(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 109933539 A
(43)申请公布日 2019.06.25

(21)申请号 201910298450.6

(22)申请日 2019.04.15

(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 何海涛 任家东 张旭 胡昌振

(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)

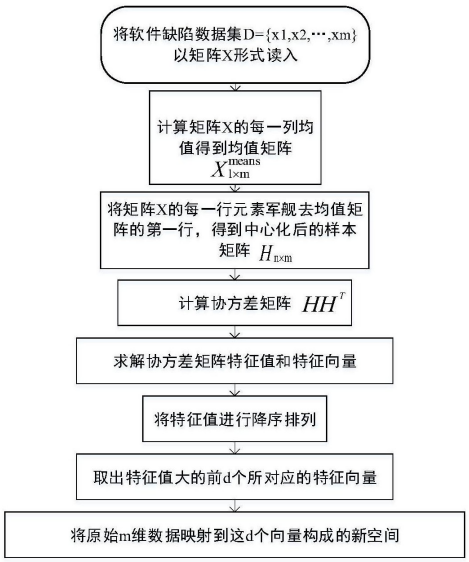
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种基于主成分分析和组合采样的软件缺陷预测方法

(57)摘要

本发明公开了一种基于主成分分析和组合采样的软件缺陷预测方法，包括如下步骤：步骤S1：对软件缺陷数据利用融合特征选择降维去噪；步骤S2：对降维后的数据执行SMOTE过采样和分层随机采样相结合进行采样，其中过采样是指通过增加少数类样本的数量，从而使得数据集中类样本达到相对平衡，分层随机采样通过划分类进行分层，在每层内采用无放回随机采样；步骤S3：对处理后的数据选取分类器并对分类器参数进行调优。本发明选择随机森林分类器，其随机选择特征子集的特性，从而进一步达到对树的随机化目的，避免了分类器过拟合问题的出现，最终提升了软件缺陷预测性能以及预测效率，为现实中预测有缺陷软件提供了良好的理论和实验依据。



CN 109933539 A

1.一种基于主成分分析和组合采样的软件缺陷预测方法,其特征在于,包括如下步骤:

步骤S1:对软件缺陷数据利用融合特征选择降维去噪;

步骤S2:对降维后的数据执行SMOTE过采样和分层随机采样相结合进行采样,其中过采样是指通过增加少数类样本的数量,从而使得数据集中类样本达到相对平衡,分层随机采样通过划分类进行分层,在每层内采用无放回随机采样;

步骤S3:对处理后的数据选取分类器并对分类器参数进行调优。

2.根据权利要求1所述的软件缺陷预测方法,其特征在于,步骤S1利用主成分分析法去除软件缺陷数据集中无关和冗余属性进行降维去噪,其中,主成分分析方法将m维特征通过线性变换映射到新的d维正交特征上,其中d<m,同时保留原始特征的绝大部分信息,并将重新构造出来的d维特征称为主元,从而使得数据由原来的m个特征降低到d个特征,具体包括如下步骤:

步骤S101:将软件缺陷数据集以矩阵$X_{n\times m}$形式输入$X_{m\times n}$,其中矩阵的行数n表示软件缺陷数据集中样本的个数,列数m表示每个样本的特征数目;

步骤S102:按列计算矩阵$X_{n\times m}$的均值,从而得到均值矩阵$X_{1\times m}^{means}$,并将矩阵$X_{n\times m}$中的每一行元素均减去$X_{1\times m}^{means}$得到进行中心化的样本$H_{n\times m}$;

步骤S103:计算中心化后样本$H_{n\times m}$的协方差矩阵HH^T,并对协方差矩阵进行特征值分解,求得对应的m个特征向量ω,然后将m个特征值λ进行降序排序,排序结果为$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_m$,最后,通过计算贡献率$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{m}\lambda_i}\geq 0.95$,并为其设定一个阈值为0.95,确定满足贡献率不小于此阈值的d值,取最大的d个特征值所对应的特征向量$\omega_1,\omega_2,\ldots,\omega_d$,从而得到投影矩阵$W^*=(\omega_1,\omega_2,\ldots,\omega_d)$,其中d为通过主成分分析法降维后的维度。

3.根据权利要求1所述的软件缺陷预测方法,其特征在于,所述分层随机采样采用无放回的随机采样,通过对无放回分层随机采样中采样率的调整使得子样本在保证类别分布不变的同时也减少了样本数量,从而克服SMOTE算法中训练模型效率降低,相应提高预测准确率。

4.根据权利要求3所述的软件缺陷预测方法,其特征在于,步骤S2具体包括:

对数据采用SMOTE算法合成少类样本,其中参数k表示在合成少类样本时,需从与当前所选样本距离相对较近的k少类样本中随机选取一个样本来进行新样本的合成,其中k的值为weka中的默认值5,采样倍率设为100%,不断进行迭代,直至有无缺陷样本达到相对平衡;

对相对平衡后的数据集进行无放回分层采样,采样率的设定为[0.1,1.0],步长为0.1,采样率为0.8。

5.根据权利要求1所述的软件缺陷预测方法,其特征在于,所述步骤S3采用经过网格搜索算法调参的随机森林分类器进行分类,具体包括如下步骤:

步骤S301:需要对随机森林分类器的两个参数决策树数目以及分裂属性数设定相应的范围和步长,并分别以这两个参数作为横纵坐标轴,建立二维坐标系,通过在坐标系中不断取点得到二维网格;

步骤S302:将网格中节点的每一数对分别作为参数的取值构建随机森林,并采用交叉验证估计分类误差;

步骤S303:选择分类误差最小的最优参数组合,直至分类误差或者步长满足要求为止,输出此组合,否则,尝试缩短步长,重复步骤S301。

一种基于主成分分析和组合采样的软件缺陷预测方法

技术领域

[0001] 本发明涉及缺陷预测方法，尤其涉及到一种基于主成分分析和组合采样的软件缺陷预测方法。

背景技术

[0002] 随着互联网技术的发展，软件产品质量的可靠性已成为软件工程领域的关注性问题，在软件开发的过程中必然会伴随着软件缺陷的出现。然而，对于本身具有潜在威胁的软件，一旦投入使用就会对公司乃至个人造成巨大的经济损失。为了有效解决这一问题，必须准确快速的预测软件可能存在的缺陷模块，从而提高软件系统的可靠性。

[0003] 目前，相关的软件缺陷预测方法主要是利用不同类型的机器学习技术。其主要考虑的是整体数据的预测准确率，虽然在此方面取得了较大的成就，但在数据预处理方面还存在很多需要改善的地方。现有技术中已有采用公开的NASA数据集对Random Forest，Naive Bayes，RPart以及SVM分类算法进行了灵敏度分析，表明不同的分类器针对不同数据集预测能力具有不确定性。考虑到这种不确定性，现有技术提出了一种新的贝叶斯组合模型，通过不断调整基模型的信用值来预测QoS，从而达到良好的预测精度。然而，这些研究并没有考虑到软件缺陷预测中的数据高维性和数据分布不均衡，缺陷类样本数通常比无缺陷类样本数高很多，导致作为多类的无缺陷样本特征掩盖少类的缺陷样本特征，使得虽整体准确率很高但针对缺陷类样本的预测性能较差；误分代价差异较大，将有缺陷倾向的模块标记为无缺陷倾向模块后，需要花费很高的代价进行更正等问题。尤其是针对少数类而言，不平衡分布使得某些机器学习方法表现效果不佳。为了解决这一问题，目前在数据层面，采用特征选择或特征提取，采样技术应用于数据集。特征选择或特征提取主要用于解决数据高维性问题，而采样方法通常是采用给少类样本随机加入高斯噪声或合成新的少类样本的方法来解决数据类不平衡问题。关于软件缺陷预测以前的研究表明，特征选择和特征提取方法确实有助于解决数据高维性这一问题。

发明内容

[0004] 为了能够在软件开发过程中及时准确地预测有缺陷的软件模块，提高软件测试资源的有效分配，针对软件缺陷预测中的数据类不平衡性和高维性问题，本发明的目的在于提供一种基于主成分分析和组合采样的软件缺陷预测方法，本发明首先通过对数据利用融合特征选择技术去除数据中无关和冗余特征以解决数据集中维度灾难问题。进而，执行SMOTE(Synthetic Minority Oversampling Technique)过采样和无放回分层随机采样方法结合来解决由于缺陷类样本数量过少，使得数据类分布不均衡而且缺陷样本信息过于缺乏，致使最终将有缺陷模块被错误预测为无缺陷模块等问题，同时通对采样率的设定在降低损失代价的同时也提高了软件缺陷预测效率。

[0005] 为实现上述目的，本发明是根据以下技术方案实现的：

[0006] 一种基于主成分分析和组合采样的软件缺陷预测方法，其特征在于，包括如下步

骤：

[0007] 步骤S1：对软件缺陷数据利用融合特征选择降维去噪；

[0008] 步骤S2：对降维后的数据执行SMOTE过采样和分层随机采样相结合进行采样，其中过采样是指通过增加少数类样本的数量，从而使得数据集中类样本达到相对平衡，分层随机采样通过划分类进行分层，在每层内采用无放回随机采样；

[0009] 步骤S3：对处理后的数据选取分类器并对分类器参数进行调优。

[0010] 上述技术方案中，步骤S1利用主成分分析法去除软件缺陷数据集中无关和冗余属性进行降维去噪，其中，主成分分析方法将m维特征通过线性变换映射到新的d维正交特征上，其中d<m，同时保留原始特征的绝大部分信息，并将重新构造出来的d维特征称为主元，从而使得数据由原来的m个特征降低到d个特征，具体包括如下步骤：

[0011] 步骤S101：将软件缺陷数据集以矩阵$X_{n\times m}$形式输入$X_{m\times n}$，其中矩阵的行数n表示软件缺陷数据集中样本的个数，列数m表示每个样本的特征数目；

[0012] 步骤S102：按列计算矩阵$X_{n\times m}$的均值，从而得到均值矩阵$X_{1\times m}^{means}$，并将矩阵$X_{n\times m}$中的每一行元素均减去$X_{1\times m}^{means}$得到进行中心化的样本$H_{n\times m}$；

[0013] 步骤S103：计算中心化后样本$H_{n\times m}$的协方差矩阵HH^T，并对协方差矩阵进行特征值分解，求得对应的m个特征向量ω，然后将m个特征值λ进行降序排序，排序结果为$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_m$，最后，通过计算贡献率$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{m}\lambda_i}\geq 0.95$，并为其设定一个阈值为0.95，确定满足贡献率不小于此阈值的d值，取最大的d个特征值所对应的特征向量$\omega_1,\omega_2,\ldots,\omega_d$，从而得到投影矩阵$W^*=(\omega_1,\omega_2,\ldots,\omega_d)$，其中d为通过主成分分析法进行降维后的维度。

[0014] 上述技术方案中，所述分层随机采样采用无放回的随机采样，通过对无放回分层随机采样中采样率的调整使得子样本在保证类别分布不变的同时也减少了样本数量，从而克服SMOTE算法中训练模型效率降低，相应提高预测准确率。

[0015] 上述技术方案中，步骤S2具体包括：

[0016] 对数据采用SMOTE算法合成少类样本，其中参数k表示在合成少类样本时，需从与当前所选样本距离相对较近的k少类样本中随机选取一个样本来进行新样本的合成，其中k的值为weka中的默认值5，采样倍率设为100％，不断进行迭代，直至有无缺陷样本达到相对平衡；

[0017] 对相对平衡后的数据集进行无放回分层采样，采样率的设定为[0.1,1.0]，步长为0.1，采样率为0.8。

[0018] 上述技术方案中，所述步骤S3采用经过网格搜索算法调参的随机森林分类器进行分类，具体包括如下步骤：

[0019] 步骤S301：需要对随机森林分类器的两个参数决策树数目以及分裂属性数设定相应的范围和步长，并分别以这两个参数作为横纵坐标轴，建立二维坐标系，通过在坐标系中不断取点得到二维网格；

[0020] 步骤S302：将网格中节点的每一数对分别作为参数的取值构建随机森林，并采用交叉验证估计分类误差；

[0021] 步骤S303：选择分类误差最小的最优参数组合，直至分类误差或者步长满足要求

为止,输出此组合,否则,尝试缩短步长,重复步骤S301。

[0022] 本发明与现有技术相比,具有如下优点:

[0023] 本发明从数据本身出发,从源头分析了目前软件缺陷预面临的问题。将特征选择和组合采样方法完美的结合在一起,在提高缺陷模块预测准确性的同时,提高了预测效率。为企业软件质量的的测试提供了良好的理论依据。

[0024] 本发明提出的SMOTE算法和无放回分层随机采样的方法,既解决了数据类不平衡问题又保证了样本类间的原始结构,更具有实际意义。

[0025] 本发明在考虑程序运行时间的前提下,运用网格搜索算法对分类器参数进行调优,选出兼顾效率和性能的最优参数组合,同时降低手动调参所造成的组合遗漏问题。

附图说明

[0026] 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。

[0027] 图1为本发明利用主成分分析法进行降维去噪的流程示意图;

[0028] 图2为PCS-RF软件缺陷预测模型的具体流程示意图。

具体实施方式

[0029] 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。

[0030] 本发明的一种基于融合特征选择和组合采样的软件缺陷预测方法,包括如下步骤:

[0031] 步骤S1:对软件缺陷数据利用融合特征选择降维去噪;

[0032] 步骤S2:对降维后的数据执行SMOTE过采样和分层随机采样相结合进行采样,其中过采样是指通过增加少数类样本的数量,从而使得数据集中类样本达到相对平衡,分层随机采样通过划分类进行分层,在每层内采用无放回随机采样;

[0033] 步骤S3:对处理后的数据选取分类器并对分类器参数进行调优。

[0034] 软件模块包含的特征属性越来越多,对于解决不同问题,这些属性的重要程度不同。甚至,有些不相关属性,有时会掩盖真正重要的分类特征,过多的维数还可能会使得算法的性能降低,进而降低模型的有效性。通过特征选择(Attribute Selection)或特征提取可以从已有的特征中跟据某种评价标准选择出最优属性的特征集,从而去除属性中不相关以及冗余的属性,达到数据集降维的目的。本技术采用主成分分析(PCA)方法进行特征提取,主要思想是将m维特征映射到新的d(d<m)维正交特征上。通过保留W*与样本的均值向量就可通过进行简单的向量减法和线性映射将高维空间的样本投影到低维空间。最小的特征值所对应的特征向量往往和噪声相关,当数据受到噪声影响时,由于PCA算法舍弃了对应于最小的m-d个特征值所对应的特征向量,因此,达到了去噪的效果,保证了被采样的数据质量。其具体流程如图1所示。

[0035] 如图1所示，步骤S1利用主成分分析法去除软件缺陷数据集中无关和冗余属性进行降维去噪，主成分分析方法是一种有效的特征提取方法，其主要思想是将m维特征通过线性变换映射到新的d(d<m)维正交特征上，同时保留原始特征的绝大部分信息，并将重新构造出来的d维特征称为主元，从而使得数据由原来的m个特征降低到d个特征，具体包括如下步骤：

[0036] 步骤S101：将软件缺陷数据集以矩阵$X_{n\times m}$形式输入$X_{m\times n}$，其中矩阵的行数n表示软件缺陷数据集中样本的个数，列数m表示每个样本的特征数目；

[0037] 步骤S102：按列计算矩阵$X_{n\times m}$的均值，从而得到均值矩阵$X_{1\times m}^{means}$，并将矩阵$X_{n\times m}$中的每一行元素均减去$X_{1\times m}^{means}$得到进行中心化的样本$H_{n\times m}$；

[0038] 步骤S103：计算中心化后样本$H_{n\times m}$的协方差矩阵HH^T，并对协方差矩阵进行特征值分解，求得对应的m个特征向量ω，然后将m个特征值λ进行降序排序，排序结果为$\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_m$，最后，通过计算贡献率$\frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{m}\lambda_i}\geq 0.95$，并为其设定一个阈值为0.95，确定满足贡献率不小于此阈值的d值，取最大的d个特征值所对应的特征向量$\omega_1,\omega_2,\ldots,\omega_d$，从而得到投影矩阵$W^*=(\omega_1,\omega_2,\ldots,\omega_d)$，其中d即为通过主成分分析法进行降维后的维度。

[0039] 根据本发明的一个具体实施例，分层随机采样采用无放回的随机采样，通过对无放回分层随机采样中采样率的调整使得子样本在保证类别分布不变的同时也减少了样本数量，从而克服SMOTE算法中训练模型效率降低，相应提高预测准确率。

[0040] SMOTE算法是一种经典的过采样方法，所谓过采样方法是指通过增加少数类样本的数量，从而使得数据集中类样本达到相对平衡。种过采样方法虽然在多数类和少数类数据上达到了相对平衡，但是随着样本数量的增加，训练时间也会相对增加，从而降低了模型的训练效率。考虑到本技术的高效性问题同时，为了保持数据集原本的类分布情况，考虑采用分层采样的方法对数据进行处理。分层采样是通过划分类进行分层，简单说就是分别对每个类别进行采样。而在每层内采用随机采样，同时，为了避免数据的冗余，我们考虑在每个类层中采用无放回的随机采样。此方法增大了各类中单位间的共同性，容易抽出具有代表性的子样本。通过对无放回分层随机采样方法中采样率的调整使得子样本在保证类别分布不变的同时也减少了样本数量，从而克服SMOTE算法中训练模型效率降低的问题，同时也相应的提高了预测准确率。

[0041] 步骤S2具体包括：

[0042] 对数据采用SMOTE算法合成少类样本，其中参数k表示在合成少类样本时，需从与当前所选样本距离相对较近的k少类样本中随机选取一个样本来进行新样本的合成，此技术中采用k值为weka中的默认值的5，采样倍率设为100％，不断进行迭代，直至有无缺陷样本达到相对平衡；其中weka是一个机器学习的平台。而k则是SMOTE这一过采样方法中涉及到的一个参数。其k表示当针对一选取的样本来合成新的少类样本时，也就是最近邻数。

[0043] 对相对平衡后的数据集进行无放回分层采样，采样率的设定为[0.1,1.0]，步长为0.1，采样率为0.8。此时，实验效果最佳。

[0044] SMOTE算法和无放回分层采样方法的结合，使得本技术在解决实际问题时更加准确快捷，对于企业来说会极大的提升企业效益，可采用性提升。

[0045] 本发明中,随机森林算法与bagging算法类似,均是基于Bootstrap方法重采样,产生多个训练集。Bootstrap采样方法是通过对原始样本进行有放回的多次抽样,每次抽样后都会通过计算得到相应的统计量和估计值,经过计算方差得到统计量的稳定性,得到样本的真实分布。随机森林算法在构建决策树的时候,采用了随机选取分裂属性集的方法。其目标就是用随机方式建立一个由多个决策树组成的森林,随机森林中的每棵决策树间是没有关联的。当测试数据进入随机森林时,本质上就是让每一棵决策树进行分类,最后取所有决策树中国分类结果最多的那类作为最终结果。

[0046] 随机森林基本思想:首先利用Bootstrap方法进行重采样,随机产生T个训练集;然后利用每个训练集,生成对应的T个决策树,在每个非叶子节点上选择属性,从M个属性中随机抽取m个属性作为当前节点的分裂属性集,并以这m个属性中最好的分裂方式对该节点进行分裂,且每棵树都必须完整生长,并不进行剪枝;其次对于测试样本,利用每棵决策树进行测试,从而得到对应的类别;最后采用投票的方式,将得到的所有决策树中输出类别最多的作为测试样本的输出类别。

[0047] 本技术在采用随机森林分类器进行分类时,考虑到分类器参数对其性能的影响采用网格搜索算法通过将变量区域网格化,遍历所有网格点,求解满足约束函数的目标函数值,最终比较选择出最优的决策树数量以及分裂属性数。在考虑建模时间花销的尽量小的情况下,实验设置决策树数量的初始范围为[10,50],步长为10,分裂属性数初始范围[1,12],步长为1。实验表明参数分别为20和4时,模型预测效果相对较好。

[0048] 其中,步骤S3采用经过网格搜索算法调参的随机森林分类器进行分类,具体包括如下步骤:

[0049] 步骤S301:需要对随机森林分类器的两个参数决策树数目以及分裂属性数设定相应的范围和步长,并分别以这两个参数作为横纵坐标轴,建立二维坐标系,通过在坐标系中不断取点得到二维网格;

[0050] 步骤S302:将网格中节点的每一数对分别作为参数的取值构建随机森林,并采用交叉验证估计分类误差;

[0051] 步骤S303:选择分类误差最小的最优参数组合,直至分类误差或者步长满足要求为止,输出此组合,否则,尝试缩短步长,重复步骤S301。

[0052] 随机森林由于在决策树的训练过程中引入了随机属性选择使得模型具有较高的抗噪能力,同时由于随机性的引入避免了模型过拟合问题的出现。

[0053] 对于分类器的最终选择取,本发明考虑到通过组合几种模型来提高机器学习的集成方法,可以提供更好的预测结果,最终选择随机森林分类器。随机森林分类器由于其随机选择特征子集的特性,从而进一步达到对树的随机化目的,避免了分类器过拟合问题的出现。此技术最终提升了软件缺陷预测性能以及预测效率,为现实中预测有缺陷软件提供了良好的理论和实验依据。

[0054] 本发明着重考虑软件缺陷数据集中由于软件度量产生的数据高维性以及软件模块的隐藏性造成的数据类不平衡问题,构建了PCS-RF软件缺陷预测模型,此模型具体流程如下图2所示。

[0055] 首先利用基于PCA的特征提取算法对类不平衡的软件缺陷数据集进行降维以及去除冗余。运用PCA进行特征提取的的目标主要是根据贡献率的大小来确定d(d<m)个新特征,

使得这些新的特征仍然可以有效的反映样本数据的主要特征，同时压缩原有数据矩阵的规模，降低计算量。进而，将新的样本数据集用于SMOTE和无放回分层随机采样相结合，保证子样本类分布不变的同时增加少类样本，通过对其采样率的设定来降低样本容量，从而生成新的样本数据集。最后，针对随机森林中决策树数目N_{tree}以及分裂属性树数$N_{feature}$采用网格搜索算法对森林分类器进行调优，有效提升了随机森林分类器性能。

[0056] 为了说明本发明所提出方法的有效性，采用多个指标对模型性能进行了评估。实验结果表明，在准确率、F-measure、AUC值、Balance值上效果都有所提升，特别表现在F-measure值和Balance值。PCS-RF方法在F-measure指标的均值上高达0.9，较同类算法相比，至少提高了15.38％，AUC值至少提高了6.74％，高达0.95。这说明PCS-RF算法削弱了类不平衡数据的对于软件缺陷预测的干扰，为软件的安全性预测分析提供了良好的理论依据。

[0057] 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。

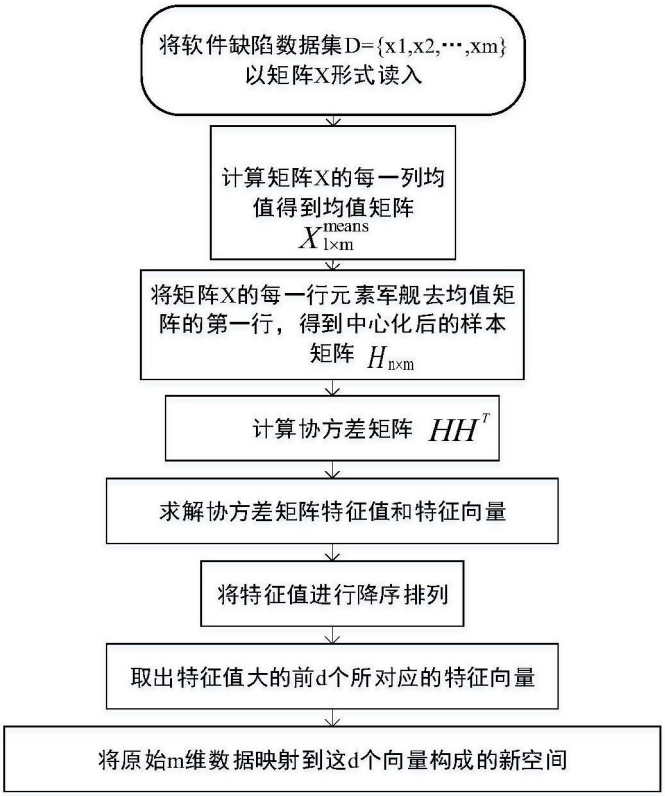

将软件缺陷数据集D={x1,x2,…,xm}以矩阵X形式读入
计算矩阵X的每一列均值得到均值矩阵 $X^{means}_{1\times m}$
将矩阵X的每一行元素军舰去均值矩阵的第一行，得到中心化后的样本矩阵 $H_{n\times m}$
计算协方差矩阵 HH^T
求解协方差矩阵特征值和特征向量
将特征值进行降序排列
取出特征值大的前d个所对应的特征向量
将原始m维数据映射到这d个向量构成的新空间

图1

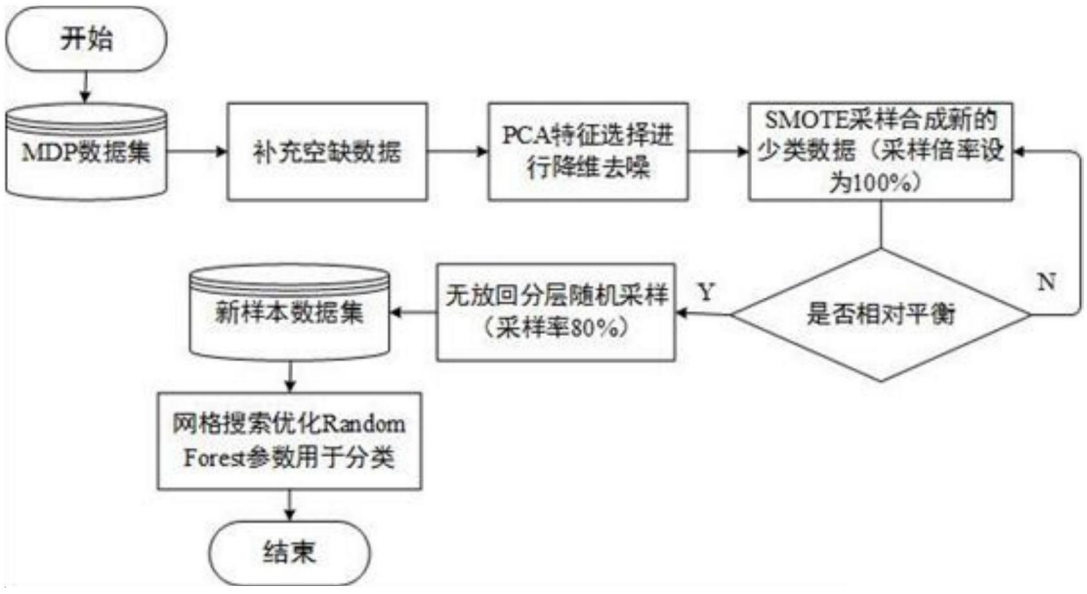

开始
MDP数据集
补充空缺数据
PCA特征选择进行降维去噪
SMOTE采样合成新的少类数据（采样倍率设为100%）
是否相对平衡
Y
N
无放回分层随机采样（采样率80%）
新样本数据集
网格搜索优化Random Forest参数用于分类
结束

图2