

基于抽象解释和数值熵的数值程序分析方法

王正谦, 刘久富, 陈哲

(南京航空航天大学 自动化学院, 江苏 南京 210016)

摘要:在高度依赖软件的信息时代,程序的正确性验证问题需要深入研究。文中提出了基于抽象解释和数值熵的数值程序正确性分析方法。抽象解释理论为程序静态分析提供了一个通用框架,在编译时能够自动地推导程序的动态性质。数值信息熵能够反映变量的值范围,通过熵值的大小可以判断变量取值是否在规定范围内。通过一个C程序对该方法进行了验证,该数值程序分析方法可以做到对程序正确性的验证,并且较单纯地抽象解释近似分析,正确性、可靠性更高。

关键词:数值程序分析;正确性;抽象解释;数值信息熵

中图分类号:TP301.2

文献标识码:A

文章编号:1673-629X(2014)04-0057-03

doi:10.3969/j.issn.1673-629X.2014.04.014

Value Range Analysis Method Based on Abstract Interpretation and Value Entropy

WANG Zheng-qian, LIU Jiu-fu, CHEN Zhe

(College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract:In the information age, software is highly dependent, thus the correctness of the program validation issues need to be further study. In the paper, introduce the value range analysis method based on abstract interpretation and value information entropy. The abstract interpretation as an important method of the static analysis, uses testing framework to deduce the program's dynamic property automatically. The value information entropy can reflect the range of a variable, through the entropy, can judge whether variable values within the prescribed scope. Through validating this method with C program, find the method can validate the correctness of the running program. Compared with the only abstract interpretation analysis, the method is higher in validity and reliability.

Key words:value range analysis; correctness; abstract interpretation; value information entropy

0 引言

程序中数值变量的取值范围对于编译优化、错误检查至关重要^[1],值范围越接近实际运行的情况,检测结果的准确率越高,然而精确地获取变量的值范围分析在理论上往往是不可能的^[2]。

形式化方法是分析软件和硬件系统中 bug 的一种有效手段。模型检验、定理证明是能够自动化实现的典型形式化方法^[3-4]。但是模型检验方法难以解决状态空间爆炸问题,机械化的定理证明过程不能够保证停机。程序正确性的不可判定性和问题的复杂性决定了在对复杂系统进行推理或计算时不得不采用某种形式的近似,忽略其中不重要的信息而仅就所关心问题进行讨论^[5]。各种形式化验证工具的本质区别在于实现近似的方法不同,P. Cousot 和 R. Cousot 提出的抽

象解释理论^[6]即试图在一个统一的框架下对近似思想进行形式化,可以有效地用于静态分析程序变量的取值范围分析。文献[7]提出了基于抽象解释和通用单调数据流框架的值范围分析框架,包括精确的定义、分析和完整的正确性证明。到目前为止,基于抽象解释的值范围分析方法一般针对于数值型变量,值范围采用单一区间进行描述,并且存在条件判断分支中变量值区间计算效率比较低等问题。区间抽象域^[8]是抽象解释理论中的一种典型的、比较准确的抽象域。它将程序中的数值变量值范围抽象表示为区间,定义了区间上的拓宽和收窄算子,并给出了程序不同类型语句节点处的值范围区间计算方法。

为了保证数值程序的正确性,并对抽象解释这一静态分析方法进行验证,引入熵值理论。熵可以反映

收稿日期:2013-06-03

修回日期:2013-09-18

网络出版时间:2014-01-28

基金项目:国家自然科学基金资助项目(60674100);南京航空航天大学基本科研业务费专项科研项目(NS2010069)

作者简介:王正谦(1987-),男,硕士研究生,研究方向为软件测试技术;刘久富,博士,硕士生导师,研究方向为软件测试技术与软件质量工程。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140128.1144.028.html>

程序的复杂性程度,对程序节点的信息进行量化比较。文中针对数值程序分析,提出将一种数值信息熵作为验证程序正确性分析的度量标准。利用抽象解释理论将程序中各个节点的数值变量范围抽象为区间,根据区间内的变量取值可以得到数值信息熵区间,接着运行该数值程序获取实际取值,以此计算程序运行时的数值熵值,最后通过比较该熵值是否在熵区间内,达到对程序的正确性验证分析的目的。

1 抽象解释

1.1 基本理论

抽象解释是一种针对计算机系统语义模型的近似理论,同时也是一个非常成熟而可靠的数学方法^[9-10]。它针对给定的程序语言赋予具体和抽象两种语义,将不可计算语义定义成具体语义,将可计算的语义定义成抽象语义,然后建立二者间的正确性关联,用抽象语义去代替具体语义,再利用对抽象语义的求解,从而达到安全地计算具体语义的目的。

完备格、Galois 连接、加宽操作和收窄操作是抽象解释理论中的基本概念^[3]。

完备格:称偏序集 (L, \subseteq) 是一个完备格当且仅当 L 中所有的子集都有上确界和下确界。特别地,称 $\perp = \cap L = \cup \emptyset$ 为 L 的最小元素, $T = \cup L = \cap \emptyset$ 为 L 的最大元素。完备格 (L, \subseteq) 一般表示为 $(L, \subseteq, \cup, \cap, \perp, T)$ 。

Galois 连接:完备格 (L, \subseteq) 与完备格 (L^*, \subseteq) 满足 Galois 连接是指存在抽象函数 $\alpha: L \rightarrow L^*$ 以及具体函数 $\gamma: L^* \rightarrow L, \forall x \in L, \forall y \in L^*,$ 满足关系 $\alpha(x) \subseteq^* y \Leftrightarrow x \subseteq \gamma(y)$ 。Galois 连接一般可形式化表示为 (L, α, γ, L^*) 。

由 Galois 连接的定义可知其具有如下性质:

- (1) $\forall x \in L, x \subseteq \gamma(\alpha(x))$;
- (2) $\forall y \in L^*, \alpha(\gamma(y)) \subseteq^* y$;
- (3) α 和 γ 是单调递增函数。

加宽操作:设 (L, \subseteq) 是一个完备格,二元操作符 $\nabla: L \times L$ 是加宽操作符,当且仅当:

- (1) $\forall l_1, l_2 \in L, l_1, l_2 \subseteq l_1 \nabla l_2$, 即 ∇ 是一个上界操作符;
- (2) 对所有的递增序列 $(l_n)_n, (l_n^\nabla)_n$ 最终稳定,其中 l_n^∇ 定义为:如果 $n = 0$, 则 $l_n^\nabla = l_n$; 否则, $l_n^\nabla = (l_{n-1}^\nabla) \nabla l_n$ 。

收窄操作:设 (L, \subseteq) 是一个完备格,二元操作符 $\Delta: L \times L$ 是收窄操作符,当且仅当:

- (1) $\forall l_1, l_2 \in L$, 若 $l_2 \subseteq l_1$, 则 $l_2 \subseteq l_1 \Delta l_2 \subseteq l_1$;
- (2) 对所有的递增序列 $(l_n)_n, (l_n^\Delta)_n$ 最终稳定,其中 l_n^Δ 定义为:如果 $n = 0$, 则 $l_n^\Delta = l_n$; 否则, $l_n^\Delta =$

$(l_{n-1}^\Delta) \Delta l_n$ 。

1.2 理论框架

抽象解释理论框架是建立在语义层次体系的部分构造过程上的。程序通常可以表示为迁移系统 $\tau = (\Sigma, \sum_i, t)$ 的一个三元数组,其中 Σ 为系统状态集, \sum_i 为系统的初态集, t 为 τ 上的状态集。对迁移关系进行抽象解释可以得到 4 种语义:部分踪迹语义、自反闭包语义、可达性语义、区间语义^[5]。文中所涉及的是区间语义。

区间语义,若迁移系统的状态集构成一个最小元素为 $-\infty$ 和最大元素为 $+\infty$ 的完备格,则可对可达性语义进一步抽象:对任意可达状态集 X ,仅考虑其最大和最小边界。定义 $\alpha''(X) = [\min X, \max X]$, 其中 $\min X$ 为下确界, $\max X$ 为上确界,即把一个可达状态集抽象为一个区间 $[l, h]$ 。对应的,定义 $\gamma''([l, h]) = \{s \in \Sigma \mid l \leq s \leq h\}$ 。这样就消除了加宽操作在计算过程中的过度近似,使得分析结果得到了精化。

2 数值信息熵计算

2.1 熵理论

信息熵在信息和信号处理方面有着广泛应用。它描述了消息中信息量的不确定程度。Shannon 在 1948 年第一次提出了熵的概念^[11],并以此作为信息的度量,后来熵成功地应用于软件复杂度的测量。

在信息论中信源输出是随机量,因而其不确定性可以用概率分布来度量。

$$H(X) = \sum_i p(x_i) \log p(x_i)$$

这里 $p(x_i), i = 1, 2, \dots, n$ 为信源取第 i 个符号的概率。 $\sum_i p(x_i) = 1, H(X)$ 称为信源的信息熵。一般来说,熵值越大,信息源所含信息量也就越大。

2.2 抽象解释下的数值信息熵

实数区间构成的完备格上有无穷递增链致使区间分析可能不终止,因此将区间算数与抽象解释框架相结合提出了区间抽象域。区间抽象域有简单易用、计算效率高以及可扩展性强等特点^[12]。区间抽象域可以表示单个变量的性质,以区间的形式为程序中每个程序点产生数值不变式,即计算所有数值变量在程序点处的所有可能取值,比如一个可达状态集抽象为一个区间 $[l, h]$ 。

将抽象区间 $[l, h]$ 所对应数据提取为 $U = \{u(i), i = 0, 1, \dots, N\}$, 类似于近似熵^[13]的算法,预先给定模式维数 m 和相似容限 r 的值(计算中取参数 $m = 2$ 或 $3, r = 0.26 * \text{std}(U)$), 则数值信息熵可以通过下面的步骤计算得到。

- ① 将序列 $\{u(i)\}$ 按顺序组成 m 维矢量 $\mathbf{X}(i)$, 即 $\mathbf{X}(i) = [u(i), u(i+1), \dots, u(i+m-1)]$, $i = 1, 2, \dots, N-m+1$
- ② 对每一个 i 值计算矢量 $\mathbf{X}(i)$ 与其余矢量 $\mathbf{X}(j)$ 之间的距离:
- $$d[\mathbf{X}(i), \mathbf{X}(j)] = \max_{k=0 \cdots m-1} |u(i+k) - u(j+k)|$$
- ③ 按照给定的阈值 $r(r>0)$, 对每一个 i 值统计 $d[\mathbf{X}(i), \mathbf{X}(j)] < r$ 的数目及此数目与总的矢量个数 $N-m+1$ 的比值, 记作 $C_i^m(r)$, 即
- $$C_i^m(r) = \{d[\mathbf{X}(i), \mathbf{X}(j)] < r \text{ 的数目}\} / (N-m+1)$$
- ④ 先将 $C_i^m(r)$ 取对数, 再求其对所有 i 的平均值, 记作 $\Phi^m(r)$, 即

⑤ 增加模式维数 m 为 $m+1$, 重复 1~4 的过程, 得到 $\Phi^{m+1}(r)$ 。从而, 此系列的数值熵为:

$$\text{En}(m, r) = \lim_{N \rightarrow \infty} [\Phi^m(r) - \Phi^{m+1}(r)]$$

在实际工作中 N 不可能为 ∞ , 当 N 为有限值时, 按照上面的步骤得出的是序列长度为 N 时数值信息熵的估计值。记作

$$\text{En}(m, r) = \Phi^m(r) - \Phi^{m+1}(r)$$

这样可以得到抽象域区间所对应的数值信息熵, 简称数值熵, 进而形成熵值区间。熵区间从概率学的角度反映了程序实际运行时变量正确取值的情况, 实际熵值在熵区间内, 确保了程序静态分析的精确度, 进而验证了数值程序的正确性。

3 实例验证

3.1 系统设计

程序设计语言的语义标识了值(比如状态、双精度实数等)的集合 V , 说明了程序 p 的执行是把一个值 v_1 迁移到另一个值 v_2 : $p \perp v_1 \rightarrow v_2$ 。程序分析标识了特性的集合 L , 说明了程序 p 的“抽象”执行是把一个特性 l_1 迁移到另一个特性 l_2 : $p \perp l_1 \triangleright l_2$ (比如状态的形状、双精度实数的上下界等)^[14]。

由于通常程序分析要求一定的准确性和可靠性, 在静态测试工具 C++Test 的基础上, 设计如下基于抽象区间域和数值熵的程序分析器, 其结构如图 1 所示。

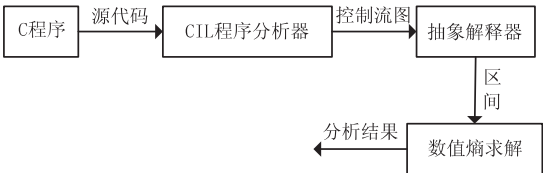


图1 基于抽象区间域和数值熵的程序分析器

3.2 实验结果与验证

为了验证数值程序检查的正确性, 利用上面的数值程序分析器分析了一个具有代表性的程序。将构造的数值程序分析器运行于 Linux Redhat 操作系统下, 待测程序的片段如图 2 所示。

```
#include <stdio.h> 1
void main() { 2
    int i=0; 3
    int j=1; 4
    while (i<10) { 5
        if(i<1) 6
            j++; 7
        else 8
            j+=2; 9
        i++; 10
    } 11
} 12
```

图2 待测程序片段

根据图 1 设计的系统结构, 对图 2 中程序进行抽象解释分析, 通过预处理分析可以将“#include <stdio.h>”忽略, 对剩余的程序进行词法、语法分析得到抽象语法树; 然后, 依据抽象域图的构造方法形成抽象域图, 得到变量 i 和 j 的抽象区间为 $[1, 11]$ 和 $[2, 20]$; 最后根据数值熵的算法求得其熵区间分别为 $[0, 0.213]$ 和 $[0, 0.315]$ 。

设定程序的实验次数后, 运行该程序, 记录每次实验中变量 i, j 的变量值, 形成变量的实际值范围区间。根据区间内的数值, 计算变量 i, j 的平均熵值为 0.102 和 0.236, 分别在区间 $[0, 0.213]$ 和 $[0, 0.315]$ 之间, 说明对变量 i 和变量 j 的静态分析是准确的。由此, 对程序中所有变量进行验证分析, 最终可以确定完整数值程序静态分析方法的正确性、可靠性。

4 结束语

文中提出的基于抽象解释和数值熵的数值程序分析方法, 可以对数值程序变量进行区间解释和检查验证。抽象解释理论可以将变量的取值抽象到区间域, 而后利用数值熵对静态分析方法进行检验, 并应用到实际小型数值程序的验证中。然而, 对于大规模软件和硬件系统的自动分析与验证还存在一定的问题, 需要进一步的研究。

考虑到大型程序的数据结构复杂性更高, 数值变量的调用会更错综复杂, 在文中的基础上, 将来可以将数值熵理论和其他区间域(如八边形区间域、多面体域等)相结合, 最后形成适合数值程序变量静态检测的自动化验证工具。

具有旋转不变性,但是检索出来的未旋转的云图的效果较好,并且具有较高相似度的云图在成像时间上与待检索云图十分接近,符合云在时序范围内连续运动的规律。

4.2.3 网格内切圆法的检索结果与分析

文中采用网格内切圆法,继续用从 2011 年的云图库中选取的 5 月 6 日 3 时文件名为“EILY0630. AWX”的兰勃托投影下的红外卫星云图,对 2011 年的兰勃托投影下的红外云图库进行检索。其前 16 幅最相似的检索结果如图 6 所示,对检索结果分析可知,网格内切圆法不但具有一定的旋转不变性,而且检索的可靠性也较好,具有较高相似度的云图在成像时间上与待检索云图十分接近,符合云在时序范围内连续运动的规律,该方法明显比前两种方法的检索效果要好。

5 结束语

文中采用几种常用形状区域特征对卫星云图进行相似性检索,并介绍了详细的检索流程。文中值得注意的地方是:首先,在进行云地分离前先进行直方图均衡化,可较好地对整个云图库进行云地分离,即便是不同类别的云图,也可以取一个普遍适用的阈值;其次,对比了几种形状区域特征的优缺点,提出了一种改进的方法。实验结果表明,文中方法能有效提取云图的特征信息,能够较好地对历史卫星云图进行相似性检索,具有一定的应用前景。在未来的工作中,将重点研究如何将形状特征与纹理特征进行融合,提高检索的可靠性。

(上接第 59 页)

参考文献:

- [1] Harrison W H. Compiler analysis of the value ranges for variables[J]. IEEE transactions on software engineering, 1977, 3(3):243-250.
- [2] Nielson F, Nielson H R, Hankin C. Principle of program analysis[M]. Berlin: Springer Verlag, 1999:211-282.
- [3] 李梦君, 李舟军, 陈火旺. 基于抽象解释理论的程序验证技术[J]. 软件学报, 2008, 19(1):17-26.
- [4] 张幸儿. 计算机编译原理: 编译程序构造实践[M]. 北京: 科学出版社, 2009.
- [5] 苏青琴, 刘久富, 陈 魁, 等. 基于抽象解释的非函数依赖不变量的检测方法[J]. 计算机技术与发展, 2012, 22(4):5-8.
- [6] Cousot P, Cousot R. Abstract interpretation: An unified lattice model for static analysis of programs by construction or approximation of fix points[C]//Proc of the 4th POPL. Los Angeles: ACM Press, 1977:17-19.
- [7] 姬孟洛, 王怀民, 李梦君, 等. 一种基于抽象解释和通用单

参考文献:

- [1] 孙学金, 王晓蕾, 李 浩, 等. 大气探测学[M]. 北京: 气象出版社, 2009.
- [2] 温泉彻, 彭 宏, 黎 琼. 基于内容的图像检索关键技术研究[J]. 微计算机信息, 2007, 23(1-3):278-280.
- [3] 陈渭民. 卫星气象学[M]. 北京: 气象出版社, 2005.
- [4] Loncaric S. A survey of shape analysis techniques[J]. Pattern recognition, 1998, 31(8):983-1001.
- [5] 孙君顶, 赵 珊. 图像低层特征提取与检索技术[M]. 北京: 电子工业出版社, 2009.
- [6] 李俊山, 李旭辉. 数字图像处理[M]. 北京: 清华大学出版社, 2010.
- [7] Zhang D S. Image retrieval based on shape[D]. Australia: Monash University, 2002.
- [8] 王水璋. 基于纹理的图像检索技术研究[D]. 太原: 太原理工大学, 2008:140-150.
- [9] 杨政武, 方 涛. 基于 Zernike 矩的图像归一化技术的研究[J]. 计算机工程, 2004, 30(12):34-36.
- [10] 李金泉, 王建伟, 陈善本, 等. 一种改进的 Zernike 正交矩亚像素边缘检测算法[J]. 光学技术, 2003, 29(4):500-503.
- [11] Li S Z. Matching: Invariant to translations, rotations and scale changes[J]. Pattern recognition, 1992, 25(6):583-594.
- [12] Lu G J, Sajjanhar A. Region-based shape representation and similarity measure suitable for content-based image retrieval[J]. Multimedia system, 1999, 7(2):165-174.
- [13] Zhang D S, Lu G J. Evaluation of similarity measurement for image retrieval[C]//Proc of IEEE international conference on neural networks & signal processing. Nan Jing, China: [s. n.], 2003:928-931.

调数据流框架的值范围分析方法[J]. 计算机研究与发展, 2006, 43(11):2020-2026.

- [8] Cousot P. Abstract interpretation based formal methods and future challenges[C]//Informatics 10 years back, 10 years ahead. London: Springer Verlag, 2001:138-156.
- [9] 王 伟, 刘久富, 娄坚波, 等. 基于多 Agent 的软件测试系统设计[J]. 计算机技术与发展, 2011, 21(4):37-39.
- [10] 赵修伟. 基于抽象解释的实时软件 WCET 研究[D]. 大连: 大连理工大学, 2009.
- [11] Shannon C E. A mathematical theory of communication[J]. Bell system technical journal, 1948, 27:379-423.
- [12] 陈立前, 王 戟, 刘万伟. 基于约束的多面体抽象域的弱接合[J]. 软件学报, 2010, 21(11):2711-2724.
- [13] Pincus S M. Approximate entropy (ApEn) as a complexity measure[J]. Chaos, 1995, 5(1):110-117.
- [14] Miné A. The octagon abstract domain[J]. Higher-order and symbolic computation, 2006, 19(1):31-100.

基于抽象解释和数值熵的数值程序分析方法

作者：[王正谦](#)，[刘久富](#)，[陈哲](#)，[WANG Zheng-qian](#)，[LIU Jiu-fu](#)，[CHEN Zhe](#)
作者单位：[南京航空航天大学 自动化学院, 江苏 南京, 210016](#)
刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

ISTIC

年，卷(期)：2014(4)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201404014.aspx