基于支持向量机的软件缺陷预测模型——王涛

摘要

软件缺陷预测在软件系统开发的各个阶段发挥着极为重要的作用。利用机器学习的相关方

法建立更好的预测模型已经被广泛研究。文章分析了支持向量机 SVM 作为二值分类模型应用到软

件缺陷预测中的实现方法，构造了基于 SVM 的可迭代增强的缺陷预测模型 SVM-DP。在 13 个基准数

据集上开展比较实验，定量地分析了应用各种核函数对 SVM-DP 模型性能的影响。实验结果显示，应

用线性内积核函数的 SVM-DP 具有最优的预测性能。同时，在与 J48 的比较实验中，最高超过 J48 预

测模型 20% 的性能进一步证明了 SVM-DP 模型应用于软件缺陷预测的有效性。

引言

软件测试是一项极其耗费成本的工作。通常认

为缺陷在软件系统中的分布是不均匀的，20% 的模

块包含超过 80% 的缺陷。因此， 软件缺陷预测的目

的，是在软件系统开发生命周期的各个阶段鉴别可

能存在缺陷的模块， 为测试人员提供缺陷分布等信

息，以此指导软件质量确保工作，这一流程已经成为

事实上的工业标准［1］ 。机器学习的方法将缺陷预

测看作二值分类问题。模型构造完成后， 对分析目

标( 可能是文件、模块、方法等) 进行分类预测， 从而

在一定条件下做出判决。B． Turhan 等人［2］ 提出了

一种多变量方法与贝叶斯理论结合的预测模型， 并

且强调了利用特征选取技术对静态代码属性进行筛

选的重要性。实验结果显示， 其预测模型有很好的

性能。T． Gyimothy 等人［3］ 分析并使用了逻辑回归、

线性回归、决策树及神经网络等四种方法构造预测

模型，并且试图找出最优的代码静态属性。在二值

分类方面支持向量机 SVM 一直被广泛采用［4］ ， SVM

是建立在统计学习理论的 VC 维理论和结构风险最

小原理基础上的， 根据有限的样本信息在模型的复

杂性和学习能力之间寻求平衡， 以期获得最好的泛

化能力［5］ 。本文分析了利用 SVM 构造缺陷预测模

型的理论方法， 构造了基于 SVM 的预测模型 SVMDP。分析了利用不同核函数对 SVM-DP 性能的影

响，并且与已知的其它方法进行了比较。

软件缺陷预测技术∗ 王青

摘要

软件缺陷预测技术从 20 世纪 70 年代发展至今,一直是软件工程领域最活跃的内容之一,在分析软件  
质量、平衡软件成本方面起着重要的作用.研究和讨论了软件缺陷预测技术的起源、发展和当前所面临的挑战,  
对主流的缺陷预测技术进行了分类讨论和比较,并对典型的软件缺陷的分布模型给出了案例研究.

引言

所谓缺陷(defect),目前为止,学术界、产业界有很多相关的术语和定义,比如故障、缺陷、 bug、错误、失误、  
失效、失败等.根据 ISO 9000 对缺陷的定义“未满足与预期或者规定用途有关的要求”,缺陷是软件中已经存在的一个部分,可以通过修改软件而消除.另一个重要的概念是失效(failure).当系统或者软件运行时,出现不正确的输出,则称为失效.严格地说,失效可能由软件缺陷引起,也可能由其他诸如人为因素、硬件故障等引起.如果我们每观察到一次失效,就可以发现一个或者多个软件缺陷,那么纠正这些缺陷就可以避免类似失效的重复出现.软件已经成为影响国民经济、军事、政治乃至社会生活的重要因素.高可靠和复杂的软件系统非常依赖于其采用的软件的可靠性.软件的缺陷是导致相关系统出错、失效、崩溃甚至机毁人亡的潜在根源.例如,1996 年6 月,欧洲“阿丽亚娜”号航天飞机因导航系统的计算机软件出现故障,致使航天飞机坠毁,造成了数亿美元的巨大损失;2005 年 3 月 31 日,欧空局的 SMART-1 月球探测器和 NASA 的“雨燕”太空望远镜的使用状态均因软件故障而受到了很大影响;2005 年 4 月,软件失灵、继而导航失误,导致耗资 1.1 亿美元的 NASA 自主交会任务DART 实验失败.

然而软件技术发展至今,任何检验、验证手段都不可能发现并排除所有的缺陷,软件作为一种无形的产物,

虽然不会磨损用坏,却随时可能因为我们不易查知的原因出现故障甚至失效.

CeBASE(Center for Empirically-Based Software Engineering)[1]是美国国家科学基金支持的经验软件工程中心,拥有世界范围著名的专家和学术机构的合作组织.2002 年,CeBASE组织了几次著名的网上研讨,并在加拿大渥太华的第 8 届度量大会(METRICS 2002)上组办了名为“What we have learned about fighting defects”的研讨会.会后根据大家研讨的结果,整理出了软件缺陷所带来的影响及缺陷分布和检测问题:软件发布后发现和修复缺陷的成本大幅增加(例如,对于一个严重问题,其成本较之需求和设计阶段增加了 100 倍);软件项目返工工作量比例居高不下,但随着过程成熟度的提高而减少;软件工程中返工工作量与缺陷、软件失效与缺陷,以及缺陷在软件模块中的分布都是符合 2-8 原则的;同行评审、评审准备工作等有助于发现更多的缺陷;有经验的开发人员可以显著减少缺陷的引入率.上述讨论结果表明,缺陷对软件质量甚至对软件经济有重要影响;同时也说明缺陷分布问题的复杂性和差异性,以及现有的缺陷预测技术在解决实际问题上的不足等.事实上,从第一个软件诞生,就伴随出现软件缺陷的检测和预测技术.检测技术在于发现缺陷,而预测技术则在于预测还未发现的缺陷.20 世纪 70 年代,出现了利用统计学习技术,根据历史数据以及已经发现的缺陷等软件度量数据预测软件系统的缺陷数目及类型.缺陷预测技术的目的在于统计计算软件系统的缺陷数、没有发现但还可能存在的缺陷数,以决定系统是否可以交付使用.缺陷预测技术为软件质量的提高和保证起着非常重要的作用,同时,也促进了软件工程技术向前大大地发展了一步.

纵观软件缺陷预测技术的发展,从 20 世纪 70 年代起,软件缺陷预测技术大体上分为静态和动态两种缺陷

预测技术,如图 1 所示.静态预测技术,主要是指基于缺陷相关的度量数据,对缺陷的数量或者分布进行预测的技术;而动态技术则是基于缺陷或者失效产生的时间,对系统缺陷随时间的分布进行预测的技术