(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 109977028 A
(43)申请公布日 2019.07.05

(21)申请号 201910274407.6

(22)申请日 2019.04.08

(71)申请人 燕山大学
地址 066004 河北省秦皇岛市海港区河北大街西段438号

(72)发明人 王倩 李亚洲

(74)专利代理机构 北京挺立专利事务所(普通合伙) 11265
代理人 刘阳

(51)Int.Cl.
G06F 11/36(2006.01)
G06N 3/00(2006.01)
G06N 3/12(2006.01)

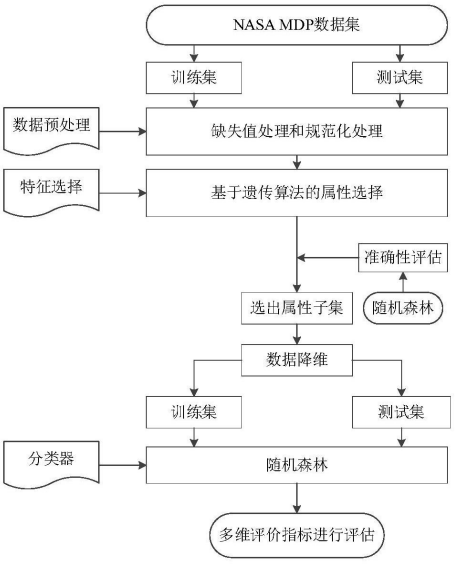
权利要求书2页 说明书7页 附图1页

(54)发明名称

一种基于遗传算法和随机森林的软件缺陷预测方法

(57)摘要

本发明公开了一种基于遗传算法和随机森林的软件缺陷预测方法,包括以下步骤:对软件缺陷数据集的各个子集进行数据预处理;基于遗传算法和随机森林算法进行特征选择;构建随机森林分类器;软件缺陷预测,利用处理后的软件缺陷数据集训练随机森林分类器,经过多测实验得到分类效果较优的随机森林分类器,然后将经过处理后的软件缺陷测试集输入到训练好的分类器中,最终获得测试集的分类结果。本发明很好地适应有差异性和类别不平衡的软件缺陷数据集;将遗传算法和随机森林算法相结合用于特征选择,达到很好的降维效果。使用基于决策树的集成算法,各自独立地学习并做出预测,将这些预测结果结合起来得到最终的预测结果。



CN 109977028 A

1.一种基于遗传算法和随机森林的软件缺陷预测方法,其特征在于,包括以下步骤:

步骤S1:通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作;

步骤S2:将遗传算法和随机森林组合分别用于缺陷特征选择和准确性评估,实现缺陷数据集的特征优化组合选择;

步骤S3:在经过特征优化选择的数据集上,基于基尼指数构建的分类回归决策树建立随机森林分类器,用于有效的软件缺陷分类预测;

步骤S4:软件缺陷预测,利用处理后的软件缺陷数据集训练随机森林分类器,经过多测实验得到分类效果较优的随机森林分类器,然后将经过处理后的软件缺陷测试集输入到训练好的分类器中,最终获得测试集的分类结果。

2.根据权利要求1所述的方法,其特征在于,所述步骤S1包括通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作,其中,缺失值处理是对数据集中缺失的属性值使用平均值进行填充;规范化处理是采用最小-最大规范化方法对数据集中除了缺陷的类别标记以外的每一条记录相对应的特征属性值均进行规范化处理,使之落入[0,1]的范围内,其中最小-最大规范化方法采用下述公式:

$$y=\frac{y-M_{\min}}{M_{\max}-M_{\min}} \tag{1}$$

其中,y为将要规范化的数值;M_{min}为其中一维中最小的数;M_{max}为其中一维中最大的数。

3.根据权利要求1所述的方法,其特征在于,所述步骤S2中的遗传算法包括以下步骤:

步骤S21:个体的编码,使用二进制基因位表示所选特征子集中的一个特征,每一个染色体就是由定长的二进制串构成,表示一个最优特征子集;

步骤S22:初始群体的设定,随机生成具有N个染色体的初始种群,N的取值范围为大于等于1且小于等于缺陷属性总数的任意整数;

步骤S23:设计适应度函数,其表示为:

$$f(X)=Accuracy(X)=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

其中X表示待评估的染色体,Accuracy(X)表示分类器使用该染色体中的特征进行分类后的精度,其中,TP表示缺陷被正确分类的数目,TN表示缺陷被错误分类的数目,FP表示非缺陷被正确分类的数据,FN表示非缺陷被错误分类的数目;

步骤S24:设计遗传算子,通过应用遗传算子生成新的染色体种群;

步骤S25:选择设定遗传的次数或者收敛到所有染色体相同作为终止条件,这样染色体是特征的最佳子集。

4.根据权利要求3所述的方法,其特征在于,步骤S24中的设计遗传算子通过选择算子、交叉算子、变异算子三个操作进行设计,其中选择算子采用轮盘赌选择方法,交叉算子采用单点交叉算子或者双点交叉算子。

5.根据权利要求4所述的方法,其特征在于,所述步骤S2中的随机森林算法构建过程如下:

(1)从原始训练集中使用随机有放回采样选出m个样本,共进行n次采样,生成n个训练集,m和n为自然数;

(2)对于n个训练集,分别训练n个决策树模型;

(3)对于单个决策树模型,那么每次分裂时根据基尼指数选择最好的特征进行分裂;

(4)每棵树都一直这样分裂下去,直到该节点的所有训练样例都属于同一类;

(5)将生成的多棵决策树组成随机森林,对于分类问题,按多棵决策树分类器投票决定最终分类结果。

6.根据权利要求5所述的方法,其特征在于,所述步骤S2中特征选择包括如下步骤:

步骤S201:将数据集分为训练集和测试集,缺陷属性的数量是特征的数量;

步骤S202:随机生成具有N个染色体的初始种群,其中,N的取值范围为大于等于1且小于等于缺陷属性总数的任意整数;

步骤S203:计算每条染色体的适应度,其中通过随机森林算法获得适应度函数的准确性;

步骤S204:选择操作员通过轮盘赌方法确定群体中N个染色体的去除和保留;

步骤S205:交叉和突变是通过交叉群体中剩余的染色体形成新的染色体,并根据交叉概率P_{cross}和突变概率$P_{variation}$进行突变;

步骤S206:终止条件是设定为G的遗传后代的数量,即迭代次数,并且重复步骤S203～步骤S205直到形成第G代染色体生成,此时特征组合被视为最佳特征子集。

7.根据权利要求6所述的方法,其特征在于,所述步骤S3包括:

步骤S301:根据数据集处理降维,得到最优特征子集,其中,再次使用随机森林模型,从降维后的训练集中,随机抽样t个训练集,一次采样的终止条件是每个叶包含不超过2个样本,t为自然数;

步骤S302:t个训练集构造出t个基于基尼指数构建的分类回归决策树,基尼指数索引用作树决策策略;

步骤S303:t个CART树构成随机森林,最终分类由多棵决策树的联合投票决定。

一种基于遗传算法和随机森林的软件缺陷预测方法

技术领域

[0001] 本发明涉及计算机领域，尤其涉及到一种基于遗传算法和随机森林的软件缺陷预测方法。

背景技术

[0002] 软件缺陷是软件失效的源头和影响软件可靠性的重要因素，尽早地预测软件中存在的缺陷，以合理分配测试验证资源并保证软件质量在软件工程领域尤为重要。软件缺陷预测技术已经成为软件工程领域的重要研究方向，主要分为静态预测和动态预测。静态预测是基于缺陷相关的度量数据，对缺陷的数量和分布进行预测；动态预测则是对故障随时间的分布进行预测。其中静态预测技术是早期进行缺陷预测使用更为普遍的技术，其使用的历史数据是采用多种方法度量对软件模块计算得到的，获取数据耗时且结构复杂，采用常用的分类技术，效果并不理想。此外，软件中有缺陷的模块数量远远少于无缺陷的模块数量，称为“类不平衡”问题，会影响缺陷预测分类的准确性。

[0003] 近年来，机器学习算法被广泛应用于各个研究领域，并取得了进展，已有学者将机器学习方法应用到软件缺陷预测中，如支持向量机、贝叶斯、决策树、关联规则等，并取得了较好的效果。Elish等人将SVM应用于软件缺陷预测，证明了SVM作为描述软件特征与易出错模块之间复杂的非线性关系的预测器具有良好的性能。

[0004] 现有与软件缺陷预测相关的专利主要集中通过数据采样或不同算法分类提高软件缺陷预测精度。现有技术中，已有解决软件度量数据冗余、缺陷预测精度不高，SVM参数选择难、消耗时间过久等问题。其中申请号为201710571098的发明专利“一种基于数据筛选和数据过采样的跨项目缺陷预测方法”设计了合理的数据筛选和数据不平衡处理策略，利用层次聚类算法筛选出真正和本项目模块数据相似的跨项目历史软件模块数据，使跨项目软件缺陷预测模型避免受到不相关跨项目历史软件模块数据的影响，然后利用过采样方法增加有缺陷的软件模块数据得到分类相对平衡的新数据集，使跨项目软件缺陷预测模型避免受到不平衡的训练数据集的影响。

[0005] 由于软件缺陷数据集的差异性比较大，而且类别极不平衡，使得机器学习算法在不同的评价指标下表现效果不是很好，同时很多现有技术也没有很好地处理类别不平衡性的问题，而使得软件缺陷预测的性能不好。

发明内容

[0006] 为了提高软件缺陷预测在多种评价指标上的性能，采用机器学习的组合的思想，本发明提出了一种基于遗传算法和随机森林相结合的软件缺陷预测方法，充分利用了遗传算法特征选择上的优势和随机森林强分类器的优势，是一种融合的软件缺陷预测方法。

[0007] 本发明首先对数据进行了缺失值和归一化的数据预处理操作，避免数据的不完整性和不一致性。进而，通过遗传算法进行特征选择，识别并移除特征空间中的无关特征和冗余特征，最终达到降低数据集的维数、缩小训练集的规模、缩短训练时间以及提高分类器的

性能。最后，通过随机森林多决策树集成且各决策树之间互相独立的特点，得到比单分类器更为客观和准确的分类预测结果。与上述的其他发明相比，本发明能更好得对软件缺陷数据集进行特征提取，同时也能更好地处理软件缺陷类别不平衡性的问题。从而在准确率、召回率、精度和AUC值等多维评价指标中表现出较好性能。

[0008] 为实现上述目的，本发明是根据以下技术方案实现的：

[0009] 一种基于遗传算法和随机森林的软件缺陷预测方法，其特征在于，包括以下步骤：

[0010] 步骤S1：通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作；

[0011] 步骤S2：将遗传算法和随机森林组合分别用于缺陷特征选择和准确性评估，实现缺陷数据集的特征优化组合选择；

[0012] 步骤S3：在经过特征优化选择的数据集上，基于基尼指数构建的分类回归决策树建立随机森林分类器，用于有效的软件缺陷分类预测；

[0013] 步骤S4：软件缺陷预测，利用处理后的软件缺陷数据集训练随机森林分类器，经过多测实验得到分类效果较优的随机森林分类器，然后将经过处理后的软件缺陷测试集输入到训练好的分类器中，最终获得测试集的分类结果。

[0014] 上述技术方案中，所述步骤S1包括通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作，其中，缺失值处理是对数据集中缺失的属性值使用平均值进行填充；规范化处理是采用最小-最大规范化方法对数据集中除了缺陷的类别标记以外的每一条记录相对应的特征属性值均进行规范化处理，使之落入[0,1]的范围内，其中最小-最大规范化方法采用下述公式：

[0015] $$y=\frac{y-M_{\min}}{M_{\max}-M_{\min}} \tag{1}$$

[0016] 其中，y为将要规范化的数值；M_{min}为其中一维中最小的数；M_{max}为其中一维中最大的数。

[0017] 上述技术方案中，所述步骤S2中的遗传算法包括以下步骤：

[0018] 步骤S21：个体的编码，使用二进制基因位表示所选特征子集中的一个特征，每一个染色体就是由定长的二进制串构成，表示一个最优特征子集；

[0019] 步骤S22：初始群体的设定，随机生成具有N个染色体的初始种群，N的取值范围为大于等于1且小于等于缺陷属性总数的任意整数；

[0020] 步骤S23：设计适应度函数，其表示为：

[0021] $$f(X)=Accuracy(X)=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

[0022] 其中X表示待评估的染色体，Accuracy(X)表示分类器使用该染色体中的特征进行分类后的精度，其中，TP表示缺陷被正确分类的数目，TN 表示缺陷被错误分类的数目，FP表示非缺陷被正确分类的数据，FN表示非缺陷被错误分类的数目；

[0023] 步骤S24：设计遗传算子，通过应用遗传算子生成新的染色体种群；

[0024] 步骤S25：选择设定遗传的次数或者收敛到所有染色体相同作为终止条件，这样染色体是特征的最佳子集。

[0025] 上述技术方案中，步骤S24中的设计遗传算子通过选择算子、交叉算子、变异算子

三个操作进行设计，其中选择算子采用轮盘赌选择方法，交叉算子采用单点交叉算子或者双点交叉算子。

[0026] 上述技术方案中，所述步骤S2中的随机森林算法构建过程如下：

[0027] (1)从原始训练集中使用随机有放回采样选出m个样本，共进行n 次采样，生成n个训练集，m和n为自然数；

[0028] (2)对于n个训练集，分别训练n个决策树模型；

[0029] (3)对于单个决策树模型，那么每次分裂时根据基尼指数选择最好的特征进行分裂；

[0030] (4)每棵树都一直这样分裂下去，直到该节点的所有训练样例都属于同一类；

[0031] (5)将生成的多棵决策树组成随机森林，对于分类问题，按多棵决策树分类器投票决定最终分类结果。

[0032] 上述技术方案中，所述步骤S2中特征选择包括如下步骤：

[0033] 步骤S201：将数据集分为训练集和测试集，缺陷属性的数量是特征的数量；

[0034] 步骤S202：随机生成具有N个染色体的初始种群，其中，N的取值范围为大于等于1且小于等于缺陷属性总数的任意整数；

[0035] 步骤S203：计算每条染色体的适应度，其中通过随机森林算法获得适应度函数的准确性；

[0036] 步骤S204：选择操作员通过轮盘赌方法确定群体中N个染色体的去除和保留；

[0037] 步骤S205：交叉和突变是通过交叉群体中剩余的染色体形成新的染色体，并根据交叉概率P_{cross}和突变概率$P_{variation}$进行突变；

[0038] 步骤S206：终止条件是设定为G的遗传后代的数量，即迭代次数，并且重复步骤S203～步骤S205直到形成第G代染色体生成，此时特征组合被视为最佳特征子集。

[0039] 上述技术方案中，所述步骤S3包括：

[0040] 步骤S301：根据数据集处理降维，得到最优特征子集，其中，再次使用随机森林模型，从降维后的训练集中，随机抽样t个训练集，一次采样的终止条件是每个叶包含不超过2个样本，t为自然数；

[0041] 步骤S302：t个训练集构造出t个基于基尼指数构建的分类回归决策树，基尼指数索引用作树决策策略；

[0042] 步骤S303：t个CART树构成随机森林，最终分类由多棵决策树的联合投票决定。

[0043] 本发明与现有技术相比，具有如下优点：

[0044] 1.本发明能够很好地适应有差异性和类别不平衡的软件缺陷数据集。

[0045] 2.本发明将遗传算法和随机森林算法相结合用于特征选择，达到很好的降维效果。

[0046] 3.本发明使用基于决策树的随机森林算法，各自独立地学习并做出预测。将这些预测结果结合起来得到最终的预测结果，因此，得到比单分类器更为客观和准确的分类预测结果。

[0047] 4.本发明能够有效地处理有软件缺陷数据集的差异性和类不平衡性。在准确率、召回率、精度和AUC值等多维评价指标中表现出较好性能。

附图说明

[0048] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它附图。

[0049] 图1为本发明的软件缺陷预测算法流程图；

[0050] 图2为本发明的单点交叉示意图；

[0051] 图3为本发明的双点交叉示意图；

[0052] 图4为本发明的变异过程示意图。

具体实施方式

[0053] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。

[0054] 本发明针对软件缺陷预测研究，细算法框架及流程图如图1所示：

[0055] 一种基于遗传算法和随机森林的软件缺陷预测方法，包括以下步骤：

[0056] 步骤S1：通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作；

[0057] 步骤S2：将遗传算法和随机森林组合分别用于缺陷特征选择和准确性评估，实现缺陷数据集的特征优化组合选择；

[0058] 步骤S3：在经过特征优化选择的数据集上，基于CART决策树 (Classification And Regression Tree)建立随机森林分类器，用于有效的软件缺陷分类预测；

[0059] 步骤S4：软件缺陷预测，利用处理后的软件缺陷数据集训练随机森林分类器，经过多测实验得到分类效果较优的随机森林分类器，然后将经过处理后的软件缺陷测试集输入到训练好的分类器中，最终获得测试集的分类结果。

[0060] 本发明中，步骤S1包括通过数据规约化方法对软件缺陷数据集进行缺失值补充等数据预处理操作，其中，缺失值处理是对数据集中缺失的属性值使用平均值进行填充；规范化处理是采用最小-最大规范化方法对数据集中除了缺陷的类别标记以外的每一条记录相对应的特征属性值均进行规范化处理，使之落入[0,1]的范围内，其中最小-最大规范化方法采用下述公式：

[0061] $$y=\frac{y-M_{\min}}{M_{\max}-M_{\min}} \qquad (1)$$

[0062] 其中，y为将要规范化的数值；M_{min}为其中一维中最小的数；M_{max}为其中一维中最大的数。

[0063] 数据规范化处理是数据挖掘的一项基础工作，不同评价指标往往具有不同的量纲，数值间的差别可能很大，不进行处理可能会影响到数据分析的结果。为了消除指标之间的量纲和取值范围差异的影响，需要对各个数据集进行规范化处理。

[0064] 步骤S2中，将遗传算法和随机森林组合分别用于缺陷特征选择和准确性评估，实现缺陷数据集的特征优化组合选择，算法描述分别如下：

[0065] 遗传算法作为一种启发式搜索策略，主要分为五大要素：个体的编码、初始群体的设定、适应度函数的设计、遗传算子的设计和终止条件的选择。遗传算法从一组随机产生的初始解，称为“种群(Population)”开始搜索过程。种群中的每个个体是问题的一个解，称为“染色体”。这些染色体在后续迭代中不断进化，称为遗传。后代是由前一代染色体通过交叉或者变异运算形成的。新一代形成中，根据“适应值”的大小选择部分后代，淘汰部分后代。适值高的染色体被选中的概率较高。据此，经过若干代之后，算法收敛于最好的染色体，它很可能就是问题的最优解。

[0066] (1)个体编码

[0067] 在遗传算法中，每条染色体都提供了一个可能的解决方案。在特征选择时一般情况使用二进制基因位表示所选特征子集中的一个特征，这样，每一个染色体就是由定长的二进制串构成，它表示一个可能的最优特征子集。如特征数量为5的染色体可能表示为<11010>，1表示被选取的特征，0则表示弃掉的特征。

[0068] (2)设定初始种群

[0069] 具有N个染色体的初始种群是随机生成的，没有任何约束条件。大量种群提供了更多的遗传多样性，但收敛速度会随着种群的增多而减慢。相反，种群如果太小，收敛速度会加快，但可能导致陷入局部最优。

[0070] (3)设计适应度函数

[0071] 适应度函数应保证优良的染色体(即优秀的特征子集)具有较高的适应度值，因此，所选特征子集应对分类具有较大的贡献，使分类的结果越精确越好，同时，特征空间的每一维都会增加分类的代价，特征子集中包含的特征项应尽可能少。提出的适应度函数如下：

[0072] $$f(X)=Accuracy(X)=\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

[0073] 其中X表示待评估的染色体，Accuracy(X)表示分类器使用该染色体中的特征进行分类后的精度。其中，TP表示缺陷被正确分类的数目，TN 表示缺陷被错误分类的数目，FP表示非缺陷被正确分类的数据，FN表示非缺陷被错误分类的数目

[0074] (4)设计遗传算子

[0075] 运用于求解最优特征子集的三个操作算子分别为：选择算子、交叉算子、变异算子。通过应用遗传算子生成新的染色体种群。遗传算子的设计可以增加种群的多样性，以便选取更优的特征子集。

[0076] ①选择算子

[0077] 遗传算法中选择算子常采用轮盘赌选择方法。轮盘赌选择又称比例选择算子，其基本思想是：各个个体被选的概率与其适应度函数值大小成正比。设群体大小为N，个体x_i的适应度为$f(x_i)$，则个体x_i的选择概率为：

[0078] $$P(X_i)=\frac{f(X_i)}{\sum_{j=1}^{N}f(X_j)} \quad (3)$$

[0079] 轮盘赌选择法可用如下过程模拟来实现：

[0080] (1)在[0,1]内产生一个均匀分布的随机数r。

[0081] (2)若If R≤Q_1,则染色体X_1被选中。

[0082] (3)若Q_{k-1}≤R≤Q_k(2≤k≤N),则染色体X_k被选中。

[0083] 其中的Q_i称为染色体X_i(i＝1,2,…,n)的积累概率,其计算公式为:

[0084] $$Q_i = \sum_{j=1}^{i} P(X_j) \quad (4)$$

[0085] ②交叉算子

[0086] 交叉运算是指对两个相互配对的染色体依据交叉概率按某种方式相互交换其部分基因,从而形成两个新的个体。交叉算子可以采用单点交叉算子,也可以采用双点交叉。交叉过程如图2和图3所示。

[0087] ③变异算子

[0088] 变异运算是指改变个体编码串中的某一位或几位基因值,从而形成新的个体。交叉运算和变异运算的相互配合,共同完成对搜索空间的全局搜索和局部搜索。一般情况下,变异算子操作首先对种群中所有个体设定的变异概率判断是否进行变异。变异过程如图4所示。

[0089] (5)终止条件的选择

[0090] 终止条件有两种情况,第一是设定遗传的次数,第二是收敛到所有染色体相同。认为这两种情况下的染色体是特征的最佳子集。

[0091] 随机森林是基于决策树的集成算法,它的原理是生成多个分类器模型,各自独立地学习并做出预测。这些预测最后结合起来得到预测结果,因此和单独分类器相比,结果会更理想。步骤S2中随机森林构建过程如下:

[0092] (1)从原始训练集中使用随机有放回采样选出m个样本,共进行n 次采样,生成n个训练集,m和n为自然数;

[0093] (2)对于n个训练集,分别训练n个决策树模型;

[0094] (3)对于单个决策树模型,那么每次分裂时根据基尼指数选择最好的特征进行分裂;

[0095] (4)每棵树都一直这样分裂下去,直到该节点的所有训练样例都属于同一类;

[0096] (5)将生成的多棵决策树组成随机森林,对于分类问题,按多棵树分类器投票决定最终分类结果。

[0097] 特征选择包括如下步骤:

[0098] 步骤S201:将数据集分为训练集和测试集,缺陷属性的数量是特征的数量;

[0099] 步骤S202:随机生成具有N个染色体的初始种群,其中,N的取值范围为大于等于1且小于等于缺陷属性总数的任意整数;

[0100] 步骤S203:计算每条染色体的适应度,其中通过随机森林算法获得适应度函数的准确性;

[0101] 步骤S204:选择操作员通过轮盘赌方法确定群体中N个染色体的去除和保留;

[0102] 步骤S205:交叉和突变是通过交叉群体中剩余的染色体形成新的染色体,并根据交叉概率P_{cross}和突变概率$P_{variation}$进行突变;

[0103] 步骤S206:终止条件是设定为G的遗传后代的数量,即迭代次数,并且重复步骤S203～步骤S205直到形成第G代染色体生成,此时特征组合被视为最佳特征子集。

[0104] 步骤S3的构建随机森林分类器包括：

[0105] 步骤S301:根据数据集处理降维,得到最优特征子集,其中,从降维后的训练集中,随机抽样t个训练集,一次采样的终止条件是每个叶包含不超过2个样本,t为自然数；

[0106] 步骤S302:t个训练集构造出t个CART树,基尼指数索引用作树决策策略；

[0107] 步骤S303:t个CART树构成随机森林,最终分类由多棵决策树的联合投票决定。

[0108] 本发明以软件缺陷预测研究中心广泛应用的由美国航空航天局 (NASA)公布的NASA IV&V Facility Metrics Data Program(MDP)数据集为实验数据,已经通过实验,软件缺陷预测分类效果较为理想。和设计的预期一致。

[0109] 本发明构建的基于遗传算法和随机森林相结合的软件缺陷预测方法与第2条所属的最好的现有技术进行对比,结果表明了本发明在平均准确率、平均召回率和平均精度上均有明显提高分别达到了93.5％、93％和95％,在平均F1-score和AUC这些综合评价指标上分别达到了94％和82.7％,同样要优于现有的软件缺陷预测方法。

[0110] 本发明显示了基于遗传算法和随机森林算法的软件缺陷预测方法在软件缺陷预测中的有效性。

[0111] 以上对本发明的具体实施例进行了描述。需要理解的是,本发明并不局限于上述特定实施方式,本领域技术人员可以在权利要求的范围内做出各种变化或修改,这并不影响本发明的实质内容。在不冲突的情况下,本申请的实施例和实施例中的特征可以任意相互组合。

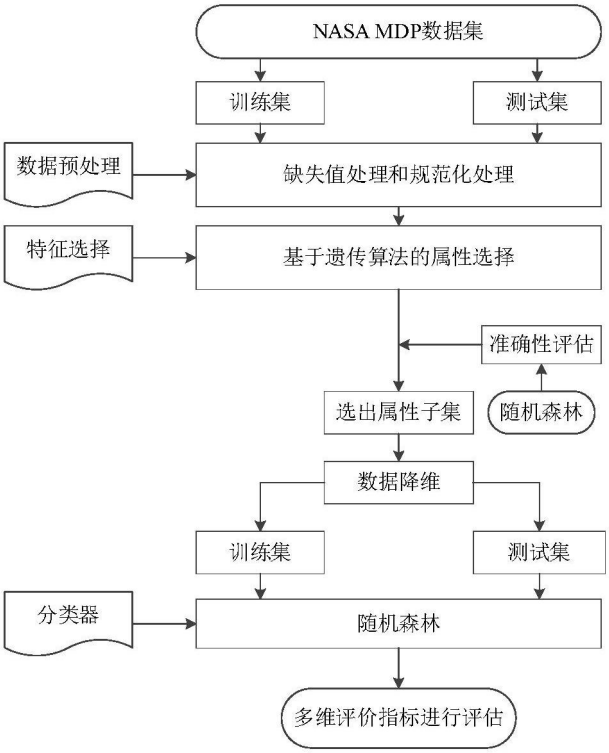

NASA MDP数据集
训练集
测试集
数据预处理
缺失值处理和规范化处理
特征选择
基于遗传算法的属性选择
准确性评估
选出属性子集
随机森林
数据降维
训练集
测试集
分类器
随机森林
多维评价指标进行评估

图1

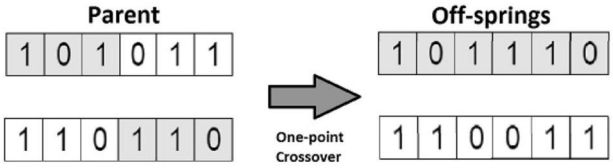

Parent
Off-springs
1 0 1 0 1 1
1 1 0 1 1 0
One-point Crossover
1 0 1 1 1 0
1 1 0 0 1 1

图2

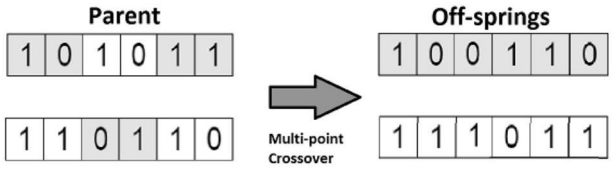

Parent
Off-springs
1 0 1 0 1 1
1 1 0 1 1 0
Multi-point Crossover
1 0 0 1 1 0
1 1 1 0 1 1

图3


Mutation
1 0 1 1 1 0
1 0 0 1 1 0

图4