软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于机器学习的 C 程序内存泄漏智能化检测方法*

朱亚伟, 左志强, 王林章, 李宣东

(计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

通讯作者: 王林章, E-mail:lzwang@nju.edu.cn

摘 要: 内存泄漏在采用显式内存管理机制的 C 语言中是一种常见的代码缺陷.内存泄漏的检测方法目前主要是静态分析与动态检测.动态检测开销大且高度依赖测试用例;静态分析目前被学术界和工业界广泛应用,但是存在大量误报,需要人工对检测结果进行确认.内存泄漏静态分析的误报通常是由于对指针、分支语句和全局变量分析的不准确性导致.本文提出一种内存泄漏的智能化检测方法,通过使用机器学习算法学习程序特征与内存泄漏之间的相关性,构建机器学习分类器并应用机器学习分类器进一步提高内存泄漏静态分析的准确性.本文首先构建机器学习分类器,然后通过静态分析方法构建从内存分配点开始的 Sparse Value Flow Graph(SVFG),并从中提取内存泄漏相关特征,再使用规则和机器学习分类器进行内存泄漏的检测.实验结果显示,我们的方法在分析指针、分支语句和全局变量时是有效的,能够提高内存泄漏检测的准确性,降低内存泄漏检测结果的误报和漏报.最后,我们对未来研究的可行性以及面临的挑战进行了展望.

关键词: 内存泄漏检测;静态分析;机器学习

Memory Leak Intelligent Detection Method for C Programs Based on Machine Learning

ZHU Ya-Wei, ZUO Zhi-Qiang, WANG Lin-Zhang, LI Xuan-Dong

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Memory leak is a common code bug for C programs which uses explicit memory management mechanisms. At present, static and dynamic program analysis have been used to detect the memory leaks. Dynamic program analysis has huge overhead and it is highly dependent on test cases. Static analysis, which always contains a large number of false positives and requires manual confirmation. The false positives of static program analysis of memory leaks are often caused by the inaccuracy in analyzing of pointers, branch statements, and global variables. Aiming at the weakness of static program analysis of memory leaks, we use machine leanring techniques to improve the accuracy of static analysis. In this paper, we propose an intelligent method for memory leak. We use machine learning algorithms to learn the correlation between program's features and memory leaks, and train a machine learning classifier, so that the classifier can improve the accuracy of static analysis of memory leaks. Firstly, we train a machine learning classifier. Then, the Sparse Value Flow Graph(SVFG) beginning at a memory allocation should be constructed by using the static program analysis, we can extract the features related to memory leaks from the SVFG. Lastly, detecting the memory leaks by using rules and machine learning classifier. The experimental results show that our method is effective in analyzing pointers, branch statements and global variables, and the method can reduce the false positives and false negatives of memory leak detection. At the end of this paper, we look ahead to the feasibility of future research and the upcoming challenges.

Key words: memory leak detection; static analysis; machine learning

内存泄漏会严重降低软件的性能,甚至造成软件在运行时崩溃.在 C 语言中,内存的分配与释放都是人为控制的,随着计算机软件的规模和复杂度的不断增加,人为的疏忽极易导致内存泄漏.由于 C 语言在计算机领域的广泛应用,C 程序中的内存泄漏问题不可忽视.目前内存泄漏的检测方法主要是静态分析与动态检测.

动态检测方法[10-12,21,22]需要执行程序,在程序运行的过程中对内存的分配,使用以及释放进行动态跟踪.由

于动态检测能够针对当次的运行结果得到一个明确的结论,因此动态检测相对静态分析,其结果更加准确,但是动态检测的准确性受限于测试用例,无法分析程序执行中不可达位置的错误.目前成熟的内存泄漏动态检测工具有 Purify[13]、Valgrind[14]等,都存在内存开销较高和可扩展性较差的问题.

静态分析是在不实际运行程序的情况下对程序代码及其结构进行分析.针对 C 语言的内存泄漏静态分析方法主要通过分析内存的分配点以及从内存分配点开始的不同路径,在相应的路径中查找与内存分配点对应的内存释放点,验证是否所有路径都存在正确的内存释放.目前,有许多内存泄漏的静态分析工作 [1-6],也有一些成熟的静态分析工具,如:Fortify[7]、Coverity[8]、Klocwork[9],这些工具在工业界的软件开发中应用广泛.静态分析方法又可根据流敏感、上下文敏感以及域敏感等进行分类,实现这些静态分析方法可以在一定程序上提高内存泄漏检测的准确率,但会极大降低检测效率.

内存泄漏的静态分析目前主要缺点是:当内存泄漏中存在一些特殊案例时,会降低静态分析的准确性,导致内存泄漏的检测出现误报或者漏报.其主要原因在于:

(1) 对分支条件缺少分析,无法识别不可达路径;对全局变量和指针的重定向未做细致分析;忽略指针偏移问题.可以通过流敏感、上下文敏感等静态分析方法在一定程度上解决以上问题,但会造成检测效率大幅降低.

(2) 动态数组的分配问题;链表的分配释放不匹配问题;循环或递归中指针分析的准确性问题,限制循环或递归的展开次数会导致指针分析不准确.流敏感、上下文敏感等目前已有的静态分析方法无法解决上述问题.

针对静态分析的上述不足,本文利用机器学习算法以提高内存泄漏静态分析的准确性,获得更加准确的内存泄漏检测结果.本文利用已有的内存泄漏案例,即含有标签为误报和真实内存泄漏的两类数据集,通过使用机器学习算法进行训练,构建内存泄漏分类器以分析程序特征与内存泄漏的相关性.本文的主要贡献在于:

(1) 本文提出一种基于机器学习的 C 程序内存泄漏智能化检测方法,该方法基于静态分析提取内存泄漏相关特征,能够提高内存泄漏检测的准确性,减少误报和漏报.

(2) 我们实现了内存泄漏的智能化检测工具 I_Mem.该工具使用多种机器学习算法构建内存泄漏检测模型,并基于静态分析提取内存泄漏相关特征,能够提高内存泄漏检测的准确性.

(3) 我们在 LLVM-4.0.0 上实现了 I_Mem,并利用 4 个开源的 C 程序(2KLOC)进行实验评估.I_Mem 总共发现了 114 个内存泄漏,漏报为 4 个,误报为 13 个,准确率为 85.6%.

本文第 1 节主要介绍背景知识.第 2 节介绍基于机器学习的 C 程序内存泄漏智能化检测方法,包括方法的基本框架、内存泄漏模型的构建、特征提取与缺陷检测.第 3 节对本文实现的工具 I_Mem 进行实验和评估.第 4 节主要介绍相关工作,包括内存泄露的静态分析方法、内存泄露的动态检测技术和基于机器学习的缺陷检测.第 5 节进行总结和展望.

1 背景

Table 1 Six Types of Statements

表 1 六种语句类型

名称	句法
取址	*p = &v*
赋值	*p = q*
Load	*p = *q*
Store	**p = q*
函数调用	*p = F(...,q,...)*
返回	*return p*

针对内存泄漏的程序静态分析需要在获取所有内存分配点后,关注内存的定义、使用以及释放.此外,C 语

言中,由于函数可以返回指针,内存泄漏检测需要对调用点作分析.因此,Saber[2,3]关注 C 程序中的 6 六种语句,如同表 1 所示.在表 1 中,p 和 q 是变量,v 表示变量或者堆对象,F 是函数.

Saber 中使用的静态分析方法是 Full-Sparse Value-Flow Analysis,通过构建 SVFG 来检测内存泄漏.其中,VFG[28,29](Value Flow Graph)表示的是句法上的语义等价,它关注变量的定义、使用,能够表示程序中变量的价值流向.VFG 不同于 CFG(Control Flow Graph)和 DFG(Data Flow Graph),CFG 表示的是控制程序逻辑执行的先后顺序,一个程序的 CFG 被用来确定对变量的一次赋值可能传播到程序中的哪些位置;DFG 是在 CFG 基础上实现的,它描述的是程序运行过程中数据的流转方式及其行为状态.在 VFG 中,每一个节点表示变量的定义,每条边表示变量的 def-use 关系.Saber 针对 VFG 进行改进构建了 SVFG.SVFG 的构建主要分为三个步骤:

(1) 预分析:根据内存分配相关的 API(如 malloc)确定内存位置.使用域敏感、调用点敏感、流和上文不敏感的安德森指针分析获取 C 程序的指针信息.

(2) 全稀疏 SSA(Static Single Assigment):在 SSA 形式中,每个被使用的变量都有唯一的定义,这可以确保精确地 def-use 关系链.针对所有内存位置,构造每个函数的 SSA 形式.关注内存位置的间接访问,如 load,store 以及函数调用操作.用程序内流敏感的指针分析对预分析的指针信息进一步稀疏提高指针分析的准确性.

(3) SVFG:基于全稀疏 SSA,获取程序内所有内存位置的 def-use 关系链和 value-flow 并构建 SVFG.每个 def-use 边会有一个"警卫"来获取分支条件.

2 基于机器学习的 C 程序内存泄漏智能化检测方法

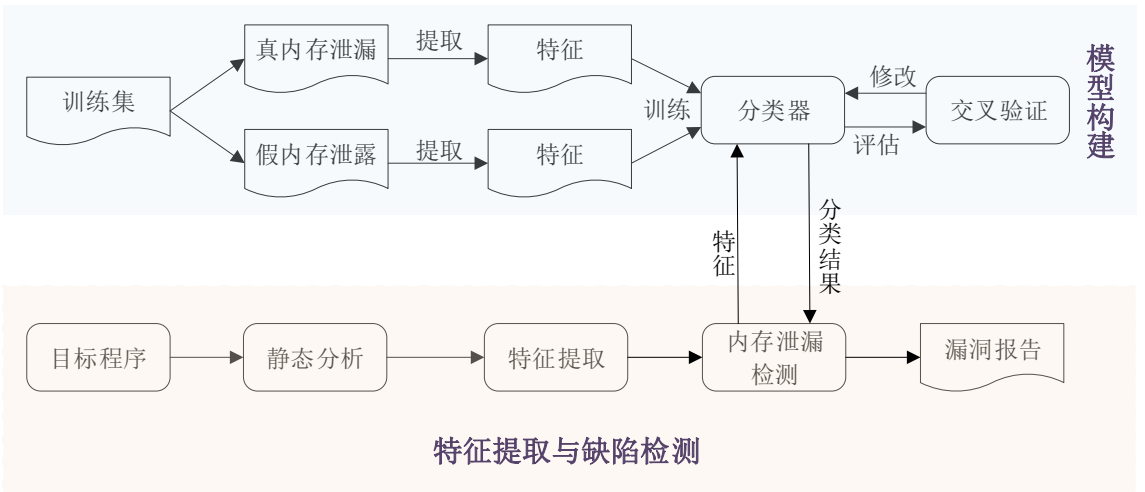


Fig.1 Framework of our work

图 1 方法框架

本节我们主要介绍基于机器学习的 C 程序内存泄漏智能化检测方法的基本框架,本方法是在内存泄漏静态分析的基础上进行改进,利用机器学习技术提高了内存泄漏静态分析的准确性.图 1 是基于机器学习的 C 程序内存泄漏智能化检测方法的主要框架,该方法主要可分为两个步骤:模型构建阶段以及特征提取与缺陷检测阶段.

(1) 模型构建阶段:首先根据已有的内存泄漏构建两个数据集,一个是包含真正的内存泄漏的数据集,另一个是包含虚假内存泄漏的数据集,然后从两个数据集中分别提取内存泄漏特征,将内存泄漏特征输入机器学习的分类器进行训练,并应用交叉验证对分类器进行评估,然后修改分类器类型及参数,选取分类效果最好的作为内存泄漏检测模型.

(2) 特征提取与缺陷检测阶段:首先利用静态分析方法分析源程序,获取所有的内存位置(内存分配点 *o*),并构建从 *o* 开始的 SVFG,提取 SVFG 中每条路径对应的内存泄漏特征并构成数据集.根据内存泄漏特征,我们可将数据集分成两部分,一部分可根据规则直接判断是否为内存泄漏,另一部分需要输入到内存泄漏检测模型中进行判断.最后将内存泄露检测结果与静态分析阶段的分配点的信息相结合,得到漏洞报告.

2.1 模型构建阶段

针对内存泄漏的静态分析中存在的不足,我们提取了 16 个内存泄漏特征.本文根据从内存分配点开始的 SVFG 提取内存泄漏特征,每一个内存分配点对应一个内存泄漏特征.提取的内存泄漏特征信息(内存分配点 *o*,内存释放点指针 *p*)包括三类:类型信息、分支信息、释放信息.类型信息包括:数组(判断 *o* 是否为数组元素)、结构体(判断 *o* 是否为结构体元素)、链表(判断 *o* 是否为链表元素).分支信息包括:循环分配(*o* 是否在循环内部分配)、循环释放(*p* 是否在循环内部释放)、循环匹配(循环分配与释放次数是否一致)、同一循环(分配与释放是否在同一循环内)、链表匹配(链表的分配与释放次数是否一致)、分支条件(若当前分支中不存在对 *o* 的释放,则判断分支条件是否为真).释放信息包括:函数间距(从 *o* 出发的 SVFG 最多经过的函数数量)、释放之前为空(*p* 释放之前是否为空指针)、*p* 指向对象数目(释放 *p* 时,*p* 指向的对象数目)、别名数目(释放 *p* 时,指向 *o* 的指针数目)、指针偏移量(释放 *p* 时,*p* 与 *o* 的偏移量)、全局指针(判断在 SVFG 中有全局变量指向 *o*)、释放(指针 *p* 是否释放). 特征如表 2 所所示.

Table 2 Features of Memory Leak

表 2 内存泄漏特征

类别	序号	特征	类型	描述
类型信息	1	数组	布尔值	*o*是否为数组元素
	2	结构体	布尔值	*o* 是否为结构体元素
	3	链表	布尔值	*o* 是否为链表元素
分支信息	4	循环分配	布尔值	o 是否在循环内部分配
	5	循环释放	布尔值	p 是否在循环内部释放
	6	循环匹配	布尔值	循环分配次数与循环释放次数是否一致
	7	同一循环	布尔值	*o* 的分配与 *p* 的释放是否在同一循环中
	8	链表匹配	布尔值	链表的分配与释放次数是否一致
	9	分支条件	布尔值	当前分支中不存在对 *o* 的释放,判断分支条件是否为真
释放信息	10	函数间距	整型	当前路径中,从分配点*o*到释放 *p* 所经过的函数数量
	11	释放之前为空	布尔值	*p* 释放之前是否为空指针
	12	*p* 指向对象数目	整型	释放 *p* 时,*p* 指向的对象数目
	13	别名数目	整型	释放 *p* 时,指向 *o* 的指针数目
	14	指针偏移量	整型	释放 *p* 时,*p* 与 *o* 的偏移量
	15	全局指针	布尔值	SVFG 中,是否有全局变量指向 *o*
	16	释放	布尔值	指针 *p* 是否释放

本文针对内存泄漏共提取 16 个特征,其中部分特征与内存泄漏静态分析的不足存在如下对应关系:

(1) 特征 9 用于对分支条件进行判断;特征 11、12、15 用于判定全局变量以及指针重定向问题;指针偏移量问题使用特征 14 进行判断.

(2) 针对内存泄漏有关的数组、链表、循环或递归问题,我们使用特征 1、3、4、5、6、7、8 进行判断.

(3) 其余特征用于对内存泄漏的一些复杂情况进行辅助判断.

确定内存泄漏相关特征后,我们需针对各种内存泄漏构建训练集.我们使用的是开源的C程序源码(包括一些大型程序和小程序),在原程序中我们会插入各种内存泄漏以构建尽量丰富的训练集,然后从中区分真正的与虚假的内存泄漏.在人工对训练数据进行标记时,我们发现如下规律:

(1) 特征15、16都为假(即在SVFG中,不存在全局变量指向内存分配点,也不存在内存释放点),通常是真正内存泄漏.

(2) 特征11为真(即 p 释放之前为空指针)、或者特征12大于1(即 p 指向的对象数目超过1)、或者特征14不为0(存在指针偏移)、或者特征16为假(即未进行内存释放),通常是真正内存泄漏.

在对训练数据进行人工标记后,我们获取每个训练样本的内存泄漏特征,并将这些作为训练集输入分类器进行模型构建.现阶段主流的机器学习算法有SVM、决策树、朴素贝叶斯分类、隐马尔可夫、随机森林、循环神经网络、长短期记忆(LSTM)与卷积神经网络等.本文采用三种机器学习算法:SVM[23]、随机森林(RF)[30]和决策树[31].下面介绍三种算法的优缺点.

(1) 决策树指的是C4.5算法,优点是产生的分类规则易于理解,训练时间复杂度较低,准确率较高;缺点是针对连续属性值的特征时计算效率低,且容易过拟合.

(2) 随机森林的优点是训练速度快,易并行化,能够处理高维度的数据,适合做多分类问题;缺点是在处理噪音较大的数据集上容易过拟合,过于随机导致无法控制模型内部运行.

(3) SVM可以通过计算数学函数将训练数据分开,这些函数被称为核函数.常用的核函数有四种:线性、多项式、径向基核函数(RBF)和Sigmoid函数.根据文献[18,24,25,26],使用RBF核函数将每个特征向量映射到高维空间,这样可以防止过拟合.缺点是对大规模训练样本难以实施,解决多分类问题存在困难.

在构建模型时,需要确定分类器类型及参数,我们使用分类的准确率作为评估标准,然后进行迭代确定最优的分类器类型及参数,步骤如下:

(1) 首先选定第一个分类器类型以及参数并进行训练,使用五折交叉验证得到的准确率作为基线.

(2) 多次修改分类器参数(根据分类器类型可自行调整),记录五折交叉验证的准确率超过基线的分类器类型、参数以及准确率.

(3) 修改分类器类型,并重复第(2)步.

(4) 选取准确率最高的分类器(类型及参数)作为本文内存泄漏检测模型.

2.2 特征提取与缺陷检测阶段

特征提取与缺陷检测主要分为以下三个步骤:程序静态分析、特征提取、内存泄漏检测.程序静态分析就是第1节所介绍的Full-Sparse Value-Flow Analysis.

2.2.1 特征提取

本文通过构建SVFG并获取特征,算法1阐述了如何通过SVFG来获取内存泄漏特征,该算法的输入是内存位置(内存分配点)集合src以及SVFG.基本思想是:分析内存分配点的信息获取内存泄漏的部分类型和指针信息,分析内存分配点到释放点的路径信息获取内存泄漏的其余信息.

首先在第1行至第4行,我们初始化存放特征信息的向量V,然后遍历内存分配点的集合src,并从内存分配点开始向前遍历SVFG获取当前内存分配点的部分类型信息和分支信息(如内存分配点是否为数组或者结构体,内存分配点是否在循环中等);同时记录内存释放点集合dst.

然后在第5行至第7行,遍历内存释放点集合dst,获取每个内存释放点的信息以及从内存分配点到内存释放点的完整路径信息.

最后将所有路径的特征信息存入V中.

算法1. SVFG遍历获取内存泄漏特征

输入:内存分配点的集合src

SVFG

Begin

1: 初始化存放特征信息的向量 V

2: **for**(i=0;i<src.size();i++)

3: 获取 src[i]的部分特征信息

4: 从内存分配点开始,向前遍历 SVFG，获取路径信息 //记录经过的内存释放点集合 dst

5: **for**(j=0;j<dst.size();j++)

6: 获取 dst[j]信息以及该条路径的完整信息

7: **end for**

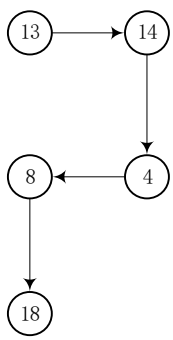
8: 将每条路径的特征信息存入 V

9: **end for**

End

```
1   void f(){
2       bool t = false;
3       char **buf = readBuf();
4       char * tmp = *buf;
5       if (t)
6           printf();
7       else
8           freeBuf(buf);
9   }
10  char** readBuf(){
11      int **Buf = malloc(10);
12      for (n = 0; n < 10; n++)
13          Buf[n] = malloc();
14      return Buf;
15  }
16  void freeBuf(char **Buf){
17      for (j = 0; j < 9; j++)
18          free(Buf[j]);
19      free(Buf);
20  }
```

(a) 输入程序



(b) SVFG

类别	序号	特征	属性值
类型信息	1	数组	TRUE
	2	结构体	FALSE
	3	链表	FALSE
分支信息	4	循环分配	TRUE
	5	循环释放	TRUE
	6	循环匹配	FALSE
	7	同一循环	FALSE
	8	链表匹配	FALSE
	9	分支条件	FALSE
释放信息	10	函数间距	3
	11	释放之前为空	FALSE
	12	p指向对象数目	1
	13	别名数目	1
	14	指针偏移量	0
	15	全局指针	FALSE
	16	释放	TRUE

(c) 特征

Fig.2 The Example of Feature Extraction

图 2 特征提取示例

如图 2 所示,我们给出了一个具体示例展示如何从源程序中提取特征.图 2a 是 C 程序源码.图 2b 是从源码第 13 行的内存位置构建的 SVFG.图 2b 中的数字表示源码的行号,这里的 SVFG 是一个精简的过程,我们只对应了行号,并未画出内存位置的所有 def-use 关系.图 2c 则是根据 SVFG 获取的特征以及对应的属性值.根据图 2a 和图 2b 可知,第 13 行的内存分配点所对应的特征 1、4、5、16 为 TRUE,表明该内存位置在循环中分配与释放,但是分配与释放次数不一致;特征 10 的值为 3,表示从内存分配点到释放位置经过 3 个函数.

2.2.2 内存泄漏检测

获取所有内存泄漏特征的数据集后,我们首先根据判定规则依次对数据集进行筛选,判定规则如下:

(1) 若特征 15、16 全部为假(即在 SVFG 中,没有全局变量指向该内存分配点,也不存在内存释放语句),则判定为疑似内存泄漏.

(2) 若特征 15 为真、特征 16 为假(即在 SVFG 中,存在全局变量指向该内存分配点,不存在内存释放语句),则无法判断是否为内存泄漏,视为警报.

(3) 对于不满足第 1 条规则且不满足第 2 条规则的数据提交给内存泄漏检测模型,得到内存泄漏检测结果.内存泄漏模型中检测结果分为两类：疑似内存泄漏、可能非内存泄漏.

根据判定规则(1)和(2),若在 SVFG 中, 没有全局变量指向该内存分配点,也不存在内存释放语句, 则可直接判断为疑似内存泄漏;若在 SVFG 中,存在全局变量指向该内存分配点,但不存在内存释放语句,由于全局变量可能在任何地方释放,我们不做判断,视为警报.因此根据规则(1)和(2),我们可直接判断数据对应的分类结果,无需使用内存泄漏检测模型.根据第 3 条规则,图 2c 所示的特征应该输入内存泄漏检测模型.在图 3 中,图 3b 是图 3a 中源码对应的内存泄漏特征.在图 3b 中,*o* 和 *o'*分别表示图 3a 中第 3 行和第 4 行的内存分配点.根据第 1 条规则,*o* 是疑似内存泄漏;根据第 2 条规则,*o'*无法判断是否泄露.

提取数据集中判定结果为疑似内存泄漏和警报的分配点,并结合内存位置信息,给出漏洞报告.报告中标明疑似内存泄漏和警报,并给出每个泄漏点的文件名、行号以及分配语句.

```
1    char *p;
2    void f(){
3        char *p1 = readBuf();
4        char *p2 = readBuf();
5        p = p2;
6    }
7    char* readBuf(){
8        int *Buf = malloc();
9        return Buf;
10   }
11   void freeBuf(char *Buf){
12       free(Buf);
13   }
```

(a)输入程序

类别	序号	特征	*o*的属性值	*o'*的属性值
类型信息	1	数组	FALSE	FALSE
	2	结构体	FALSE	FALSE
	3	链表	FALSE	FALSE
分支信息	4	循环分配	FALSE	FALSE
	5	循环释放	FALSE	FALSE
	6	循环匹配	FALSE	FALSE
	7	同一循环	FALSE	FALSE
	8	链表匹配	FALSE	FALSE
	9	分支条件	TRUE	TRUE
释放信息	10	函数间距	2	2
	11	释放之前为空	FALSE	FALSE
	12	p指向对象数目	1	1
	13	别名数目	1	2
	14	指针偏移量	0	0
	15	全局指针	FALSE	TRUE
	16	释放	FALSE	FALSE

(b)特征

Fig.3 Different Cases of Memory Leak

图 3 内存泄漏的不同情况

3 工具与评估

在实验中,我们基于 LLVM 编译器(版本 4.0.0)实现我们的基于机器学习的机器学习的 C 程序内存泄漏智能化检测工具 I_Mem,每个 C 程序的源文件都由 Clang 编译成 LLVM bitcode 文件格式,再由 LLVM Gold Plugin 进行合并生成整个程序的 bc 文件.在模型构建阶段,我们使用的 SVM 分类器是 libSVM[27],随机机森林和决策树使用的是机器学习工具 weka[32].我们的实验分为两部分,一是在模型构建阶段对我们的模型进行评估;二是在特征提取与缺陷检测阶段对内存泄漏的检测结果进行评估.

3.1 模型构建阶段

我们从开源的 C 程序中提取真正的与虚假的内存泄漏实例来构建机器学习分类器,我们提取内存泄漏示例的开源 C 程序主要有 icecast-2.3.1、cluster-3.0、droplet-3.0、wine-0.9.24 以及 SPEC2000 中的 ammp、equak,然后通过以下步骤获取真正的与虚假的内存泄漏.

(1) 关注源码中所有的内存分配点,通过添加或者注释内存释放点来获取真正的与虚假的内存泄漏.

(2) 仿照已经获取的内存泄漏实例, 在源码中插入各种内存泄漏以构建丰富的训练集,尽量满足各个特征的属性.

通过以上两个步骤,我们可以获取大量的内存泄漏实例用于机器学习分类器的构建.模型构建阶段总过获取了 1728 个内存泄漏实例(396 个虚假内存泄漏,1332 个真正内存泄漏).构建模型中,我们采用五折交叉验证来确定准确率最高的分类器类型和参数.准确率是正确分类的样本数与总样本数之比.

我们使用五折交叉验证,在 1728 个实例中进行模型训练与评估,对于 SVM、随机森林和决策树三种机器学习算法,我们选取每种机器学习算法在训练过程中准确率最高的进行展示,如表 3 所示.

Table 3 Results of Classification

表 3 分类结果

机器学习算法	分类准确率
SVM	97.7%
随机森林	99.6%
决策树	99.6%

如表 3 所示,SVM 的分类准确率为 97.7%,随机森林与决策树的准确率 99.6%.因为决策树容易出现过拟合现象,因此我们选取随机森林作为我们内存泄漏检测模型.实验结果表明我们构建的随机森林模型在对真实与虚假的内存泄漏进行分类时是有效的.

3.2 特征提取与缺陷检测阶段

我们实验分为两部分,第一部分的实验数据来自基准程序 Siemens[33].共有 4 个程序,这些实验对象的规模都比较小,存在的内存泄漏不多,因此我们在源码中手工植入内存泄漏,特别是植入与数组、循环、链表等有关的内存泄漏,以验证内存泄漏模型在检测各种类型内存泄漏时的有效性. 第二部分数据来源于原生 SPEC2000[34],我们选取了 3 个程序, 以验证内存泄漏模型在检测大规模程序时的有效性.实验对象见表 4.

Table 4 Experimental Subjects

表 4 实验对象

来源	名称	代码行数(KLOC)
Siemens	tcas	0.17
	replace	0.56
	print_tokens	0.72
	print_tokens2	0.57
SPEC2000	vpr	17.0
	mesa	49.7
	vortex	52.7

我们的方法是在 SVFG 的基础上提取内存泄漏特征并利用机器学习技术进行内存泄漏检测,因此我们选取 Saber 来比较本文方法和静态分析方法的效果.Saber 通过构建源码的 SVFG 来判断内存泄漏.在表 5 中展示了本文方法与 Sabe 在 Siemens 程序上的对比实验结果.

根据表 5 中的结果,我们可以得到如下结论:

(1) 两种内存泄漏检测方法的漏报数目都比较低.

(2) 针对内存泄漏的一些特殊案例时,如内存泄漏的分配、使用或者释放出现了循环、递归、链表等情况时,Saber 误报较多,例如 print_tokens 程序中的 34 个内存分配点,Saber 的误报是 10 个,本文只有 5 个.

Table 5 Experimental Results of Siemens

表 5 Siemens 实验结果

程序	内存泄漏路径/植入	Saber 漏报	Saber 误报	Saber 准确率	I_Mem 漏报	I_Mem 误报	I_Mem 准确率
tcas	12/6	0	3	75%	0	0	100%
replace	26/12	0	8	69.2%	0	3	88.5%
print_tokens	34/15	0	10	70.6%	0	5	85.3%
print_tokens2	46/20	1	14	67.4%	1	5	87.0%
合计	118/53	1	35	69.5%	1	13	88.1%

表 6 为 SPEC2000 的实验结果.总结两部分实验,我们可以得到如下结论:

(1) 本文方法在静态分析的基础上利用机器学习算法提高了内存泄漏检测在特殊案例上的准确性.

(2) 本文的内存泄漏模型在检测大规模程序时是有效的.

Table 6 Experimental Results of SPEC2000

表 6 SPEC2000 实验结果

程序	内存泄漏路径	I_Mem 漏报	I_Mem 误报	Saber 漏报	Saber 误报
vpr	134	0	1	0	3
mesa	76	0	1	0	4
vortex	9	0	0	0	4
合计	219	0	2	0	11

3.3 讨论

首先,本文方法确实有助于提高内存泄漏检测的准确率.在模型构建阶段,我们对构建的内存泄漏检测模型进行五折交叉验证,交叉验证结果准确率高达 95%以上,实验结果表明我们构建的内存泄漏检测模型在对真实与虚假的内存泄漏进行检测时是有效的.在特征提取与缺陷检测阶段,在 Siemens 的实验数据上,我们的方法与 Saber 进行对比实验,Saber 的平均准确率只有 69.5%,我们的方法准确率高达 88.1%;在 SPEC2000 的实验数据上，总共 184 内存分配点，我们的误报只有 2 个.

综上所述,我们的实验结果表明:基于机器学习的 C 程序内存泄漏智能化检测方法在针对数组、循环等相关的内存泄漏时能够得到更加准确的检测结果.在静态分析的基础上利用机器学习算法,使得内存泄漏检测结果更加准确.此外,实验中存在一些不足需要注意:目前我们只针对 C 语言内存泄漏进行检测,并不支持 C++;实验中我们所选取的是简单的基准程序并在源码中插入一些内存泄漏特殊案例,并不保证该方法对内存泄漏其他特殊案例检测都具有较高的准确率.但是我们相信,实验的结果确实表明了本文的方法在检测内存泄漏上的可行性以及准确性.

4 相关工作

在本节中,我们主要通过以下 3 个方面来介绍和讨论相关工作.第一,内存泄漏的静态分析;第二,内存泄漏的动态检测;第三,基于机器学习的缺陷检测.

4.1 内存泄漏的静态分析

内存泄漏通常是由于人为的对程序中动态内存的管理不当造成的,内存泄漏会导致内存空间被消耗且在程序运行期间无法回收和重新利用.内存泄漏十分隐蔽,在程序运行初期不易被发现,但是对于长期运行的程序,特别是服务器,影响十分显著,它会降低程序性能,甚至导致程序在运行时崩溃.特别是在 C 语言中,内存的分配与释放都是人为控制的,内存释放这一步骤极容易被忽略,从而导致内存泄漏.

静态分析主要是根据特定的错误模式来查找内存泄漏或者建立内存状态模型来进行内存泄漏检测.Cherem 等人[1]通过构建数据流图进行数据流分析,分析数值从内存分配点到内存释放点的路径中是否正确

传播来检测内存泄露.Saber 通过构建 C 程序的 SVFG 检测内存泄漏.Orlovich 等人[4]首先假设内存泄漏存在,然后进行反向数据流分析,验证该假设是否成立.RL_Detector[5]基于控制流图(CFG)进行数据流分析, 利用静态符号执行对于每个资源(包括内存泄露)收集所有路径约束,通过路径约束计算数据流条件并检测资源泄露.Heine 等人[6]开发了一个描述指针隶属关系的模型,在该模型中,每个内存对象只能被一个拥有指针指向,因此该指针是唯一的且具有传递性,基于此对程序生成一种约束,用来检测内存泄漏.

内存泄漏静态分析的优点是能够自动化运行,检测速度快;缺点是误报较多.目前也有一些工作是在静态分析的基础上,对内存泄漏进行修复.内存泄漏修复需要首先定位内存泄漏位置,因此,内存泄漏修复的准确性首先取决于内存定位的准确性,其次是插入释放语句的准确性.目前主要的内存泄漏修复工作有:Leakfix[15]基于指针分析和数据流分析,判断内存泄露位置并进行内存泄漏的修复.Leakfix 能够保证为一个内存泄露生成多个修复程序,但不能保证生成的修复程序完全解决了该内存泄漏.AutoFix[16]是根据已有的静态分析警报,通过指针分析,构建 VFG,再根据活性分析,判断内存泄漏位置,进行内存泄漏的修复.Autofix 通过代码插桩进行内存泄漏的修复并在一个沙箱中运行程序进行检查,确保修复的安全性.内存泄漏的修复首先需要确保内存泄漏检测的准确性,不能对误报进行修复;其次内存泄漏修复需要保证修复的正确性.Leakfix 在修复之后需要使用内存泄漏检测工具进行检测,AutoFix 则是自动的在修复之后运行程序进行安全性检查.因此,内存泄漏的修复受限于规模,如何提高内存泄漏修复的可扩展性以及准确性依然是一个难题.

本文的主要工作是在静态分析的基础上利用机器学习技术进行内存泄漏检测,减少了静态分析的漏报和误报,提高了静态分析的准确性,尤其在针对内存泄漏的特殊案例时,能够显著提高内存泄漏检测的准确性.

4.2 内存泄漏的动态检测

内存泄漏的动态检测方法需要运行源代码,在程序运行过程中对内存分配、使用以及释放进行动态跟踪.LeakPoint[10]基于污点传播的思想监控内存对象的状态,追踪内存最后的使用位置以及失去引用的位置.DOUBLETAKE[11]将程序执行拆分为多个块,在每个块运行开始之前保存程序状态,在该块之行结束之后检查程序状态,判断内存是否发生错误.Snipe[12]利用处理器的监视单元(PMU)进行指令采样来跟踪对于堆内存的访问指令,然后通过离线模拟器分析指令,计算堆对象的陈旧度并重新执行相关指令捕获程序执行期间的内存泄漏.Omega[21]和 Maebe[22]主要采用指针计数的思想记录内存对象的引用计数.

动态检测相对于静态分析更加准确,但是动态检测无法分析程序执行中不可达位置的错误.动态检测最大的缺点是效率低,耗时长.目前,代码规模和复杂度日渐增加,内存泄漏的动态检测效率远远无法满足工业界的需求,并且随着静态分析技术的发展,内存泄漏静态分析结果的准确率逐渐提高,静态分析的使用更加广泛.因此,本文的方法是基于静态分析,利用规则和机器学习技术进行内存泄漏检测,具有高效率和高可靠性.

4.3 基于机器学习的缺陷检测

目前,机器学习技术已被广泛运用于程序分析中以检测程序缺陷.Alatwi 等人[17]实现了一种检测安卓应用程序是否属于恶意软件的方法,其主要思想是将 apk 反汇编为源码,然后利用静态分析方法提取代码代征,并选取可用于模型预测的最佳特征组合,构建 SVM 分类器进行检测.Tac[18]主要利用机器学习算法来消除 typestate 和指针分析的差距,它从源码中提取 35 个特征,并训练 SVM 分类器以检测 Use-After-Free.Nagano 等人[19]对执行文件进行静态分析,然后利用机器学习的分类算法以及自然语言处理技术来识别恶意软件.Grieco 等人[20]从二进制文件中提取程序静态特征,从程序执行中提取动态特征,然后利用机器学习技术训练模型来检测内存冲突.

本文的主要工作是通过静态分析提取内存泄漏特征,然后利用规则和机器学习模型检测内存泄漏.我们关注的重点是内存泄漏的特殊案例,因此在检测内存泄漏上准确性更高.

5 总结与展望

本文提出了一种基于机器学习的 C 程序内存泄漏智能化检测方法,首先构建机器学习模型,然后在内存泄

漏静态分析的基础上提取内存泄漏特征,并利用规则和机器学习模型进行内存泄漏的检测.实验结果表明,我们构建的内存泄漏检测模型在对真实与虚假的内存泄漏进行检测时是有效的.相对于目前的内存泄漏静态分析方法,本文方法在针对数组、循环等相关的内存泄漏时检测结果更加准确.

在本文方法的实验中,我们也遇到了一些问题和挑战,这也是我们未来的研究方向: (1)目前本文在特征选取与缺陷检测阶段选取的实验数据为简单程序且人为插入了各种内存泄漏,下一步需要选取一些开源的大型程序进行实验.(2)本文目前关注的重点是 C 语言中与循环、结构体、数组以及链表有关的内存泄漏,可以扩展至 C++中,关注类以及各类容器有关的内存泄漏问题.(3)当前的工作可以扩展到其他内存缺陷中,目前提取的特征适用于构建内存泄漏的分类器,我们可以提取更多的程序特征,构建多种内存缺陷检测的分类器.(4)本文目前的静态分析方法是对每个内存分配点提取一个内存泄漏特征,我们可以进行细化,对于每个内存分配点的每条路径提取一个内存泄漏特征,这样检测结果会更加准确,检测报告会更加详细.

References:

[1] Cherem S, Princehouse L, Rugina R. Practical memory leak detection using guarded value-flow analysis. Acm Sigplan Notices, 2007, 42(6):480-491.

[2] Sui Y, Ye D, Xue J. Static memory leak detection using full-sparse value-flow analysis. In: Proc. of the 2012 Int'l Symp. on Software Testing and Analysis. New York: ACM Press, 2012. 254−264.

[3] Sui Y, Ye D, Xue J. Detecting memory leaks statically with full-sparse value-flow analysis. IEEE Trans. on Software Engineering, 2014,40(2):107−122.

[4] Orlovich M, Rugina R. Memory Leak Analysis by Contradiction. In static Analysis Symposium, 2006, 405-424.

[5] Ji X, Yang J, Xu J, et al. Interprocedural path-sensitive resource leaks detection for C programs. In Proceedings of the Fourth Asia-Pacific Symposium on Internetware, ACM, 2012. 1–9.

[6] Heine DL, Lam MS. Static detection of leaks in polymorphic containers. In: Proc. of the 28th Int'l Conf. on Software Engineering. New York: ACM Press, 2006. 252−261.

[7] HP Fortify. 2016. http://www8.hp.com/us/en/software-solutions/application-security/index.html

[8] Coverity. The coverity static analysis tools. 2016. http://www.coverity.com/

[9] Klocwork. The Klocwork static analysis tool. 2001~2016. http://www.klocwork.com/

[10] Clause J, Orso A. LEAKPOINT: Pinpointing the causes of memory leaks. In: Proc. of the 32nd ACM/IEEE Int'l Conf. on Software Engineering. New York: ACM Press, 2010. 515−524.

[11] Liu T, Curtsinger C, Berger E D. DoubleTake: Fast and Precise Error Detection via Evidence-Based Dynamic Analysis. In Proceedings of the 38th International Conference on Software Engineering, 2016:911-922.

[12] Jung C, Lee S, Raman E, Pande S. Automated memory leak detection for production use. In: Proc. of the 36th Int'l Conf. on Software Engineering. New York: ACM Press, 2014. 825−836.

[13] Hastings R, Joyce B. Purify: Fast detection of memory leaks and access errors. In: Proc. of the USENIX Conf. Berkeley: Usenix Association, 1992. 125−138.

[14] Nethercote N, Seward J. Valgrind: A framework for heavyweight dynamic binary instrumentation. ACM SIGPLAN Notices, 2007, 42(6):89−100.

[15] Gao Q, Xiong Y, Mi Y, Zhang L, Yang WK, Zhou ZP, Xie B, Mei H. Safe memory-leak fixing for C programs. In: Proc. of the 37th Int'l Conf. on Software Engineering, Vol.1. Piscataway: IEEE Press, 2015. 459−470.

[16] Yan H, Sui Y, Chen S, et al. AutoFix: an automated approach to memory leak fixing on value-flow slices for C programs. Acm Sigapp Applied Computing Review, 2017, 16(4):38-50.

[17] Alatwi H A, Oh T, Fokoue E, et al. Android Malware Detection Using Category-Based Machine Learning Classifiers. Conference on Information Technology Education. ACM, 2016:54-59.

[18] Yan H, Sui Y, Chen S, et al. Machine-Learning-Guided Typestate Analysis for Static Use-After-Free Detection. Computer Security Applications Conference. ACM, 2017:42-54.

[19] Nagano Y, Uda R. Static analysis with paragraph vector for malware detection. The, International Conference. 2017:1-7.

[20] Grieco G, Grinblat G L, Uzal L, et al. Toward Large-Scale Vulnerability Discovery using Machine Learning. ACM Conference on Data and Application Security and Privacy. ACM, 2016:85-96.

[21] Omega: An instant leak detector tool for valgrind. 2016. http://www.brainmurders.eclipse.co.uk/omega.html.

[22] Cherem S, Princehouse L, Rugina R. Practical memory leak detection using guarded value-flow analysis. ACM SIGPLAN Notices, 2007,42(6):480−491.

[23] CorinnaCortesandVladimirVapnik.1995. Support-vectornetworks. Machine learning 20,3(1995),273–297.

[24] S Sathiya Keerthi and Chih-Jen Lin. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural computation 15, 7(2003), 1667–1689.

[25] Hsuan-Tien Lin and Chih-Jen Lin. 2003. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical Report. Department of Computer Science, National Taiwan University.

[26] VladimirVapnik.2013. The nature of statistical learning theory. Springerscience &businessmedia.

[27] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 3 (2011), 27.

[28] Steffen B, Knoop J, Rüthing O. The value flow graph: A program representation for optimal program transformations. In Proceedings of the 3rd European Symposium on Programming (ESOP'90), 1990:389-405.

[29] Rugina R, Rugina R, Rugina R. Practical memory leak detection using guarded value-flow analysis. ACM Sigplan Conference on Programming Language Design and Implementation. ACM, 2007:480-491.

[30] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[31] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems Man & Cybernetics, 2002, 21(3):660-674.

[32] The University of Waikato. Weka. http://www.cs.waikato.ac.nz/~ml/weka/, 2013.

[33] Siemens. 2016. http://sir.unl.edu/

[34] SPEC: Standard Performance Evaluation Corporation. http://www.spec.org, September 2000.