基于特征选择的软件缺陷预测方法

蒋　帅
(郑州工业应用技术学院 信息工程学院,河南 郑州 451100)

摘　　要:软件缺陷对软件功能的实现具有不可预知的危险,是软件产品的固有成分,提高软件的可靠性,关键在于降低软件缺陷出现的概率,而如何利用已有缺陷数据构建预测系统框架是研究的重点. 针对传统软件测试技术虽然能够有效发现软件缺陷,但需要消耗大量的时间和精力,制约软件开发效率的缺点,提出基于特征选择的软件缺陷预测方法,算法对软件缺陷模型的经验数据集进行多特征选择,进而克服数据集之间的冗余性移除无关特征,得到缺陷模型的分类,最终实现软件缺陷的精确预测. 实验表明,基于特征选择的软件缺陷预测方法具有较好的预测效果和较高的应用价值.

关　键　词:特征选择;软件缺陷;缺陷模式
中图分类号:TP311.55　　**文献标识码**:A　　**文章编号**:1673 - 1670(2019)05 - 0040 - 04

随着数字化时代的到来和计算机技术的飞速发展,计算机软件的数量、规模以及更新迭代都在发生着日新月异的变化. 快速预测并发现软件缺陷,可以大大提高工作效率,进而节省人力和资源成本[1]. 但是由于计算机软件的复杂程度在不断增加,软件缺陷受各方面因素的相互影响,使得传统的软件测试技术难以有效地处理多层次的因果关系,具有很大的不确定性. 另一方面,软件数量、规模发展的同时,积累了大量缺陷数据,如何利用大规模的软件缺陷数据集构建预测系统模型,总结缺陷模式,提高软件的可靠性是研究的重点[2]. 基于特征选择的软件缺陷预测方法,一方面可以利用不同数据集之间的互补性提高软件缺陷预测效率,另一方面可以克服数据集之间的冗余性,移除无关特征,进而克服维数灾难,降低计算量,提高效率[3].

1　软件缺陷预测

1.1　软件缺陷定义及成因

对于软件缺陷没有一个统一的说法,广泛意义上只要没有满足用户的期望或者给用户带来不便,都可以理解为软件缺陷[4]. 软件缺陷存在于软件开发的任一时期,是软件失效、故障以及错误的根源. 由于软件规模的庞大和复杂程度的增加,在软件开发的各个阶段,很多原因都会引起软件缺陷,如:没有完全了解用户的意图进而做出错误的需求分析,进行自以为是的软件设计;编程不规范;团队成员之间模块的开发不能很好地进行衔接;软件模块的增加不能很好地与系统进行兼容,维护不便;系统算法不合理,系统整体稳定性较差等[5]. 软件错误、缺陷和故障之间的关系如图 1 所示.

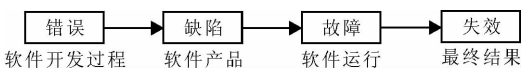

图 1　软件缺陷关系

1.2　软件缺陷模式

模式为抽象概念,软件缺陷模式为相似的或规律性的软件缺陷重复性出现,总结的规律. 已有的软件缺陷检测方法是基于整个软件的,是从缺陷的性质进行分类,其目的是为了强化对软件缺陷的认识进而在以后的软件开发中规避同类问题的出现,但是这种方法无法针对庞大的软件体系,没有统一的评判标准[6]. 软件缺陷模式是软件缺陷检测的基础,在这个基础上可以抽取归纳相应的缺陷模式. 软件缺陷模式的获取主要分为两个步骤:缺陷数据的收集和缺陷模式的分类. 在同等算法下缺陷模式越多,可提供的缺陷预测信息就越多,检测能

收稿日期:2019 - 07 - 11
基金项目:河南省科技攻关计划项目(162102210119)
作者简介:蒋　帅(1984—　),男,河南省新郑市人,硕士,郑州工业应用技术学院信息工程学院高级实验师,主要从事软件工程方面的研究.

力就越强[7]. 软件缺陷模式获取如图 2 所示.

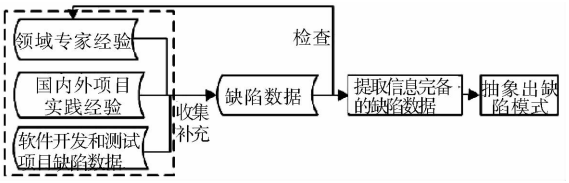

图 2　软件缺陷模式获取

1.3　软件可靠性准则

软件可靠性是系统稳定的主要因素,影响软件可靠性的原因主要有:软件缺陷、编码不规范和未知因素带来的应变能力缺失,即系统鲁棒性差. 为了提高软件可靠性,需要进行软件缺陷预测(SDP),避免软件运行故障[8]. 在进行软件缺陷预测(SDP)时,首先要了解软件可靠性准则,即模块化设计、简单原则、严格按照标准流程排除开发人员个人因素、可实现性设计,同时随着软件规模的不断升级,还要有冗余设计,即余量设计[9]. 因此,在进行软件开发时,应严格按照软件可靠准则进行设计,尽可能降低软件复杂度,保证数据之间转换的准确性,同时在数据传递过程中减少数据的丢失,进而保证系统的稳定性[10].

2　软件数据的特征选择

2.1　面向缺陷预测的软件度量

软件缺陷预测主要有 4 个步骤:首先,标记缺陷模块类别,将软件模块划分为两个集合,即有缺陷集和无缺陷集;其次,根据软件度量方法得到有缺陷模块的属性信息;继而,利用得到的缺陷模块属性信息,通过机器学习训练分类器;最后,利用训练得到的分类器进行软件缺陷预测,得到新模块的属性. 这里面涉及两个重要因素,即软件度量和通过机器学习特征训练得到分类器,进行新模块的属性预判. 为了得到高质量的训练集进行机器学习,需要得到更具有判别性和规律性的度量元集合作为训练特征,常用的训练特征有:类的加权方法数、类之间的耦合性、类的响应次数、属性个数等[11]. 为了评估互信息特征选择在软件缺陷测试中的性能,采用均方根误差作为回归模型的度量指标:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{pred,i} - x_{actual,i})^2}{n}}. \quad (1)$$

式(1) 中,第 i 个模块的缺陷个数估计值用 $x_{pred,i}$ 表示,模块缺陷个数的实际值则用 $x_{actual,i}$ 表示.

2.2　特征选择

特征选择是从原始特征空间 N 个特征中,抽取 m 个特征子集,其中 $m < N$,m 个特征子集类间差越大越能代表系统的完整性,特征选择是为了在减少计算量的同时得到高质量的训练集,进而进行机器学习,从而利用训练得到的分类器对未知模块进行缺陷预测[12]. 特征选择首先需要对已有的缺陷数据进行聚类和标记;继而选择合适的度量元,特征的选择关系到分类器的准确性,对预测模型的建立具有十分重要的意义,度量元则是特征选择的关键,选择信息量大的度量元可以提高模型的预测能力,进而准确地对缺陷进行预测;最终通过机器学习建立预测模型[13]. 特征选择的步骤如图 3 所示.

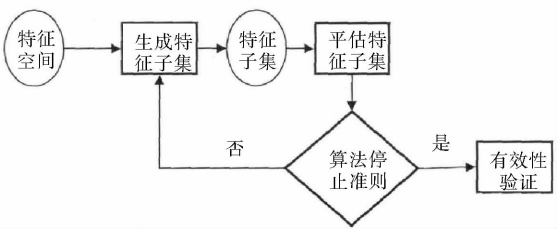

图 3　特征选择步骤

3　基于互信息特征选择的软件缺陷预测

为了提高软件质量,需要及时高效地预测软件缺陷,由于各个缺陷模块之间的冗余性,完整缺陷数据集并不能很好作为软件缺陷预测的依据[14]. 特征选择算法依据相应的准则从原始特征空间选择出代表性的特征子集作为分类依据,提高软件缺陷预测的准确性[15]. 最优子集是软件缺陷预测的重点,如何得到最优子集关系到算法的优劣,在此基础上提出基于互信息的特征选择,一方面去除子集之间的冗余性,另一方面最大限度地保留子集之间的互补性[16]. 子集之间的冗余性和互补性通过信息熵表示,信息熵是指随机变量的不确定性测度,数学表示为:

$$S = k\log W. \quad (2)$$

式(2) 中,S 为信息熵,W 为状态数,k 为玻尔兹曼常数.

子集之间的冗余性和互补性可以利用联合熵表示,联合熵表示两个以上随机变量具有共同信息量的部分,即冗余性. 假设存在两个随机变量 X 和

Y,联合熵 $H(X,Y)$ 的数学表示为:

$$H(X,Y) = -\sum_{(x,y)\in\chi\times\gamma} p(x,y)\log p(x,y) . \quad (3)$$

式(3)中,$\chi\times\gamma$ 为随机变量的值域空间,$p(x,y)$ 为随机变量的概率分布函数.

互信息是指两个或多个随机变量之间的依赖关系,即数据的互补性. 互信息的值代表多个随机变量含有数据集信息的量,假设存在两个随机变量 X 和 Y,互信息 $I(X,Y)$ 的数学表示为:

$$I(X,Y) = \sum_{x\in\chi}\sum_{y\in\gamma} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} . \quad (4)$$

式(4)中,$p(x)$,$p(y)$ 为随机变量 X 和 Y 边缘概率分布函数,$p(x,y)$ 为联合概率密度函数.

将信息熵和互信息通过数学方法结合起来,可以表示为:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) . \quad (5)$$

通过式(5)可知,随机变量 X,Y 互信息 $I(X,Y)$ 为随机变量 X 的熵加上随机变量 Y 的熵,同时去除随机变量 X 和 Y 的联合熵,且 $H(X)+H(Y)\geqslant H(X,Y)$. 通过信息熵和互信息可以确定随机变量之间的冗余性和互补性,进而得到最优子集,继而对分类器进行训练,从而实现软件缺陷预测.

具体算法流程为:

1)对采集的缺陷样本数据进行预处理,并进行参数初始化;

2)通过公式(4)计算每个特征与分类特征之间的互信息,即样本特征和分类特征的类间距离;

3)计算样本特征与分类特征之间的互信息,并比较它们的值,得出互信息最大的样本特征,将其加入最优子集;

4)重复步骤 3),计算样本中每一个特征与分类特征之间的互信息,直到最优子集数量满足算法设定的符合要求精度的阈值,进而进行分类器训练;

5)利用分类器实现对未知软件模块准确预测.

4 实验与仿真

实验仿真主要由两步组成:第一步,选择实验数据集,然后对实验数据集进行预处理,用互信息特征选择算法从实验数据集中筛选出最优特征子集,笔者从 PROMISE 开源的数据集中选取 4 种实验数据集,软件缺陷特征则使用数据集 CK metrics 的 10 个缺陷特征;第二步,利用第一步得到的最优子集,通过机器学习对分类器进行训练,建立软件预测模型,然后对数据集进行软件缺陷预测,进而验证本文算法的有效性. 通过表 1 可知实验数据集的具体信息.

表 1 实验数据集

数据集	模块数	缺陷模块数	缺陷占比/%
Ivy	103	61	59.2
Ant	120	17	14.1
Xalan	612	121	19.8
Jedit	485	23	4.74

实验结果与分析:对进行预处理的数据集利用互信息特征选择算法,筛选最优特征子集,然后利用 CK metrics 的 10 个缺陷特征对分类器进行训练,进而建立软件缺陷预测模型,继而对 PROMISE 数据集中的缺陷模块数进行识别,来验证算法的有效性. 为了增加算法的广度和深度,将本文算法与其他算法进行比较,具体结果如表 2 所示.

表 2 不同算法缺陷检测精确度 %

算法	数据集			
	Ivy	Ant	Xalan	Jedit
本文算法	83.64	81.32	82.61	79.31
基于关联度特征选择算法(CFS)	78.64	76.31	77.14	75.23
决策树回归算法(DTR)	79.62	77.35	78.24	74.21

通过表 2 可知,基于互信息的特征选择算法与基于关联度特征选择算法和决策回归算法相比,在不同数据集上对缺陷检测的精度都具有明显优势,有一定的广度和深度,但在缺陷检测的时效性上仍需做进一步验证.

5 结论

随着软件规模和数量的增加,软件缺陷也在不断增加,已有的软件缺陷检测技术难以在精度和效率上满足人们的期望,基于此提出基于特征选择的软件缺陷检测方法. 利用大量的软件缺陷特征,通过互信息和信息熵进行特征选择,得到最优特征子集,继而对分类器进行训练,建立预测模型. 实验分析可知,算法在精确度上要优于其他检测算法,而扩充完善缺陷模式,提高缺陷自动检测的可靠性,

合理有效地为软件开发服务是下一步工作的重点.

参考文献:

[1]王青,伍书剑,李明树. 软件缺陷预测技术[J]. 软件学报, 2008, 19(7):1565 - 1580.

[2]聂林波, 刘孟仁. 软件缺陷分类的研究[J]. 计算机应用研究, 2004, 21(6):84 - 86,98.

[3]尹相乐, 马力, 关昕. 软件缺陷分类的研究[J]. 计算机工程与设计, 2008, 29(19):4910 - 4913.

[4]陈翔, 顾庆, 刘望舒,等. 静态软件缺陷预测方法研究[J]. 软件学报, 2016, 27(1):1 - 25.

[5]杨朝红,宫云战,肖庆,等. 基于软件缺陷模型的测试系统[J]. 北京邮电大学学报(自然科学版), 2008, 31(5):1 - 4.

[6]姜慧研,宗茂,刘相莹. 基于 ACO - SVM 的软件缺陷预测模型的研究[J]. 计算机学报,2011, 34(6):1148 - 1154.

[7]刘望舒,陈翔,顾庆,等. 一种面向软件缺陷预测的可容忍噪声的特征选择框架[J]. 计算机学报, 2018, 41(3):506 - 520.

[8]刘海,郝克刚. 软件缺陷原因分析方法[J]. 计算机科学, 2009, 36(1):242 - 243,251.

[9]王斌,吴太文,胡培培. 软件缺陷分类和分析研究[J]. 计算机科学, 2013, 40(9):16 - 20,24.

[10]蒋乐天,徐国治. 软件缺陷及软件可靠性技术[J]. 计算机仿真, 2004, 21(2):141 - 144.

[11]李宁,李战怀. 软件缺陷数据处理研究综述[J]. 计算机科学, 2009, 36(8):21 - 25,78.

[12]王涛,李伟华,刘尊,等. 基于支持向量机的软件缺陷预测模型[J]. 西北工业大学学报,2011, 29(6):864 - 870.

[13]LIU S ,CHEN X ,LIU W,et al. FECAR:A Feature Selection Framework for Software Defect Prediction[C]. Vasteras:38th Annual International Computer Software and Applications Conference, 2014.

[14]XU Z,LIU J, YANG Z J,et al. The Impact of Feature Selection on Defect Prediction Performance:An Empirical Comparison[C]. Ottawa:27th International Symposium on Software Reliability Engineering Workshops,2016.

[15]XU Z,XUAN J F,LIU J,et al. MICHAC:Defect Prediction via Feature Selection based on Maximal Information Coefficient with Hierarchical Agglomerative Clustering [C]. Osaka:IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering,2016.

[16]刘望舒, 陈翔, 顾庆,等. 软件缺陷预测中基于聚类分析的特征选择方法[J]. 中国科学:信息科学, 2016, 46(9):1298 - 1320.

(责任编辑:王彦江)

Research on Software Defect Prediction Method Based on Feature Selection

JIANG Shuai

(School of Information Engineering, Zhengzhou University of Industry Technology, Zhengzhou, Henan 451100,China)

Abstract:Software defects have unpredictable dangers to the implementation of software functions, and are inherent components of software products. The key to improving the reliability of software lies in reducing the probability of software defects,and how to use the existing defect data to construct the prediction system framework is the focus of research. Although the traditional software testing technology can effectively find software defects, it needs a lot of time and effort so as to limit software development efficiency. So this paper proposes a software defect prediction method based on multi-feature selection. The algorithm performs multi-feature selection on the empirical data set of the software defect model, and overcomes the redundant removal irrelevant features between the data sets, and obtains the classification of the defect model, and finally realizes the accurate prediction of software defects. Experiments show that the software defect prediction method based on multi-feature selection has better prediction effect and higher application value.

Key words:feature selection,software defect,defect mode