基于期望的重要抽样方法研究

2010-12-28 14:24:02

周泓/邱月

【内容提要】

传统的Monte Carlo方法仿真稀有事件需要较长的时间，而重要抽样技术可以有效地缩短仿真时间，提高仿真效率。文章提出一种新的重要抽样实现方法，用来估计仿真模型中的稀有事件的概率；利用期望寻找最优重要抽样分布函数，并与传统的Monte Carlo算法进行比较。仿真结果显示了该方法在估计稀有事件概率方面的有效性。

【关 键 词】稀有事件/重要抽样/期望/似然比

引言

稀有事件是一种发生概率低但后果严重的事件，如亚洲金融危机，东南亚大海啸等。上世纪六、七十年代，美国的核电厂在进行安全性分析时，最先涉及到对稀有事件的风险评估[1]。随后在化学工业、环境保护、航天工程、医疗卫生、交通运输、经济等领域得以推广和应用[4～7]。虽然，稀有事件的发生概率比较低，但一旦发生就会造成严重的后果，对人类的生产和生活产生巨大的影响。这些成了推动稀有事件相关学科兴起和发展的主要动力，关于稀有事件的决策问题的研究受到很高的重视。在针对稀有事件问题的决策中，一个非常关键的问题就是对事件发生的可能性加以科学估计，它是决策的基础和前提。目前，大部分稀有事件风险评估多集中于定性分析层面，由于稀有事件发生机理的复杂性和发生模式的多样性，定量化分析的难度很大。因此稀有事件的定量化分析具有重要的理论和现实意义[1～3]。

应用仿真技术可以更加直观地分析不确定性因素的表现形式和后果。也就是说，通过在仿真过程中模拟影响重大的稀有事件的发生及其发生后的系统运行情况，监测运行中性能测度的变化，能够有效地评估此类稀有事件对系统性能的影响，进而寻找可行的解决办法。传统的Monte Carlo仿真(MCS)方法在估计稀有事件发生的概率时，需要进行数目庞大的仿真实验，只有这样才有可能得到有价值的结果，而这往往超过了计算机的承受能力[3，8]。重要抽样(Importance Sampling，IS)技术可以很好地解决这个问题[3]。重要抽样的主要思想是通过尺度变换(Change of Measure，CM)来修改决定仿真输出结果的概率测度，使本来发生概率很小的稀有事件频繁发生，从而加快仿真速度，在较短的时间内得到稀有事件[2]。

本文提出一种基于期望的解决稀有事件概率估计的方法——应用判断抽样密度函数与最优重要抽样分布函数的期望是否为1来实现重要抽样的方法。这种方法通过极小化抽样密度函数与最优重要抽样分布函数的期望与1之间的距离，进而从分布族{f(•；v)}中确定参数向量v来选取一个密度函数，使 $g_{opt}(x)$ 与f(•；v)距离最近，最后得到稀有事件的概率估计。本文还将该方法应用在项目进度管理问题中，并与Monte Carlo仿真(MCS)方法进行比较。

一、重要抽样

重要抽样技术利用修改了的概率密度函数进行抽样，得到以较高概率出现的样本，然后通过对其输出结果加权来补偿由修改密度函数带来的偏差。按这种思路，可以在较短的时间内得到稀有事件[9]。

（一）稀有事件

假定 $\{f(\cdot;v)\}$ 是概率密度函数族，其中 v 是参数向量。X 是具有密度函数 $f(\cdot;U)$ 的随机变量，s 是一个实值函数，要估计事件 $\{s(x)\geqslant r\}$ 概率，如果 $l=\mathrm{P}\{s(x)\geqslant r\}$ 很小，例如 $l<10^{-5}$，就称 $\{s(x)\geqslant r\}$ 是一个稀有事件。传统的 Monte Carlo 仿真方法就是根据密度函数 $f(\cdot;U)$，随机产生 N 个独立同分布的样本 $X_1,\cdots,X_N$，若样本数足够大，则可以用

$$\hat{l}=\frac{1}{N}\sum_{i=1}^{N}I_{\{s(X_i)\}}=\frac{1}{N}\sum_{i=1}^{N}I_{\{s(X_i)\geqslant r\}} \tag{1}$$

作为 l 的无偏估计量。其中 $I_{\{s(X_i)\}}$ 是指标函数，满足

$$I_{\{s(X_i)\geqslant r\}}=\begin{cases}1 & s(X_i)\geqslant r\\ 0 & s(X_i)<r\end{cases} \tag{2}$$

但是，当 $\{s(x)\geqslant r\}$ 是一个稀有事件时，(1) 式右端的指示函数 $I_{\{s(X_i)\geqslant r\}}$ 大部分为零。为了精确估计稀有事件概率 l，需要的仿真次数就会相当地多。

(二) 重要抽样分布函数

重要抽样是用一个新的概率密度函数 $g(x)$ 代替 $f(\cdot;U)$，并满足

$$l=\int I_{\{s(X_i)\}}f(\cdot;U)\mathrm{d}x=\int_{\{s(x)\geqslant r\}}f(\cdot;U)\mathrm{d}x=\int_{\{s(x)\geqslant r\}}L(x)g(x)\mathrm{d}x \tag{3}$$

其中 $L(x)$ 为一个加权函数，$L(x)=\dfrac{f(\cdot;U)}{g(x)}$ 称为似然比函数。这里 $g(x)$ 可以是满足上式的任意概率密度函数，称为重要抽样分布函数。

在估计稀有事件概率时，首先根据密度函数 $g(x)$ 随机产生 R 个独立同分布的样本 $X_1,\cdots,X_R$，然后利用(3) 式得出 l 的重要抽样估计量，如 (4) 式所示。

$$\hat{l}_{IS}=\frac{1}{R}\sum_{i=1}^{R}I_{\{s(X_i)\geqslant r\}}L(X_i)=\frac{1}{R}\sum_{i=1}^{R}I_{\{s(X_i)\geqslant r\}}\frac{f(X_i;U)}{g(X_i)} \tag{4}$$

由以上过程可以看出，通过对 $f(\cdot;U)$ 分布的随机变量采用 Monte Carlo 抽样得到的稀有事件发生概率估计 $\hat{l}$，与对 $g(x)$ 分布的随机变量采用重要抽样得到的稀有事件发生概率估计 $\hat{l}_{IS}$，二者的数学期望相同，均为 l 的无偏估计，但是它们的方差不同，所以选择 $g(x)$ 时就要极为谨慎。重要抽样密度函数 $g(x)$ 的选择是重要抽样仿真成功的关键。如果选择了合适的 $g(x)$，则只需很少的样本数便可得到的一个可靠估计；否则重要抽样过程的效率可能很差。

二、基于期望的重要抽样

式 (4) 中 $\hat{l}_{IS}$ 的方差可由下式估算：

$$\mathrm{Var}[\hat{l}_{IS}]=\frac{1}{R}\mathrm{Var}_{g(x)}[I_{\{s(X_i)\}\geqslant r}L(x)]$$

不难看出，如果将重要抽样密度函数选为

$$g(x)=g_{opt}(x)=I_{\{s(X_i)\geqslant r\}}\frac{f(x;U)}{l} \tag{5}$$

则 $\hat{l}_{IS}(x)$ 的方差为零，与仿真的次数无关，称 $g_{opt}(x)$ 为最优重要抽样分布函数。但由于 l 是一个未知量，所以想得到最优的重要抽样密度函数是不可能的。

由以上分析可见，最优重要抽样分布估计量的有效性完全依赖于最优重要抽样分布函数的选择，一种有效的处理办法是在分布族 $\{f(\cdot;v)\}$ 中通过确定参数向量 v 选取一个密度函数使其逼近最优重要抽样密度函数，即使 $v(\cdot;v)$ 与 $g_{opt}(x)$ 的距离最近。

命题 1 假设随机变量 X 的密度函数为 $f(x)$，$g(x)$ 是任意的概率密度函数，则 $E\left[\dfrac{f(X)}{g(X)}\right]=1$ 的充分必要条件是 $f(x)=g(x)$。

证明：若 $f(x)=g(x)$，显然 $E\left[\dfrac{f(X)}{g(X)}\right]=1$ 成立。

下面证明，若 $E\left[\dfrac{f(X)}{g(X)}\right]=1$，可得 $f(x)=g(x)$。

$$E\left[\frac{f(X)}{g(X)}\right]=\int\frac{f(x)}{g(x)}f(x)\mathrm{d}x=\int\frac{f^2(x)}{g(x)}\mathrm{d}x+\int g(x)\mathrm{d}x-\int g(x)\mathrm{d}x=\int\left[\frac{f^2(x)}{g(x)}+g(x)\right]\mathrm{d}x-1$$

$$\geqslant 2\int\sqrt{\frac{f^2(x)}{g(x)}}g(x)\mathrm{d}x-1\geqslant 2\int f(x)\mathrm{d}x=1 \tag{6}$$

因为 $g(x)$ 是概率密度函数，所以可得 $\int g(x)\mathrm{d}x=1$。根据柯西不等式当且仅当 $\dfrac{f^2(x)}{g(x)}=g(x)$ 时，等号成立。因为 $g(x)$ 和 $g(x)$ 是概率密度函数，所以可得 $f(x)=g(x)$。

由前所述，当 $\dfrac{f(X)}{g(X)}$ 期望为 1 时，随机变量 X 的概率密度函数 $f(x)$ 等于 $g(x)$。所以极小化 $g_{opt}(x)$ 与 $f(\cdot;v)$ 之间的距离就等价于将 $g(x)$ 选取为最优重要抽样分布函数 $g_{opt}(x)$，当根据密度函数 $f(\cdot;v)$ 产生 X 时，$\dfrac{f(X;v)}{g_{opt}(X)}$ 的期望为 1，也即 $\left|E\left[\dfrac{f(X;v)}{g_{opt}(X_i)}\right]-1\right|$ 最小，来确定参数 v。将(5) 式中 $g_{opt}(x)$ 的表达式代入，得到下面的极小化问题

$$\min_{v}\left|E_v\left[\frac{lf(X;v)}{I_{\{s(X_i)\geqslant r\}}f(X;U)}\right]-1\right| \tag{7}$$

再运用一次重要抽样技巧，得到

$$\min_{v}\left|E_w\left[\frac{lf(X;v)}{I_{\{s(X_i)\geqslant r\}}f(X;U)}W(X;v,w)\right]-1\right| \tag{8}$$

其中，$W(X;v,w)=\dfrac{f(X;v)}{f(X;w)}$ 是似然比函数。

从前面叙述可见，算法的关键在于事件 $\{s(X)\geqslant r\}$ 的概率不能太小，而对于稀有事件来说，这个概率恰恰是很小的，所以指示函数 $I\{s(X_i)\geqslant r\}$ 大部分很小。为保证事件发生次数达到分析所需要的样本量，在上面的算法过程中，同时建立了两个更新参数序列 $\{v_t,t\geqslant 0\}$ 和 $\{r_t,t\geqslant 1\}$，使得

$p_{v_{t-1}}\{s(x) \geqslant r_t\} \geqslant \rho(\rho \in (0,1))$ 为事先定义的数。

式（8）中，指示函数处在分母的位置，所以将（2）式中定义的指示函数 $I_{\{s(X_i)\geqslant r\}}$ 修改定义为

$$I_{\{s(X_i)\geqslant r\}} = \begin{cases} 1 & s(X_i) \geqslant r \\ 0 & s(X_i) < r \end{cases}$$

l 是一个待估的量，采用迭代方法来解决这个问题，即用一个更新参数序列 $\{l_t, t>0\}$ 来解决这个问题，亦即

$$l_t = \frac{1}{N}\sum_{i=1}^{N} I_{\{s(X_i)\geqslant r\}} W(X_i; u, v_{t-1}) \qquad (9)$$

综上，本文所建立的基于期望的重要抽样算法过程如下：

①赋初值 $v_0 = u$，$\rho = 0.1$；

②从重要抽样密度函数 $f(\cdot; v_{t-1})$ 中生成样本 $X_1, \cdots, X_N$，计算样本的 $(1-\rho)$ 分位数并赋值给 r_t，当 $r_t < r$ 时，更新 r，按照（9）式更新 l_t；

③使用②中的样本 $X_1, \cdots, X_N$ 求解问题（8）得到新解 v_t；

④当 $r_t < r$ 时，迭代次数增加 1，转②；否则，结束循环，求出问题的解 v^*；

⑤利用（4）式计算稀有事件概率。

三、仿真示例

下面以项目管理中的一类稀有事件为例来说明本文算法的有效性。在某项工程中，关键路线上共有5个关键活动。令x表示每一个活动的工时，s(x)表示这项工程总的完工期，所以s(x)是5个变量的和。本文所考虑的问题是总的完工期大于一个给定时间r的事件(s(x))＞r)的概率，即需要估计l=P(s(x)＞r)。当r＞r*时，(s(x)＞r)是一个稀有事件，此处r*=arg(P(S(x)＞r)= 10^{-5}。在不影响算法有效性的前提下，考虑每项活动的工时分别服从均值为μ=25的指数分布的情况。本文对给定的两个r值500和800进行仿真，表1、表2显示了仿真结果。表中 $\hat{l}$ 是l的估计量，N是仿真次数，90%H.W.表示90%置信区间半长。使用基于期望的重要抽样方法估计l=P(s(x)＞r)，其结果列于表2。本文同时用标准的Monte Carlo仿真方法来估计l=P(s(x)＞r)，结果列于表1。由于仿真试验具有随机性，所以对每种情况独立做了100次仿真试验，表中数据均取自100次独立试验的平均值。

由表1和表2的对比不难看出，当r=500时，传统的Monte Carlo仿真算法需要进行200000次仿真实验才可以得到置信区间半长为3.2800e-006的估计量；而基于期望的重要抽样方法在进行1000次仿真实验后就可以得到置信区间半长为5.4580e-006的估计量，计算时间大大减少。当r=800时，传统的Monte Carlo仿真算法进行5000000次仿真实验后，仿真不到稀有事件。而本文提出重要抽样方法在分别进行3000次仿真实验后，得到了令人满意的结果。可见本文提出的方法对于小概率事件的计算精度高于其他几种方法，所以对于稀有事件来说，本文提出的方法的计算结果更可靠。

表1　基于Monte Carlo方法估计稀有事件的概率

r	$\hat{l}$	90%H.W.	N
500	1.0000e－006	3.2800e－006	200000
800	－	－	5000000

注：表中的“－”表示经过了5000000次抽样后，稀有事件没有发生，以至于无法建立有效的区间估计。

表2　基于期望的重要抽样方法估计稀有事件的概率

r	$\hat{l}$	90%H.W.	N
500	1.2556e－006	5.4580e－006	1000
800	4.0538e－012	3.2506e－012	3000

以上分析显示出本文算法与传统的Monte Carlo仿真算法相比，仿真效率有明显提高。很显然，对于小概率事件传统的Monte Carlo仿真方法无能为力，但是本文算法仍然可以在较少的仿真次数下得到理想的结果。由此验证了本文提出的方法对于计算稀有事件概率这类问题的有效性。

四、结论

本文通过判断待确定的抽样密度函数与最优重要抽样密度函数的比值期望是否为1的思想，提出了一种新的基于期望的重要抽样方法。该算法将期望与重要抽样算法相结合，通过极小化抽样密度函数与最优重要抽样密度函数的比值的期望与1之间的距离，来选取重要抽样分布函数，再根据最优重要抽样分布函数生成样本，得到稀有事件概率的估计量。并将仿真结果与标准的Monte Carlo仿真方法进行比较，结果显示出本文算法在估计稀有事件概率方面的有效性。

稀有事件仿真方法的研究主要有交叉熵方法和极小化方差的方法[5]，因此在理论方面还存在着广阔的发展空间。理论方面如初值的选取和如何在仿真中应用方差衰减技术都需要进行更深入的研究。另外，该方法可用于经济危机和金融预警等领域的分析。在方法的改进上，可以考虑和其他的方差衰减技术相结合，以期构成更加有效的稀有事件仿真方法。

【参考文献】

[1]大亚湾核电站PRA项目组.大亚湾核电站概率风险分析[J].中南工学院学报，1999,13(2).

[2]Hammersley, J, M, and D. C, Handscomb. Moute Carlo Methods[M]. Methuen, London, 1964.

[3]M Hsieh. Adaptive Importance Sampling for Rare Event Simulation of Queuing Networks[D]. Ph. D. thesis, Stanford University, California, 1997.

[4]Hong ZHOU, Yue QIU, Yue-qin WU. An Early Warning System for Loan Risk Assessment Based on Rare Event Simulation[J]. Asia Simulation Conference, 2007.

[5]Kroese, D. P. and R. Y. Rubinstein. The Transform Likelihood Ratio Method for Rare Event Simulation with Heavy Tails[J]. Queueing Systems, 2004,(46).

[6]Hall P. Beck, William D. Davidson. Establishing and Early Warning System Predicting Low Grades in College Students from Survey of Academic Orientations Scores[J]. Research in Higher Education, 2001,42(6).

[7]Cohen, I., B. Golany, and A. Shtud. Managing Stochastic Finite Capacity Multi-Project Systems Through the Cross-Entropy Method[J]. Annals of Operations Research, 2005, 134.

[8]JA Bucklew. Introduction to Rare Event Simulation[M]. Springer, New York, 2004.

[9]周泓，邱月，吴学静.基于重要抽样技术的稀有事件仿真方法[J].系统仿真学报，2007,(18).ˆ

【原文出处】《统计与决策》(武汉)2008年21期第4～6页

【作者简介】周泓，邱月，北京航空航天大学经济管理学院（北京　100083）

文档附件：

隐藏评论

用户昵称：　(您填写的昵称将出现在评论列表中) 匿名

请遵纪守法并注意语言文明。发言最多为2000字符（每个汉字相当于两个字符）

3393

发表