

2000, 2(4)

可靠性工程

广义遗传算法

抽样理论

计算机科学 2000Vol 27No 4

基于广义遗传算法的自适应重要抽样理论^{*}An Adaptive Importance Sampling Theory
Based on The Generalized Genetic Algorithm

董 聪 郭晓华

(清华大学 北京 100084) (中科院模式识别国家重点实验室 北京 100080)

Abstract In the present paper, using the generalized genetic algorithm, the problem of finding out all design points in the case of generalized multiple design point is solved, establishing recursion-type bound-and-classification algorithm, the problem of reducing and synthesizing generalized multiple design points is also solved. The present paper shows that the adaptive importance sampling theory based on the generalized genetic algorithm is an more efficient tool for the reliability simulation of nonlinear systems.

Keywords Generalized genetic algorithm, Bound-and-classification algorithm, Adaptive importance sampling theory, Nonlinear system reliability simulation

1 引言

随着计算技术和软件工具的迅猛发展,具有良好普适性的数值模拟方法在许多领域开始受到了广泛的重视。由于 Harbitz^[1]、Melchers^[2]、Wu^[3]和本文作者^[4,5]的一系列工作,基于重要抽样理论的非线性系统可靠性分析技术取得了很大的进展。尽管如此,基于重要抽样理论的非线性系统可靠性分析技术中一些带有根本性的理论问题并没有彻底解决。突出表现是,对于广义多设计点问题,寻找所有设计点的求解算法不仅没有提出,甚至没有提出过合理的求解思路。造成这一局面的原因是,多设计点问题的求解需以非连通域非线性系统全局优化问题的解决为前提,而后一问题在传统优化理论的框架下难以解决。以 Morgan 的基因理论(1920)为生物基础,以 Eldridge 与 Gould 的间断平衡理论(1972)为进化策略,本文作者创立了用于求解非连通域非线性系统全局优化问题的广义遗传算法^[6-8]。在理论上,广义遗传算法采用了和 Holland 的经典遗传算法^[9]完全不同的定向进化模式,并通过隔离机制,使进化进程实现了严格的并行化。在工程应用方面,采用广义遗传算法,我们成功地解决了目前世界上单跨最长的铁路、公路双用桥—香港青马大桥传感器群的最优布点设计问题^[10,11]。这表明,基于广义遗传算法的全局优化方法在理论上是先进的,在大型工程中是可以实现的^[12]。采用广义遗传算法,本文解决

了广义多设计点问题中所有设计点的寻找问题;通过建立递归型的约界-归类算法,解决了广义多设计点的压缩和综合问题。本文的工作表明,非线性系统可靠性仿真中普遍存在的广义多设计点问题,其所有设计点的寻找、压缩和综合这一长期以来悬而未决的难题,可在基于广义遗传算法的自适应重要抽样理论的框架下得到有效的解决。

2 自适应重要抽样原理

一般地说,对于复杂的概率问题,作为统计计算的蒙特卡罗法在获得近似解方面是很有用的。蒙特卡罗法精度高,通用性强,随着计算机应用技术的发展与进步,以蒙特卡罗法为基础的各类数值模拟方法在非线性系统可靠性分析与评估领域的作用正在引起高度的重视。

定义随机设计变量矢量 $\bar{X}$ 的联合概率密度函数为 $f_{\bar{X}}(\bar{x})$, 则系统失效概率 P_f 为

$$P_f = \int_{\Omega_f} f_{\bar{X}}(\bar{x}) d\bar{x} \quad (1)$$

其中, Ω_f 为失效域。

对于复杂的概率问题,式(1)中的 P_f 通常不存在解析描述。因此,数值计算和数值模拟成为一种主要的计算手段。

引入指示函数 $I[\cdot]$, 其定义为:

$$I[\bar{x}] = \begin{cases} 1 & \bar{x} \in \Omega_f \\ 0 & \bar{x} \in \bar{\Omega}_f \end{cases} \quad (2)$$

^{*}国家自然科学基金(59505011, 59778039)、863 计划(863-2-443)和模式识别国家重点实验室基金资助项目。

其中, $\Omega_i = \bigcup_{j=1}^m \{G_j \leq 0\}$, G_i 表示第 i 个失效模式的安全裕量方程, $G_i < 0$ 表示第 i 个失效模式的失效域, $G_i > 0$ 表示安全域, $G_i = 0$ 表示安全边界, $\bar{\Omega}_i$ 是 Ω_i 的余集, m 是系统所含的失效模式总数。

采用以上符号, 式(1)可改写为:

$$P_i = \int_{\Omega_i} f_{\bar{X}}(\bar{x}) d\bar{x} \quad (3)$$

其中, Ω 表示 $\bar{X}$ 的定义域, $\Omega_i \subset \Omega$ 。

蒙特卡罗法的基本思想是按照设计变量的概率密度函数进行抽样, 将多次模拟试验中落入失效域 Ω_i 中的点数 N_i 与总抽样数 N 之比作为失效概率 P_i 的无偏估计 $\hat{P}_i$, 即

$$\hat{P}_i = \frac{N_i}{N} = \frac{1}{N} \sum_{j=1}^N I[\bar{x}_j] \quad (4)$$

$$\begin{cases} E(\hat{P}_i) = P_i \\ D(\hat{P}_i) = \frac{P_i(1-P_i)}{N} \approx \frac{\hat{P}_i(1-\hat{P}_i)}{N} \end{cases}$$

其中, $\bar{x}_j$ 是按概率密度函数 $f_{\bar{X}}(\bar{x})$ 抽取的第 j 组随机设计变量矢量 $\bar{X}$ 的观察值。

当 P_i 值很小时, $\bar{x}_j$ 落入 Ω_i 域中的几率很小, 也就是说抽样效率很低。

重要抽样法的基本思想是通过修改抽样过程, 用重要抽样密度函数代替原来的抽样密度函数, 使对 P_i 贡献大的抽样出现的几率增加, 从而提高抽样效率。其公式为

$$P_i = \int_{\Omega_i} \int_{\Omega'} I[\bar{v}] \frac{f_{\bar{X}}(\bar{v})}{h_{\bar{V}}(\bar{v})} h_{\bar{V}}(\bar{v}) d\bar{v} \quad (5)$$

其中, $h_{\bar{V}}(\bar{v})$ 是新选的重要抽样密度函数, Ω' 是 $\bar{V}$ 的定义域。 $\bar{V}$ 的定义域 Ω' 可以和 $\bar{X}$ 的定义域 Ω 不同, 只需满足限制条件 $\Omega_i \subset \Omega'$ 即可。

作为概率密度函数, $h_{\bar{V}}(\bar{v})$ 应满足以下条件:

$$\begin{cases} h_{\bar{V}}(\bar{v}) > 0 \\ \int_{\Omega'} h_{\bar{V}}(\bar{v}) d\bar{v} = 1 \end{cases} \quad (6)$$

P_i 的模拟均值和方差分别为:

$$\begin{cases} E(\hat{P}_i) = P_i \approx \hat{P}_i = \frac{1}{N} \sum_{j=1}^N \left\{ I[\bar{v}_j] \frac{f_{\bar{X}}(\bar{v}_j)}{h_{\bar{V}}(\bar{v}_j)} \right\} \\ D(\hat{P}_i) \approx \frac{1}{N-1} \left\{ \frac{1}{N} \sum_{j=1}^N I[\bar{v}_j] \left[\frac{f_{\bar{X}}(\bar{v}_j)}{h_{\bar{V}}(\bar{v}_j)} \right]^2 - \hat{P}_i^2 \right\} \end{cases} \quad (7)$$

在系统失效概率计算中, 重要抽样密度函数的选择应当兼顾这样两个方面: ①增加有效抽样比例; ②增加有效抽样中对 P_i 值贡献大的抽样出现的几率。

增加有效抽样比例的方法很多, 最常用的方法就是选择合适的抽样域。由于失效域 Ω_i 和抽样域 Ω' 之间存在关系 $\Omega_i \subset \Omega'$, 因此, Ω' 愈接近 Ω_i , 则有效抽样比例也就愈大。基于试探方式的递归型自适应优化算法可用来寻找 Ω' 。

• 2 •

Melchers 建议以系统的安全裕量方程按 H-L 算法求得的设计点为采样中心, 以 N 维无关正态概率密度函数 φ 作为统一的重要抽样密度函数。Melchers 的建议无疑受到了 Shinozuka 关于设计点为最可几失效点 (the Most Probable Point, MPP) 这一流行观点的影响。从下面的论述中可以看出, Shinozuka 的观点并不严格正确。

我们认为, 对于蒙特卡罗模拟来说, 将采样中心选在失效面上既不充分, 也不必要。在重要抽样法中, 本文建议初始采样中心按以下方式确定:

$$\begin{cases} f_{\bar{X}}(\bar{x}^*) = \max f_{\bar{X}}(\bar{x}) \\ g(\bar{x}^*) \leq 0 \end{cases} \quad (8)$$

当且仅当 $\bar{X}$ 的所有分量 X_i 均服从正态分布且统计无关时, 式(8)中的 $\bar{x}^*$ 为传统意义上按 H-L 算法求得的设计点。目前流行的按 H-L 算法求解 $\bar{x}^*$ 的方式等价于式(9)。

$$\begin{cases} f_{\bar{X}}(\bar{x}^*) = \max f_{\bar{X}}(\bar{x}) \\ g(\bar{x}^*) = 0 \end{cases} \quad (9)$$

采样中心的选择对有效抽样所占的比例及有效抽样中对 P_i 贡献大的子样所占的比例均会产生重要的影响。当选 N 维无关正态概率密度函数作为重要抽样密度函数时, 重要抽样法必须解决的另一个问题是如何确定 N 维正态概率分布的方差。在方差选择上, 通常可取原始方差的某一个倍数作为重要抽样密度函数中相应随机变量的方差。另外一种可以采用的采样中心和方差的自适应迭代算法由式(10)表述。

$$\begin{cases} E(\bar{V}) = E(\bar{X} | \bar{X} \in \Omega_i) \\ D(\bar{V}) = D(\bar{X} | \bar{X} \in \Omega_i) \end{cases} \quad (10)$$

其中, $E(\bar{V})$ 为采样中心。

不难看出, 式(8)是以增加有效抽样中对 P_i 值贡献大的抽样出现的几率为目标来确定采样中心, 而式(10)则是以增加有效抽样比例为目标来确定采样中心, 可以推知, 最优采样中心应在两者之间。

我们的研究表明, 采用由式(8)确定的 $\bar{x}^*$ 为采样中心, 使用原始方差即可获得高效率的模拟计算结果。如果需要的话, 可进一步利用式(10), 采用自适应迭代的方式寻找最优采样中心。

在重要抽样法中, 有 4 种因素是可以经过迭代计算的方式进行优选的。它们是: ①重要抽样密度函数的类型; ②抽样域; ③采样中心; ④方差。很明显, 对任一因素或其组合均可建立相应的自适应优化算法。

期望估计法是重要抽样法中最特殊的一种, 其修改抽样函数为有限域 Ω_0 内的均匀分布函数, 计算程序如下:

1) 确定设计点 $\bar{x}^*$ 为抽样区域的参考点。目前 $\bar{x}^*$ 通常取为 $\bar{X}$ 的均值点或由 H-L 方法确定的设计点, 本文建议 $\bar{x}^*$ 由式(8)确定;

2) 定义一个由 $[x_i^* - k\sigma_i, x_i^* + k\sigma_i]$ ($i=1, 2, \dots, n$) 构成的 n 维有界闭区域 Ω_0 为抽样域, 抽样区域的广义体积记为 A 。通常取 $k=3\sim 5$;

3) 在 Ω_0 中抽取均匀分布的随机矢量 $\bar{z}_j, j=1, 2, \dots, N$;

4) 按下式近似估算 P_j :

$$\hat{P}_j = \frac{A}{N} \sum_{i=1}^N f(\bar{z}_i) f_{\bar{x}}(\bar{z}_i) \quad (11)$$

从上面的计算过程可以清楚地看出: $\hat{P}_j$ 的计算正确性和准确性依赖于以下 3 个因素: ① Ω_j 是否有界。②若 Ω_j 无界, 则 P_j 是否可用 $f_{\bar{x}}(\bar{x})$ 在一有限闭区域 Ω_k 内的积分来逼近。③ Ω_0 和 Ω_j 之间的相互关系。

在实际问题中, Ω_j 通常是无界的。因此, 我们必须假定 P_j 可用 $f_{\bar{x}}(\bar{x})$ 在一有限域内的积分来逼近。这一假定成立的条件是什么, 目前缺乏应有的研究。大量的研究工作用在改进模拟结果的收敛速率上, 而忽略了其收敛解是否为真值这个最关键的问题。

改进期望估计法计算效率的关键在于选择合适的采样中心。很明显, 最优采样中心必为失效域内的某一点, 通常情况下不会位于失效面上。采样中心的迭代可采用自适应算法来实现, 基于经验贝叶斯推理的自适应优化算法是一个值得注意的发展方向^[12]。

需要指出的是, P_j 不仅可用数值模拟的方法, 也可用数值计算的方法直接计算。其中, 以数论中的佳格点集为采样点的拟蒙特卡罗法可能会成为一个有前途的研究方向。

3 广义多设计点的选择与合成方法

前面已叙及, 可由广义遗传算法求得广义多设计点问题中的各设计点, 接下来我们研究采用下文给出的约界-归类算法解决广义多设计点的压缩与合成问题。

设系统共有 n 个设计点, 这 n 个设计点可分成 m 组。首先将 n 个设计点按其可靠度指数 β_j 由小到大的顺序排列, 并将其序号构成的集合定义为 B 。将第 i 组设计点构成的集合定义为 A_i , Φ 定义为空集, ϵ_j 定义为设计点 i 和 j 在正则化空间中的欧几里得距离, 则分组可按 ([13]) 方式进行:

① $\beta_{\min} = \min(\beta_j), \forall j \in B$ 。

② $\forall j \in B$, 如果 $\beta_j > c_p \beta_{\min}$, 将 j 从 B 中删除, 否则, 不做处理。

③ 置 $i=1$ 。

④ $\beta_{\min}^i = \min(\beta_j), \forall j \in B$ 。

⑤ 将 $\beta_{\min}^i$ 所对应的设计点代码赋给 k 。

⑥ $\forall j \in B$, 若 $\epsilon_{kj} < \epsilon_c(\beta_k, \sigma_k)$, 则将设计点归入集合 A_i , 同时将 j 从集合 B 中删除。

⑦ 若 $B=\Phi$, 则转入 ⑧, 否则 $i=i+1$, 转入 ④。

⑧ $m=i$, 分组完毕。

⑨ $\forall j \in A_i$, 如果 $\beta_j > c_p \beta_{\min}^i$, 将 j 从 A_i 中删除, 否则, 不做处理。

以上约界-归类算法中, ④和⑤完成初步约界任务, 最大限度地减少了有效设计点的数目。③~⑧完成归类任务, 最大限度地减少了抽样中心的数目。从逻辑上讲, 分组完毕之后还可以采用约界算法对组内的设计点进行二次压缩, 过程如⑨所述。约界-归类-约界的结合使实际抽样空间压缩至最小。 c_p 为约界参数, 通常取 $c_p=1.5$, $\epsilon_c(\beta_k, \sigma_k)$ 是一个阈值, 其取值原则上与 β_k 和 σ_k 有关。在实际运算中, 可以采用以下算式:

$$\epsilon_c(\beta_k, \sigma_k) = c \left(\frac{\beta_k}{\beta_{\min}} \right)^q \sigma_k, 0 < q < 1 \quad (12)$$

其中, c 和 q 为常数, 通常取 $c=2\sim 3, q=0.5\sim 0.8$ 。

设系统的所有设计点已按上述方法分为 m 组, 按照确定的规则从每组中找出一个设计点作为此组的代表点, 通过自适应迭代算法得到最优抽样中心, 则系统的重要抽样密度函数选择方法和系统综合失效概率由下式给出。

$$P_f = \sum_{i=1}^m a_i \int_{\Omega_i} \dots \int_{\Omega_i} \frac{f_{\bar{x}}(\bar{v}_i)}{\sum_{j=1}^m a_j h_{v_j}(\bar{v}_i)} h_{v_j}(\bar{v}_i) d\bar{v}_i \quad (13)$$

a_i 由下式计算:

$$a_i = \frac{P_{f_i}}{\sum_{i=1}^m P_{f_i}} \approx \frac{\Phi(-\beta_i)}{\sum_{i=1}^m \Phi(-\beta_i)} \quad (14)$$

其中, Φ 表示标准正态分布的累计分布函数。

式 (13) 中的积分域是 Ω_f 而不是 Ω_j , 这一结果使得并行处理十分困难。为实现并行处理, 我们可根据上述分组情况将失效域 Ω_f 划分成 m 个互不相交的子区域 $\Omega_{f_i}, i=1, 2, \dots, m, \Omega_f = \bigcup_{i=1}^m \Omega_{f_i}$, 通过在 $f_{\bar{x}}(\bar{x})$ 值较大的子区域中抽取较多的子样, 来达到提高抽样效率和减小方差的目的。计算公式为:

$$\begin{cases} P_f = \int_{\Omega_f} f_{\bar{x}}(\bar{x}) d\bar{x} = \sum_{i=1}^m \int_{\Omega_{f_i}} f_{\bar{x}}(\bar{x}) d\bar{x} = \sum_{i=1}^m Int_i \\ Int_i = \int_{\Omega_{f_i}} f_{\bar{x}}(\bar{x}) d\bar{x} \end{cases} \quad (15)$$

在每个子域 Ω_{f_i}, Int_i 采用上文介绍的自适应重要抽样算法进行计算。每个子域 Ω_{f_i} 中的抽样点数 N_i , 由式 (16) 确定:

$$N_i = N \times a_i \quad (16)$$

4 自适应重要抽样原理的应用前景

我们发现, 当安全裕量方程呈现强非线性时, 国际

结构安全性联合委员会(JCSS)推荐的、用于模式失效概率计算的 JC 算法(即 R-F 算法)的计算精度得不到保证,且在非正态变量向正态变量转换的迭代计算中,迭代结果有时会出现不收敛或收敛于非真值的现象。即使对于一些比较简单的非线性问题,JC 算法的计算精度也得不到保证,比如,取安全裕量方程为常见的二次型: $Z=C-X_1X_2<0, X_1, X_2>0; X_1\sim N(0,1), X_2\sim N(0,1), X_1$ 和 X_2 相互独立,当 C 取 2.50, 2.75, 3.00, 3.50, 3.75 时,由 JC 算法计算的模式失效概率 P_f 及

相应的真值由表 1 所列。可见,JC 算法的计算误差相当大。对同一问题,采用上文介绍的自适应重要抽样算法 500 次抽样的模拟计算结果由表 1 所列,可见计算误差很小。大量的模拟计算结果表明,随着安全裕量方程非线性程度的增加,JC 算法的计算精度愈来愈差,而自适应重要抽样算法的模拟精度几乎不发生变化。因此,在可靠性工程领域,以自适应重要抽样算法为代表的数值模拟算法更适合作为处理强非线性问题的基础。

表 1 自适应重要抽样算法和 JC 算法的精度比较

安全裕量方程 $Z=C-X_1X_2$	模式失效概率计算与模拟结果				
	JC 算法		自适应重要抽样算法		真值
	计算结果	误差(%)	模拟结果	误差(%)	
$Z=2.50-X_1X_2$	0.6210×10^{-2}	45.57	0.4307×10^{-2}	0.96	0.4266×10^{-2}
$Z=2.75-X_1X_2$	0.2980×10^{-2}	45.01	0.2073×10^{-2}	0.88	0.2055×10^{-2}
$Z=3.00-X_1X_2$	0.1350×10^{-2}	44.54	0.9337×10^{-3}	-0.03	0.9340×10^{-3}
$Z=3.50-X_1X_2$	0.2326×10^{-3}	43.85	0.1570×10^{-3}	-2.90	0.1617×10^{-3}
$Z=3.75-X_1X_2$	0.8842×10^{-4}	43.56	0.6201×10^{-4}	1.15	0.6159×10^{-4}

致谢 在论文完成过程中,作者之一董聪的硕士研究生丁李梓、董树利、陈林和李鹏进行了大量的模拟计算,他们的工作为本文有关算法的完善和定型起了非常重要的作用。在广义遗传算法创立之初,作者的工作得到过航空科学基金(95B51062)、中国博士后基金、CIMS 科研基金、清华大学智能技术与系统国家重点实验室基金和攀登计划重大项目的部分资助,在此谨向以上机构及个人的帮助和支持表示衷心的感谢。

参考文献

- 1 Harbitz A. An Efficient Sampling Method for Probability of Failure Calculation. *Struct. Safety*, 1986, 3(1): 109~115
- 2 Melchers R E. Importance Sampling in Structural System. *Struct. Safety*, 1989, 6(1): 3~10
- 3 Wu Y T. Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis. *AIAA Journal*, 1994, 32(8): 1717~1723

- 4 董聪. 非线性系统可靠性分析的重要抽样法. *强度与环境*, 1996(2): 60~64
- 5 董聪, 刘西拉. 非线性结构系统可靠性理论及其模拟方法. *土木工程学报*, 1998, 31(1): 33~43
- 6 董聪. 广义遗传算法. *大自然探索*, 1998, 17(1): 33~37
- 7 董聪, 郭晓华. 广义遗传算法的逻辑结构与全局收敛性的证明. *计算机科学*, 1998, 25(5): 38~42
- 8 董聪, 郭晓华. 广义遗传算法的数学结构. *中国科学基金*, 1999, (2): 77~80
- 9 Holland J H. *Adaptation in Natural and Artificial Systems*. MIT Press, 1992
- 10 董聪, 秦权, 陈一珉. 基于广义遗传算法的青马桥传感器群最优布点设计. *CCNS'97*, 南京: 人民邮电出版社, 1997: 479~484
- 11 董聪, 郭晓华. 基于广义遗传算法的全局优化方法. *计算机科学*, 1999, 26(6)
- 12 李虎军. 桥梁诊断与遗传算法. *科学时报*, 1999-4-26
- 13 董聪. 可靠性工程中的经验 Bayes 方法(I). *北京航空航天大学学报*, 1996, 22(4): 460~464