计算机科学
COMPUTER SCIENCE

基于特征加权的反事实解释方法:以信贷风控场景为例

王宝财, 吴国伟

引用本文

王宝财, 吴国伟. 基于特征加权的反事实解释方法:以信贷风控场景为例[J]. 计算机科学, 2024, 51(12): 259-268.

WANG Baocai, WU Guowei. Feature-weighted Counterfactual Explanation Method:A Case Study in Credit Risk Control Scenarios [J]. Computer Science, 2024, 51(12): 259-268.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300047

基于特征加权的反事实解释方法:以信贷风控场景为例

王宝财　吴国伟
大连理工大学软件学院　辽宁 大连 116000
(wangbaocai.dlut@163.com)

摘　要　机器学习技术在金融领域的应用越来越多,为用户提供可解释的机器学习方法已成为一个重要的研究课题。近年来,反事实解释引起了广泛关注,它通过提供扰动向量来改变分类器得到的预测结果,从而提高机器学习模型的可解释性。但现有方法存在生成的反事实实用例缺乏可行性和可操作性的问题。文中提出了一种新的反事实解释框架,通过引入特征变量代价权重矩阵的概念,考虑不同特征变量改变的难易程度,使得反事实结果更符合实际情况并更具可行性。同时,通过专家预定义特征变量代价权重矩阵的方式,提出了一种计算特征变量代价权重的可行方法,并允许用户根据实际情况进行个性化调整。定义的目标函数综合考虑了特征加权距离、稀疏性和接近性3个指标,确保了反事实结果的可行性、简洁性和接近原始样本集的性质。采用遗传算法来求解问题,进而生成最佳的行动方案。通过对真实数据集进行实验,证实了所提方法相比现有的反事实方法能够生成可行性和可操作性更强的反事实用例。

关键词:机器学习;可解释性;反事实解释;权重矩阵;遗传算法
中图分类号　TP391

Feature-weighted Counterfactual Explanation Method:A Case Study in Credit Risk Control Scenarios

WANG Baocai and WU Guowei
School of Software Technology,Dalian University of Technology,Dalian,Liaoning 116000,China

Abstract　The application of machine learning technology in the financial field is becoming more and more prevalent,and providing interpretable machine learning methods to users has become an important research topic. In recent years,counterfactual explanation has attracted widespread attention,which improves the interpretability of machine learning models by providing perturbation vectors to change the predicted results obtained by classifiers. However,existing methods face feasibility and operability issues in generating counterfactual instances. This paper proposes a new counterfactual explanation framework that introduces the concept of feature-variable cost weight matrix,considering the ease of changing different feature variables to make the counterfactual results more realistic and feasible. At the same time,by predefining the feature-variable cost weight matrix by experts,a feasible method for calculating the cost weight of feature variables is proposed,allowing users to make personalized adjustments according to actual situations. The defined objective function comprehensively considers three indicators:feature-weighted distance,sparsity,and proximity,ensuring the feasibility,simplicity,and closeness to the original sample set of counterfactual results. Genetic algorithms are used to solve the problem and generate the optimal action plan. Through experiments on real datasets,it is confirmed that our method can generate feasible and actionable counterfactual instances compared to existing counterfactual methods.

Keywords　Machine learning,Interpretability,Counterfactual explanation,Weight matrix,Genetic algorithm

1　引言

随着金融业的不断发展,信用评估已成为金融机构关注的焦点之一。信用评估是评估客户是否会违约的关键步骤,被视为一个重要的非线性二分类问题。评估模型的输入包括类别数据(如人口统计学信息、个人信用记录)和数值数据(如收入、贷款总额),而输出是对用户违约概率的预测。信用评分的目的是将申请人分为两类:信誉良好的人和信誉不良的人[1]。通常认为信誉良好的人不会发生违约,信誉不良的人极有可能发生违约。根据中国银监会的数据,商业银行的不良贷款余额巨大[2],这凸显了信贷风险管理的重要性。商业银行和金融科技公司致力于构建强大的信用评估模型,以降低不良贷款的风险。传统的统计模型在处理信用评估时存在局限性,因此人工智能方法备受研究人员的青睐。支持向量机、深度神经网络和集成学习等技术被广泛应用于信贷风控领域,以提高模型的准确性和预测能力[3-8]。尽管这些基于

到稿日期:2024-03-07　返修日期:2024-06-16
通信作者:吴国伟(wgwdut@dlut.edu.cn)

机器学习的模型表现出色，但由于其固有的不透明性，无法向用户解释模型的预测结果，因此实施起来很困难，故需要构建更为准确、可靠、可解释的信用评估模型。

当前已有很多关于机器学习模型可解释性的研究，从解释方法的特点出发，可解释性方法主要分为3类：依赖于模型的方法、独立于模型的方法和因果解释方法。依赖于模型的方法指模型本身具有可解释性，在训练过程中就能理解模型的决策过程。这类方法通常采用简单易懂的模型，如朴素贝叶斯、线性回归、决策树、基于规则的模型[9-14]。独立于模型的方法则是在模型训练之后进行解释，主要通过解释方法或解释模型来揭示模型的工作机制和决策依据。根据解释的目的和对象不同，独立于模型的方法可分为全局解释和局部解释，分别对应全局解释方法和局部解释方法。独立于模型的方法的重点在于设计高保真的解释方法或构建高精度的解释模型[15-19]。

因果解释方法着重于解释模型在不同输入或参数下的决策，因此对用户来说是更友好的。Pearl提出了3个层次的可解释性：统计相关的解释、因果干预的解释和基于反事实的解释。其中，反事实解释被认为是实现最高层次可解释性的方法，因为它探讨了介入和关联问题[20]。反事实解释是一种基于实例的因果解释方法，它并没有明确回答模型为什么会作出这样的决策，主要侧重的是“改变什么条件，模型的决策结果会发生改变”。例如，在一个信用贷款案例中，如果一个人的贷款申请被拒绝，那么他想知道的不仅是为什么会被拒绝，还想了解改变什么条件后申请才会被批准，这对申请人来说是更有效的。总的来说，反事实解释被认为是实现最高层次可解释性的方法，因为它关注于因果关系和实例级别的解释，而不仅仅是模型的决策结果。

现有的反事实解释的研究主要集中在提高反事实的质量上，很少关注反事实在特定领域的可行性，现有研究还集中在增加所生成反事实解释的多样性或合理性上。在Rodriguez等的研究中，提出的模型利用多样性强化损失来在学习过程中发现不同的解释，通过在分离的潜在空间中学习扰动[21]。Delser等提出通过平衡可信度、变化强度和对抗能力这3个目标来生成反事实解释[22]。Poyiadzi等采用了用户标注的方式判断所生成的解释是否可行[23]。Kanamori等试图通过确定反事实例子是否为离群值来生成属于数据集分布的解释[24]。然而，现有方法没有从工程应用的角度和结合特定领域知识进行可解释的研究。特定领域的某些特征变量是无法改变的或者很难改变的，例如年龄、性别、学历等，然而传统的反事实方法未考虑到特征变量改变的代价，例如信贷风控领域。一个反事实用例是：该用户的年龄减少10岁即可获得贷款或者改变该用户的性别可以获得贷款。但年龄是不能减少的，性别是不能改变的。再例如：改变年收入和学历特征变量的代价是不同的，年收入从10万提升到11万也许很容易，然而学历却很难从专科提升到本科；或者，年收入从10万提升到11万很容易，然而提升到20万却很难。因此，生成反事实解释时要考虑到不同特征变量值改变的代价。

另外，现有的反事实方法没有考虑到用户的定制化需求，对于属性相同的实例，输出的反事实解释都是相同的。例如，如果被拒贷的用户A和被拒贷的用户B的信息是一样的，现有反事实方法对于用户A和用户B的反事实解释结果是一样的，并没有考虑到用户的个体差异。对于不同的用户来说，同一个特征变量的改变代价是不同的。例如，对于用户A来说，年收入从10万提升到11万很容易，然而对于用户B来说，年收入很难从10万提升到11万；对于用户A来说，最近一个月消费从1000元提升到3000元很容易，然而对于用户B来说，最近一个月消费很难从1000元提升到3000元。因此，生成反事实解释时要考虑用户的定制化需求。

为了解决上述问题，本文提出了特征加权的反事实方法，并以信贷风控场景为例进行了说明和实验，证明了本文方法的可行性。本文方法的整体框架如图1所示。本文的主要贡献有：

1）提出了特征变量代价权重矩阵的概念，确保了结果的可行性。

2）给出专家预定义特征变量代价权重矩阵的方法，使得本文方法具有可操作性。

3）在专家预定义特征变量代价权重矩阵的基础上，用户可以自定义特征变量代价权重矩阵，从而满足不同用户的个性化需求。

4）权衡多个指标，综合考虑用户和企业的需求，定义了一个新的目标函数，并使用遗传算法求解，生成行动方案。

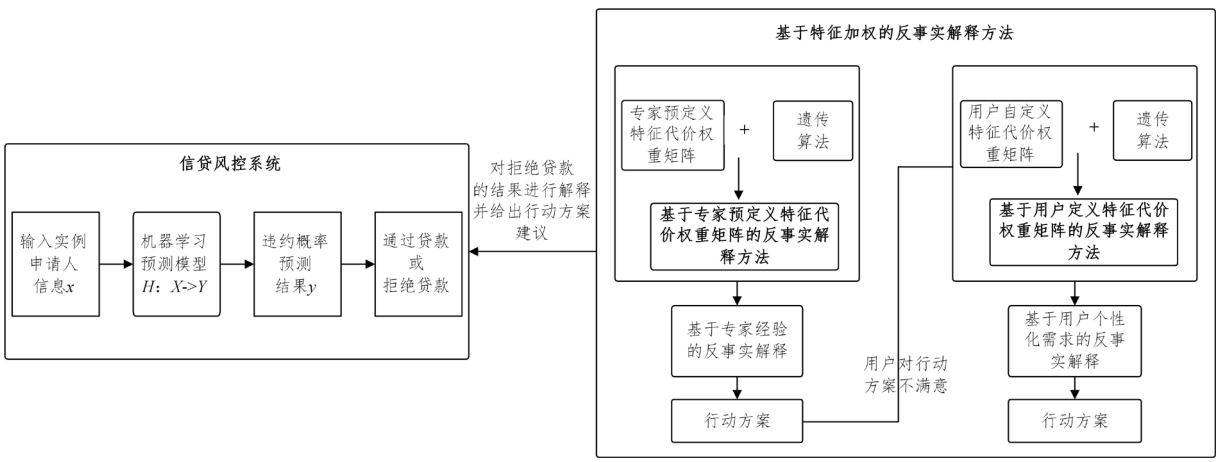

图1 基于特征加权的反事实解释方法的框架图

Fig. 1 Frame diagram of counterfactual interpretation method based on weighed features

2 准备工作及问题定义

2.1 符号定义

每个样本$(\boldsymbol{x},y)$包含了用户的输入信息和相应的真实标签。通常，输入信息 x 由人口统计学信息、个人信用记录和个人财务信息等组成。真实标签 y 属于$\{0,1\}$，其中“1”表示信誉不良的人，会发生违约，“0”表示信誉良好的人，不会发生违约。因此，本文将二分类问题视为预测任务，这对于反事实解释来说是足够的。对于多类分类问题，可以将其简化为目标类别和其他类别之间的二分类问题。本文将输入和输出分别表示为 $\boldsymbol{X}=X_1\times\cdots\times X_D\subseteq R^D$ 和 $\boldsymbol{Y}=\{0,1\}$。向量 $\boldsymbol{x}=(x_1,\cdots,x_D)\in\boldsymbol{X}$ 表示一个实例，信用评估建模的目标是根据由多个用户样本构成的数据集，构建并训练评估模型，从而预测用户的违约概率，进而判断是否给用户发放贷款。本文使用一个分类器 $H:\boldsymbol{X}\to\boldsymbol{Y}$ 来表示该评估模型。通常情况下，H 是一个非线性函数。

2.2 行动和行动集

对于分类器 $H:\boldsymbol{X}\to\boldsymbol{Y}$，以及满足 $H(x)=1$ 的实例 $\boldsymbol{x}\in\boldsymbol{X}$，本文定义一个扰动向量 $\boldsymbol{a}\in R^D$ 来表示采取的行动，使得 $H(\boldsymbol{x}+\boldsymbol{a})=0$。行动集合 $\boldsymbol{A}=A_1\times\cdots\times A_D$ 是一个有限的可行动集合，其中 $0\in A_d$ 并且对于每个 $\boldsymbol{d}\in[\boldsymbol{D}]$，有 $A_d\subseteq\{\boldsymbol{a}\in R\mid x_d+a_d\in X_d\}$。

可以根据分类器 H 的类型和特征变量 $d\in[\boldsymbol{D}]$ 自动确定每个 A_d[25-26]。例如，如果 x_d 表示“年龄”这个特征，那么对于任何 $a_d\in A_d$，都有 $a_d\in N\cup\{0\}$。如果 x_d 是一个不可变特征（如性别），那么 $A_d=\{0\}$。

2.3 分类器

本文分类器表示为 $H:\boldsymbol{X}\to\boldsymbol{Y}$，$\boldsymbol{X}$ 为输入，$\boldsymbol{Y}$ 为输出。本文使用人工神经网络（ANN）模型作为分类器，ANN 的原理示意图如图 2 所示。

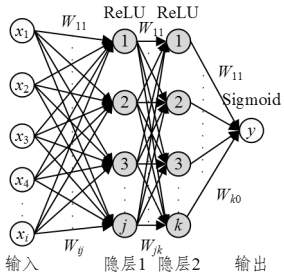

图 2 ANN 模型

Fig. 2 ANN model

ANN 模型由一系列相互连接的节点（神经元）组成，这些节点模仿了生物神经元之间的连接方式。每个神经元都接收来自其他神经元的输入，并通过激活函数处理这些输入，最终生成相应的输出。

ANN 模型由多个层次构成，典型的包括输入层、隐藏层和输出层。输入层接收原始数据，隐藏层处理数据并进行特征提取，而输出层生成模型的预测结果。信息通过神经元之间的连接进行传递，每个连接都具有一个相关的权重。ANN 模型的训练过程通常包括反向传播算法，该算法根据模型预测的误差来调整连接权重，以优化模型的性能。

2.4 代价函数

代价函数的定义如式(1)所示，其中，$x\in X$ 表示样本集中的一个实例，$\tilde{x}$表示反事实的一个实例；$H(\cdot)$是已经训练好的二分类预测模型，$H(\tilde{x})$是实例$\tilde{x}$在预测模型 $H(\cdot)$下的输出结果；代价函数值越低，表示相应反事实用例的效果越好。

$$cost(\tilde{x}\mid x)=\begin{cases}\lambda_1 dist(x,\tilde{x})+\lambda_2 lof(\tilde{x})+\lambda_3 spr(\tilde{x}), & \\ \qquad\qquad H(\tilde{x})\neq H(x) & \\ +\infty \qquad H(\tilde{x})=H(x)\end{cases}\tag{1}$$

其中，$dist(x,\tilde{x})$为距离函数，表示反事实用例的合理性，如式(2)所示；$lof(\tilde{x})$为反事实用例的接近性，即尽量接近原始样本集，如式(4)－式(6)所示；$spr(\tilde{x})$表示反事实用例的稀疏性，即特征变量值发生改变的数量，如式(7)所示。λ 为权重系数，用来权衡 3 个优化目标的权重，满足约束：$0\leqslant\lambda\leqslant1$ 且 $\lambda_1+\lambda_2+\lambda_3=1$。当 $H(\tilde{x})=H(x)$时，表示$\tilde{x}$不是可行解，因此给代价函数赋值$+\infty$。代价函数由 3 个子项加权组成，接下来将详细解释每个子项的具体含义。

2.4.1 特征加权距离函数

欧氏距离（Euclidean Distance）是在欧几里得空间中计算两个点之间的直线距离。在 n 维空间中，欧氏距离的公式可以表示为两个向量之间的欧几里得范数，即 $d(x,\tilde{x})=\sqrt{\sum_{i=1}^{j}(a_i-b_i)^2}$，其中，$a_i$和$b_i$是两个向量在第 i 维的坐标值。这个公式衡量了两个点之间的直线距离，是最常用的距离度量方法之一。

式(1)中的第一项 $dist(x,\tilde{x})$为特征的加权距离函数，是结合用户设定的特征权重和欧氏距离得到的，表示生成的反事实用例的合理性，如式(2)所示：

$$dist(x,\tilde{x})=\sqrt{\sum_{i=1}^{j}w_i(\Delta x_i)(x_i-\tilde{x}_i)^2}\tag{2}$$

其中，$\Delta x_i=\tilde{x}_i-x_i$ 表示特征变量的改变量；$w_i(\Delta x_i)\in\boldsymbol{W}$ 为用户对特征变量 x_i 值改变所付出的代价而设置的权重，$w_i(\Delta x_i)$越大，表示特征变量 x_i 值改变所付出的代价越大，特征代价权重矩阵 $\boldsymbol{W}$ 的定义如式(3)所示。该距离函数考虑到了不同特征变量值改变代价是不同的，例如改变年龄和年收入的代价是不同的，年龄增长 1 岁和增长 10 岁的代价也是不同的。矩阵 $\boldsymbol{W}$ 的每一行表示不同的特征变量对应的代价权重，每一列表示特征变量变化的区间对应的代价权重。Δd_i 表示特征变量 x_i 变化的区间步长。

$$\boldsymbol{W}=\begin{bmatrix}w_1(-\infty,-n\times\Delta d_1) & w_1(-n\times\Delta d_1,-(n-1)\times\Delta d_1) & \cdots & w_1((-\Delta d_1,0)) & w_1(0,\Delta d_1) & w_1(\Delta d_1,2\times\Delta d_1) & \cdots & w_1(n\times\Delta d_1,+\infty)\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\ w_j(-\infty,-n\times\Delta d_j) & w_j(-n\times\Delta d_j,-(n-1)\times\Delta d_j) & \cdots & w_j(-\Delta d_j,0) & w_j(0,\Delta d_j) & w_j(\Delta d_j,2\times\Delta d_j) & \cdots & w_j(n\times\Delta d_j,+\infty)\end{bmatrix}\tag{3}$$

其中，$w_i(\Delta x_i)=w_i((k-1)\times\Delta d_i, k\times\Delta d_i)$，if $\Delta x_i\in((k-1)\times\Delta d_i, k\times\Delta d_i]$。

2.4.2 离群因子(LOF)

式(1)的第二项表示生成的反事实用例的接近性，即生成的反事实用例要尽量接近原始的样本集，通过将反事实建立在输入数据集分布中，可以实现反事实的可操作性和现实性[27]。通过最小化给定实例的 lof 分数来考虑反事实的可行性，这是由 Breunig 等首次提出的[28]。反事实必须在可行性范围内，这意味着反事实应该公平地代表真实数据集的分布。否则，反事实可能只是一个离群值或一种在现实生活中不可能的情况。本研究使用了 lof，与传统的离群检测方法不同，传统的离群检测方法只是简单地将一个实例标记为是否是给定数据集中的一个离群值，而 lof 显示了实例在其局部区域中作为离群值的程度。此外，反事实的一个用途是检测离群值和所谓的错误，但因为本文中的反事实旨在提供一个可行的、可操作的解释，所以生成的反事实应该接近原始数据的分布。如果生成的反事实是一个离群点，那么预测模型基于离群点计算到的预测结果将是不准确的。以本文信贷场景为例，将会增加违约风险，因此生成的反事实必须接近原始数据的分布。lof 基于局部密度计算分数，局部性由 k 个邻居定义，为了考虑数据集中存在的更广泛的局部特征，本研究将 k 设置为 20。为了获得式(6)中的 lof，计算了可达距离(见式(4))和局部可达密度(见式(5))。

用局部异常因子 lof 来表示反事实用例在其局部区域内的异常值程度，lof 越小越好；lof 根据局部密度计算其得分，局部性由其 k 个邻居定义；可达距离的定义如式(4)所示，局部可达密度如式(5)所示；lof 的计算式如式(6)所示。通常情况下，lof 值大于 1，则表示该实例很有可能是异常点，因为相比其相邻实例其局部密度更小。

$$rd_k(o,p)=\max\{k\text{-}distance(o),d(o,p)\} \tag{4}$$

$$lrd_k(p)=\frac{|N_k(p)|}{\sum_{o\in N_k(p)}rd_k(p,o)} \tag{5}$$

$$lof_k(p)=\frac{\sum_{o\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \tag{6}$$

其中，$k\text{-}distance$ 表示点 o 和它 k 个邻居的平均距离，$d(o,p)$ 表示点 o 和点 p 的空间距离，$rd_k(o,p)$ 为点 p 到点 o 的第 k 可达距离，即在点 p 与点 o 的距离、距离点 o 最近的 k 个点中取距离较大的一个。$N_k(p)$ 表示 p 附近距离 $k\text{-}distance$ 以内的邻居的数量，$lrd_k(p)$ 表示点 p 的局部可达密度，是基于点 p 的 k 个最近邻的平均可达距离的倒数，距离越大，密度越小。

2.4.3 特征改变数量

式(1)的第三项表示生成的反事实用例的稀疏性，即特征变量值发生改变的数量；改变的特征变量数量越少，反事实用例就越简洁，越容易理解。

$$spr(\tilde{x})=\text{特征改变的数量} \tag{7}$$

2.5 问题定义

本文的目标是找到一个行动 $a\in A$，使得代价 $cost(a|x)$ 最小化，从而生成可行的反事实用例。这个问题的定义如下。

定义 1 给定一个分类器 $H:\boldsymbol{X}\rightarrow\boldsymbol{Y}$，输入实例 $\boldsymbol{x}\in\boldsymbol{X}$，使得 $H(\boldsymbol{x})=1$，给定特征代价权重矩阵 $\boldsymbol{W}$、权重系数 λ 和实例集合 $\boldsymbol{X}$，其中 $0\leqslant\lambda\leqslant1$ 且 $\lambda_1+\lambda_2+\lambda_3=1$。我们的目标是找到一个行动 $\boldsymbol{a}^*\in\boldsymbol{A}$，它是式(8)优化问题的最优解。

$$\underset{a\in A}{\text{minimize}}\ cost(\boldsymbol{a}|\boldsymbol{x})\quad \text{s.t. } \boldsymbol{H}(\boldsymbol{x}+\boldsymbol{a})=0 \tag{8}$$

3 基于遗传算法和特征加权的反事实方法

3.1 遗传算法

本文采用遗传算法(GA)来求解上述问题，这种算法能够生成近最优解。GA 是一种无梯度方法，适用于任何模型，具有模型无关性。相比其他优化技术，GA 既采用探索性方法又采用穷举法，避免陷入局部最小值，还能处理多目标优化问题。此外，Dandl 等的研究也使用了类似的遗传算法来生成解释模型的规则或反事实用例[29-30]。为将改变特征的代价纳入解释过程，本文提出了特征代价权重矩阵概念，用来加权不同特征改变所需要的代价，并结合遗传算法求解反事实用例，以更好地解释模型。

遗传算法从一个包含一定数量基因的染色体集开始，以“进化”的方式对集中的种群进行操作，以搜索最佳染色体。每个染色体代表一个解决方案，因此一组染色体-种群表明了一组可能的解决方案。在这种情况下，每个染色体表示一个可能的反事实候选，解决方案根据上面声明的适应度函数进行评估。遗传算法的操作包括选择、突变和交叉，以生成种群的一代，这使得算法同时使用开发和探索方法进行搜索。

3.2 基于遗传算法和特征加权的反事实方法

生成反事实用例可以被定义为一个优化问题，通过在特征空间中寻找近似于原始特征但产生不同决策结果的扰动，来找到一个行动方案，即反事实用例。本文提出了一种基于遗传算法的特征加权的反事实用例生成方法，将特征改变的代价权重加入到评价函数中，使遗传算法生成反事实用例。并以信贷风控系统的决策结果的反事实解释为案例，给出用户如何改进可以获得贷款的解释。

使用多目标优化方法，可以考虑多个目标的权衡，可以生成更好的反事实用例，从而提高所生成的反事实用例的可理解性。首先，自定义特征变量改变的代价权重矩阵 $\boldsymbol{W}$，权重 w_{ij} 的区间为[0,1]，当某个特征变量不允许改变时，该特征变量的代价权重设置为 $+\infty$，权重越大表示该特征对于该用户来说改变的代价越高，在生成反事实用例时优先考虑改变权重小的特征。其次，要考虑生成的样本的数据分布和改变的特征的数量。生成的反事实用例优先使用特征改变代价小的特征，并且改变的特征数量越少，越容易被用户理解和接受。如果生成的反事实用例使用了大量的不相关的特征，对于用户来说是不易于理解的。除了特征代价这个指标外，还需要考虑稀疏性和与原始样本的接近性，即合理性、稀疏性和接近性是评价一个好的反事实样本的关键指标。本文中，使用这 3 个指标的加权值作为评价函数，如式(1)所示。

本文的目的是找到满足多目标的、用户满意的、实际可操作的反事实用例。本文使用遗传算法来寻找最优解，流程图如图 3 所示。

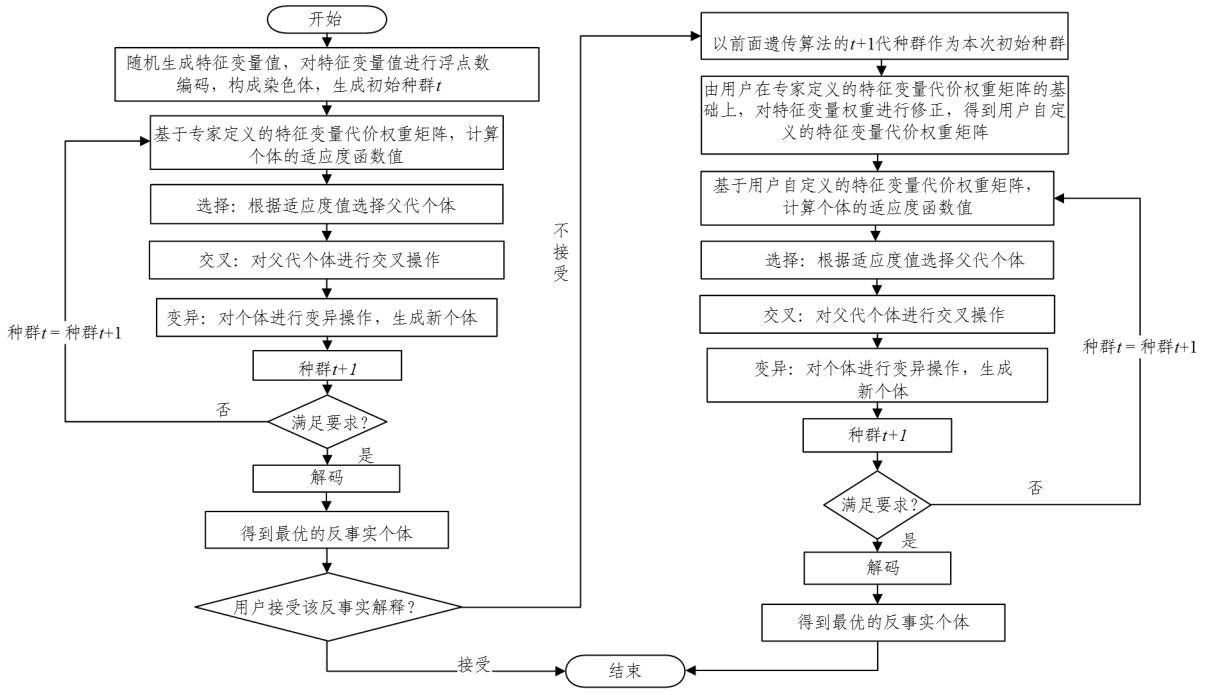

图 3　基于遗传算法和特征加权的反事实算法流程

Fig. 3　Process of counterfactual algorithm based on genetic algorithm and weighted features

在种群初始化阶段，遗传算法生成一个初始种群开始搜索。初始种群是由在遗传算法操作之前声明的一个特定数量的基因随机生成的染色体组成。大量的种群通过扩大搜索空间来引入更多的多样性，但收敛速度较慢，而较小的种群收敛速度更快，但问题的搜索空间可能不足以找到近最优解。图 4 给出了实验中的染色体编码方案。每个基因表示反事实的特征值，并被设计为在 0 和 1 之间的实数值。染色体的基因是反事实用例的特征值，因此染色体的长度等于特征的数量。在本文的案例中，染色体的长度为 28，与预测模型的输入特征数量相同。在实验中，种群大小设置为 1 000，并在过程中保持不变。为了初始化种群，染色体通过在 0～1 的范围内随机设置基因（即特征）的新值来生成。在设计了种群初始化后，交叉和突变运算符被用来保持种群的多样性。

在通过适应度函数评估染色体后，它会使用遗传算法运算符进化到下一代。在选择过程中，使用轮盘赌选择法、锦标赛选择法和排名选择法等方法根据染色体的适应度选择染色体。例如，在锦标赛方法中，会随机选择几个染色体进行几个锦标赛，并根据每个染色体的适应度值选择获胜者，这与锦标赛比赛相似。锦标赛大小表示锦标赛中染色体的数量（即参赛者），随后举行几场比赛以选择具有最高适应度值的染色体。在实验中，使用了锦标赛方法，因为它可以轻松实现和调整。实验中锦标赛的大小设置为 4。

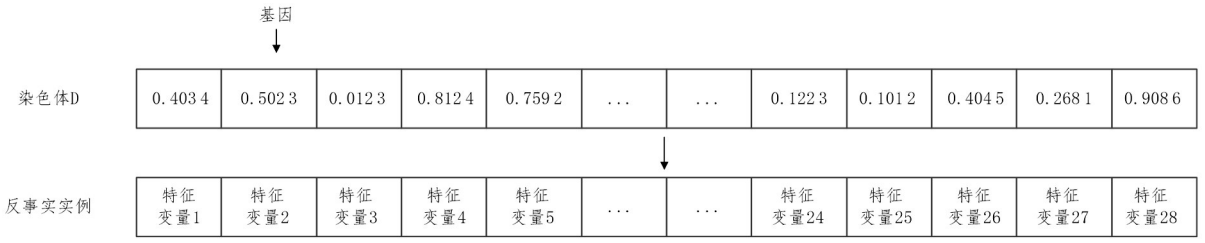

图 4　染色体编码示意图

Fig. 4　Schematic diagram of chromosome encoding

在交叉阶段，选定的染色体成对地去生成新的染色体。它是通过对两个染色体（即父代染色体）进行操作来创建新的染色体（后代）。可以使用 k 点交叉法来进行操作，该方法随机选择 k 个点固定在父代染色体上，然后交换这些点之间的基因。本文采用了两点交叉，在实验中交叉概率被设置为 0.7，并且在操作过程中保持不变。在没有交叉操作的情况下，相同的染色体会被传递到下一代，当后代适应度值超过父母时，后代可以取代种群。

在变异阶段，算法以给定概率随机变异染色体中的基因。在本实验中，通过随机重置基因的值来对染色体进行变异。这是遗传算法操作中的一种探索性方法，可以防止陷入局部最优解。高变异概率可能延迟收敛，但可以防止陷入局部最优解。低变异概率可能导致过早收敛。该过程返回适应度评估步骤并重复，直到满足遗传算法的停止标准。在实验中，变异概率被设置为 0.1，在操作过程中保持不变。

算法的步骤如下：

步骤 1　由领域专家预定义特征变量代价权重矩阵 $\boldsymbol{W}_c$。

步骤 2　执行遗传算法，根据步骤 1 的特征变量代价权

重矩阵 $\boldsymbol{W}_e$ 计算各个个体的代价函数值，直到得到满足条件的最优个体为止，输出反事实用例，作为当前结果的解释，如果用户接受该解释，则停止，否则执行步骤 3。

步骤 3　在专家预定义的特征变量代价权重矩阵 $\boldsymbol{W}_e$ 的基础上由用户对特征变量代价权重进行调整，得到用户自定义的特征变量代价权重矩阵 $\boldsymbol{W}_u$。

步骤 4　执行遗传算法，根据步骤 3 的特征变量代价权重矩阵 $\boldsymbol{W}_u$ 计算各个个体的代价函数值，直到得到满足条件的最优个体为止，输出反事实用例，作为当前结果的解释，算法停止。

4　实验过程及结果分析

本文采用的数据集是来自某移动公司提供的真实客户样本的用户信用数据集。该数据集共包含 50000 个用户的样本，每个样本有 28 个属性。这些属性涵盖了用户的基本信息、通信支出、历史消费记录、用户话费敏感度、观看电影频率、去商场次数以及网上购物次数等多种类型的用户属性，具体如表 1 所列。

本文采用 ANN 模型作为是否拒绝客户贷款的预测模型，ANN 模型的具体参数如表 2 所列。本文采用的遗传算法的具体参数如表 3 所列。特征变量代价权重矩阵定义如表 4 所列。

首先定义每个特征变量的改变步长，然后基于步长定义改变的各个区间特征变量的代价权重，代价权重为[0,1]，如果特征变量在某个区间是不可改变的，则代价权重定义为 $+\infty$。

表 1　特征变量描述

Table 1　Description of feature variables

特征变量	特征变量名称	特征变量	特征变量名称
v_1	用户实名制是否通过核实	v_{15}	是否是经常逛商场的人
v_2	用户年龄	v_{16}	近 3 个月月均商场出现次数
v_3	是否是大学生客户	v_{17}	当月是否逛过福州仓山万达
v_4	是否是黑名单客户	v_{18}	当月是否到过福州山姆会员店
v_5	是否是 4G 不健康客户	v_{19}	当月是否看电影
v_6	用户网龄(月)	v_{20}	当月是否游览景点
v_7	用户最近一次缴费距今时长(月)	v_{21}	当月是否到体育场馆消费
v_8	缴费用户最近一次缴费金额(元)	v_{22}	当月网购类应用使用次数
v_9	用户近 6 个月平均消费值(元)	v_{23}	当月物流快递类应用使用次数
v_{10}	用户账单当月总费用(元)	v_{24}	当月金融理财类应用使用总次数
v_{11}	用户当月账户余额(元)	v_{25}	当月视频播放类应用使用次数
v_{12}	缴费用户当前是否欠费缴费	v_{26}	当月飞机类应用使用次数
v_{13}	用户话费敏感度	v_{27}	当月火车类应用使用次数
v_{14}	当月通话交往圈人数	v_{28}	当月旅游资讯类应用使用次数

表 2　ANN 模型参数

Table 2　Parameters of ANN model

参数	参数值
隐藏层数量	2
隐藏层神经元数量	12,6
转换函数	ReLU
输出层转换函数	Sigmoid
优化器	SGD
学习率	0.01
最大迭代次数	10000

表 3　遗传算法参数

Table 3　Parameters of genetic algorithm

参数	详情
适应度函数	$fitness(\tilde{x}\mid x)=-cost(\tilde{x}\mid x)$ $\lambda_1=0.7,\lambda_2=0.2,\lambda_3=0.1$
初始种群数量	1000
最大迭代次数	100
选择	Tournament(size=4)
交叉率	0.7(two-point crossover)
变异率	0.1

表 4　专家预定义的特征代价权重矩阵

Table 4　Feature cost weight matrix predefined by experts

特征变量	特征改变步长	特征改变代价权重							
	Δd	$(-\infty,-3\Delta d]$	$(-3\Delta d,-2\Delta d]$	$(-2\Delta d,-\Delta d]$	$(-\Delta d,0]$	$(0,\Delta d]$	$(\Delta d,2\Delta d]$	$(2\Delta d,3\Delta d]$	$(3\Delta d,+\infty)$
v_1	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0	0	0	0
v_2	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.1	1	$+\infty$	$+\infty$
v_3	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
v_4	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
v_5	1	1	1	1	1	1	1	1	1
v_6	6	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.6	0.8	1	$+\infty$
v_7	1	0	0	0	0	0.1	0.6	0.7	1
v_8	300	0	0	0	0	0.1	0.15	0.2	1
v_9	100	1	0.8	0.5	0	0.5	0.8	1	1
v_{10}	1000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{11}	1000	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.1	0.2	0.3	1
v_{12}	1	0	0	0	0	0.1	0.1	0.1	0.1
v_{13}	1	1	0.7	0.6	0.5	0.5	0.6	0.7	1
v_{14}	20	0.25	0.2	0.15	0.1	0.1	0.15	0.2	0.25
v_{15}	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
v_{16}	10	0.1	0.1	0.1	0.1	0.1	0.2	0.3	1

（续表）

特征变量	特征改变步长	特征改变代价权重							
	Δd	$(-\infty,-3\Delta d]$	$(-3\Delta d,-2\Delta d]$	$(-2\Delta d,-\Delta d]$	$(-\Delta d,0]$	$(0,\Delta d]$	$(\Delta d,2\Delta d]$	$(2\Delta d,3\Delta d]$	$(3\Delta d,+\infty)$
v_{17}	1	0	0	0	0	0	0	0	0
v_{18}	1	0	0	0	0	0	0	0	0
v_{19}	1	0	0	0	0	0	0	0	0
v_{20}	1	0	0	0	0	0	0	0	0
v_{21}	1	0	0	0	0	0	0	0	0
v_{22}	1 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{23}	500	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{24}	1 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{25}	10 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{26}	500	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{27}	50	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{28}	100	1	0.3	0.2	0.1	0.1	0.2	0.3	1

以某个拒贷用户为例，该用户的原始特征变量值及各反事实方法的结果如表5所列。

表5 实验结果的对比

Tabel 5 Comparison of experimental results

特征变量	拒贷实例 x	TCE	ECCE	UCCE	特征变量	拒贷实例 x	*TCE*	*ECCE*	*UCCE*
v_1	1	1	1	1	v_{15}	0	0	0	0
v_2	69	**15**	69	69	v_{16}	0	0	0	0
v_3	0	0	0	0	v_{17}	0	0	0	0
v_4	0	0	0	0	v_{18}	0	0	0	0
v_5	0	0	0	0	v_{19}	0	0	0	0
v_6	1	1	1	1	v_{20}	0	0	0	0
v_7	0	0	0	0	v_{21}	0	0	0	0
v_8	0	9.8	0	0	v_{22}	20	20	20	20
v_9	41.66	41.66	**195.45**	41.66	v_{23}	0	0	0	0
v_{10}	41.66	41.66	41.66	41.66	v_{24}	6	6	6	6
v_{11}	50	50	50	50	v_{25}	42	42	42	42
v_{12}	0	0	0	0	v_{26}	0	0	0	0
v_{13}	5	5	5	5	v_{27}	0	0	0	0
v_{14}	1	1	1	**58**	v_{28}	0	0	0	0

具体步骤如下：

步骤1 由专家预定义特征代价权重矩阵，权重值代表特征变量值改变的代价大小，如表4所列。

步骤2 按照图3的流程执行遗传算法，首先随机生成遗传算法的初始种群，对特征变量值进行归一化处理，将特征变量值缩放为[0,1]之间的实数，按照图4所示进行基因浮点编码。根据式(1)计算种群中个体的适应度函数值。然后执行选择、交叉、变异操作，直到满足停止条件为止，根据适应度函数值得到最优的反事实个体。根据图3所示的流程，如果用户对该反事实解释满意，则停止，否则进入步骤3。

步骤3 在专家预定义的特征代价权重矩阵的基础上由用户对特征代价权重进行调整，用户只需根据反事实结果和专家预定义的权重矩阵对个别特征的代价权重进行调整。如表6所列，用户对特征变量 v_9 和 v_{14} 的代价权重进行了调整。

步骤4 将前面遗传算法最后一代种群作为本次算法的初始种群，使用表6中的用户自定义的特征变量代价权重矩阵计算适应度函数值。然后执行选择、交叉、变异操作，直到满足停止条件为止，根据适应度函数值得到最优的反事实个体。

表6 用户自定义的特征代价权重矩阵

Table 6 Feature cost weight matrix defined by users

特征变量	特征改变步长	特征改变代价权重							
	Δd	$(-\infty,-3\Delta d]$	$(-3\Delta d,-2\Delta d]$	$(-2\Delta d,-\Delta d]$	$(-\Delta d,0]$	$(0,\Delta d]$	$(\Delta d,2\Delta d]$	$(2\Delta d,3\Delta d]$	$(3\Delta d,+\infty)$
v_1	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0	0	0	0
v_2	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.1	1	$+\infty$	$+\infty$
v_3	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
v_4	1	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
v_5	1	1	1	1	1	1	1	1	1
v_6	6	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.6	0.8	1	$+\infty$
v_7	1	0	0	0	0	0.1	0.6	0.7	1
v_8	300	0	0	0	0	0.1	0.15	0.2	1
v_9	100	1	0.8	0.5	0	**0.7**	**0.9**	1	1
v_{10}	1 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{11}	1 000	$+\infty$	$+\infty$	$+\infty$	$+\infty$	0.1	0.2	0.3	1
v_{12}	1	0	0	0	0	0.1	0.1	0.1	0.1
v_{13}	1	1	0.7	0.6	0.5	0.5	0.6	0.7	1
v_{14}	20	**0.1**	**0.1**	**0.1**	**0.1**	0.1	**0.1**	**0.1**	**0.1**

（续表）

特征变量	特征改变步长 Δd	特征改变代价权重 (−∞,−3Δd]	(−3Δd,−2Δd]	(−2Δd,−Δd]	(−Δd,0]	(0,Δd]	(Δd,2Δd]	(2Δd,3Δd]	(3Δd,+∞)
v_{15}	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
v_{16}	10	0.1	0.1	0.1	0.1	0.1	0.2	0.3	1
v_{17}	1	0	0	0	0	0	0	0	0
v_{18}	1	0	0	0	0	0	0	0	0
v_{19}	1	0	0	0	0	0	0	0	0
v_{20}	1	0	0	0	0	0	0	0	0
v_{21}	1	0	0	0	0	0	0	0	0
v_{22}	1 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{23}	500	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{24}	1 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{25}	10 000	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{26}	500	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{27}	50	1	0.3	0.2	0.1	0.1	0.2	0.3	1
v_{28}	100	1	0.3	0.2	0.1	0.1	0.2	0.3	1

将表5的结果转换为表7的行动方案。传统反事实方法(TCE)[31]的行动方案是：用户的年龄由69岁降低为15岁，并且最近一次缴费金额由0元提升为9.8元，那么该用户将通过贷款，不具备可行性，用户年龄不能降低。本文提出的专家预定义的特征代价权重矩阵的反事实方法(ECCE)的行动方案是：用户近6个月的平均消费由41.66元提升为195.45元，即可通过贷款，具备可行性。本文提出的用户自定义的特征代价权重矩阵的反事实方法(UCCE)的行动方案是：如果用户当月通话交往人数从1人增加到58人，那么该用户将通过贷款，具备可行性。

表7 行动方案

Tabel 7 Action plan

反事实方法	变化的特征变量	行动方案	*cost*	*dist*	*spr*	*lof*
TCE	v_2(用户年龄)	69−＞15(−54)	∞	∞	2	1.74
	v_8(最近一次缴费金额)	0−＞9.8(+9.8)				
ECCE	v_9(近6个月平均消费值)	41.66−＞195.45(+153.79)	1.34	1.38	1	1.80
UCCE	v_{14}(当月通话交往圈人数)	1−＞58(+57)	1.26	1.27	1	1.66

由表5可证明本文方法给出的解释是具有实际可行性的，不仅能够解释用户贷款被拒绝的原因，也能够告诉用户如何做出改变，从而能够获得贷款。

本文针对5名被拒贷的用户，使用不同的反事实方法并比较结果。其中，*dist* 表示用户满意度，鉴于用户操作的便利性，*dist* 值越小满意度越高；*spr* 是特征变化数量，*spr* 值越小越好。*lof* 反映反事实用例是否为离群点，是企业关注的重要指标，*lof* 值越小，违约风险越低。*cost* 值则综合考虑了 *dist*，*spr* 和 *lof*，用于最终评估反事实方法，*cost* 值越小，说明方法越优。

当 *dist* 为∞时，表示无法给出用户满意的行动方案，由表8可见，对于用户 x_1，x_3，x_4，传统反事实方法(TCE)无法给出可行解。

表8 实验结果对比

Table 8 Comparison of experimental results

实例	反事实方法	*cost*	*dist*	*spr*	*lof*
x_1	TCE	∞	∞	2.00	1.74
	ECCE	1.34	1.38	1.00	1.80
	UCCE	**1.26**	1.27	1.00	1.66
x_2	TCE	0.59	0.35	1.00	1.43
	ECCE	0.59	0.35	1.00	1.43
	UCCE	**0.52**	0.03	2.00	1.01
x_3	TCE	∞	∞	2.00	1.09
	ECCE	1.13	1.07	1.00	1.80
	UCCE	**1.00**	0.71	2.00	1.01
x_4	TCE	∞	∞	1.00	1.04
	ECCE	0.41	0.15	1.00	1.03
	UCCE	**0.30**	0.00	1.00	1.03
x_5	TCE	4.16	5.19	2.00	1.27
	ECCE	1.30	1.28	1.00	2.06
	UCCE	**0.82**	0.31	1.00	4.07

针对5个被拒贷的用户，专家代价函数反事实方法(ECCE)和用户自定义代价函数的反事实方法(UCCE)均能给出可行解。通过 dist 指标，我们可以看到 TCE 方法在信贷风控场景下失效，很难给出用户满意的解释。ECCE 方法在不需要用户调整代价权重的前提下能够得到用户较为满意的解释。UCCE 可以根据用户自定义的特征变量代价权重矩阵得到更符合用户实际情况的个性化的反事实结果。TCE，ECCE 和 UCCE 这3种方法的 *spr* 指标差别不大，均在用户可接受范围内。

前4组实验中，TCE，ECCE 和 UCCE 的 *lof* 值相近，都符合数据分布。但在第5组，ECCE 和 UCCE 的 *lof* 值显著升高，意味着反事实用例为离群点，增加了违约风险。尽管对于用户 x_5，*dist* 和 *spr* 指标优秀，行动方案可行，但企业需权衡 *lof* 值带来的风险。若企业接受一定风险，可采用本文实验参数，使得 UCCE 方案对用户和企业都可行。若风险容忍度低，则需调整参数 λ_2。为了权衡 *dist*，*spr* 和 *lof* 指标，我们使用 *cost* 指标，如式(1)所示，可以根据用户和企业的实际情况调节 λ_1，λ_2，λ_3 参数的值，进而满足不同的信贷风控场景需求，确保反事实解释提供的改变特征的行动方案能够满足预期。

结束语 针对现有反事实解释方法的不足，本文提出了一种基于遗传算法的特征加权的反事实方法。首先，提出了特征变量代价权重的概念，考虑到了不同特征变量改变的难易程度，使得到的结果更符合现实情况，并且是可行的。其次，提出了特征变量代价权重矩阵的概念，通过由专家预定义特征变量代价权重矩阵的方式，给出了一种计算特征变量

代价权重的可行方法。然后，在专家定义的特征变量代价权重矩阵的基础上，可以由用户对特征变量代价权重矩阵进行修正，得到用户自定义的特征变量代价权重矩阵，从而可以满足不同用户对结果的差异化需求。我们定义的目标函数同时考虑到了特征加权距离、稀疏性和接近性 3 个指标。特征加权距离指标保证了反事实结果是可行的，稀疏性指标保证了反事实结果的简洁性，接近性保证了反事实用例尽量接近原始的样本集，避免反事实用例成为一个离群值或一个在现实生活中不可能的情况。最后，使用遗传算法对问题进行求解，并给出行动方案。实验结果表明，本文方法可以对信贷风控的结果进行很好的解释，同时具有可行性和可操作性。虽然本文通过提出特征变量代价权重矩阵的概念使得反事实解释具有可行性和可操作性，并以信贷风控场景为例进行了说明，但怎样将不同领域的特征代价权重矩阵与专家知识相结合，进而给出普适、通用的反事实框架是我们未来的一个研究方向。另外，本文只针对金融领域特定的数据集进行了实验，未来将考虑在更多类型的数据集上进行实验，进一步验证本文方法的普适性和鲁棒性。对于用户来说，操作特征代价权重矩阵有一定的难度，后续我们将研究更友好的交互方式，从而增加本文方法的实用性。本文通过调节式(1)中的 λ_1，λ_2，λ_3 参数的值来获取满足用户和企业需求的反事实结果，在后续的研究中，我们将深入探讨这些参数值如何影响反事实的结果，并对其因果关系进行深入探讨。对于用户敏感数据的隐私保护问题以及反事实解释可能对用户隐私产生的影响，本文暂未考虑。在未来的研究中，我们将进一步研究方法的可扩展性、实用性、因果解释力、模型透明度以及伦理隐私问题，进一步提升反事实方法的有效性和适用性。

参 考 文 献

[1] ALA'RAJ M, ABBOD M. Classifiers consensus system approach for credit scoring[J]. Knowledge-Based Systems, 2016, 104: 89-105.

[2] ZHANG M Y. Research on Credit Risk Management in Banking Under the New Situation[J]. Chinese Journal of Business Management, 2016, 10(14): 15-16.

[3] LEE T S, CHEN I F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines[J]. Expert Systems with Applications, 2005, 28(4): 743-752.

[4] KANG S, CHO S. Approximating support vector machine with artificial neural network for fast prediction[J]. Expert Systems with Applications, 2014, 41(10): 4989-4995.

[5] MOHAMMADI N, ZANGENEH M. Customer credit risk assessment using artificial neural networks[J]. International Journal of Information Technology and Computer Science, 2016, 8(3): 58-66.

[6] LIU X Y, QU Y W, ZHOU Q Y. Self-attention credit assessment model[J]. Chinese Journal of Computer Engineering and Applications, 2019, 55(13): 36-41.

[7] YU L, WANG S Y, LAI K K. Credit risk assessment with a multistage neural network ensemble learning approach[J]. Expert Systems with Applications, 2008, 34(2): 1434-1444.

[8] ZHOU M X. Study on User Profiling Based on Deep Neural Networks[D]. Changsha: Hunan University, 2018.

[9] MELIS D A, JAAKKOLA T. Towards robust interpretability with self-explaining neural networks[C]// Proceedings of the 32nd Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2018: 7775-7784.

[10] POULIN B, EISNER R, SZAFRON D, et al. Visual explanation of evidence with additive classifiers[C]// Proceedings of the 18th Conf on Innovative Applications of Artificial Intelligence. Palo Alto, CA: AAAI Press, 2006: 1822-1829.

[11] KONONENKO I. An efficient explanation of individual classifications using game theory[J]. Journal of Machine Learning Research, 2010, 11(Jan): 1-18.

[12] HAUFE S, MEINECKE F, GÖRGEN K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging[J]. NeuroImage, 2014, 87: 96-110.

[13] HUYSMANS J, DEJAEGER K, MUES C, et al. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models[J]. Decision Support Systems, 2011, 51(1): 141-154.

[14] BRESLOW L A, AHA D W. Simplifying decision trees: A survey[J]. The Knowledge Engineering Review, 1997, 12(1): 1-40.

[15] KONG X W, YANG H. A defense method against adversarial examples based on the interpretability of deep neural network models: China, CN112364885A[P]. 2021-02-12.

[16] 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. USA: ACM Press, 2016: 1135-1144.

[17] ZHOU Z H, JIANG Y, CHEN S F. Extracting symbolic rules from trained neural network ensembles[J]. AI Communications, 2003, 16(1): 3-15.

[18] LIN K, GAO Y. Model interpretability of financial fraud detection by group SHAP[J]. Expert Systems with Applications, 2022, 210: 118354.

[19] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: High-precision model-agnostic explanations[C]// Proceedings of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018.

[20] PEARL J. Theoretical impediments to machine learning with seven sparks from the causal revolution[J]. arXiv: 1801.04016, 2018.

[21] RODRIGUEZ P, CACCIA M, LACOSTE A, et al. Beyond Trivial Counterfactual Explanations with Diverse Valuable Explanations[C]// Proceedings of the International Conference on Computer Vision (ICCV). 2022: 1036-1045.

[22] DEL SER J, BARREDO-ARRIETA, DÍAZ-RODRÍGUEZ N, et al. On generating trustworthy counterfactual explanations [J]. Information Sciences, 2024, 655: 119898.

[23] POYIADZI R, SOKOL K, SANTOS-RODRIGUEZ, et al. FACE: Feasible and Actionable Counterfactual Explanations [C]// Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). New York: ACM, 2020.

[24] KANAMORI K, TAKAGI T, KOBAYASHI, et al. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization[C]// Proceedings of the IJCAI International Joint Conference on Artificial Intelligence. 2020: 2855-2862.

[25] BERK U, ALEXANDER S, YANG L. Actionable recourse in linear classification[C]// Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019: 10-19.

[26] CUI Z C, CHEN W L, HE Y J, et al. Optimal action extraction for random forests and boosted trees[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. USA: ACM, 2015: 179-188.

[27] VERMA S, DICKERSON J, HINES K. A Review of Counterfactual Explanations for Machine Learning [J]. arXiv: 2020: 1-13.

[28] BREUNIG M M, KRIEGEL H, NG R T, et al. LOF: Identifying Density-Based Local Outliers [C] // Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. USA: ACM, 2000: 4-23.

[29] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. Factual and Counterfactual Explanations for Black Box Decision Making [J]. IEEE Intelligent Systems, 2019, 34(6): 14-23.

[30] DANDL S, MOLNAR C, BINDER M, et al. Multi-Objective Counterfactual Explanations[C]// Proceedings of the International Conference on Parallel Problem Solving from Nature. 2020: 448-469.

[31] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal of Law & Technology, 2018, 31(2): 842-887.

WANG Baocai, born in 1988, Ph.D candidate. His main research interests include machine learning interpretability and intelligent credit risk control systems.

WU Guowei, born in 1973, Ph.D, professor, Ph.D supervisor. His main research interests include advanced computing and intelligent systems.

(责任编辑:喻藜)