(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107957944 A
(43)申请公布日 2018.04.24

(21)申请号 201711195641.7

(22)申请日 2017.11.24

(71)申请人 浙江大学
地址 310013 浙江省杭州市西湖区余杭塘路866号

(72)发明人 温盈盈 尹建伟 吴朝晖 邓水光 李莹

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
代理人 胡红娟

(51)Int.Cl.
G06F 11/36(2006.01)

权利要求书1页 说明书4页 附图2页

(54)发明名称

面向用户数据覆盖率的测试用例自动生成方法

(57)摘要

本发明公开了一种面向用户数据覆盖率的自动生成测试用例的方法，包括：(1)获得某软件产品的用户使用数据，组成数据集D_{origin}，并清洗所述数据集D_{origin}得到数据集D_{clean}；(2)采用机器学习方法处理所述数据集D_{clean}，获得测试用例集TC；(3)利用所述测试用例集TC对所述某软件产品进行测试，修复已经出现的漏洞，重复测试直至无严重漏洞，并将修复后的某软件产品重新投入使用；(4)记录重新投入使用的某软件产品的用户使用数据，组成数据集D_{add}，合并所述数据集D_{add}和所述数据集D_{origin}，得到作为下一轮测试用的数据集D_{origin}'。该方法可以提高测试过程的效率。

CN 107957944 A

1.一种面向用户数据覆盖率的自动生成测试用例的方法，包括以下步骤：

(1)获得某软件产品的用户使用数据，组成数据集D_{origin}，并清洗所述数据集D_{origin}得到数据集D_{clean}；

(2)采用机器学习方法处理所述数据集D_{clean}，获得测试用例集TC；

(3)利用所述测试用例集TC对所述某软件产品进行测试，修复已经出现的漏洞，重复测试直至无严重漏洞，并将修复后的某软件产品重新投入使用；

(4)记录重新投入使用的某软件产品的用户使用数据，组成数据集D_{add}，合并所述数据集D_{add}和所述数据集D_{origin}，得到作为下一轮测试用的数据集D_{origin}'。

2.如权利要求1所述的面向用户数据覆盖率的自动生成测试用例的方法，其特征在于，所述清洗所述数据集D_{origin}得到数据集D_{clean}包括：

判断所述数据集D_{origin}中数据量是否足够，

若是，直接删除掉所述数据集D_{origin}中的异常数据，得到所述数据集D_{clean}；

若否，对所述数据集D_{origin}中的重复数据、关键字段缺失数据进行初步清洗，并格式规整初步清洗完的数据，获得所述数据集D_{clean}。

3.如权利要求1所述的面向用户数据覆盖率的自动生成测试用例的方法，其特征在于，所述步骤(2)包括：

(2-1)基于所述数据集D_{clean}中的N个数据特征，将所述数据集D_{clean}映射到N个数据特征上，组成数据特征集$P=\{P_i, 1\leq i\leq N\}$，并获得每个数据特征上数据的映射值，P_i表示第i个数据特征；

(2-2)根据数据特征P_i上的数据分布特点，确定所述数据特征P_i上数据划分的数量k_i，并采用聚类算法对所述数据特征P_i上数据自动聚类成k_i类，

(2-3)基于自动分类结果，将所述数据特征P_i上数据划分成k_i类，每类数据用$C_{i,j}$表示，并基于$C_{i,j}$中包含数据实例的个数，为$C_{i,j}$赋予权重$W_{i,j}$，其中，$1\leq j\leq k_i$；

(2-4)计算$C_{i,j}$包括的数据的均值，将所述均值作为$C_{i,j}$的代表性中心点$O_{i,j}$；

(2-5)将N个数据特征上的所有代表中心点$O_{i,j}$进行交叉合成，获得多个测试用例取值组合；

(2-6)根据权重$W_{i,j}$计算每个测试用例取值组合的综合权重，选取综合权重排在前50%～75%大的测试用例取值组合组成测试用例集TC。

4.如权利要求3所述的面向用户数据覆盖率的自动生成测试用例的方法，其特征在于，所述权重$W_{i,j}$的计算过程为：

统计所述数据集D_{clean}中数据实例的个数N_{total}，统计$C_{i,j}$中包含数据实例的个数$n_{i,j}$，则权重$W_{i,j}$为：

$$W_{ij}=\frac{n_{i,j}}{N_{total}}\ (1\leq i\leq N,\ 1\leq j\leq k_i)。$$

5.如权利要求3所述的面向用户数据覆盖率的自动生成测试用例的方法，其特征在于，所述综合权重的获取过程为：

将每个测试用例取值组合包含的所有代表中心点$O_{i,j}$对应$C_{i,j}$的权重W_{ij}相乘，获得该测试用例取值组合的综合权重。

面向用户数据覆盖率的测试用例自动生成方法

技术领域

[0001] 本发明属于数据处理领域，具体涉及一种面向用户数据覆盖率的自动生成测试用例的方法。

背景技术

[0002] 测试是软件开发过程中必不可少的环节，对软件质量度量的一种方式，用以判断软件实际的运行结果是否与预期的一致。测试用例是测试步骤中的关键元素。测试用例作为被测试程序的输入，用以观察程序的表现和结果，由此发现程序中的错误和缺陷。

[0003] 测试用例的生成，长期以来依靠软件测试人员的经验和专业素养产生，手工完成。近期测试用例自动生成算法逐渐获得了许多研究者的关注，并产生了大量的成果。测试用例自动生成算法，从程序本身的结构出发，使生成出的测试用例能够最大限度地覆盖程序分支，从而尽量排除每个代码块中的漏洞。目前，生成算法追求的目的，除了算法本身的效率之外，将注意力集中在程序结构本身的正确性上。

[0004] 但是每个代码块的使用频率不尽相同，如果将测试力度平均分配到每个代码块中，那么难以集中精力发现软件中用户常用代码块的漏洞。用户使用软件时，最常使用的部分，如果出现程序漏洞，将会极大影响用户的体验，以及软件产品的质量。从程序分支覆盖程度的传统评价指标出发，无法考虑到用户的实际使用情况。目前，尚未有从用户实际使用的角度出发，进行测试用例自动生成的方法发明。

发明内容

[0005] 针对传统测试用例自动生成方法中，只考虑程序结构，未考虑用户实际使用情况的不足，本发明提出了一种面向用户数据覆盖率的自动生成测试用例的方法。

[0006] 面向用户数据覆盖率的自动生成测试用例的方法，包括以下步骤：

[0007] (1)获得某软件产品的用户使用数据，组成数据集D_{origin}，并清洗所述数据集D_{origin}得到数据集D_{clean}；

[0008] (2)采用机器学习方法处理所述数据集D_{clean}，获得测试用例集TC；

[0009] (3)利用所述测试用例集TC对所述某软件产品进行测试，修复已经出现的漏洞，重复测试直至无严重漏洞，并将修复后的某软件产品重新投入使用；

[0010] (4)记录重新投入使用的某软件产品的用户使用数据，组成数据集D_{add}，合并所述数据集D_{add}和所述数据集D_{origin}，得到作为下一轮测试用的数据集D_{origin}’。

[0011] 作为优选，所述清洗所述数据集D_{origin}得到数据集D_{clean}包括：

[0012] 判断所述数据集D_{origin}中数据量是否足够，

[0013] 若是，直接删除掉所述数据集D_{origin}中的异常数据，得到所述数据集D_{clean}；

[0014] 若否，对所述数据集D_{origin}中的重复数据、关键字段缺失数据进行初步清洗，并格式规整初步清洗完的数据，获得所述数据集D_{clean}。

[0015] 作为优选，所述步骤(2)包括：

[0016] (2-1)基于所述数据集D_{clean}中的N个数据特征，将所述数据集D_{clean}映射到N个数据特征上，组成数据特征集$P=\{P_i, 1\leqslant i\leqslant N\}$，并获得每个数据特征上数据的映射值，$P_i$表示第i个数据特征；

[0017] (2-2)根据数据特征P_i上的数据分布特点，确定所述数据特征P_i上数据划分的数量k_i，并采用聚类算法对所述数据特征P_i上数据自动聚类成k_i类，

[0018] (2-3)基于自动分类结果，将所述数据特征P_i上数据划分成k_i类，每类数据用$C_{i,j}$表示，并基于$C_{i,j}$中包含数据实例的个数，为$C_{i,j}$赋予权重$W_{i,j}$，其中，$1\leqslant j\leqslant k_i$；

[0019] (2-4)计算$C_{i,j}$包括的数据的均值，将所述均值作为$C_{i,j}$的代表性中心点$O_{i,j}$；

[0020] (2-5)将N个数据特征上的所有代表中心点$O_{i,j}$进行交叉合成，获得多个测试用例取值组合；

[0021] (2-6)根据权重$W_{i,j}$计算每个测试用例取值组合的综合权重，选取综合权重排在前50％～75％大的测试用例取值组合组成测试用例集TC。

[0022] 聚类是一个把数据对象集划分成多个组或簇的过程，使得簇内的对象具有很高的相似性，但与其他簇中的对象很不相似。聚类是一种数据挖掘工具，有多种不同的算法可供选择，可以根据实际的数据特点进行具体算法的选择。

[0023] 作为优选，所述权重$W_{i,j}$的计算过程为：

[0024] 统计所述数据集D_{clean}中数据实例的个数N_{total}，统计$C_{i,j}$中包含数据实例的个数$n_{i,j}$，则权重$W_{i,j}$为：

[0025] $$W_{ij}=\frac{n_{i,j}}{N_{total}}\ (1\leq i\leq N,\ 1\leq j\leq k_i)$$。

[0026] 在步骤(2-5)中，将N个数据特征上的所有代表中心点$O_{i,j}$进行交叉合成的过程，将每个数据特征的可能取值限定为数据特征的代表性点的值，共k_i种取值，一个测试用例包含N个数据特征，每个数据特征的取值，从k_i种中选取，共可生成$k_1\times k_2\times\cdots\times k_N$种不同取值组合的测试用例。

[0027] 在步骤(2-6)中，每个测试用例取值组合的综合权重的计算中，将每个测试用例取值组合包含的所有代表中心点$O_{i,j}$对应$C_{i,j}$的权重W_{ij}相乘，获得该测试用例取值组合的综合权重，综合权重越大的取值组合，说明其出现的可能性较大，需要成为重点的测试对象。生成出最具代表性的测试用例，在有限测试资源的情况下，提高对实际使用情况的覆盖率。

[0028] 本发明具有的有益效果为：

[0029] 充分利用实际使用产品所产生的数据，从使用数据的角度出发，通过人工智能的方式生成测试用例，改变现有测试用例针对程序覆盖率而非用户实际使用模块覆盖的现状。以数据覆盖率作为全新的测试用例生成标准，提升测试过程的效率和针对性。为产品的测试方式带来革新。

附图说明

[0030] 图1是本发明实施例提供的面向用户数据覆盖率的自动生成测试用例的方法的流程框图；

[0031] 图2是本发明实施例提供的使用机器学习分析生成测试用例的详细方法流程图。

具体实施方式

[0032] 为了更为具体地描述本发明，下面结合附图及具体实施方式对本发明的技术方案进行详细说明。

[0033] 本实施例利用机器学习算法，提出了一种通过对用户实际使用产品的数据进行自动分析，生成测试用例，用以提高测试过程效率的方法。

[0034] 图1是本发明实施例提供的面向用户数据覆盖率的自动生成测试用例的方法的流程框图。参见图1，本实施例提供的方法包括以下步骤：

[0035] S101，获得某软件产品的用户使用数据，组成数据集D_{origin}，并清洗所述数据集D_{origin}得到数据集D_{clean}。

[0036] 本步骤中，清洗所述数据集D_{origin}得到数据集D_{clean}的具体过程为：

[0037] 判断所述数据集D_{origin}中数据量是否足够，

[0038] 若是，直接删除掉所述数据集D_{origin}中的异常数据，得到所述数据集D_{clean}；

[0039] 若否，对所述数据集D_{origin}中的重复数据、关键字段缺失数据进行初步清洗，并格式规整初步清洗完的数据，获得所述数据集D_{clean}。

[0040] S102，采用机器学习方法处理所述数据集D_{clean}，获得测试用例集TC。

[0041] S102的具体过程如图2所示，参见图2，该步骤具体包括：

[0042] S201，基于所述数据集D_{clean}中的N个数据特征，将所述数据集D_{clean}映射到N个数据特征上，每个数据特征对应一个数据维度，组成数据特征集$P=\{P_i, 1\leqslant i\leqslant N\}$，并获得每个数据特征上数据的映射值，$P_i$表示第i个数据特征，也表示第i维数据；

[0043] S202，根据数据特征P_i上的数据分布特点，确定所述数据特征P_i上数据划分的数量k_i，并采用聚类算法对所述数据特征P_i上数据自动聚类成k_i类，

[0044] S203，基于自动分类结果，将所述数据特征P_i上数据划分成k_i类，每类数据用$C_{i,j}$表示，并基于$C_{i,j}$中包含数据实例的个数，为$C_{i,j}$赋予权重$W_{i,j}$，其中，$1\leqslant j\leqslant k_i$；

[0045] 本步骤中，所述权重$W_{i,j}$的计算过程为：

[0046] 统计所述数据集D_{clean}中数据实例的个数N_{total}，统计$C_{i,j}$中包含数据实例的个数$n_{i,j}$，则权重$W_{i,j}$为：

[0047] $$W_{ij}=\frac{n_{i,j}}{N_{total}}\ (1\leq i\leq N,\ 1\leq j\leq k_i)$$

[0048] S204，计算$C_{i,j}$包括的数据的均值，将所述均值作为$C_{i,j}$的代表性中心点$O_{i,j}$；

[0049] S205，将N个数据特征上的所有代表中心点$O_{i,j}$进行交叉合成，获得多个测试用例取值组合；

[0050] S206，根据权重$W_{i,j}$计算每个测试用例取值组合的综合权重，按照从大道小的顺序排列综合权重，并选取综合权重排在前50%～75%(本实施例选60%)的测试用例取值组合组成测试用例集TC；

[0051] 本步骤中，将每个测试用例取值组合包含的所有代表中心点$O_{i,j}$对应$C_{i,j}$的权重W_{ij}相乘，获得该测试用例取值组合的综合权重，综合权重越大的取值组合，说明其出现的可能性较大，需要成为重点的测试对象。生成出最具代表性的测试用例，在有限测试资源的情况下，提高对实际使用情况的覆盖率。

[0052] S103,利用所述测试用例集TC对所述某软件产品进行测试,修复已经出现的漏洞,重复测试直至无严重漏洞,并将修复后的某软件产品重新投入使用;

[0053] S104,记录重新投入使用的某软件产品的用户使用数据,组成数据集D_{add},合并所述数据集D_{add}和所述数据集D_{origin},得到作为下一轮测试用的数据集D_{origin}'。

[0054] 以上所述的具体实施方式对本发明的技术方案和有益效果进行了详细说明,应理解的是以上所述仅为本发明的最优选实施例,并不用于限制本发明,凡在本发明的原则范围内所做的任何修改、补充和等同替换等,均应包含在本发明的保护范围之内。


获得某软件产品的用户使用数据，组成数据集D_{origin}，并清洗所述数据集D_{origin}得到数据集D_{clean}
S101
采用机器学习方法处理所述数据集D_{clean}，获得测试用例集TC
S102
利用所述测试用例集TC对所述某软件产品进行测试，修复已经出现的漏洞，重复测试直至无严重漏洞，并将修复后的某软件产品重新投入使用
S103
记录重新投入使用的某软件产品的用户使用数据，组成数据集D_{add}，合并所述数据集D_{add}和所述数据集D_{origin}，得到作为下一轮测试用的数据集D_{origin}'
S104

图1

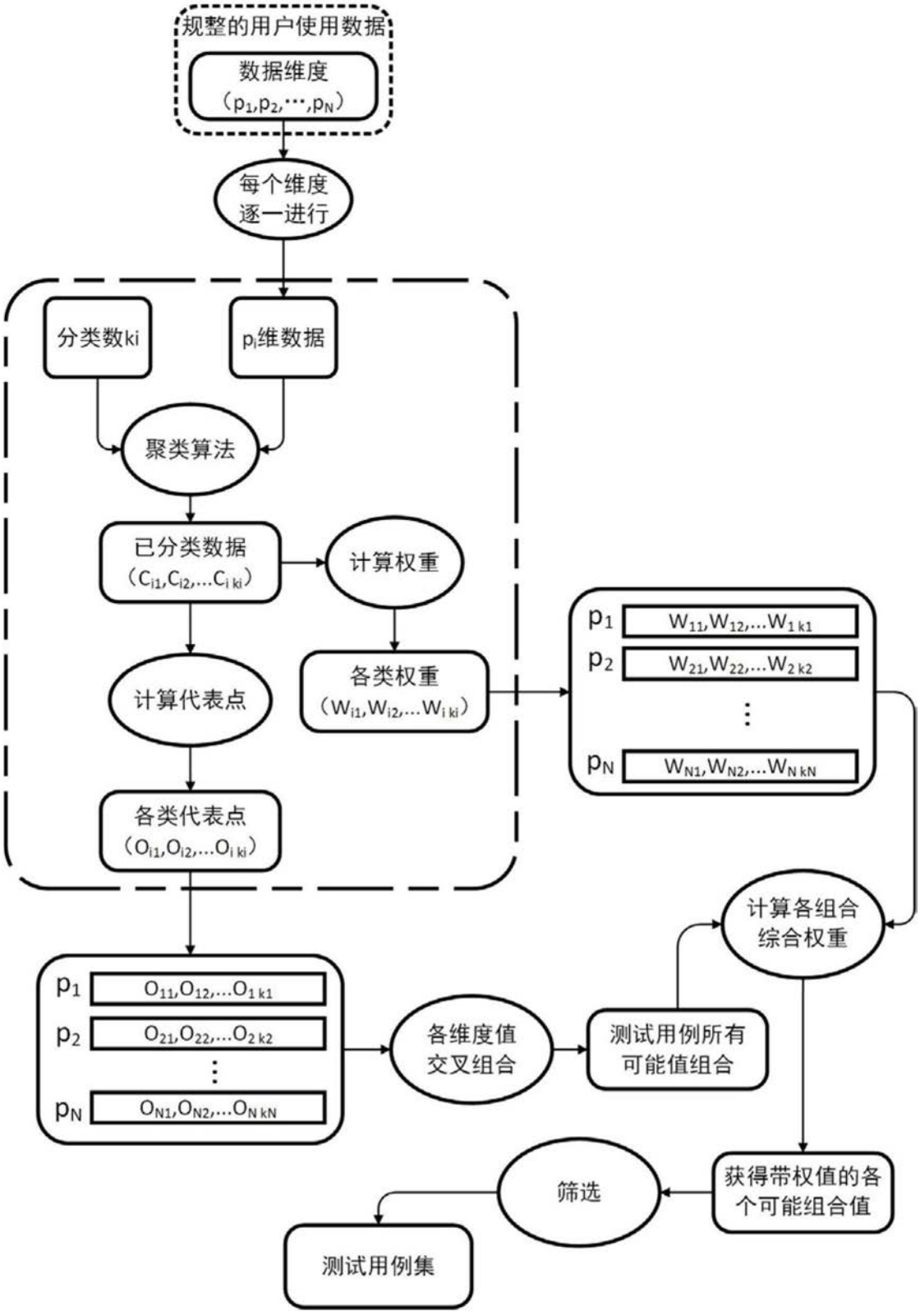
规整的用户使用数据
数据维度
（p1,p2,…,pN）
每个维度
逐一进行
分类数ki
pi维数据
聚类算法
已分类数据
（Ci1,Ci2,...Ci ki）
计算权重
计算代表点
各类权重
（Wi1,Wi2,...Wi ki）
各类代表点
（Oi1,Oi2,...Oi ki）
p1 W11,W12,...W1 k1
p2 W21,W22,...W2 k2
pN WN1,WN2,...WN kN
p1 O11,O12,...O1 k1
p2 O21,O22,...O2 k2
pN ON1,ON2,...ON kN
各维度值
交叉组合
测试用例所有
可能值组合
计算各组合
综合权重
获得带权值的各
个可能组合值
筛选
测试用例集

图2