**摘要**

软件缺陷定位通过定位程序中的缺陷所在位置为调试提供关键信息，从而提高软件调试和修复的效率，是一种重要的软件质量保障技术。在此前的研究中，基于频谱、变异和切片的软件缺陷定位方法已取得了较好的效果。然而，上述方法都存在大量语句可疑度相同的问题，需要测试人员人工确认，影响缺陷定位的效率和精确度。例如：使用Ochiai和DStar方法对Defects4J中6个项目的395个存在缺陷的版本进行缺陷定位，都有近70%的版本存在多条语句可疑度并列排名第1。针对上述问题，本文提出了一种通过语句复杂度分析提升缺陷定位精确度的方法CBFL（Complexity-based fault localization）。首先，使用SBFL、MBFL等传统方法进行缺陷定位，得到程序中语句的可疑度排名；然后，针对排名靠前的语句进行复杂度分析，提取特征以及类别标签，构建数据集；最后，构造基于语句复杂度特征的缺陷倾向性分类模型，预测排名靠前的语句中更有可能存在缺陷的位置。基于上述方法，本文实现了基于语句复杂度分析的缺陷定位工具，并在真实的缺陷数据集 Defects4J 项目上进行了仿真实验，分别与基于频谱的缺陷定位方法DStar和基于变异的缺陷定位方法MUSE进行了对比，仿真结果表明本文所提出的方法具有更好的缺陷定位效果。

Software fault localization, which is an important software quality assurance technology, provides the location of the faults in software to improve the efficiency of debugging and repairing. In previous research, software fault localization techniques, such as spectrum-based**,** mutation-based, and program slicing, have been widely used and achieved good results. However, many statements could have same suspicious values by using these techniques, which will consume large amount of manual effort to confirm and affect the accuracy of fault localization. For example, using Ochiai or DStar to locate 395 faulty versions of 6 projects in Defects4J, nearly 70% of the faulty versions have more than one suspicious statement are ranked as tied 1. To address the above problem, this paper proposes a complexity-based fault localization (CBFL) technique to improve the precision of fault localization. First, it uses traditional techniques such as spectrum-based fault localization (SBFL) and mutation-based fault localization (MBFL) to locate faults, and gets suspicious values of statements in the program; then, it performs complexity analysis for the suspicious statements ranked are top tied 1, extracts features to build the model of statements; finally, it trains a classification model of fault statements based on complexity features of statements, and predicts the position of faults in the statements that are more likely to have faults. Based on the CBFL, this paper implements a fault localization tool, and conducts simulation experiments on the Defects4J dataset. In comparing with the DStar and MUSE, the simulation results show that the precision of CBFL outperforms DStar and MUSE in terms of *Einspect*@*n* and *Exam*.

1. **引言**

近年来，随着互联网与信息技术的飞速发展，各种软件产品的数量和规模与日俱增。各行各业对软件产品的需求越来越迫切，也越来越复杂。当软件产品的数量和规模日趋庞大，发现和修复软件中缺陷的难度也就逐步增加。而软件中包含的缺陷数目与软件产品的质量息息相关，包含的缺陷数量越多，软件可靠性面临的挑战就越大。为了保证软件的可靠性，在软件项目正式发布上线之前，通常要经过科学且系统的软件测试。软件开发及测试人员需要精准找到缺陷所在位置并进行修复，直至达到上线标准才可正式发布，这个过程往往会消耗掉大量的人力、物力与时间，而且人工修复错误率较高。

In recent years, with the rapid development of the Internet and information technology, the number and scale of various software products have increased day by day. The demand for software products in all walks of life is becoming more and more urgent and more complex. As the number and scale of software products become larger, the difficulty of finding and repairing software faults gradually increases. The number of faults contained in software is closely related to the quality of software products. The more faults contained, the greater the challenge to software reliability.(The more faults a software contains, the more challenges it faces to reliability.)To ensure the reliability of the software, before the software project is officially released, it usually undergoes scientific and systematic software testing.(Before the software project is officially released and launched, in order to ensure the reliability of the software, it is necessary to go through scientific and systematic software testing.) Software developers and testers need to accurately find the location of the fault and fix it until it reaches the online standard before it can be officially released.(,only when it reaches the online standard can it be officially released.) This process often consumes a lot of manpower, material resources and time, and (on the other hand) the error rate of manual repair is high. (the error rate of manual repair of faults is high.)

软件缺陷定位通过定位程序中的缺陷所在位置为调试提供关键信息，从而提高软件调试和修复的效率，是一种重要的软件质量保障技术。在此前的研究中，基于频谱、变异和切片的软件缺陷定位方法已取得了较好的效果。然而，这些方法都面临一个十分严峻的挑战，即缺陷定位的效率和精度不高。例如，使用Ochiai和DStar方法对Defects4j中6个项目的395个存在缺陷的版本进行缺陷定位，都有近70%的版本存在多条语句可疑度并列排名第1的问题~~，并且~~~~DStar生成的并列第1的语句平均数量是22~~。这时候的定位结果仍然需要耗费大量人力进行人工确认，而这无疑加重了软件开发过程中的成本开销。因此，研究出有效提升软件缺陷定位精确度的方法，用于提升软件缺陷定位工作的效率是十分必要的。

Software fault location provides key information for debugging by locating the location of the defect in the program, thereby improving the efficiency of software debugging and repairing. It (Software fault location)is an important software quality assurance technology. In previous research, software fault location methods based on frequency spectrum, mutation, and slice have achieved good results.(better positioning effect.)However, these methods all face a very serious challenge, that is that the efficiency and accuracy of fault location are not high. (inefficient and inaccurate.)For example, using Ochiai or DStar to locate 395 faulty versions of 6 projects in Defects4J, nearly 70% of the faulty versions have more than one suspicious statement are ranked as tied 1.~~The average number of statements for 1 is 22.~~The positioning result at this time still requires a lot of manpower for manual confirmation, which undoubtedly increases the cost of the software development process. Therefore, it is very necessary to develop a method to effectively improve the accuracy of software fault location, which is used to improve the efficiency of software fault location work.

在本文中，提出了一种通过语句复杂度分析提升缺陷定位精确度的方法CBFL （Complexity-based fault localization）。通常复杂度被应用于数学和计算机导论中，用于计算可执行程序运行时所需要的时间资源和内存资源。而在本论文中，我们要提出一种新的语句复杂度，以程序中每一行代码为单位，根据所提度量元，度量该语句的复杂度作为判断缺陷语句的条件。据我们所知，这样的度量元在之前的研究中还未被提出过（尤其在缺陷定位的背景下）。

In this article, a method CBFL (Complexity-based fault localization) is proposed to improve the accuracy of fault location through statement complexity analysis. Complexity is usually used in mathematics and computer introduction (Introduction to Computers)to calculate the time resources and memory resources required by the executable program to run. In this thesis, we will propose a new statement complexity, which takes each line of code in the program as a unit, and measures the complexity of the statement as a condition for judging faulty statements according to the proposed metric. As far as we know, such a metric has not been proposed in previous researches (especially in the context of fault location).

A.解决方案（PROPOSED SOLUTION）

本文提出了一种通过语句复杂度分析提升缺陷定位精确度的方法CBFL。首先，使用传统方法进行缺陷定位，得到程序中语句的可疑度排名；然后，针对排名的语句进行复杂度分析，提取特征以及类别标签，构建数据集；最后，构造基于语句复杂度特征的缺陷倾向性分类模型，预测排名靠前的语句中更有可能存在缺陷的位置。本文实现了基于语句复杂度分析的缺陷定位方法，并在真实的缺陷数据集 Defects4J上进行了仿真实验，对每个经过缺陷定位的程序结果进行筛选，并通过使用Einspect@n和EXAM作为评价标准与原方法对比，印证了本文所提出方法的有效性。

This paper proposes a method CBFL to improve the accuracy of fault location through statement complexity analysis. First, use traditional methods to locate faults and get the suspiciousness ranking of statements in the program; then, analyze the complexity of the ranked sentences, extract features and category labels, and build a data set; finally, it trains a classification model of fault statements based on complexity features of statements, and predicts the position of faults in the statements that are more likely to have faults.

This paper implements a fault location method based on statement complexity analysis, and conducts a simulation experiment on the real fault data set Defects4J. Using CBFL to screen the results of each fault location program, and uses Einspect@n and EXAM as evaluation criteria Compared with the original method, it confirms the effectiveness of the method proposed in this article. paper.

*B.* CONTRIBUTIONS

本文的主要贡献可以总结如下。

**·** 提出一种新的方法CBFL，通过语句复杂度分析提升缺陷定位精确度。

**·** 提出一种新的语句复杂度计算方法，以程序中每一行代码为单位，根据本文所提出度量元，提取语句特征并构造缺陷倾向性分类模型

**·** 在Defects4j基准测试上对CBFL与SBFL和MBFL进行的实验评估。

本文的其余部分组织如下。第2节介绍了与本文中提出主题相关的工作。第3节概述了基于语句复杂度分析的缺陷定位方法CBFL的设计及步骤。第4节对所提出的研究问题设计实验，介绍评估标准，并根据评估标准与对照实验进行比较，分析实验结果。展示了本文提出的方法如何提高了缺陷定位结果的有效性。第5节论述了方法的有效威胁和未来发展。最后，我们在第6节做出对本文工作的总结。

1. **相关工作（RELATED WORK）**

软件缺陷定位通过定位程序中的缺陷所在位置为调试提供关键信息，从而提高软件调试和修复的效率，是一种重要的软件质量保障技术。在这一节，我们将介绍与本文提出缺陷定位方法相关的传统软件缺陷定位工作。

**2.1基于频谱的软件缺陷定位（SBFL）**

SBFL运用程序谱和测试结果分析软件中的语句，从而定位到缺陷语句所在位置。程序谱包含软件测试时程序实体的覆盖信息，测试结果是指测试用例是否执行通过。当一条语句没有被失败的测试用例覆盖时，则认为该语句为缺陷的可疑度极低，相反，如果一条语句频繁地被失败的测试用例覆盖，则认为这条语句的可疑度非常高。通常SBFL具有计算方法易操作以及消耗资源少的特点，并且该技术的定位效果较好，所以是目前软件缺陷定位问题的研究热点。

研究人员已经提出了很多基于频谱的缺陷定位方法，其中较为经典的有：由 Jone 等人最先提出的Tarantula法，以及Abreu等人随后提出的Jaccard方法和 Ochiai方法。这三种基于频谱的缺方法在缺陷定位工作中都取得了不错的定位效果。而随着研究的不断深入，Wong 等人通过进一步分析测试用例在程序实体上，成功执行的数量对缺陷定位结果的影响，得出了被成功测试用例覆盖次数越多的程序实体，其覆盖次数对可疑值的贡献度越小这一结论，从而提出了Wong1(s)、Wong2(s)和Wong3(s)三个公式。随后为了突出在成功用例数量较多时，失败用例对定位结果的影响，又提出了DStar方法，并且通过实验对比发现 DStar 方法的缺陷定位效果优于其他同类缺陷定位方法。

**2.2基于变异的软件缺陷定位（**MBFL**）**

MBFL是基于变异的缺陷定位技术。这里的变异是一种基于错误注入的软件缺陷定位方法，将一个程序采用某种规则对程序中的一处或多处进行改变，得到一个修改后的可执行的程序副本。其中使用的规则称为变异算子，得到的程序称为变异体。具体来讲，MBFL可以通过变异程序中的缺陷来修复程序，其关注的是当变异一个有缺陷的语句会对程序带来什么影响。首先，MBFL会产生多个突变算子，通过用变异算子去替换或修改程序中的语句，突变算子可以是语句，操作，变量，常量等。然后，将突变后的变异体（程序）使用测试用例执行，收集变异体覆盖信息，最后计算变异体的可疑度，一般选择可疑度最大的变异体作为对应语句的可疑度值。通常来说，如果一个正确的语句被变异，那么程序执行通过的测试用例会减少。相比之下，变异有缺陷的语句，会比变异正确的语句通过更多的测试用例。

MBFL是近几年出现的新兴缺陷定位方法，其中被提出的经典方法有：由文章[]提出的一种MBFL技术MUSE和文章[]提出的Metallaxis，MUSE和Metallaxis中被kill的定义是不一样的。在MUSE中，失败的测试用例必须变为通过才能算作一个突变体。在Metallaxis中，一个失败的测试用例只需要生成一个不同的输出(可能仍然失败)就可以算作为一个突变体。而经过研究表明，MUSE和Metallaxis这两个方法是性能最好的两种MBFL技术[]，能够有效的提高软件缺陷的定位效率，从而达到减小缺陷定位的成本的目标。

**2.3其他缺陷定位方法**

目前还有其他的缺陷定位方法，例如基于程序切片的缺陷定位方法[]，在定位工作过程中，分析程序的输出变量以及执行路径，将输出语句或者与变量有所关联的语句做程序切片，从而达到缩小程序代码规模的效果。早在1984年，程序切片用于软件缺陷定位首次引起了研究人员的注意[]。Lin等[]将动态切片与正向跟踪技术相结合，提出了一种新的缺陷定位方法，该方法在前向跟踪过程中利用了来自设计规范的信息，增强上下文以减小所构造动态切片的大小[]。谢晓媛等[]使用变质切片的概念开发了一种可扩展SBFL的框架，该框架在不需要测试预言机的情况下，可提供与现有SBFL技术相近的性能。此外，董俊华[]将程序频谱与程序切片相结合，提出了一种程序切片缺陷定位方法，并与Tarantula、Jaccard、Ochiai三种频谱方法进行对比，实验结果再次证明了程序切片缺陷定位方法能够有效提高软件缺陷的定位效率，从而达到减小缺陷定位的成本的目标。

1. **方法设计**

根据本文第一节内容，现有缺陷定位方法中最突出的问题是缺陷定位效率不高，定位结果往往存在大量语句可疑度相同，难以区分的问题。因此本文提出基于语句复杂度分析的方法，旨在区别可疑度相同的程序语句，提高现有缺陷定位方法的定位效率，减少人工审查的代码量，从而降低软件项目的成本。其步骤如图1所示。

1. 数据提取与度量元度量：根据一个已知真实的缺陷数据集进行数据提取，对程序中每条语句进行是否是真实缺陷的标记。并根据提出的语句复杂度度量元对定位到的每条语句进行度量，提取每条语句的度量元信息。每条语句的度量元都形成一个能够输入机器学习模型的特征向量。
2. 构建模型：将标记好的特征向量输入到有监督学习的软件缺陷倾向性预测模型。软件缺陷倾向性预测对应机器学习中的二分类问题，此步骤即是利用机器学习中的分类算法构建一个能够用于二分类的分类器。所构建的分类器用于判断待测语句是否有缺陷。
3. 缺陷预测：利用构建好的分类器对待测语句进行缺陷预测。（这里要不要对待测语句进行度量元度量）将待测语句输入构建好的软件缺陷倾向性预测模型，输出得到测试语句的预测结果，查看其是否有缺陷。

接下来将概述基于语句复杂度分析的缺陷定位方法CBFL的设计及步骤，并且详细介绍本文提出的度量元，在此之前的缺陷定位研究中，语句复杂度的度量元还没有被提出过。

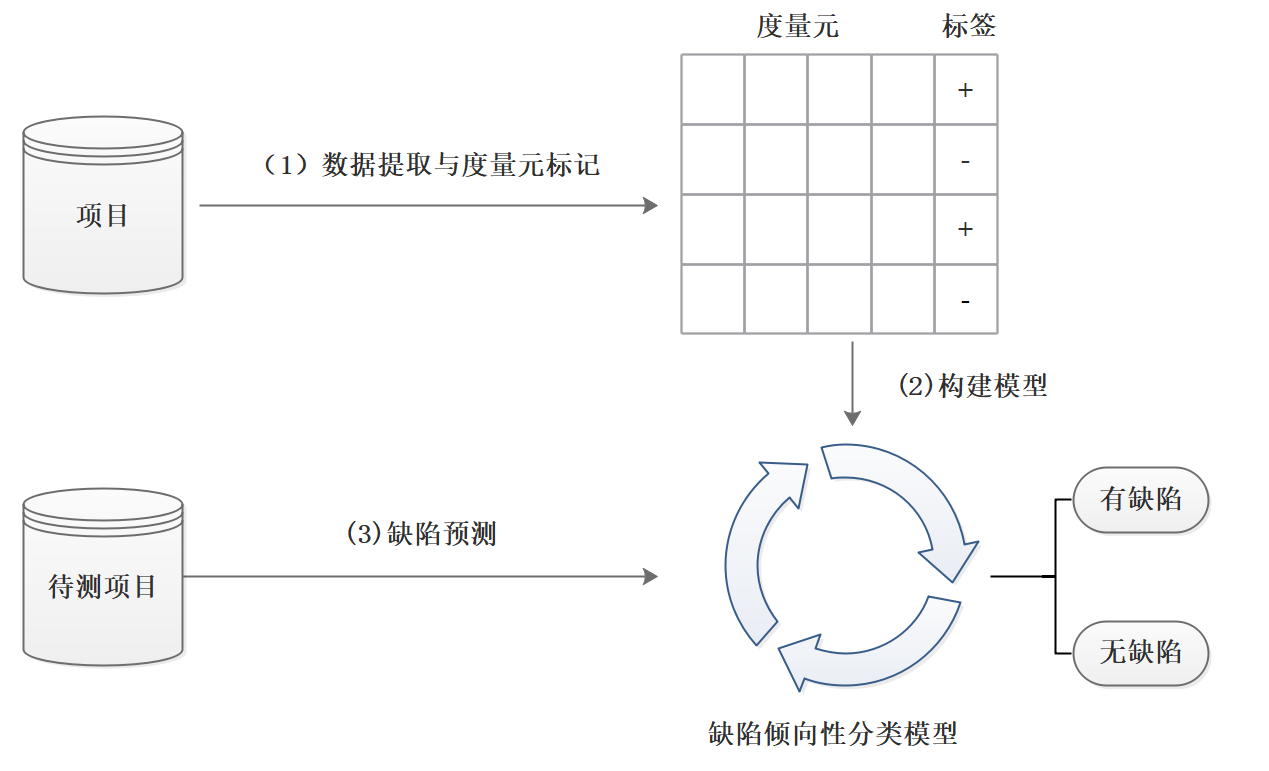


图 1软件缺陷倾向性预测技术流程图

3.1软件缺陷倾向性预测模型的构建工作

本文是在基于语句复杂度分析的基础上进行一次缺陷倾向性预测。首先要对一个已知的缺陷数据集进行数据提取，提取到程序中的每条语句，并对每条语句是否是真实缺陷进行标记。然后，根据提出的语句复杂度度量元对定位到的每条语句进行度量，提取每条语句的度量元信息。并将每条语句的度量元都形成一个能够输入机器学习模型的特征向量。如图1中（2）步骤上的表格，每一行代表一条语句，在每一行的最后标注着此语句的标签，“+”为真实缺陷，“-”为错误缺陷。每一列表示一个度量元，表示当前语句中是否含有此度量元，或含有此度量元的数量。最后，将标记好的特征向量输入到有监督学习的软件缺陷倾向性预测模型。软件缺陷倾向性预测对应机器学习中的二分类问题，此步骤即是利用机器学习中的分类算法构建一个能够用于二分类的分类器。目标是所构建的分类器能够用于判断待测语句是否有缺陷。

由于模型构建的过程对数字类型的训练数据相对友好，故本文在确定度量元属性后，根据已知数据集提供的项目版本，获取到相应的源程序语句集。利用AST树和正则表达式等方法对各个代码文件进行处理，并且得到每条语句信息，信息包括语句的所在项目文件以及行号等信息。考虑到并不是每条语句都能取到所有属性值，故而本文在度量元数据统计过程中将不存在的特征值记为 0。

3.2软件缺陷倾向性预测模型的度量元设计

在软件工程的算法中有复杂度的概念，比如用来定性描述算法运行时间的复杂度，被称为时间复杂度（Time complexity）。本文的研究对象为程序语句，并且将复杂度的概念引入到程序语句中。根据本文提出的一种新的语句复杂度度量元确定语句的复杂程度，而后根据程序语句的复杂度来判断其缺陷情况。

本文提出一种新的语句复杂度以程序中每一行代码为单位，根据度量元度量该语句的复杂度作为判断缺陷语句的条件。据我们所知，在缺陷定位的背景下这样的度量元在此前研究中还未被提出过。度量元的质量是决定缺陷预测性能最重要的因素之一。缺陷预测领域经常使用的度量元有CK度量元、McCabe度量元等，主要关注代码复杂度以及面向对象特征两方面。本文针对语句的缺陷预测设计了度量元集SC度量元(statement complexity)。

我们在设计语句复杂度的度量元时，首先考虑了程序语句中的变量、变量类型、常量、常量类型、关键字、运算符、语句长度等常见代码语句特征，又考虑到语句在整体代码中的嵌套层数也会影响语句的复杂度，所以加入了代码深度这一特征。本文使用Eclipse的jdt库将java文件转换为AST树，通过遍历AST树的方式查找指定代码语句中包含的变量、类、函数和代码深度等特征；通过编写语法分析器，解析代码语句中的括号数量、括号深度和运算符等特征。最终我们设计出了24种语句复杂度的度量元，并在多次调试后，选择了其中的12种，将其作为模型的度量元评估语句的复杂度，如表3所示。

表3 语句复杂度度量元（Metric of statement complexity）

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | implication | | employ |
| 变量（variable） |  |  |  |
| 变量数量（Number of variables） |  |  | √ |
| 变量类型（Variable type） |  |  |  |
| 变量种类（Variable types） |  |  |  |
| 常量（Constant） |  |  |  |
| 常量数量（） |  |  |  |
| 常量类型 |  |  |  |
| 运算符 |  |  |  |
| 运算符数量 |  |  | √ |
| 数组/列表 |  |  | √ |
| 关键字 |  |  |  |
| 关键字数量 |  |  | √ |
| 函数 |  |  |  |
| 函数数量 |  |  | √ |
| 类 |  |  |  |
| 类数量 |  |  | √ |
| 逻辑运算 |  |  | √ |
| 逻辑运算数量 |  |  |  |
| 代码长度（字符串） |  |  | √ |
| 代码元素数量 |  |  | √ |
| 注释（//,/\*\*/） |  |  |  |
| 括号深度 |  |  | √ |
| 括号总数 |  |  | √ |
| 代码深度 |  |  | √ |

这些度量元包括变量，运算符，数组，括号深度，括号总数，关键字，函数，类，逻辑，元素数，字数（字符串的长度），深度。其中变量varTotal表示当前语句中的变量总数；optTotal表示当前语句中运算符的数量；array表示当前语句中是否包含数组；bracketDepth表示当前语句中括号的层数；bracketTotal表示当前语句中括号的总数；keywordTotal表示当前语句中关键字的数量；methodTotal表示当前语句中包含的函数数量；typeTotal表示当前语句中包含类的数量；Logic表示当前语句中是否含有逻辑运算；lengthEle表示当前语句中包含元素的数量；lengthWorld表示当前语句的长度，包括所有空格，字符；depth表示当前语句在代码中的层数。

1. **实验与结果**

**A.**

DStar和MUSE是两种最流行的缺陷定位方法，通过检查前10个可疑元素，DStar可以定位大约43%的故障。MUSE可以通过检查前10个可疑元素定位约19%的故障。因此，我们选择DStar作为频谱方法的代表技术，MUSE作为变异方法的代表技术，使用两种方法产生的缺陷定位结果作为下一步数据集的组成。

我们提出CBFL方法实现了CDStar和CMUSE两种方法，方法分别结合了DStar和Muse。在我们的实证研究中，我们选择0.3%、0.5%、0.7%、1%、1.3%、1.7%、2%的语句作为重新排名的候选，以评估我们的方法。（为了证明以这个比值是有效的，画个图，解释以前百分之几作为数据集，能够有效的降低类不平衡问题）

为了评估该方法的有效性，本研究旨在回答以下研究问题:

RQ1: CBFL能否单独使用？怎么使用会更好？

RQ2: CBFL是否提高了缺陷定位结果的精确度？效果如何？

RQ3:与SBFL和MBFL相比，CBFL提升了多少？

**B.实验步骤**

**1）数据集（Datasets）**

为了回答研究问题，我们使用DStar和MUSE在Defects4j数据集上进行了缺陷预测。Defects4j是一个真实的故障数据集，由6个开源Java项目组成，是一个可重现错误的集合。对于每个缺陷，Defects4j提供了一个buggy版本源码和一个fix版本的源码，以及开发人员编写的测试用例集合。Defects4j数据集已经被应用到多个缺陷定位研究中，作为一个被众所周知且广泛使用的数据集，能够确保我们的结果能够具有普遍性。

表1总结了Defects4j基准的细节数据，它包含来自6个开源项目的395个缺陷版本:Chart(26)、Closure (133)、Lang(65)、Math(106)、Mockito (38)和Time(27)。在表2中，第一列是介绍项目名称，第二列是介绍该项目下有多少个版本的缺陷，第三列是每个项目的平均代码行数。第四列是每个项目的平均测试用例数量，每个缺陷版本包含一个或多个缺陷。

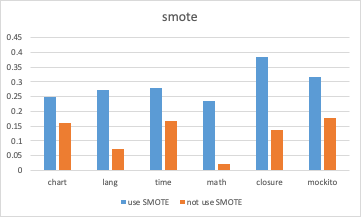
表1 **Defects4J(V1.4.0)基准的细节数据**

|  |  |  |  |
| --- | --- | --- | --- |
| Projects | Faults | LOC | Number of test cases |
| Chart | 26 | 7057 | 428 |
| Closure | 133 | 30660 | 351 |
| Lang | 65 | 1731 | 88 |
| Math | 106 | 7036 | 305 |
| Mockito | 38 | 4252 | 964 |
| Time | 27 | 3959 | 738 |
| Total | 395 | 48748 | 2874 |

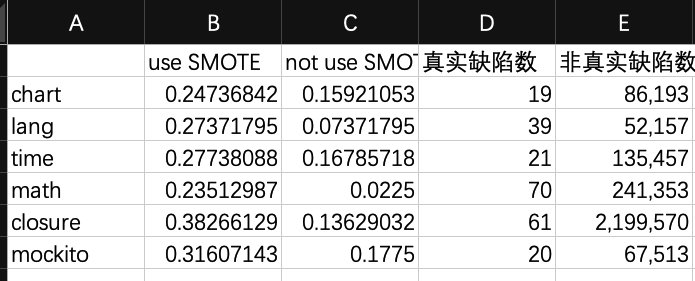
**2）**类不平衡问题处理

在现实的分类学习任务中，经常会遇到类不平衡问题。类不平衡问题是指在分类问题中，各个类别的样本数据量差距较悬殊。在不平衡数据集中，少数类的数据量占比仅为数据总量的10-20%，甚至还会更少。用类别不平衡的数据集训练所得的模型，往往会对少数类别的预测精度(例如precision和recall等)较低，在数据严重不平衡状态下训练出来的模型对于少数类而言有效性极低。在数据类别不平衡的问题上，常用的解决办法大致分为两种，即对原始数据进行过采样或者欠采样。其中，过采法（Oversampling）一般是指增加原始数据中少数类的样本数量，反之，欠采样（Undersampling）一般指通过外部干预，减少原始数据中多数类的样本数量。

在构建数据集的过程中，我们发现缺陷预测数据集存在非常严重的类不平衡现象，为缓解类不平衡问题，本方法参考文献[]中使用软件缺陷预测领域中常见的采样方法对每个缺陷版本的预测结果进行处理。具体来讲，本方法中使用SMOTE方法对缺陷预测数据集进行过采样处理。



**图3 使用SMOTE/不使用SMOTE的正负样本占比对比**

SMOTE(Synthetic minority oversampling technique)算法是实现过采样的一种方式，又叫做合成少数过采样算法。其工作原理不同于随机过采样，并不只是对原始样本中的少数类进行复制，而是通过分析该类样本数据的特点，而后根据现有样本 特性模拟形成新样本，并将模拟的新样本添加到数据集中。由图3所示，当没有使用SMOTE时，6个项目的正负样本占比最高为0.1678，最低占比的甚至不足0.023。而使用SMOTE后，正负样本占比能够达到平均约0.289，最高占比为将近0.4，已经达到足以构建模型的要求.

|  |  |
| --- | --- |
| 0.28872164 | 0.122845997 |

3）预测模型构建及度量

本文参考文献[ ]，使用软件缺陷预测领域中常用的5种二分类模型分别作为CBFL的缺陷倾向性预测模型。使用的缺陷倾向性预测模型有：高斯贝叶斯分类器（GNB）、K邻近分类器（KNNC）、决策树分类器（DTC）、随机森林分类器（RFC）和支持向量机分类器（SVC）。

实验的测试集选取，本文选择在生成的缺陷预测排名中分别取前0.3%、0.5%、0.7%、1%、1.3%、1.7%、2%作为测试数据集。对实验结果使用F1-score[]、EXAM、Einspect@n作为评估指标。

为了计算recall，需要使用如表2示的混淆矩阵对样本的预测结果进行评价，其中正样本被正确预测称为TP，负样本被错误预测称为FP，负样本被正确预测称为TN，正样本被错误预测称为FN。

表2 二分类问题的混淆矩阵

|  |  |  |
| --- | --- | --- |
| 预测样本  真实样本 | 正 | 负 |
| 正 | TP | FN |
| 负 | FP | TN |

根据表2中的混淆矩阵，召回率Recall的计算公式如公式（）所示。

精确率Precision的计算公式如公式（）所示。

本文采用了EXAM，Einspect@n来衡量SBFL，MBFL及CBFL的性能。在实际中，开发人员往往更关注排名靠前的可疑语句，当他们刚开始检查时，逐条尝试可疑度更高的语句，而当连续检查多个不是真实缺陷的语句时，程序员可能就会失去耐心并且开始跳跃着检查，甚至放弃检查。因此使用EXAM，Einspect@n作为评估指标更贴合实际。

EXAM是缺陷定位领域中常用的度量标准，它被定义为在第一个错误语句出现之前需要检查的语句的百分比。这里我们的主要目的是能够尽早的为开发人员提供真实的缺陷，所以当需要检查的语句越少，EXAM值越低，缺陷预测的效果就越好。

文章[]使用Einspect@n来评估缺陷预测方法的有效性，Einspect@n是通过检查排名前n个语句，并计算成功定位到的缺陷总数。在n越小的情况下，Einspect@n的值越大，则缺陷预测更有效。据Kochhar等人[]的研究统计，通常开发人员在进行缺陷定位时仅会检查怀疑度列表中分数最高的一些程序语句。其中，9.43%的开发人员只检查TOP-1程序语句，而 73.58%的开发人员会检查TOP-5程序语句，接近98%的开发人员认为检查TOP-10程序语句是比较合适的。TOP-1，因此我们将选择Einspect@n(n=1,5,10)作为具体的评价标准。

上述的实验均采用8折交叉验证。具体而言，将原始的据集等分为8份，使用其中7份作为该轮模型的训练集，剩余的1份作为测试集，记录本轮模型的预测结果。如此重复8次，每次选择不同的测试，以此保证每份数据被作为过测试集1次。

一张展示用10折交叉验证和8折交叉验证后真实数据量过少的情况的图。

通常使用10折交叉验证，但根据本实验数据集实际情况，（如图\*\*\*\*），使用10折交叉验证会导致数据过于集中，真实缺陷占比过少，并且无法使用过采样算法解决问题。在单独计算recall和precision时，数据过于单一，无法在实验中确定模型优异，因此我们选择降低等分的数量，来提高真实缺陷在等分后数据集中的占比。

C．EXPERIMENTAL RESULTS AND ANALYSIS（实验结果与分析）

RQ1: CBFL的缺陷数量预测模型性能比较？（比较分类器的效果）

~~为了研究在Solidity智能合约缺陷数量预测中不同度量元集和不同模型之间的性能差异，~~为了研究在CBFL方法下，模型之间的效果差异，实验将本文提出度量元构成的缺陷数量预测数据集，分别应用5种分类模型进行训练，最终得到这些模型不同的预测结果，并使用F1和recall这两个指标进行度量。实验的结果。

图\*是使用不同模型对Defect4J中6个项目前0.3%数据进行缺陷数量预测情况使用F1进行评估的结果。根据分类模型将\*图分为6组，每组为Defect4J中的一个项目，每组使用5种颜色代表不同的分类方法。

在图中，由于数据集的特点，不同分类模型在不同项目中F1评价指标上存储较大差异，并不能很好的确定分类模型的优异，我们将6个项目中的F1评价指标平均计算以统一结果，展示不同分类模型的综合效果。

图\*是将6个项目的F1评价指标平均后的结果。更具分类模型将图\*分为5个矩形，每个矩形为一种分类方法。

使用不同分类模型对缺陷数量预测情况使用所有项目平均recall进行评估的结果。根据分类模型将\*图分为5个矩形，每个矩形为一种分类方法。

为综合比较评测指标，将图\*结果进行排序，结果F1值从高到低依次为：决策树（0.163837484），贝叶斯（0.162034731），KNN（0.143978826），随机森林（0.050877193）SVC（0.115063172），由结果可知，决策树、贝叶斯和KNN效果相近，为确定最优分类模型，且在项目背景下，目的是确定更多真实缺陷，引入recall作为评测指标

图\*

图\*是使用不同模型对Defect4J中6个项目前0.3%数据进行缺陷数量预测情况使用recall进行评估的结果。根据分类模型将\*图分为6组，每组为Defect4J中的一个项目，每组使用4种颜色代表不同的分类方法。

从图\*中可以看出，在使用贝叶斯对缺陷数量进行预测时，在Defect4J的六个项目中，recall值较高，分别为：closure（0.935483871）、mockito（0.8）、time(0.761905)、chart（0.684210526）、lang（0.487179487）、math（0.428571429），

~~而使用随机森林时，平均recall值最低，分别为：mockito（0.2）、math（0.090909091）、closure（0）、chat（0）、lang（0）、time（0）~~

其他方法分别为：决策树、K近邻、支持向量机（都要列举一下么？）

以上可能还需要加点分析

由于Defect4J不同项目中，不同模型在平均recall的评价指标上存在差异，我们将6个项目中的recall评价指标平均计算以统一结果，展示不同分类模型的综合效果。，如图\*

图\*

图\*是使用不同模型对缺陷数量预测情况使用所有项目平均recall进行评估的结果。根据分类模型将\*图分为5个矩形，每个矩形为一种分类方法

从图\*中可以看出，贝叶斯的平均recall值最高，为0.68289172其余依次为支持向量机（ 0.341080393）

为综合比较评测指标，将图\*结果进行排序，结果为，bayes排名最高，为0.0484，其余依次为：SVC（0.341080393），KNN（0.310760207），tree（0.234926053），randomForest（0.048484848）

综合以上分析，我们对RQ1得出以下结论：对于缺陷数量预测，在综合考虑数据选取和SMOTE的使用后，贝叶斯的效果最好。

综合以上分析，我们对RQ1得出以下结论：对于Solidity缺陷数量预测，SC-Sol在MAE评估指标中性能最好，COOP-SC-Sol在FPA评估指标中性能最好。此外，综合考虑MAE和FPA指标的排名，RFR模型的平均性能最好，使用RFR模型时，COOP-SC-Sol的预测性能优于SC-Sol。

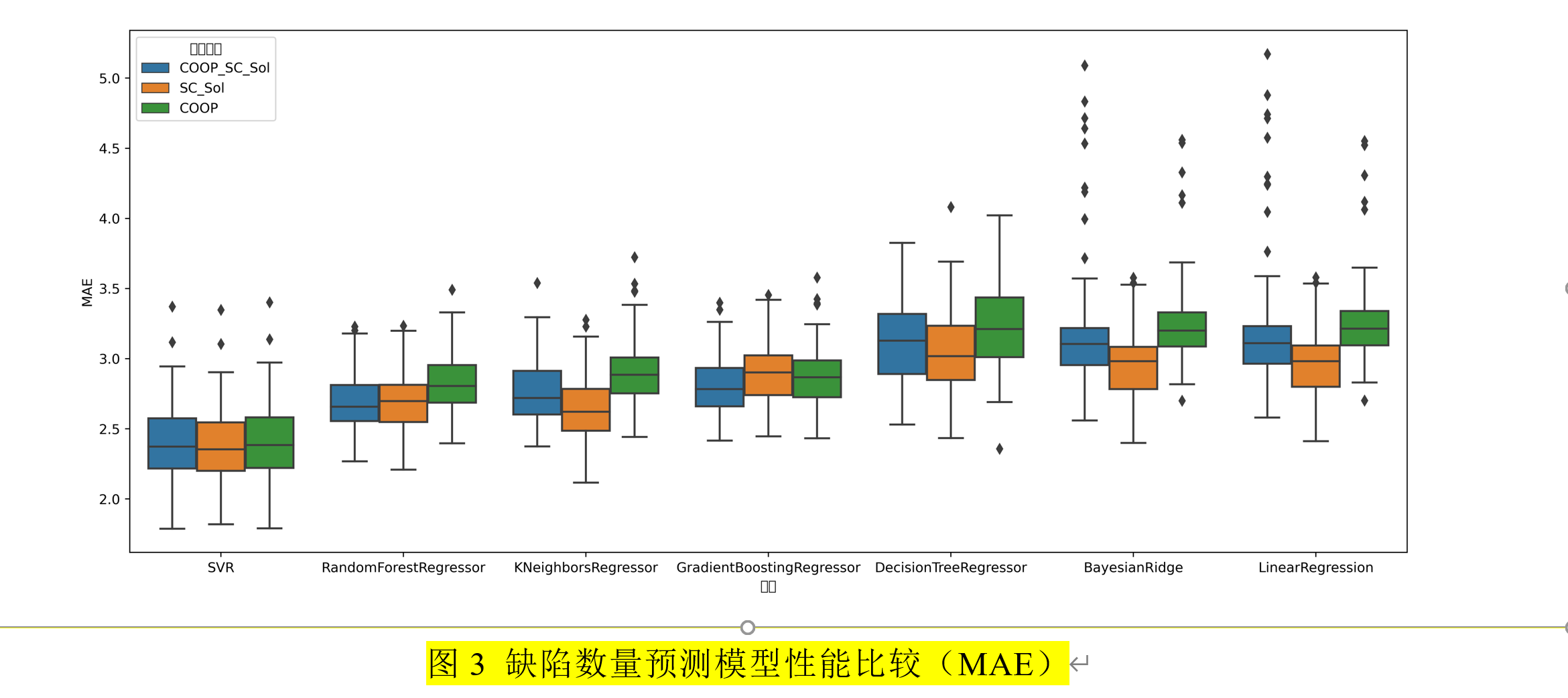
综合以上分析，我们对RQ1得出以下结论：缺陷数量预测，贝叶斯在MAE评估指标中性能最好，COOP-SC-Sol在FPA评估指标中性能最好。此外，综合考虑MAE和FPA指标的排名，RFR模型的平均性能最好，使用RFR模型时，COOP-SC-Sol的预测性能优于SC-Sol。

（这里看看要不要引入新的评估方法，做一个和本图类似的结果评估）

RQ2: CBFL能否单独使用？怎么使用会更好？

RQ3:与SBFL和MBFL相比效果如何？CBFL提升了多少？

介绍数据集对5种分类器进行了训练，并得到了结果，进行比较，使用EXAM和@n进行比较，



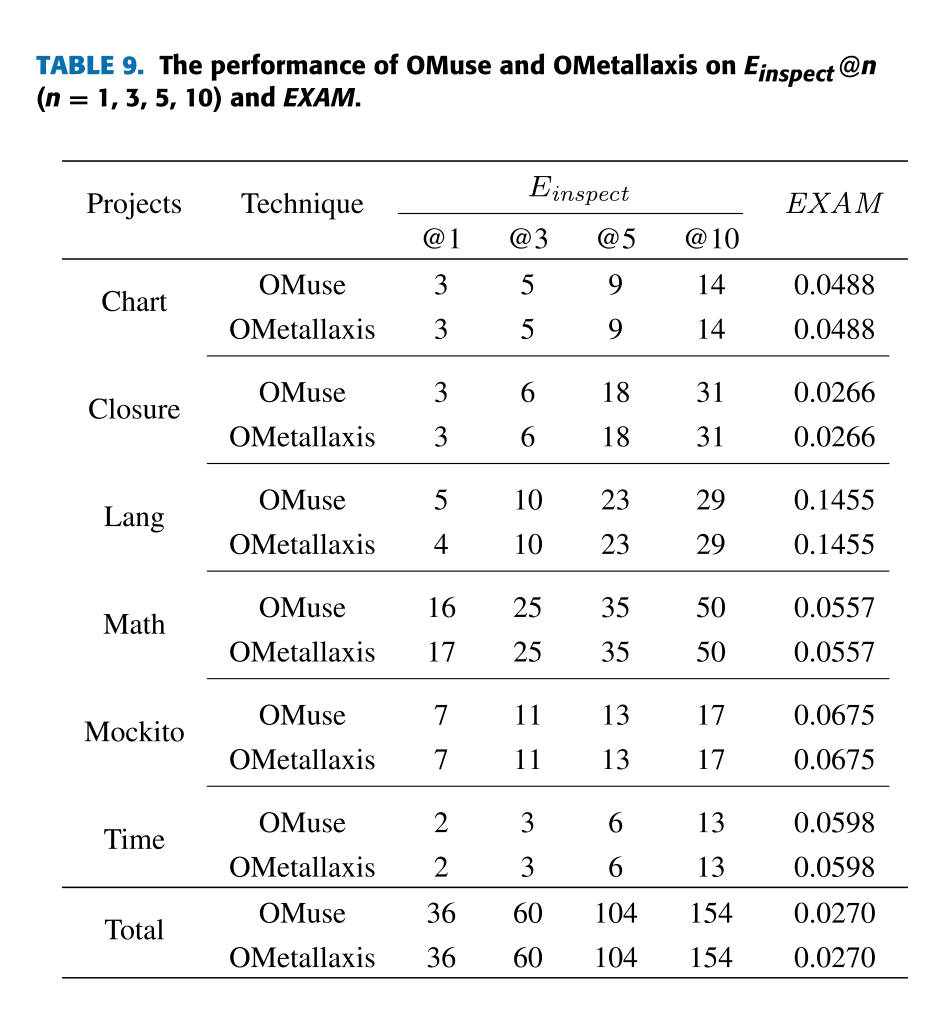
根据文章提出的结论，在缺陷结果中有多条语句具有相同的可疑度是很常见的，对于defects4j的6个项目，可疑度并列第1的语句平均数量是22，具体来说，平均有22条语句被DStar认为可疑度相同。

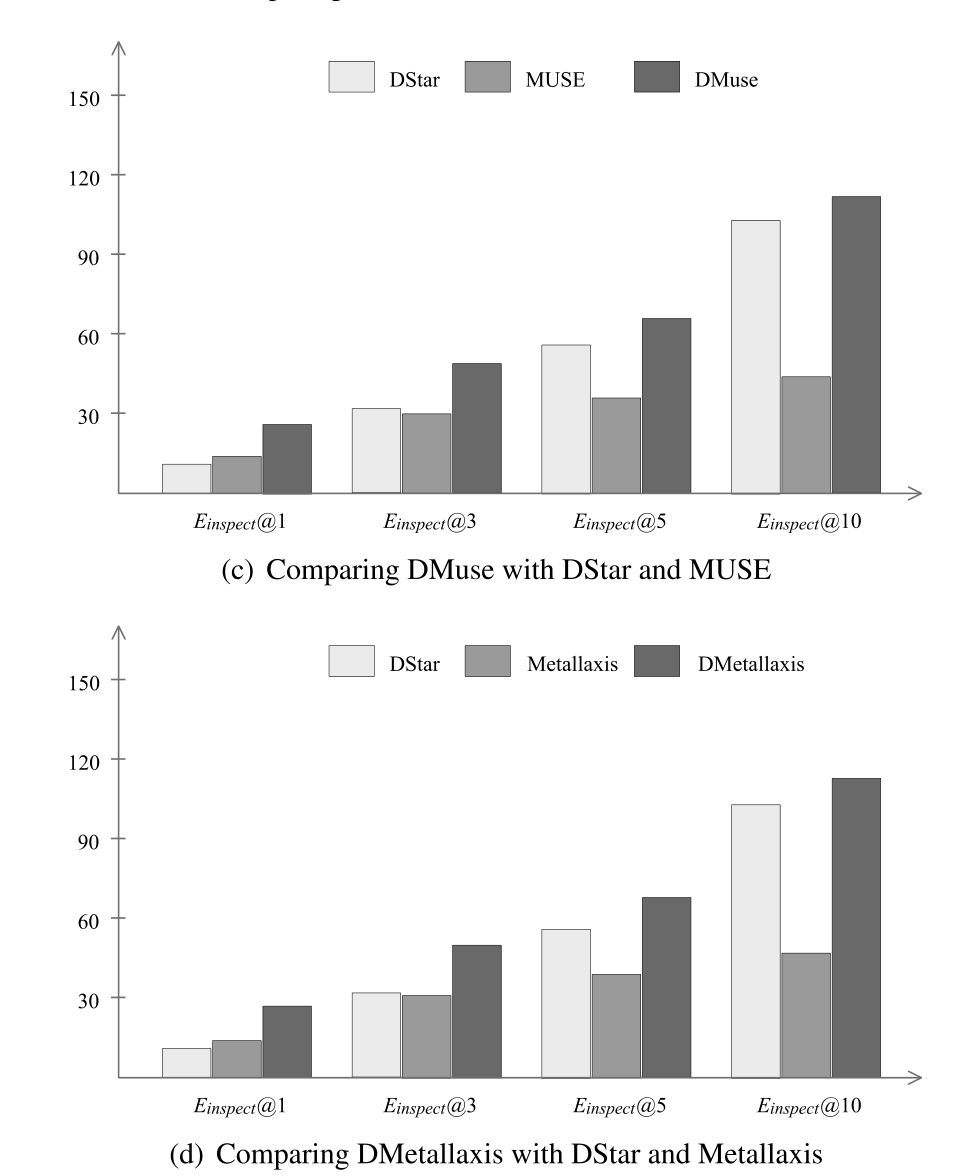
如表所示，（RQ2展示单独使用cbfl的实验结果，放出EXAM和Einspect@n）

（解释表格的信息）

RQ2未定

RQ3，比较DStar，MUSE和CDStar，CMUSE使用柱状图和表格的形式，比较，得到





CDStar，CMUSE，

Answer for RQ1:回答RQ1.

1. **讨论（DISCUSSION）**

我们的方法CBFL对于提高缺陷定位的精确度和提高程序员调试缺陷效率是有用的。本节讨论了我们研究方法的意义，以及对我们研究有效性的威胁和我们方法的局限性。

**基于已有传统缺陷定位技术优化其定位结果**

根据研究结果表明，CBFL对缺陷定位结果的优化是有效的， 把CBFL作为补充测试技术，与多种软件缺陷定位方法结合，提高缺陷定位的结果，并且。。。**（还要补充2个小节）**

5.1威胁有效性**Threats to Validity**

**External Validity（外部有效）**

我们的研究使用了Defects4J缺陷，一个真实的缺陷的集合。我们使用这个广泛使用的现实缺陷基准，旨在确保我们的结果具有普遍性。我们对Defects4J的6个项目均进行了测试，并均有提升效果（介绍具体数据）以此证明我们方法的适用性。此外，我们还随机检查了这些被定位的语句，确定它们是真实的缺陷。

**数据集的局限性**

因为训练模型的数据还不够多，尤其是目前使用了SMOTE来过采样，如果我们能够得到一个真实的类平衡的数据集去训练，是不是结果会更好。举例说明，数据集的提高对实验结果的影响时很大的。得出我们的模型可以更好。

1） 训练数据集来源：本文研究过程所用到的训练数据集是单一的，在后续研究中可以尝试多元化。使用其它的现有公开数据集或者针对特定项目进行训练数据提取。

2） 可疑度相同的语句：本文主要关注了可疑度并列第一并且其中包含真正缺陷的语句，后续研究中还可以考虑将可疑度按梯队划分，如能够分开考虑并列第一、并列第二或其它情况。

3） 模型的普适性：本文提出的模型只在 DStar 方法上进行了实验，后续研究可以考虑将模型迁移至其它已有的缺陷定位方法中，判定对于除 DStar 以外的现有软件缺陷定位方法是否具有提升效率的作用。

**5.3未来工作**

未来是否可以把这个思想投入到APR工作中去，

换用新的数据集再次训练模型，尝试得到更好的实验结果。

1. **结尾（CONCLUSION）**

典型的基于频谱的缺陷定位方法 DStar，定位结果中往往存在大量可疑值相同的语句，人工审查代码成本相对较高。针对这一问题，我们提出了CBFL，这是一种识别软件缺陷根本原因的新方法，是对现有测试和调试工具的补充。首先， 根据DStar方法在Defects4J数据集里6个项目上的定位结果，提取可疑值并列第一的程序语句，对其进行特征提取与类别标记，构造原始训练数据集。随后运用预处理后的数据集训练产生分类决策树模型，用于对程序语句进行缺陷倾向性预测，并对所构建的模型进行评估。最后，将理论研究过程所构建的缺陷预测模型，用于软件缺陷定位系统的开发，最终实现了一款用于缺陷定位工作的系统。综上所述，本文主要贡献如下：

1）提出了一组用于衡量程序语句复杂度的特征属性。能够有效分析现有缺陷定位方法中，具有相同可疑度的程序语句复杂度。

2）构建出了通过程序语句复杂度进行缺陷倾向性预测的模型。能够根据程序语句的复杂度信息，对程序语句进行分类，给出有效的缺陷分析结果。

3）设计并实现了一款用于缺陷定位工作的系统。将上述研究应用于实践，运用语句复杂度分析的方法辅助软件缺陷定位工作，减少了人工审查代码过程的工作量，提升了现有缺陷定位方法的精确度，一定程度上降低了软件项目消耗的成本。