

获奖理由：苏东庄教授作为重要成员参加我国第一代计算机（104 机）的研制工作，主持编著我国最早的具有重要影响力的计算机系统结构教材，率先开展中文海量信息全文检索研究并在产业化方面做出了突出贡献。

据 CCF 微信公众号