软件可靠性估计与计算复杂性的关系浅析*

朱 鸿

(南京大学计算机软件研究所 南京 210093)
(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘要 软件可靠性估计是软件可靠性研究的重要问题之一.提出一种根据软件随机测试的结果进行软件可靠性估计的方法.它使被测软件的复杂性成为估计公式中的一个因素,从而克服了现有同类方法把软件看作是黑箱的缺点.它以软件的邻域集合的伪维数作为软件复杂性度量,从而使根据软件的可能错误空间进行可靠性估计成为可能.还证明了算法的计算复杂性与软件所可能计算的函数集合的伪维数之间存在着一定的联系.

关键词 软件可靠性,可靠性估计,随机测试,函数集合伪维数,复杂性分析.

中图法分类号 TP311

软件可靠性估计是软件可靠性研究的重要问题之一.随着计算机软件在安全性至关重要的应用领域中的迅速推广,软件可靠性日益显示出其重要性.所谓软件可靠性是指,软件在一定的运行环境下,在一定的运行时间内,软件不发生故障的概率.软件可靠性有多种度量,其中较常用的有软件可用率(Availability)、调用成功率(Probability of Success on Demand)等等.

本文探讨如何根据软件随机测试的结果进行软件可靠性估计.所谓随机测试是指,软件测试的测试实例集是根据软件输入空间上的某一概率分布由随机取样而获得的.如果该概率分布是软件实际使用中输入数据的分布,则称该随机测试是代表性随机测试,否则,称为非代表性随机测试.虽然应用非代表性随机测试的结果数据可以估计软件中存在错误的个数,但文献中尚无据此估计软件可靠性的直接方法.而代表性随机测试则直接与软件可靠性相联系.因此,本文仅讨论代表性随机测试.下文中软件随机测试均指代表性随机测试.

应用代表性随机测试的结果数据进行软件可靠性估计,较成熟的主要工作包括Howden的假设检验方法[1]、Parnas等人的二项分布方法[2]、Miller等人的分域估计方法[3]等等.这些方法都基于同一软件可靠性的模型,即把软件看作是一个从输入数据到输出数据的函数,其可靠性定义为软件从输入数据计算出正确的输出数据的概率.这一模型是软件可靠性调用成功率度量的数学抽象.这些工作都声称是保守的,即所作出的可靠性估计以很高的概率保证高于软件的实际可靠性.然而,这些可靠性估计方法都面临着一种类似悖论的尴尬局面.

假设按照某一可靠性估计公式,根据软件随机测试的结果数据可以得到软件可靠性估计值r.如果我们根据对软件的知识,可以把软件的输入空间分解成一系列子空间,那么,应用同一可靠性估计公式和相同随机测试数据,我们可以得到一个新的可靠性估计值r'.因为可靠性估计r是保守的,应用关于软件的知识应该使我们能够改进可靠性估计,使之更加接近软件的实际可靠性.所以,我们期待r'大于等于r.但是,应用现有的可靠性估计方法都只能得到小于r的r'.例如,应用随机测试方法测试某核电站安全保护系统中识别冷媒丢失故障的软件时,可把故障模式分成11种不同的情形,针对每一种情形选择5 000个随机测试数据.假设这11种故障发生的概率p_n是相同的,那么,如果测试时软件不发生故障,那么,按照*Miller*等人的分域估计方法,系统的失败率估计为$\hat{\theta}=\sum_{n=1}^{11}p_n\frac{a}{t_n+a+b}=1/5002$,其中$a,b\geqslant 1$为对软件中可能包含的错误的先验假设.如果我们不把故障模式分成11种模式,而把整个软件系统的输入空间看作一个整体,那么,使用5000×11个随机测试数据来测试该软件时,如果测试时软件不发生故障,那么,应用同一个失败率估计公式,我们可以得到该软件系统的失败率估计为$\hat{\theta}'=\frac{a}{t+a+b}=1/55002$.注意,这里$\hat{\theta}>\hat{\theta}'$.

---

* 本文研究得到国家863高科技项目基金、国家杰出青年基金、国家教委优秀年轻教师基金和国家教委留学回国人员基金资助.
作者朱鸿,1961年生,博士,教授,博士导师,主要研究领域为软件工程,软件安全性,软件测试.

本文通讯联系人:朱鸿,南京 210093,南京大学计算机软件研究所
本文1997-04-07收到原稿,1997-07-21收到修改稿

应用其他可靠性估计公式也会得到类似的结果.

我们认为,造成这一问题的主要原因是这些可靠性估计的方法都把软件看作是黑箱,忽略了软件的复杂性.从直觉上看,越复杂的软件需要的测试实例就越多.而上述这些可靠性估计方法对所有软件都要求同样多的测试实例.

我们在应用PAC(probably almost correct)归纳推理模型[4]对软件测试充分性准则的公理系统的研究中,发现软件复杂性与可靠性估计存在着一定的联系.[5,6]本文进一步分析这种联系作为软件可靠性估计方法的可能性.

1 软件可靠性估计

让 D 为非空集合,称为软件的输入数据空间.我们把程序 P 及其功能规约 S 都看作是 D 上的函数,即 $P,S:D\to D'$,其中 D' 为输出数据空间.让 Pr 为 D 上的一个给定的概率分布.程序 P 在概率分布 Pr 下相对于功能规约 S 的可靠性 $Rel_{Pr}(P,S)$ 定义为

$$Rel_{Pr}(P,S)=Pr(\{x\in D|P(x)=S(x)\}) \tag{1}$$

或等价地表示为

$$Rel_{Pr}(P,S)=\int_{x\in D}Pr(X)\times I(x)\mathrm{d}D,\quad \text{其中 } I(x)=\begin{cases}1, & \text{当 } P(x)=S(x) \text{ 时}\\ 0, & \text{当 } P(x)\neq S(x) \text{ 时}\end{cases},\ x\in D.$$

若概率分布 Pr 和功能规约 S 在上下文中已经明确,在不致造成含混的情况下,我们也将用 $Rel(P)$ 表示 $Rel_{Pr}(P,S)$,且称之为程序 P 的可靠性.

让 T 为根据 Pr 从 D 中随机抽样所得到的测试实例集.设 T 的样本大小为 m,软件 P 在 T 中 k 个数据上是正确的.若公式 γ 使得 $\gamma(m,k)\in[0,1]$,且在概率至少是 δ 的情况下,软件 P 的可靠性 $Rel(P)\geqslant\gamma(m,k)$,则称 γ 为或然度为 δ 的软件可靠性估计,简称为 δ-可靠性估计.

在文献[6]中,我们证明了下述定理.

定理1. 让 F 为一系列函数的集合,且软件 P 及其功能规约 S 均属于 F.设 T 为根据输入空间 D 上的概率分布 Pr 随机抽样而得,$0<\varepsilon,\delta<1$.若 $\dim_P(F)<\infty$,且程序 P 在 T 中的所有测试数据上都正确,则 P 的可靠性大于等于 $2\sup\{\varepsilon|\varphi_P(\varepsilon,\delta)\geqslant\|T\|\}$ 的概率至少为 δ.其中 $\dim_P(F)$ 为函数集合 F 的伪维数,函数 $\varphi_P(\varepsilon,\delta)$ 为集合 F 的样本复杂性.

函数集合的伪维数的定义如下.

定义1(伪维数). [10]F 为从集合 Z 到实数集合 R 的函数集合.F 的伪维数,记为 $\dim_P(F)$,是最大的自然数 k,使得存在两个长度为 k 的序列 $X=(x_1,x_2,\ldots,x_k)\in Z^k$ 和 $U=(u_1,u_2,\ldots,u_k)\in R^k$.对 X 的任意一个子序列 $Y=(x_{i_1},x_{i_2},\ldots,x_{i_l})$,存在 $f\in F$,使对任意 $x_i\in Y$,都有 $f(x_i)+u_i>0$,且对任意 $x_i\notin Y$,有 $f(x_i)+u_i\leqslant 0$.若这样的 k 不存在,则 $\dim_P(F)$ 为无穷大.

Haussler 在文献[9]中证明:若函数集合 F 的伪维数是有限的,则集合 F 是PAC可学习的,且集合 F 的样本复杂性具有如下上界.

$$\frac{32}{\varepsilon}\left(2\dim_P(F)\ln\left(\frac{16e}{\varepsilon}\right)+\ln\left(\frac{8}{\delta}\right)\right)$$

由此可知,对任意 $\delta(0<\delta<1)$,下述公式是 δ-可靠性估计.

$$2\sup_{\varepsilon}\left\{\|T\|\leqslant\frac{32}{\varepsilon}\left(2\dim_P(F)\ln\left(\frac{16e}{\varepsilon}\right)+\ln\left(\frac{8}{\delta}\right)\right)\right\} \tag{2}$$

这一可靠性估计公式与现有软件可靠性估计的主要区别有两个.① 它把软件放在一个邻域 F 内来考虑.函数集合 F 反映了程序员对软件的认识,刻画了软件可能的变化范围.因此,据此对软件的可靠性进行估计不是把软件完全看作是黑箱.② 它包含了反映软件复杂性的伪维数,从而使我们能够根据软件的复杂性来调整可靠性估计.

例如,文献[8,9]中证明:若函数集合 F 是一个向量空间,则该向量空间的维数就是 F 的伪维数.因此,如果我们可以判断一个软件 P 所完成的计算是 N 个输入数据的线性函数,那么,我们可以让 F 成为所有 N 个自变元的线性函数的集合,此时,$\dim_P(F)=N+1$.如果软件 P 所完成的计算是 N 个自变元的2阶函数,则可让 F 为所有 N 个自变元的2阶函数的集合,此时,$\dim_P(F)=1+N+N^2$.

在软件排错性测试中,我们把软件放在一个邻域内进行考察,要求测试实例集能够将被测软件与邻域内的其他软件区分开来.这样的一个邻域称为该软件的一个可能的错误空间.[10]应用伪维数的如下性质,我们还可以根据错误空间的复杂性进行软件可靠性估计.

引理 1.[7,8]让 p 为一个给定的从 D 到实数的函数，G 为从 D 到实数的函数的集合. 则函数集合 $\{p+g \mid g \in G\}$ 的伪维数即为 G 的伪维数.

2　与计算复杂性的关系

可靠性估计公式(2)中，用函数集合的伪维数来度量软件的复杂性. 而直接计算一个软件及其功能规约所在的函数集合的伪维数常常是十分困难的. 是否可以用人们比较熟悉的计算复杂性度量来进行可靠性估计呢？这是本节将要讨论的问题.

让函数 $|.|: D \to N$ 为输入数据的复杂性度量，即它是从 D 到自然数集合的映射，$f: N \to N$. 设功能规约 S 所要求解决的问题和程序 P 的计算复杂性均为 $O(f(n))$. 定义 $J(f(n))$ 为满足如下条件的函数 P 的集合.

(1) P 为 D 上的函数；

(2) 存在一个计算复杂性为 $O(f(n))$ 的程序 G，P 可以由 G 来计算.

为了讨论方便，我们假设计算复杂性为算法中为了完成所需的计算任务进行加法运算的次数.

由第 1 节中所给出的可靠性估计公式，我们可以根据 $J(f(n))$ 的伪维数来对软件的可靠性进行估计. 然而，对于给定的计算复杂性 $f(n)$，计算 $J(f(n))$ 是十分困难的. 下面，我们给出 $O(f(n))$ 和 $J(f(n))$ 之间的一个不等式.

定理 2. 让 $O^*(f(n))$ 为函数集合 $\{g(|x|) \mid g(n)=O(f(n))\}$. 则

$$\dim_P(J(f(n))) \geqslant \dim_P(O^*(f(n))). \tag{3}$$

为了证明上述定理，我们首先证明如下有关伪维数的性质.

引理 2. 设 P, Q 为从 Z 到 R 的函数的集合，则有

$$P \subseteq Q \Rightarrow \dim_P(P) \leqslant \dim_P(Q).$$

证明：由伪维数的定义易得. 详细证明略. □

定理 2 的证明(1)：定义集合 $V(f(n))=\{\Gamma_{h(x)} \mid h(|x|)=O(f(n))\}$，其中 $\Gamma_{h(x)}$ 为形如如下程序所计算的函数.

```
Γh(x): Begin
         input x;
         y:=x;
         for n:=1 to h(|x|) do y:=y-1
         output y;
       End
```

为了讨论方便，不失一般性，计算时间复杂性不计入循环变量 n 的改变以及循环上界 $h(|x|)$ 的计算时间，而只计入循环体内的计算时间. 此时，$\Gamma_{h(x)}$ 的计算复杂性为 $h(|x|)$. $\Gamma_{h(x)}$ 所计算的函数为 $\Gamma_{h(x)}=x-h(|x|)$. 所以

$$\begin{aligned}\dim_P(V(f(n))) &= \dim_P(\{x+h(|x|) \mid h(n)=O(f(n))\}) \\ &= \dim_P(\{h(|x|) \mid h(n)=O(f(n))\}) \qquad \text{〈由引理 1〉} \\ &= \dim_P(O^*(f(n))).\end{aligned}$$

因为 $V(f(n)) \subseteq J(f(n))$，由引理 2，$\dim_P(V(f(n))) \leqslant \dim_P(J(f(n)))$. 命题得证. □

在上述证明中，我们从具有一定计算复杂性的程序中选择了一个子集，并证明了这个子集的伪维数与所有与该计算复杂度函数同阶的函数所构成的函数集合的伪维数相同. 问题是，所选择的子集能够在多大程度上反映整个集合的复杂程度呢？目前，这仍然是一个尚未解决的问题. 下面，我们给出定理 2 的第 2 种证明，从中可以看出这个子集的性质并非是独特的，具有相同性质的子集可能有许多.

引理 3. 让 f 为实数集合 R 上的严格单调连续函数，F 为从集合 Z 到实数集合 R 的函数的集合，则 $\dim_P(F)=\dim_P(f \circ F)$，其中 $f \circ F=\{f \circ g \mid g \in F\}$.

证明：由伪维数的定义可得. 详细证明略. □

定理 2 的证明(2)：定义集合 $U(f(n))=\{\Psi_{h(x)} \mid h(|x|)=O(f(n))\}$，其中 $\Psi_{h(x)}$ 为形如如下程序所计算的函数.

```
Ψh(x): Begin
         input x;
         y:=x;
         for n:=1 to h(|x|) do y:=y+y
         output y;
       End
```

$\Psi_{h(x)}$ 的计算复杂性为 $h(|x|)$. 而 $\Psi_{h(x)}$ 所计算的函数为 $\Psi_{h(x)}=x \cdot 2^{h(|x|)}$. 所以

$$\begin{aligned}\dim_P(U(f(n)) &= \dim_P(\{x \cdot 2^{h(|x|)} \mid h(n) = O(f(n))\}) \\ &= \dim_P(\{\log_2(x \cdot 2^{h(|x|)}) \mid h(n) = O(f(n))\}) \quad \text{〈由引理 3〉} \\ &= \dim_P(\{h(|x|) + \log_2 x \mid h(n) = O(f(n))\}) \\ &= \dim_P(\{h(|x|) \mid h(n) = O(f(n))\}) \quad \text{〈由引理 1〉} \\ &= \dim_P(O^*(f(n))\end{aligned}$$

因为 $U(f(n)) \subseteq J(f(n))$，由引理 2，$\dim_P(U(f(n)) \leqslant \dim_P(J(f(n))$. □

在上述证明中，集合 $U(f(n))$ 中的函数都是指数函数，它们是所有具有相同计算复杂性的函数中增长最快的，而其伪维数并不比证明(1)中的由线性函数构成的集合 $V(f(n))$ 的伪维数大. 但是，我们目前还无法证明定理 2 中的等式成立.

还需要指出的是，集合 $O^*(f(n))$ 是以 D 为定义域的函数的集合，它不同于自然数上的函数集合 $O^+(f(n)) = \{g(n) \mid g(n) = O(f(n))\}$. 通常，$O^+(f(n))$ 的伪维数较容易得到. 例如，若 $f(n) = n^2$，则 $O^+(f(n))$ 为所有形如 $\alpha + \beta n + \gamma n^2$ 的函数的集合，这是一个维数为 3 的向量空间. 因此，其伪维数为 3.

让 F 为从集合 Z 到 R 的函数的集合，h 为从 Z' 到 Z 的一个给定的函数. 定义集合 $F \circ h = \{f \circ h \mid f \in F\}$. 因为 $O^*(f(n)) = O^+(f(n)) \circ |.|$，如下关于伪维数的性质可以使我们从 $O^+(f(n))$ 的伪维数得到 $O^*(f(n))$ 的伪维数.

定理 3. 让 F 为从集合 Z 到 R 的函数的集合，h 为从 Z' 到 Z 的一个给定的函数. 则

$$\dim_P(F) \geqslant \dim_P(F \circ h).$$

证明：让 $\dim_P(F \circ h) = K$. 由伪维数的定义，存在 $X' \in Z'^K$，$U \in R^K$，使得对 X' 的任意一个子序列 Y'，存在 $f \circ h \in F$，满足条件

$$f \circ h(x'_i) + u_i > 0, \text{ 如果 } x'_i \in Y';$$
$$f \circ h(x'_i) + u_i \leqslant 0, \text{ 如果 } x'_i \notin Y'.$$

让 $X = \langle x_1, x_2, \ldots, x_K \rangle \in Z^K$，其中 $x_n = h(x'_n)$，则 X 具有如下性质：对 X 的任意一个子序列 Y，存在 $f \in F$ 满足条件

$$f(x_i) + u_i > 0, \text{ 如果 } x_i \in Y,$$
$$f(x_i) + u_i \leqslant 0, \text{ 如果 } x_i \notin Y.$$

因此，F 的伪维数大于等于 K. 命题得证. □

定理 4. 让 F 为从集合 Z 到 R 的函数的集合，h 为从 Z' 到 Z 的一个给定的函数. 如果 h 满足如下条件，对任意 $x \in Z$，存在 $x' \in Z'$，使得 $h(x') = x$，则如下等式成立.

$$\dim_P(F) = \dim_P(F \circ h).$$

证明：由定理 3，$\dim_P(F) \geqslant \dim_P(F \circ h)$. 因此，只需证明 $\dim_P(F) \leqslant \dim_P(F \circ h)$.

让 $\dim_P(F) = K$. 由伪维数的定义，存在 $X \in Z^K$，$U \in R^K$，使得对 X 的任意一个子序列 Y，存在 $f \in F$ 满足条件

$$f(x_i) + u_i > 0, \text{ 如果 } x_i \in Y,$$
$$f(x_i) + u_i \leqslant 0, \text{ 如果 } x_i \notin Y.$$

由函数 h 的性质，存在 $X' = \langle x'_1, x'_2, \ldots, x'_K \rangle \in Z'^K$ 使得 $h(x'_n) = x_n, n = 1, 2, \ldots, K$. 显然，对任意 $f \in F$，$f(x_i) + u_i > 0$ 当且仅当 $f(h(x'_i)) + u_i > 0$. 因此，对 X' 的任意一个子序列 Y'，都存在 $f \circ h \in F \circ h$，使得

$$f \circ h(x'_i) + u_i > 0, \text{ 如果 } x'_i \in Y';$$
$$f \circ h(x'_i) + u_i \leqslant 0, \text{ 如果 } x'_i \notin Y'.$$

由伪维数的定义，$F \circ h$ 的伪维数大于等于 K. 命题得证. □

3 结束语

本文提出了一种软件可靠性估计方法. 它使被测软件的复杂性成为估计公式中的一个重要因素，从而克服了现有的根据软件随机测试的结果进行可靠性估计的方法把软件看作是黑箱的缺点. 它以软件的邻域集合的伪维数作为软件复杂性度量，从而使根据软件的可能错误空间进行可靠性估计成为可能. 本文还证明了算法的计算复杂性与软件所可能计算的函数集合的伪维数之间存在着一定的联系. 但是，这两者之间关系的研究还是十分初步的，许多问题还有待进一步研究.

参考文献

1 Howden W E. Functional Program Testing and Analysis. New York: McGray-Hill, 1987

2 Parnas D L, Van Schouwen A J, Kwan S P. Evaluation of safety-critical software. Communications of ACM, 1990,33(6):636～648
3 Miller W M, Morell L J, Noonan R E *et al*. Estimating the probability of failure when testing reveals no failures. IEEE Transactions on Software Engineering, 1992,18(1):33～43
4 Valiant L C. A theory of the learnable. Communications of ACM, 1984,27(11):1134～1142
5 Zhu Hong. An induction theory of software testing. Science in China, 1995,38(supplement):58～72
6 Zhu Hong. A formal interpretation of software testing as inductive inference. Journal of Software Testing, Verification and Reliability, 1996,6(1):3～31
7 Dudley R M. Central limit theorem for empirical measures. Annals of Probability, 1978,6(6):899～929
8 Wenocur R S, Dudley R M. Some special Vapnik-Chervonenkis class. Discrete Mathematics, 1981,33:313～318
9 Haussler D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and Computation, 1992,100(1):78～150
10 Zeil S J. Testing for perturbations of program statements. IEEE Transactions on Software Engineering, 1983,SE-9(3):335～346

Toward a Relationship Between Software Reliability Estimation and Complexity Analysis

ZHU Hong

(*Institute of Computer Software Nanjing University Nanjing* 210093)
(*State Key Laboratory for Novel Software Technology Nanjing University Nanjing* 210093)

Abstract Estimation of software reliability according to random testing is of particular importance in software reliability engineering. The author proposes a method for software reliability estimation in this paper. It reckons the complexity of the computation as an important factor of reliability estimation so that software is not considered as a black-box. It uses the pseudo-dimension of software neighbourhood as a measure of software complexity. Therefore, it can be used to estimate software reliability according to fault-based random testing. The author also proves some relationships between computational complexity and the pseudo-dimension of the set of functions that can be computed within the complexity bound.

Key words Software reliability, reliability estimation, random testing, pseudo-dimension, complexity analysis.