计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300158

生成链接树:一种高数据真实性的反事实解释生成方法

王 明 武文芳 王大玲 冯 时 张一飞
东北大学计算机科学与工程学院 沈阳 110169
(2001819@stu.neu.edu.cn)

摘 要 超大的数据规模及结构复杂的深度模型在互联网数据的处理与应用方面表现出了优异的性能,但降低了人工智能(Artificial Intelligence,AI)系统的可解释性。反事实解释(Counterfactual Explanations,CE)作为可解释性领域研究中一种特殊的解释方法,受到了很多研究者的关注。反事实解释除了作为解释外,也可以被视为一种生成的数据。从应用角度出发,文中提出了一种生成具有高数据真实性反事实解释的方法,称为生成链接树(Generative Link Tree,GLT),采用分治策略与局部贪心策略,依据训练数据中出现的案例生成反事实解释。文中对反事实解释的生成方法进行了总结并选取了其中热门的数据集来验证GLT方法。此外,提出"数据真实性(Data Fidelity,DF)"的指标,用于评估反事实解释作为数据的有效性和潜在应用能力。与基线方法相比,GLT生成的反事实解释数据的真实性明显高于基线模型所生成的反事实解释。

关键词:可解释性;填充式;反事实解释;数据真实性
中图法分类号 TP391

Generative Link Tree:A Counterfactual Explanation Generation Approach with High Data Fidelity

WANG Ming,WU Wen-fang,WANG Da-ling,FENG Shi and ZHANG Yi-fei
School of Computer Science and Engineering,Northeastern University,Shenyang 110169,China

Abstract The super large data scale and complex structure of deep models show excellent performance in processing and application of Internet data,but reduce the interpretability of AI systems. Counterfactual Explanations(CE) has received much attention from researchers as a special kind of explanation approach in the field of interpretability research. Counterfactual Explanations can be regarded as a kind of generated data in addition to being an explanation. From the viewpoint of application,this paper proposes an approach for generating counterfactual explanations with high data fidelity,called generative link tree(GLT),which uses a partitioning strategy and a local greedy strategy to construct counterfactual explanations based on the cases appearing in the training data. Moreover,it summarizes the generation methods of counterfactual explanations and select popular datasets to verify the GLT method. In addition,the metric of "Data Fidelity (DF)" is proposed to evaluate the fidelity and potential application of the counterfactual explanation as data from an experimental perspective. Compared with the baseline method,the data fidelity of the counterfactual explanation generated by the GLT method is significantly higher than that of the counterfactual explanation generated by the baseline model.

Keywords Interpretability,Filling type,Counterfactual explanations,Data fidelity

1 引言

AI系统通常会随着模型及数据规模的增大表现出愈加优异的性能。然而,过大的参数规模或大量的数据使得AI系统的透明性和可解释性[1]变差。近年来,反事实解释[2]备受关注,这是Wachter等[2]将反事实[3]引入可解释领域而提出的一种基于局部样例的事后解释方法[4]。与其他解释方法的不同之处在于,反事实解释不会给出关于问题"为什么分类器会将x分类为y"的解释,而是给出了关于"对x做出何种变化可以使得分类器将x'分类为y'"这类问题的解释。表1以一个常见的案例介绍了反事实解释的形式,加粗字体表示反事实解释相对查询样本改变的特征。表1中,对于一个低收入用户来说,反事实解释没有告知其为何被分类为低收入用户,而是给出了关于其如何才能成为高收入用户的解释(攻读一个硕士学位并从事专业性质的工作)。通过这种方式,反事实解释可以为用户提供一种实现其目标的参考方式,这对普通用户来说通常比仅仅获知原因更有吸引力。

到稿日期:2022-03-16 **返修日期**:2022-05-30
基金项目:国家自然科学基金(62172086,61872074)
This work was supported by the National Natural Science Foundation of China(62172086,61872074).
通信作者:王大玲(wangdaling@cse.neu.edu.cn)

表 1 反事实解释示例

Table 1 Example of counterfactual explanation

样例	年龄	工作类别	受教育类型	工作类型	性别	每周工作时间/h	收入类别
查询样本	24	Private	Bachelors	Service	Male	40	Low income
反事实解释	27	Private	**Master**	**Professional**	Male	40	High income

通过观察反事实解释的形式就可以看出，反事实解释本质上也是一种生成的数据，但这种数据在相关领域并没有得到应用。通过分析现有的关于反事实解释生成方法的研究发现，这些工作的关注点集中在查询样本与反事实解释之间的距离上，但这通常是与验证模型高度相关的，这就导致生成的反事实解释的泛用性差，且对原始数据的忠实度较低。

忠实度问题[5]受到了可解释性领域很多研究者的关注，保持对原始数据的忠实度[6]是应用反事实解释的前提，但反事实解释领域的相关讨论却不多。从实验的角度分析，如果反事实解释对原始数据的忠实度足够高，其应该具有如下性质。

(1)有效性

反事实解释的类别应为期望类别，在实验中依赖验证模型，即生成反事实解释所依赖的模型，其对反事实解释的分类结果应为期望类别[7]。

(2)泛适应能力

将与生成反事实解释所依赖的验证模型不同的分类模型称为“第三方模型”，对反事实解释的分类也应当具备较高的分类精度、F1-score 等指标。

本文从问题关注点、生成方式及方法类型等方面总结了近几年研究领域内的一些反事实解释生成方法，如表 2 所列。经过分析发现，尚未见研究者从应用角度提出反事实解释生成方法，相关评价指标[8]也没有考虑反事实解释的忠实度。

表 2 近期反事实解释生成方法的总结

Table 2 Summary of recent studies on generating counterfactual explanations

研究工作	关注问题	生成方法	生成类型
文献[2]	COUNTERFACTUAL EXPLANATION	ADAM optimizer	Perturbative
文献[7]	feasibility, diversity	Gradient descent	Perturbative
文献[10]	actionable	Integer programming	Perturbative
文献[11]	feasibility, actionable	Shortest path on graph over the dataset	Perturbative
文献[12]	good counterfactual	CBR	Filling
文献[13]	visual explanation	Exhaustive search & continuous relaxation	Filling
文献[14]	efficiently	Search with encoded prototypes & k-d trees	Perturbative
文献[15]	sparsity, diversity, valid	k-NN XCs	Filling
文献[16]	actionable	Minimal intervention	Perturbative
文献[17]	proximity	Two weighting strategies	Perturbative
文献[18]	efficiently, diversity	Mixed integer programming	Perturbative
文献[19]	feasibility	Partial structural causal model	Perturbative
文献[20]	model-agnostic, data-type-agnostic, distance-agnostic, diversity	Standard SMT solver	Perturbative
文献[21]	spurious associations	Manual	Perturbative
文献[22]	generate natural counterfactual visual explanation	Change attribute description text	Perturbative
文献[23]	generate visual counterfactual explanation	Heuristic algorithm to find changes	Perturbative

假设机器学习模型对原始数据具有较高的分辨能力，可以将模型对反事实解释的预测结果作为其对反事实解释忠实度的认可程度。进而，本文提出了数据真实性的概念与指标计算方式。基于此，本文将从反事实解释的数据真实性出发，希望生成的反事实解释不仅可以用作解释，同时还可以以数据的形式参与到 AI 系统的训练、应用等过程中。

本文的主要贡献如下：

(1)参考有效性与忠实度的定义，提出了关于反事实解释数据真实性的概念并定义了计算方式；

(2)提出了一种基于原始数据中的案例生成高数据真实性反事实解释的方法——生成链接树；

(3)通过 GLT 方法与基线方法生成的反事实解释的实验对比表明，GLT 方法在数据真实性上其有较大的优势。

2 相关工作

2.1 利用扰动的方式生成反事实解释

早期的反事实解释生成方法大都属于扰动式，其基本思想是在查询样本上施加扰动或在特征空间中改变查询样本的位置来生成反事实解释，主要方式是通过解优化问题减少特定的损失以得到可以施加在查询样本上的最小扰动，从而生成反事实解释。以反事实解释这一概念的提出者 Wachter 等[2]为例，他们通过 ADAM[9]解算器减少预测结果与目标标签、反事实解释与查询样本之间距离的联合损失来生成反事实解释。此外，也有研究者通过整数编码[10]、梯度下降[7]等方式结算特定的优化问题以生成反事实解释。

2.2 基于案例生成反事实解释

近年来，反事实解释的可行性愈加受到众多研究者的关注，除了在优化目标中加入针对可行性相关指标所定义的相关损失外，一些研究工作[11-13]也考虑依据原始数据中案例的特征构建反事实解释。由于特征值都是在原始数据中真实存在的，基于案例的生成方式相比直接解算优化问题所生成的反事实解释具有更高的可行性。基于案例的生成方式主要有两种思路。

(1)通过搜索与查询样本距离最近的目标类样例来获得

可施加在查询样本上的最小扰动，以生成反事实解释的扰动式方法。例如 Poyiadzi 等[11]将数据空间视为图，通过在图中寻找查询样例到目标类样例之间最短路径的方式来生成反事实解释。

(2)在原始数据中搜索目标类别案例，利用案例的特征值填充反事实解释特征空间的填充式方法。Keane 等利用基于案例推理的方法生成反事实解释[12]，首先在原始数据中构建良好的“反事实对”案例，即特征差异较少的查询样本与反事实组合。根据查询样本最近邻的“反事实对”之间的特征异同，可以选择对应的特征填充反事实解释的特征空间。

2.3　反事实解释的评价

很多研究者针对反事实解释的评价提出了一些指标，Verma 等[8]较全面地总结了反事实解释的评价指标，对反事实解释的有效性、可行性、稀疏性、数据流形封闭性、因果关联性以及方法可替代性等指标都进行了介绍。Yue 等[6]定义了反事实解释的忠实度，将其表示为反事实解释对原始数据的特征空间的归属程度。

针对生成反事实解释的应用问题，需要考虑反事实解释对原始数据的忠实度，即更高的有效性与泛适应能力。因此，本文参考填充式技术提出生成链接树方法以构建反事实解释，并提出数据真实性这一评价指标，用于评估反事实解释对原始数据的忠实度，衡量其可应用的潜力。

3　反事实解释生成方法

已有研究表明，使用原始数据中的原型样本[14]生成的反事实解释更加可行且有效[15]，因此本文选择基于原型样本生成反事实解释。基于原型样本的填充式生成方法的本质是一个搜索的过程。区别于扰动式方法[16]根据原型样本和查询样本之间的差异计算干预措施的方式，填充式方法[13]通过搜索原型样本和查询样本的特征组合生成反事实解释。

本文提出的生成链接树方法利用分治的思想，将搜索反事实解释的过程划分为若干组局部特征组合的搜索过程，构建表示局部特征组合的局部贪心树，依据预定的规则筛选出局部最优的特征组合路径。最后将所有的局部贪心树所筛选出的路径进行链接，按照完整的特征选择路径分别从原型样本与查询样本中选择特征填充反事实解释的特征空间。3.1 节及 3.2 节将分别介绍局部特征组合的筛选过程及全局的链接过程。此外，3.3 节将介绍本文提出的评估反事实解释忠实度的数据真实性指标。

3.1　局部贪心树

局部贪心树是用于表示原型样本 P 和查询样本 S 局部特征组合的树状结构，树的节点由原型样本和查询样本的特征值构成。例如，当两个样本的前 3 个特征的编码后特征向量分别为 $\ddot{x}_p=[0.6,0.89,0.49]$ 以及 $\ddot{x}_s=[0.8,0.45,0.87]$ 时，可以构造如图 1 所示的局部贪心树(Local Greedy Tree，LGT)。图 1 中，橙色节点表示原型样本 P 的特征值编码，绿色节点表示查询样本 S 的特征值编码，黄色路径表示原型样本的局部特征子集，蓝色路径表示查询样本的局部特征子集。

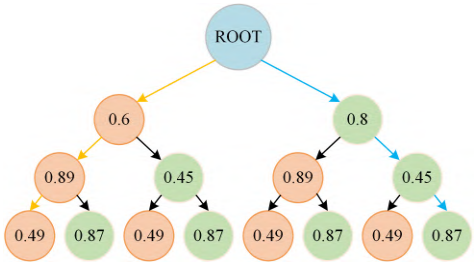


图 1　局部贪心树示例(电子版为彩图)

Fig. 1　Example of local greedy tree

局部贪心树是一棵满二叉树，由树的节点和边构成，其中树的节点可以根据来源分为原型节点和查询节点。基于此，局部贪心树可表示为 $LGT=\mathrm{Tree}(\boldsymbol{V}=(\boldsymbol{P}\cup\boldsymbol{S}),\boldsymbol{E})$，其构建过程如算法 1 所示。

算法 1　LGT 构建算法

输入：原型特征子集 p，查询特征子集 s

输出：局部贪心树 lgt

```
1. l ← length(p); /* 获取局部特征子集长度 */
2. for i ← 1 to l do:
3.    lg t ← Node("ROOT"); /* 初始化树 */
4.    for node ← ergodic(lgt [i+1]) do: /* 遍历局部贪心树的第 i 层
      节点 */
5.       node. left, node. right ← p[i], s[i];
6.    end for;
7. end for;
8. return lgt.
```

算法 1 中的 p 和 s 分别表示原型样本 P 及查询样本 S 的局部特征子集，$length(\cdot)$ 表示获取特征子集的长度，$ergodic(\cdot)$ 为遍历树结构某一层节点的方法。

建立局部贪心树的目的是筛选出相对较好的局部特征组合。为了比较不同局部特征组合的优劣，定义了局部相对相似性(Local Relative Similarity，LRS)，其计算式如式(1)所示：

$$LRS=\frac{\exp(\cos(\ddot{x}_p,\ddot{x}))}{\mathrm{sigmoid}(d(\ddot{x}_s,\ddot{x}))} \tag{1}$$

其中，$\ddot{x}_p$ 代表原型样本 P 的局部特征子集，$\ddot{x}_s$ 表示查询样本 S 的局部特征子集，而 $\ddot{x}$ 则表示局部贪心树中每条路径上的局部特征组合。LRS 的计算可以分为两部分，分子部分表示当前特征组合与原型样本特征子集之间的相似度，用于约束局部特征组合与原型样本的相似性，以保证其有效地变为期望的类别；分母部分利用当前特征组合与查询样本之间的欧氏距离来表示产生反事实解释所需的变化，用于避免为了产生反事实解释盲目向原型样本靠拢的情况。此外，本文分别利用 $\exp(\cdot)$ 与 $\mathrm{sigmoid}(\cdot)$ 控制分子和分母对相似度或相对距离的要求。为了保证生成的反事实解释的有效性，分子部分设置为增速更快的指数函数。此外，LRS 的计算方式可以灵活更改以适应对反事实解释的不同要求。根据式(1)，可以得出图 1 中局部贪心树每条路径上的局部相对相似度，如表 3 所列。

表 3 局部贪心树中每条路径上的局部相对相似度

Table 3 Local LRS of each path in LGT

路径	相似性	代价	LRS
〈0,0,0〉	1.0000	0.6148	4.1882
〈0,0,1〉	0.9682	0.4833	4.2572
〈0,1,0〉	0.9466	0.4294	4.2542
〈0,1,1〉	0.8757	0.2000	4.3658
〈1,0,0〉	0.9911	0.5814	4.2006
〈1,0,1〉	0.9729	0.4400	4.3495
〈1,1,0〉	0.9128	0.3800	4.1949
〈1,1,1〉	0.8757	0.0000	4.8013

在上面的例子中，通过计算局部贪心树每条路径上的局部特征组合的相对相似度，可以得出在这 3 个特征子集上的最佳反事实解释路径是〈1,1,1〉，即查询样本 S 的局部特征子集。

3.2 生成链接树

生成链接树是局部贪心树的链接组合，通过生成链接树构造反事实解释的过程可以归纳为 3 个步骤：划分样本特征、构建局部贪心树、拼接特征选择路径并填充反事实解释。本文将待生成的反事实解释视为一个空白的特征空间，首先对原型样本 P 和查询样本 S 的特征进行划分，构建局部贪心树以计算局部特征组合的 LRS，并得到最优的特征选择路径，然后通过将局部特征组合进行拼接得到完整的反事实解释。

在筛选特征组合的过程中，同时考虑的特征越多，越能代表样本的全局特征，但随着特征维度的增加，贪心树的复杂程度会呈指数方式增加，容易出现内存溢出或生成反事实解释所需时间过长的问题，因此需要对特征进行划分。划分特征时，既要保证局部特征在一定程度上代表样本，又要考虑生成反事实解释所需的代价。实验过程显示，当局部特征少于 3 个时，筛选时易受极端特征值的影响；反之，当局部特征数达到乃至超过 10 个时，需要的计算资源、存储资源及生成反事实解释所需的时间都迅速增加，因此局部特征设置为 3～9 个较为合适。

以表 1 中的查询样本为例，图 2 给出了生成链接树的划分与链接过程。

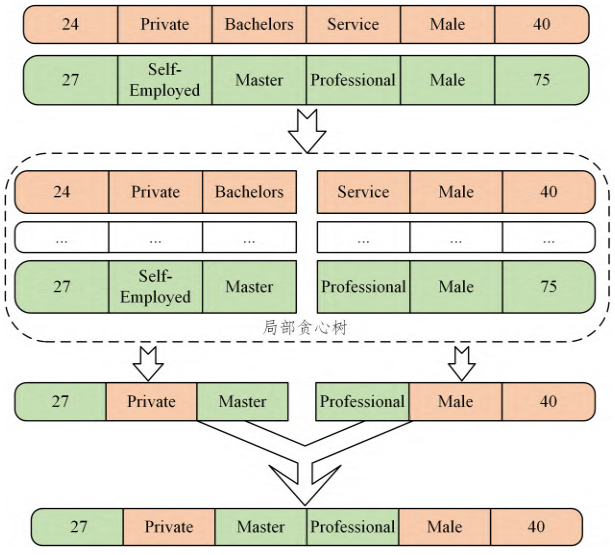


图 2 生成链接树的划分与链接

Fig. 2 Divisions and links of GLT

当特征数较少时，可以不进行划分操作，但为了展示全部的过程，图 2 将 6 个特征划分为 2 组，构建了 2 个如 3.1 节中介绍的局部贪心树，根据局部贪心策略筛选出局部最优特征组合。局部贪心树得到的局部最优特征组合路径分别为 $Path_1$=〈1,0,1〉与 $Path_2$=〈1,0,0〉。生成链接树将路径进行拼接，得到 $Path_{CE}$=〈1,0,1,1,0,0〉，然后根据路径从查询样本及原型样本中选取特征并将其填充到反事实解释的特征空间中，这样就产生了一个待验证的反事实解释。根据前文的思想，可以给出生成链接树 3 个阶段的形式化定义。

(1)特征划分，对于特征长度为 L 的数据样本 $S=\{f_1,\cdots,f_L\}$，为了减少生成反事实解释所需的资源并缩短时间，需将样本划分为 k 个长度为 l 的局部特征子集。根据经验，l 最恰当的取值区间为[3,9]，划分过程如式(2)所示：

$$lfs=\begin{cases}\{f_1,\cdots,f_l\}\\ \cdots \quad ,3\leqslant l\leqslant 9\\ \{f_{l\times(k-1)},\cdots,f_L\}\end{cases} \tag{2}$$

其中，f 表示样本中的特征，局部特征子集序列用 lfs 表示。

(2)构建局部贪心树并选择局部最优特征路径。长度为 l 的局部特征子集所构成的局部贪心树共有 2^l 条路径，筛选过程如算法 2 所示。

算法 2 局部特征选择算法

输入：原型特征子集 p，查询特征子集 s

输出：局部特征选择路径 $path_{CE}$

```
1. 根据算法 1 构建局部贪心树 lgt;
2. max_lrs, path_CE ← 0, []; /* 初始化局部相对相似度及特征选择路径 */
3. for path in lgt do:
4.    lrs ← calculate(path); /* 计算路径上的相对相似度 */
5.    if lrs > max_lrs do:
6.       max_lrs ← lrs;
7.       path_CE ← path;
8.    end if;
9. end for;
10. return path_CE.
```

算法 2 中的 calculate(·)表示计算该条路径上局部特征组合的相对相似度，计算式如式(1)所示。

(3)根据特征选择路径填充反事实解释。经过前两个阶段，可以得到 k 组局部特征选择路径，由特征选择路径生成反事实解释的过程如算法 3 所示。

算法 3 反事实解释填充算法

输入：原型样本 P，查询样本 S，特征选择路径序列 PathList

输出：反事实解释 CE

```
1. path, CE ← [], [];
2. /* 拼接特征选择路径 */
3. for p in PathList do:
4.    path ← path.push(p);
5. end for;
6. /* 填充反事实解释 */
7. for i ← 1 to Length(path) do:
8.    if path[i] == 0 do:
9.       CE.push(P[i]);
10.   else do:
```

```
11.    CE. push(S[i]);
12.  end if;
13. end for;
14. return CE.
```

算法 3 中的 push(·)操作表示向指定的列表中添加元素。通过拼接特征选择路径，控制最终填充到反事实解释中特征的来源，进而从原型样本与查询样本中选取特征进行填充。

完整的生成链接树的过程如图 3 所示，蓝色的燕尾箭头表示 GLT 的 3 个过程。对两个样本进行划分后，形成若干组局部特征子集，橙色节点和绿色节点分别表示原型样本和查询样本的特征。通过构建局部贪心树表示两个样本的局部特征组合，红色路径为局部贪心树筛选出的最优选择路径，灰绿色节点表示中间省略的若干特征。根据拼接后的局部路径可以选择出最终需要填充到反事实解释特征空间中的特征，虚线箭头表示最终的特征填充过程。

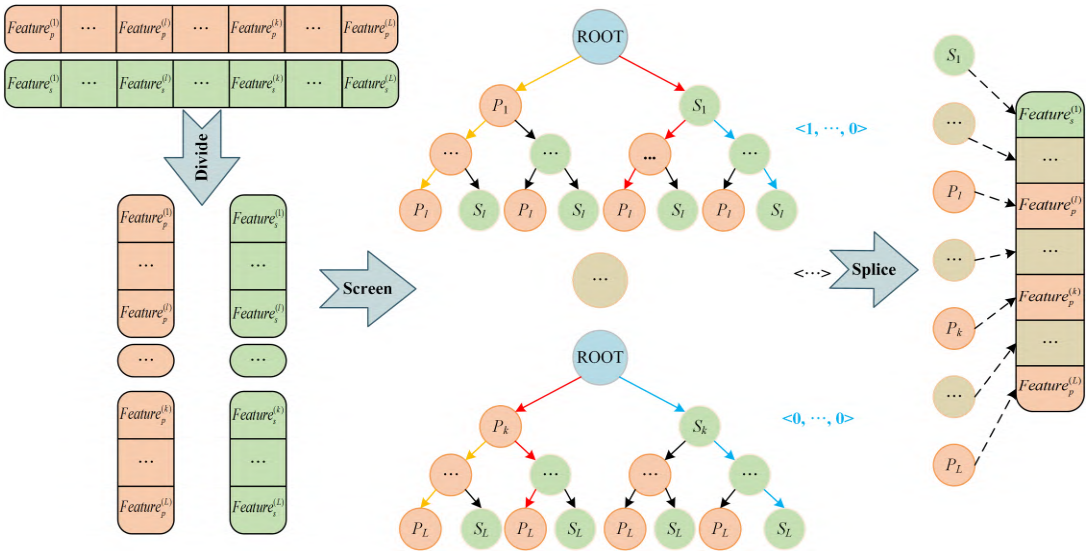


图 3　生成链接树的整体结构(电子版为彩图)

Fig. 3　Overall structure of GLT

完整的特征组合就是一个待验证的反事实解释，为了生成最终有效的反事实解释，需要依赖验证模型进行验证。若验证模型对反事实解释的分类结果与目标类别一致，则可以将之视为有效的反事实解释。

3.3　数据真实性

根据第 1 节中的分析，反事实解释要具备相对原始数据较高的忠实度，在实验中应该符合两条基本的性质。由于在生成的过程中需要依赖验证模型进行验证，筛选出验证模型分类为期望类别的反事实解释，因此在保证验证模型不变的情况下，反事实解释通常可以满足第一条性质。针对第二条性质的泛适应能力，本文提出数据真实性的评价指标，计算式如式(3)所示：

$$\frac{\sum_{i=1}^{n}(w_i \times p_i)}{\sum_{i=1}^{n} w_i} \tag{3}$$

其中，w_i 为第三方模型对应的权重，以其对原始数据分类的准确程度表示，以 F1-score 来度量；p_i 表示第三方模型对反事实解释数据真实性的认可程度，以第三方模型对反事实解释的分类准确程度表示，以 F1-score 来度量。F1-score 的计算式如式(4)所示：

$$F1\text{-}score = \frac{2 \times P \times R}{P + R} \times 100\% \tag{4}$$

其中，P 和 R 分别表示分类器的准确率(Precision)与召回率(Recall)。F1-score 通常初被用于评估分类器的性能，本文将其化用为特定权重下对反事实解释是否属于目标类别的认可程度。

4　实验结果与分析

为了验证 GLT 方法生成的反事实解释的数据真实性，本文对比了多个第三方模型对反事实解释的分类结果，并根据这些结果以及提出的数据真实性指标与基线模型进行了对比。有关本文的实验代码会在论文发表后公布在 GitHub[1)] 网站上。

4.1　实验数据集

参考了反事实解释的相关研究工作，本文选择了最热门的两个数据集：Adult income 数据集[24] 和 German credit 数据集[25]。

(1) Adult income[24]。该数据集包含基于 1994 年人口普查数据库的人口、教育等信息，可以从 UCI 机器学习库[26] 获得。本文选用了 Mothilal 等[7] 预处理过的版本，筛选了其中的 8 个特征。分类模型的任务是对每个样例的个人收入是否超过 50 000 美元进行分类。

(2) German credit[25]。该数据集中的信息是从银行获得的与个人贷款相关的信息，如个人当前特定银行的信用卡数、目前工作持续时间等信息等共 20 个特征。本文使用的是从 UCI 数据库直接获取的版本，未经过处理。分类模型的任务是根据用户的属性确定其信用类型，确定其是否具有信用风险。

1) https://github.com/NEU-DataMining/GLT

4.2 实验设置

本文随机选取查询样本，相同实验重复 10 组取均值，对样本中的分类特征进行目标编码[27]，分别使用 GLT 及基线方法生成反事实解释。以本文提出的数据真实性为评价指标，分别从为单个查询样本生成多个反事实解释、为多个查询样本生成多个反事实解释两个角度进行了多组实验。在这些实验中，Adult income 数据集中的特征被划分为 2 组，每组 4 个；German credit 数据集中的特征被划分为 4 组，每组 5 个。此外，为了探讨不同粒度的划分对数据真实性的影响，本文还在特征数较多的 German credit 数据集上进行了不同划分的实验。

本文选择了随机森林模型[28]作为反事实解释的验证模型，验证了生成的反事实解释是否为目标类别。此外，选择了常用的机器学习模型，如决策树[29]、朴素贝叶斯[30]等模型作为第三方模型，用于评价反事实解释的数据真实性。第三方模型的权重采用 10 折交叉验证[31]的方式来确定。

4.3 实验结果

如果将第三方模型对反事实解释的分类结果视为其对反事实解释数据真实性的认可程度，那么权重就可以被理解为相应“评委”的权威性。经过 10 折交叉验证，本文确定了 5 个第三方模型的权重，结果如表 4 所列。

表 4 第三方模型验证权重

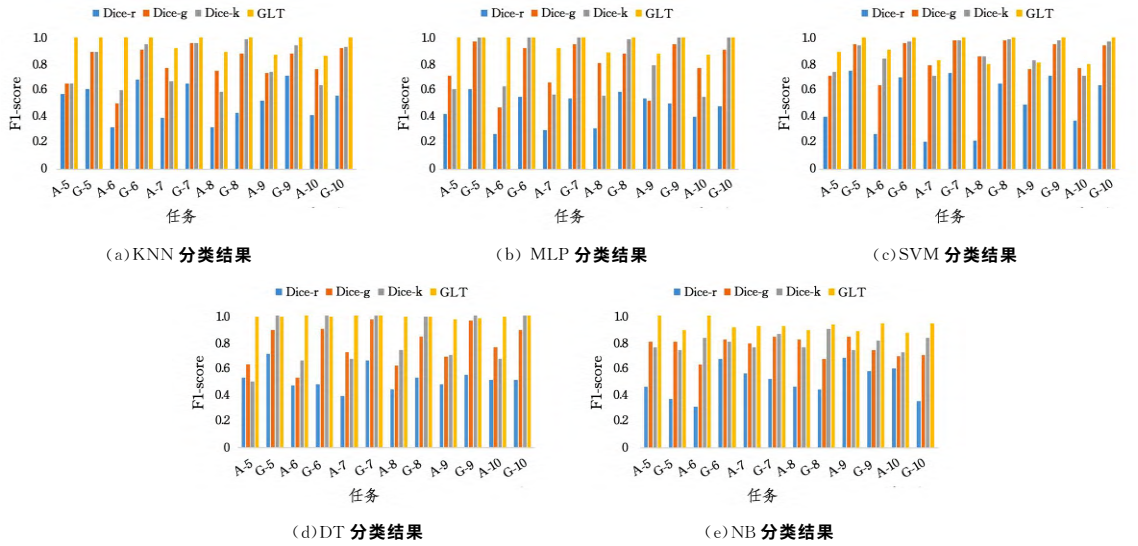
Table 4 Validation weights of the third-party model

模型	Adult 数据集	German 数据集
KNN	0.73	0.66
MLP	0.75	0.67
SVM	0.74	0.69
DT	0.69	0.65
NB	0.74	0.70

在确定了 5 个第三方模型的认证权威性之后，本文选择了 Mothilal 等[7]提出的 Dice 框架作为基线方法，从为单个查询样本生成多个反事实解释、为多个查询样本生成多个反事实解释的角度进行了对比实验，结果分别在 4.3.1 节与 4.3.2 节中进行介绍。

4.3.1 单一查询多个反事实解释

本文为单一查询样本生成了 5～10 个反事实解释，利用第三方模型验证其数据真实性，实验结果如图 4(a)—图 4(e)所示，给出了不同模型对反事实解释分类的 F1-score。



(a)KNN 分类结果 (b) MLP 分类结果 (c)SVM 分类结果

(d)DT 分类结果 (e)NB 分类结果

图 4 第三方模型的分类结果

Fig. 4 Classification results of third-party models

图 4 中，A 表示 Adult income 数据集，数字表示生成反事实解释的数目，例如“A-5”表示 Adult income 数据集中的一个查询样本生成 5 个反事实解释。Dice-r，Dice-g 及 Dice-k 分别表示 Dice 框架中提供的 3 种基于机器学习的方法。按照表 4 所列的权重，根据式(2)，得到不同方法生成反事实解释的数据真实性，如表 5 所列。

表 5 单一查询样本的多个反事实解释的数据真实性

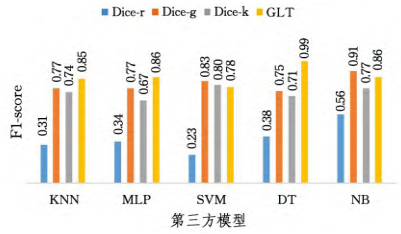
Table 5 Data fidelity of multiple counterfactual explanations for a single query sample

方法	A-5	G-5	A-6	G-6	A-7	G-7	A-8	G-8	A-9	G-9	A-10	G-10
Dice-r	0.47	0.61	0.33	0.62	0.37	0.62	0.35	0.53	0.54	0.61	0.46	0.51
Dice-g	0.70	0.90	0.55	0.90	0.75	0.94	0.77	0.85	0.71	0.89	0.75	0.87
Dice-k	0.65	0.91	0.71	0.94	0.68	0.96	0.70	0.97	0.76	0.94	0.66	0.94
GLT	**0.98**	**0.98**	**0.98**	**0.98**	**0.92**	**0.98**	**0.89**	**0.98**	**0.88**	**0.98**	**0.88**	**0.99**

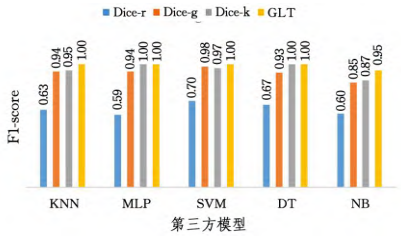
表 5 中，“A-5”“Dice-r”等与图 4 中的含义一致，加粗字体表示对应任务下最高的数据真实性。

4.3.2 多个查询多个反事实解释

本文从为多个查询样例生成反事实解释的角度进行了实验，分 10 次随机抽取了 5 个查询样本，为每个样本生成 10 个反事实解释，并验证了这些反事实解释的数据真实性。第三方模型对反事实解释进行分类的 F1-score 如图 5(a)、图 5(b)所示。



(a) Adult



(b) German

图 5　两个数据集上的分类结果

Fig. 5　Classification results on both datasets

经过加权计算后的数据真实性的结果如表 6 所列。

表 6　多样本多反事实解释的数据真实性

Table 6　Data fidelity of multiple counterfactual explanations for multiple samples

数据集	Dice-r	Dice-g	Dice-k	GLT
Adult	0.36	0.81	0.74	**0.87**
German	0.64	0.93	0.96	**0.99**

4.3.3　GLT 不同划分粒度

第一阶段样本特征划分是 GLT 的重要组成，本文在 German credit 数据集上进行了不同粒度的划分实验，即将 German credit 中的样本划分为不同大小的子集，讨论不同划分对反事实解释数据真实性的影响。不同划分粒度下的数据真实性如表 7 所列。其中，局部特征数表示划分后每个特征子集中的元素数，k 个特征子集中的元素数如式(5)所示：

$$\mathrm{card}(FSet_j)=\begin{cases} l, & j\neq k \\ L-l*(k-1), & j=k \end{cases} \tag{5}$$

其中，card(·)表示集合的基数，即其中的元素数，$FSet_j$ 表示第 j 个局部特征子集。由于在 German credit 数据集上生成的反事实解释数据的真实性都较高，为了方便比较，表 7 列出的数据真实性保留了小数点后的 5 位。此外，还增加了不同划分粒度下的反事实解释生成时间比较。

表 7　不同划分粒度下的数据真实性

Table 7　Data fidelity at different segmentation granularity

局部特征数	数据真实性	平均生成时间/s
3	0.98706	0.21
4	0.98764	0.23
5	0.98826	0.27
6	0.98746	0.33
7	0.98826	0.44
8	0.98783	0.65
9	0.98824	1.14

4.4　方法分析

本文提出了反事实解释生成方法生成链接树(GLT)，并与基线方法进行了生成反事实解释的数据真实性的对比实验。实验结果显示，GLT 方法生成的反事实解释具有更高的数据真实性。此外，随着生成反事实解释数目的增加，由 GLT 产生的反事实解释的数据真实性比基线方法更加稳定。

针对多个查询样例的生成反事实解释的实验同样表明，本文方法能生成数据真实性更高的反事实解释，这说明在多个第三方模型的验证下，反事实解释仍属于目标类别，有着较高的有效性且对原始数据有较高的忠实度，为反事实解释的应用提供了可能。

通过对不同划分粒度下生成反事实解释的数据真实性进行比较，结果表明，当局部特征子集中的元素数增加时，通常可以生成数据真实性更高的反事实解释，但生成时间会迅速增加。根据经验，局部特征子集中的元素数设置为 $l\in[3,9]$ 较为合适。

结束语　本文从反事实解释应用这样一个新研究视角出发，提出了验证反事实解释适用于应用任务的评价指标，即数据真实性。此外，本文还提出了一种用于生成高数据真实性反事实解释的生成链接树(GLT)方法，通过基于案例特征填充的方式生成反事实解释。经过实验验证，该方法生成的反事实解释具有更高的数据真实性，为反事实解释的应用提供了可能。

在未来的工作中，可以考虑从局部特征组合 LRS 的计算方法、原型样本的选取方法等方面考虑对 GLT 方法进行优化，以提高反事实解释的数据真实性及可解释性；也可以从局部贪心树的构建角度出发，讨论不同深度的局部贪心树对反事实解释的生成效率、可解释性及数据真实性的影响。此外，还可以从应用角度，将 GLT 生成的反事实解释应用于具体任务，这也是本文提出 GLT 方法的主要目的。

参 考 文 献

[1] GUNNING D, STEFIK M, CHOI J, et al. XAI—Explainable artificial intelligence[J]. Science Robotics, 2019, 4(37): eaay7120.

[2] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal of Law & Technology, 2017, 31: 841-887.

[3] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect[M]. Basic Books, 2018.

[4] MOLNAR C. Interpretable machine learning[M]. Lulu. com, 2020.

[5] VELMURUGAN M, OUYANG C, MOREIRA C, et al. Evaluating fidelity of explainable methods for predictive process analytics[C]// International Conference on Advanced Information Systems Engineering. Cham: Springer, 2021: 64-72.

[6] YUE Z, WANG T, SUN Q, et al. Counterfactual zero-shot and open-set visual recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15404-15414.

[7] MOTHILAL R K, SHARMA A, TAN C. Explaining machine learning classifiers through diverse counterfactual explanations [C]// Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020: 607-617.

[8] VERMA S, DICKERSON J, HINES K. Counterfactual explanations for machine learning: A review[J]. arXiv: 2010. 10596, 2020.
[9] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412. 6980, 2014.
[10] USTUN B, SPANGHER A, LIU Y. Actionable recourse in linear classification[C]// Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019: 10-19.
[11] POYIADZI R, SOKOL K, SANTOS-RODRIGUEZ R, et al. FACE: feasible and actionable counterfactual explanations [C]// Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 344-350.
[12] KEANE M T, SMYTH B. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai(xai)[C]// International Conference on Case-Based Reasoning. Cham: Springer, 2020: 163-178.
[13] GOYAL Y, WU Z, ERNST J, et al. Counterfactual visual explanations[C]// International Conference on Machine Learning. PMLR, 2019: 2376-2384.
[14] LOOVEREN A V, KLAISE J. Interpretable counterfactual explanations guided by prototypes[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2021: 650-665.
[15] SMYTH B, KEANE M T. A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations[J]. arXiv: 2101. 09056, 2021.
[16] KARIMI A H, SCHÖLKOPF B, VALERA I. Algorithmic recourse: from counterfactual explanations to interventions [C]// Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021: 353-362.
[17] GRATH R M, COSTABELLO L, VAN C L, et al. Interpretable credit application predictions with counterfactual explanations [J]. arXiv: 1811. 05245, 2018.
[18] RUSSELL C. Efficient search for diverse coherent explanations [C]// Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019: 20-28.
[19] MAHAJAN D, TAN C, SHARMA A. Preserving causal constraints in counterfactual explanations for machine learning classifiers[J]. arXiv: 1912. 03277, 2019.
[20] KARIMI A H, BARTHE G, BALLE B, et al. Model-agnostic counterfactual explanations for consequential decisions[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 895-905.
[21] KAUSHIK D, HOVY E, LIPTON Z C. Learning the difference that makes a difference with counterfactually-augmented data [J]. arXiv: 1909. 12434, 2019.
[22] ZHAO W, OYAMA S, KURIHARA M. Generating natural counterfactual visual explanations [C] // Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 5204-5205.
[23] GOMEZ O, HOLTER S, YUAN J, et al. Vice: Visual counterfactual explanations for machine learning models[C]// Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020: 531-535.
[24] KOHAVI R, BECKER B. Adult [EB/OL]. 2019. (1996-05-01). http://archive. ics. uci. edu/ml/datasets/Adult.
[25] HOFMANN H. Statlog(German Credit Data)[EB/OL]. (1994-11-17). http://archive. ics. uci. edu/ml/datasets/statlog+(germag+credit+data).
[26] ASUNCION A, NEWMAN D. UCI machine learning repository [EB/OL]. [2013-05-28]. http://archive. ics. uci. edu/ml.
[27] MICCI-BARRECA D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems[J]. ACM SIGKDD Explorations Newsletter, 2001, 3(1): 27-32.
[28] BIAU G, SCORNET E. A random forest guided tour[J]. Test, 2016, 25(2): 197-227.
[29] SAFAVIAN S R, LANDGREBE D. A survey of decision tree classifier methodology [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660-674.
[30] RISH I. An empirical study of the naive Bayes classifier[J]. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001, 3(22): 41-46.
[31] REFAEILZADEH P, TANG L, LIU H. Cross-validation[M]. Encyclopedia of Database Systems, 2009: 532-538.

WANG Ming, born in 1997, postgraduate, is a student member of China Computer Federation. His main research interests include interpretable machine learning and counterfactual explanation.

WANG Da-ling, born in 1962, Ph. D, professor, Ph. D supervisor, is a senior member of China Computer Federation. Her main research interests include social media processing, interpretable dialogue generation and sentiment analysis.

（责任编辑：喻藜）