文章编号：1000-8608(2016)03-0285-07

# 泊松分布下信息安全事件概率计算模型

宋明秋*，郝　岩

（大连理工大学 系统工程研究所，辽宁 大连　116024）

**摘要：** 从信息安全事件的概率分布规律出发，根据泊松分布的基本特征，通过数学证明了信息安全事件发生频数服从泊松分布，并采用国家互联网应急中心(CNCERT/CC)统计数据验证了这一理论结果. 在此基础上，基于贝叶斯定理，建立了泊松分布下的信息安全事件概率计算模型. 根据泊松分布的概率质量函数，计算了信息安全事件发生频数的先验概率分布；通过构建似然函数调整先验概率分布，得到信息安全事件发生频数后验概率分布；最后，采用CNCERT/CC统计数据验证了该模型的可行性和有效性.

**关键词：** 信息安全事件；泊松分布；贝叶斯定理；后验概率

**中图分类号：**C931　　**文献标识码：**A　　**doi：**10.7511/dllgxb201603010

## 0　引　言

信息安全事件指由于自然或者人为以及软硬件本身缺陷或故障，对信息系统造成危害，或对社会造成负面影响的事件[1]，如病毒、木马、黑客攻击和信息泄露等. 信息安全事件极有可能危害业务运行和威胁信息安全，造成巨大损失[2]. 为了对易发信息安全事件提早采取防御措施，有必要对信息安全事件发生概率展开研究，以更优的费效比对信息安全事件进行管理和控制.

目前信息安全事件概率计算的主要方法有：(1)统计分析法. Whitman 等[3-4]通过调查对信息安全事件作了分类与统计，在此基础上，Farahmand等[5-6]通过风险分析，对信息安全事件进行了概率估算. Jonsson 等[7]将信息安全攻击过程划分为3个阶段：学习阶段、标准攻击阶段和创新攻击阶段，研究发现信息安全事件在标准攻击阶段发生的概率更大. (2)概率风险评价法. Hausken[8]将信息安全事件发生的概率用一个逻辑函数来刻画. Gritzalis 等[9]提出一个基于效用理论的风险概率模型. Mukhopadhyay 等[10]提出一个基于贝叶斯信念网络(Bayesian belief network，BBN)模型来度量在线交易安全风险，采用 Copula 模型计算损失的联合概率分布，BBN模型的输出结果为信息安全事件发生的概率. Liu等[11]利用BBN模拟潜在的网络攻击路径，绘制出BBN攻击图，运用贝叶斯推理方法对信息安全事件进行概率计算. Ammann 等[12-13]在攻击图模型的基础上引入了脆弱性度量，定量地刻画了系统的信息安全风险概率. Ishiguro 等[14]提出了一种网络威胁检测方法，通过扫描指定IP端口，分析其时间序列转换频率，给出该信息安全事件再次发生的概率.

上述研究主要集中在概率估算与风险概率的实时计算，其中概率估算存在很大的不确定性，往往会造成较大误差；而在风险概率计算中有些概率计算指标无法获得准确数据，只能依据历史经验与专家评价，不能实现真正意义上的量化计算. 为了更好地从定量角度解决信息安全事件的概率计算问题，本文从信息安全事件的概率分布基本规律出发，将信息安全事件发生频数作为随机变量，通过数学证明其服从泊松分布，并通过CNCERT/CC统计数据对这一结果进行验证. 在此基础上，结合贝叶斯定理，构建信息安全事件概率计算模型，以解决信息安全事件概率计算问题.

**收稿日期：**2015-11-26；　**修回日期：**2016-03-06.
**基金项目：**国家自然科学基金资助项目(71171028)；“十二五”国家科技支撑计划资助项目(2013BAH01B03).
**作者简介：**宋明秋*(1967-)，女，副教授，E-mail：songmq@dlut.edu.cn；郝　岩(1990-)，男，硕士生，E-mail：hhuqzhao@163.com.

## 1 信息安全事件发生频数及其分布

### 1.1 信息安全事件发生频数

信息安全事件发生频数是指单位时间 $T$ 内信息安全事件发生的次数. 信息安全事件的发生通常是离散的、相互独立的,因此,可以将信息安全事件的发生频数看作离散随机变量 $X$. 常用的几种离散分布如二项分布、泊松分布、几何分布以及超几何分布,它们的定义和基本特征如表 1 所示. 与常用离散分布的定义、概率质量函数和分布特征比较,信息安全事件的发生不是一个随机抽样问题,$X$ 显然不服从超几何分布. 另外,试验中每次信息安全事件的发生概率 $P$ 是随机的,不是固定不变的,所以说 $X$ 不服从二项分布和几何分布. 从表 1 中泊松分布的定义和分布特征来看,信息安全事件应符合泊松分布.

表 1 常用离散分布[15-16]

Tab. 1 Common discrete distributions[15-16]

| 离散分布类型 | 定义 | 概率质量函数 | 分布特征 |
|---|---|---|---|
| 二项分布 | 假设做了 $n$ 次独立试验,其中每次试验成功概率为 $P$,失败概率为 $1-P$,$X$ 代表试验中成功的次数,则称 $X$ 服从二项分布 | $P(i)=\binom{n}{i}P^i(1-P)^{n-i}$; $i=0,1,\cdots,n$ | (1)每次试验只会发生两种对立的结果之一;<br>(2)每次试验成功的概率 $P$ 固定不变;<br>(3)重复试验是相互独立的 |
| 泊松分布 | 对于取值为 0,1,…的随机变量 $X$,如果对于某个 $\lambda>0$,有 $P(i)=P(X=i)=\frac{e^{-\lambda}\lambda^i}{i!}$; $i=0,1,\cdots$ 则称 $X$ 服从泊松分布 | $P(X=i)=\frac{e^{-\lambda}\lambda^i}{i!}$; $\lambda>0$ | (1)在充分小的观测单位上,$X$ 的取值最多为 1;<br>(2)$X$ 的取值只与观测单位的大小有关,而与观测单位的位置无关;<br>(3)在某个观测单位上,$X$ 的取值与其他各观测单位上的 $X$ 的取值无关 |
| 几何分布 | 进行独立试验直到出现一个结果为成功,$P$ 为成功概率,$X$ 为首次成功时所需要做的试验次数,则称 $X$ 服从几何分布 | $P(n)=(1-P)^{n-1}P$; $n=1,2,\cdots$ | (1)每次试验只会发生两种对立的结果之一;<br>(2)每次试验成功的概率 $P$ 固定不变;<br>(3)重复试验是相互独立的 |
| 超几何分布 | 设有 $N$ 件产品,其中 $M$ 件不合格. 若从中不放回地随机抽取 $n$ 件,则其中含有的不合格产品的件数 $X$ 服从超几何分布 | $P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$; $k=0,1,\cdots,r$ 其中 $r=\min\{M,n\}$ | (1)有限总体;<br>(2)抽取指定种类;<br>(3)不放回抽样 |

### 1.2 信息安全事件发生频数的泊松分布

一个随机过程 $\{N(t),t\geqslant 0\}$ 称为计数过程,$N(t)$ 是随机变量,表示时间段 $[0,t]$ 内发生的事件数. 如果满足以下条件,此计数过程称为具有速率 $\lambda(\lambda>0)$ 的泊松过程[15-16]:

(1)$N(0)=0$;

(2)$\{N(t)\}$ 是独立增量过程;

(3)对任何 $t,s\geqslant 0$,$N(s,t+s]$ 服从参数为 $\lambda t$ 的泊松分布,即

$$P(N(s,t+s]=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t};\ k=0,1,\cdots \quad (1)$$

将信息安全事件的发生频数作为随机变量 $N(t)$,其中 $t\geqslant 0$. 在 $t=0$ 时显然有 $N(0)=0$,满足条件(1).

$N(t)-N(s)(0\leqslant s<t)$ 为在 $(s,t]$ 上的增量,信息安全事件的发生是相互独立的,所以对任意正整数 $n$ 和任意选定的 $0\leqslant t_0<t_1<t_2<\cdots<t_n$,$n$ 个增量 $X(t_1)-X(t_0)$,$X(t_2)-X(t_1)$,…,$X(t_n)-X(t_{n-1})$ 相互独立,故可以称 $\{N(t),t\geqslant 0\}$ 为独立增量过程,满足条件(2).

在充分小的时间间隔 $h$ 内,只能发生一次信息安全事件,发生两次或两次以上的概率极低,可以忽略不计[17],可表示为

$$P\{N(h)\geqslant 2\}=o(h);\ h>0 \quad (2)$$

在充分小的时间间隔 $h$ 内,近似地认为只发生 1 次安全事件的概率与时间间隔 $h$ 的长度成正

比[17]，可表示为

$$P\{N(h)=1\}=\lambda h+o(h);\ h>0 \tag{3}$$

其中 $\lambda(\lambda>0)$ 为此计数过程的速率.

只需证明式(2)和(3)满足式(1)即可证明随机变量 $N(t)$ 服从泊松分布.

记 $P_n(t)=P\{N(t)=n\}$，令 $h>0$，则

$$\begin{aligned}P_0(t+h)&=P\{N(t+h)=0\}=\\&P\{N(t)=0,N(t+h)-N(t)=0\}=\\&P\{N(t)=0\}\cdot P\{N(t+h)-\\&N(t)=0\}=\\&P\{N(t)=0\}\cdot P\{N(h)=0\}\end{aligned}$$

由式(2)和(3)得

$$P_0(t+h)=P_0(t)[1-\lambda h+o(h)] \tag{4}$$

整理得

$$\frac{P_0(t+h)-P_0(t)}{h}=-\lambda P_0(t)+\frac{o(h)}{h} \tag{5}$$

令 $h\to 0$，取极限得微分方程：

$$P_0'(t)=-\lambda P_0(t) \tag{6}$$

由 $P_0(0)=P\{N(0)=0\}=1$，解得

$$P_0(t)=\mathrm{e}^{-\lambda t} \tag{7}$$

类似地，对 $n\geqslant 1$，可得微分方程：

$$P_n'(t)=-\lambda P_n(t)+\lambda P_{n-1}(t) \tag{8}$$

当 $n=1$ 时，由 $P_0(t)=\mathrm{e}^{-\lambda t}$ 和 $P_1(0)=0$ 得

$$P_1(t)=\lambda t\mathrm{e}^{-\lambda t} \tag{9}$$

由数学归纳法，并注意到 $P_n(0)=0$，得

$$P_n(t)=\frac{\mathrm{e}^{-\lambda t}(\lambda t)^n}{n!} \tag{10}$$

式(10)可证明该随机过程符合条件(3)，故可认为随机变量 $N(t)$ 服从泊松分布，即信息安全事件的发生频数服从泊松分布.

在实际事例中，当一个随机事件，例如一段时间内感染病毒的主机数量、服务器遭到网络攻击的次数、信息安全漏洞增加数量、发生数据泄露的次数、网站被篡改的数量等，以固定的平均瞬时速率 $\lambda$ 随机独立地出现时，那么这个事件在单位时间内出现的次数或个数就近似地服从泊松分布.

本文选取 CNCERT/CC 统计数据，基于 R 语言对不同种类的信息安全事件的发生频数作泊松分布的 $\chi^2$ 拟合优度检验，具体检验结果见表 2.

$P_v$ 表示样本间的差异由抽样误差所致的概率，是用于判断原始假设是否正确的重要证据. 其值小于 0.05 表示结果显著，小于 0.01 表示结果非常显著. 如表 2 结果显示，$P_v$ 均小于 $2.2\times 10^{-16}$，表示以上泊松分布 $\chi^2$ 拟合优度检验结果均非常显著. 因此，可以说信息安全事件的发生频数服从泊松分布.

表 2　泊松分布的 $\chi^2$ 拟合优度检验

Tab. 2　$\chi^2$ goodness of fit test of Poisson distribution

| 信息安全事件种类 | $\chi^2$ | $d_f$ | $P_v$ |
|---|---|---|---|
| 感染病毒 | 11 521 000.0 | 140 | $<2.2\times10^{-16}$ |
| 网站被篡改 | 92 982.0 | 139 | $<2.2\times10^{-16}$ |
| 网站被植入后门 | 434 340.0 | 141 | $<2.2\times10^{-16}$ |
| 网站页面被仿冒 | 274 770.0 | 141 | $<2.2\times10^{-16}$ |
| 信息安全漏洞 | 3 496.3 | 137 | $<2.2\times10^{-16}$ |

## 2　信息安全事件概率计算模型构建

### 2.1　贝叶斯定理描述

贝叶斯公式的离散分布形式如下：

设 $B_1,B_2,\cdots,B_n$ 是样本空间 $\Omega$ 的一个分割子集，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{i=1}^{n}B_i=\Omega$，如果 $P(A)>0,P(B_i)>0,i=1,2,\cdots,n$，则

$$P(B_i\mid A)=\frac{P(B_i)P(A\mid B_i)}{\sum_{i=1}^{n}P(B_i)P(A\mid B_i)};\quad i=1,2,\cdots,n \tag{11}$$

$\Omega$ 表示信息安全事件发生频数的样本空间，$B_i$ 为样本空间 $\Omega$ 的一个子集，即信息安全事件发生频数的一个区间. $P(B_i)$ 表示信息安全事件发生频数在区间 $B_i$ 内的先验概率分布. 事件 $A$ 称为先验概率 $P(B_i)$ 的修正元素，$P(A|B_i)$ 为似然函数，这里表示信息安全事件发生频数在区间 $B_i$ 内发生事件 $A$ 的似然性，也可表示为 $L(B_i\mid A)$. $P(B_i|A)$ 表示后验概率分布，也可称为条件概率分布，即在事件 $A$ 条件下信息安全事件发生频数在区间 $B_i$ 内的概率分布.

### 2.2　先验概率分布

若随机变量 $X$ 只取非负整数 0，1，…，且其概率分布服从

$$P(X=k)=\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!};\ \lambda>0 \tag{12}$$

则随机变量 $X$ 的分布称为泊松分布，记作 $P(\lambda)$.

在泊松分布中唯一不确定的参数 $\lambda$ 为泊松过程的速率，在本文中，其实际意义为单位时间 $T$ 内信息安全事件平均发生频数 $\overline{X}$. 某组织在一段时间内持续监测信息安全事件的发生状况，统计

不同阶段单位时间 $T$ 内信息安全事件发生频数 $X_i$，共收集 $n$ 条样本数据. 可以计算得到单位时间 $T$ 内平均信息安全事件发生频数

$$\overline{X}=\sum_{i=0}^{n}\frac{X_i}{n} \tag{13}$$

即参数

$$\lambda=\sum_{i=0}^{n}\frac{X_i}{n} \tag{14}$$

在本文中，根据 $\lambda$ 的值，将信息安全事件发生频数分成若干区间 $B_i$，$i=1,2,\cdots,n$. 假设区间 $B_i$ 内有 $X_1,X_2,\cdots,X_m$ 共 $m$ 个样本数据，根据式(12)、(14)，计算得到信息安全事件发生频数的先验概率分布为

$$P(B_i)=\sum_{j=1}^{m}P(X=X_j)=\mathrm{e}^{-\sum_{i=0}^{n}\frac{X_i}{n}}\sum_{j=1}^{m}\frac{\left(\sum_{i=0}^{n}\frac{X_i}{n}\right)^{X_j}}{X_j!} \tag{15}$$

### 2.3 似然函数

似然函数是一个关于统计模型参数的函数，这个函数中的自变量是统计模型的参数. 对于结果 $X$，在参数集合 $\theta$ 上的似然，就是在给定这些参数值的基础上，观察到的结果的概率 $L(\theta|X)=P(X|\theta)$. 也就是说，似然是关于参数的函数，在参数给定的条件下，对于观察到的 $X$ 的条件分布.

设 $A$ 为一随机事件，用一个随机变量 $X$ 来表示事件 $A$ 的发生次数，$X=0$ 表示发生，$X=1$ 表示不发生，则 $X$ 服从二点分布 $b(1,p)$，其中 $p$ 是未知的事件 $A$ 发生率. 现抽取 $n$ 个样本看事件 $A$ 是否发生，得到样本 $x_1,\cdots,x_n$，这批观测值发生的概率为

$$P(X_1=x_1,\cdots,X_n=x_n;p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i} \tag{16}$$

将式(16)看作未知参数 $p$ 的函数，用 $L(p)$ 表示，记为

$$L(p)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i} \tag{17}$$

$L(p)$ 即为区间 $B_i$ 内发生事件 $A$ 的似然函数，可用 $P(A|B_i)$ 表示，即

$$P(A\mid B_i)=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i} \tag{18}$$

根据最大似然估计的基本原理，将式(18)两端取对数并关于 $p$ 求导令其为 0，即得如下方程，又称似然方程：

$$\frac{\partial\ln L(p)}{\partial p}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0 \tag{19}$$

解之即得 $p$ 的最大似然估计为

$$\hat{p}=\hat{p}(x_1,\cdots,x_n)=\sum_{i=1}^{n}x_i/n=\overline{x} \tag{20}$$

即 $P(A|B_i)$ 可表示为

$$P(A\mid B_i)=\sum_{i=1}^{n}x_i/n=\overline{x} \tag{21}$$

### 2.4 后验概率分布

后验概率分布就是未知量作为随机变量的概率分布，并且是在基于试验或者调查所获得的信息上的条件概率分布. 这里指在事件 $A$ 给定并纳入考虑之后的条件概率分布.

根据式(11)、(15)、(21)，计算得信息安全事件发生频数的后验概率分布为

$$P(B_i\mid A)=\mathrm{e}^{-\sum_{i=0}^{n}\frac{X_i}{n}}\sum_{j=1}^{m}\frac{\left(\sum_{i=0}^{n}\frac{X_i}{n}\right)^{X_j}}{X_j!}\cdot\overline{x} \tag{22}$$

## 3 实例分析

### 3.1 数据选取

CNCERT/CC 通过对基础信息网络、金融证券等重要信息系统的自主监测，以及与合作伙伴进行数据共享，纵向统计每周信息安全事件发生次数，发布网络安全信息与动态周报. 因此，本文以周($T$)为单位选取信息安全事件统计数据，描述计算信息安全事件概率分布的具体计算过程.

设随机过程$\{N(T),T\geqslant 0\}$，每周信息安全事件的发生频数作为随机变量 $N(T)$，每周($T$)发生的信息安全事件是相互独立的，根据1.2节可同理证明 $N(T)$ 服从泊松分布. CNCERT/CC 作为我国网络安全应急体系的核心协调机构，能够确保数据的真实性和有效性. 因此，可以利用 CNCERT/CC 统计数据对以上模型进行验证.

### 3.2 信息安全事件发生频数的先验概率分布

记事件 $B$ 为单位时间内(周)信息安全事件发生频数，$B_i$ 为事件 $B$ 的一个子集，表示信息安全事件发生频数的一个子区间.

CNCERT/CC 对信息安全事件的统计包括境内感染病毒的主机数量、境内被篡改网站总数、境内被植入后门网站总数、针对境内网站的仿冒

页面数量以及新增信息安全漏洞数量，本文将其均记为信息安全事件的发生频数. 以 2013 年的统计数据为例[18]，对数据进行统计分析，根据式(13)、(14)，计算得单位时间(周)内这些信息安全事件发生频数的均值 $\lambda$ 分别为 886 000、5 857、4 061、779、154，如表 3 所示. 根据泊松分布的性质，位于期望 $\lambda$ 附近概率较大，分布较为集中，随着 $\lambda$ 的增加，分布逐渐趋于对称[19]，如图 1 所示.

表 3　信息安全事件发生频数的概率分布

Tab. 3　Probability distribution of information security incidents frequency

| 信息安全事件种类 | 信息安全事件发生频数均值 $\lambda$ | 先验概率分布 $P(B_i)$ | | | 似然函数 $P(A\|B_i)$ | | | 后验概率分布 $P(B_i\|A)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $B_1$ | $B_2$ | $B_3$ | $B_1$ | $B_2$ | $B_3$ | $B_1$ | $B_2$ | $B_3$ |
| 感染病毒 | 886 000 | (0,800 000)<br>0.196 0 | (800 000,<br>1 000 000)<br>0.699 2 | (1 000 000,<br>+∞)<br>0.104 8 | 0.714 3 | 0.636 4 | 0.888 9 | 0.206 5 | 0.656 2 | 0.137 3 |
| 网站被篡改 | 5 857 | (0,5 000)<br>0.144 2 | (5 000,<br>6 500)<br>0.673 3 | (6 500,<br>+∞)<br>0.182 5 | 0.333 3 | 0.743 6 | 1.000 0 | 0.065 7 | 0.684 7 | 0.249 6 |
| 网站被植入后门 | 4 061 | (0,3 000)<br>0.051 4 | (3 000,<br>5 000)<br>0.884 5 | (5 000,<br>+∞)<br>0.064 1 | 0.727 3 | 0.625 0 | 1.000 0 | 0.057 1 | 0.844 9 | 0.098 0 |
| 网站页面被仿冒 | 779 | (0,600)<br>0.334 8 | (600,<br>1 000)<br>0.500 4 | (1 000,<br>+∞)<br>0.164 8 | 0.357 1 | 0.869 6 | 1.000 0 | 0.166 2 | 0.604 8 | 0.229 0 |
| 信息安全漏洞 | 154 | (0,100)<br>0.100 3 | (100,<br>200)<br>0.798 9 | (200,<br>+∞)<br>0.100 8 | 0.750 0 | 0.736 8 | 0.875 0 | 0.100 0 | 0.782 7 | 0.117 3 |

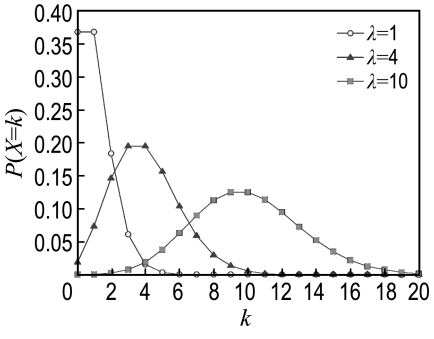


图 1　参数为 $\lambda$ 的泊松分布概率质量函数

Fig. 1　Probability mass function of Poisson distribution with parameter $\lambda$

为充分体现泊松分布的概率分布规律，根据 $\lambda$，将事件 $B$ 分为若干子区间 $B_i$，$i=1,2,3$. 根据式(15)，分别计算 $B_i$ 内的先验概率分布 $P(B_i)$. 以感染病毒的主机数量为例，根据 $\lambda=886\ 000$，将事件 $B$ 分为(0,800 000)、(800 000,1 000 000)、(1 000 000,+∞)3 个子区间. 根据式(15)，计算先验概率分布 $P(B_i)$ 分别为 0.196 0、0.699 2、0.104 8. 从表 3 $P(B_i)$ 计算结果中不难看出，信息安全事件的先验概率分布主要集中在均值 $\lambda$ 附近，偏离 $\lambda$ 越大概率迅速减小，符合泊松分布的性质.

### 3.3　信息安全事件发生频数的后验概率分布

CNCERT/CC 除了统计境内每周发生的信息安全事件，对每周境内的网络安全环境也作了统计，并将其分为优、良、中、差、危 5 个等级. 本文将网络安全等级为优和良的记为网络安全环境良好，其他记为网络安全环境恶劣. 记事件 $A$ 为网络安全环境恶劣，作为对先验概率的修正元素. $P(A|B_i)$ 为似然函数，表示信息安全事件发生频数在区间 $B_i$ 内网络安全环境恶劣的似然性. $P(B_i|A)$ 为后验概率分布，表示网络安全环境恶劣条件下信息安全事件发生频数在区间 $B_i$ 内的概率分布.

统计区间 $B_i$ 内所有数据，并记录网络安全环境是否恶劣，根据式(21)和(22)分别计算 $P(A|B_i)$ 和 $P(B_i|A)$. 以网站被篡改数量为例，根据统计分析后的区间 $B_i$ 内数据和式(21)计算得到的

似然函数 $P(A|B_i)$ 分别为 0.333 3、0.743 6、1.000 0；再根据式(22)，计算得到后验概率分布 $P(B_i|A)$ 分别为 0.065 7、0.684 7、0.249 6. 与先验概率分布 $P(B_i)$ 相比较后发现，网络安全环境恶劣条件下，信息安全事件发生频数较大区间内的概率分布会变大. 如表 3 所示的计算结果表明，大多数情况下，区间 $B_i$ 内数值越大，$P(A|B_i)$ 和 $P(B_i|A)$ 数值也会普遍越大，即信息安全事件发生频数越大的区间内网络安全环境恶劣的似然值普遍越大. 另外，也普遍存在网络安全环境恶劣条件下信息安全事件发生频数较大区间内的概率分布会变大.

## 4 结 语

本文将信息安全事件发生频数看作离散随机变量，从数学定义上证明了其服从泊松分布，选取 CNCERT/CC 数据进行泊松分布 $\chi^2$ 拟合优度检验，结果进一步表明其服从泊松分布. 结合贝叶斯定理，构建信息安全事件概率计算模型. 根据之前得到的结论，解决了信息安全事件先验概率分布难以计算的问题；然后构建似然函数 $P(A|B_i)$，求取 $B_i$ 内发生事件 $A$ 的最大似然估计，调整先验概率分布，优化概率计算结果，得到信息安全事件发生频数的后验概率分布. 实例结果表明，该模型能够准确计算信息安全事件概率分布.

## 参考文献：


[1] 中华人民共和国国家质量监督检验检疫总局. 信息安全技术 信息安全事件分类分级指南：GB/Z 20986—2007[S]. 北京：中国标准出版社，2007.
General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. Information Security Technology — Guidelines for the Category and Classification of Information Security Incidents: GB/Z 20986-2007 [S]. Beijing: China Standards Press, 2007. (in Chinese)

[2] 中华人民共和国国家质量监督检验检疫总局. 信息技术 安全技术 信息安全事件管理指南：GB/Z 20985—2007[S]. 北京：中国标准出版社，2007.
General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. Information technology — Security techniques — Information Security Incident Management Guide: GB/Z 20985-2007 [S]. Beijing: China Standards Press, 2007. (in Chinese)

[3] Whitman M E. Enemy at the gate: Threats to information security [J]. **Communications of the ACM**, 2003, **46**(8): 91-95.

[4] Keller S, Powell A, Horstmann B, *et al*. Information security threats and practices in small businesses [J]. **Information Systems Management**, 2005, **22**(2): 7-19.

[5] Farahmand F, Navathe S B, Enslow P H, *et al*. Managing vulnerabilities of information systems to security incidents [J]. **ACM International Conference Proceeding Series**, 2003, **50**: 348-354.

[6] Farahmand F, Navathe S B, Sharp G P, *et al*. Evaluating damages caused by information systems security incidents [C] // **Economics of Information Security**. New York: Springer US, 2004: 85-94.

[7] Jonsson E, Olovsson T. Quantitative model of the security intrusion process based on attacker behavior [J]. **IEEE Transactions on Software Engineering**, 1997, **23**(4): 235-245.

[8] Hausken K. Returns to information security investment: The effect of alternative information security breach functions on optimal investment and sensitivity to vulnerability [J]. **Information Systems Frontiers**, 2006, **8**(5): 338-349.

[9] Gritzalis S, Yannacopoulos A N, Lambrinoudakis C, *et al*. A probabilistic model for optimal insurance contracts against security risks and privacy violation in IT outsourcing environments [J]. **International Journal of Information Security**, 2007, **6**(4): 197-211.

[10] Mukhopadhyay A, Chatterjee S, Saha D, *et al*. E-risk management with insurance: A framework using copula aided Bayesian belief networks [J]. **Proceedings of the Annual Hawaii International Conference on System Sciences**, 2006, **6**: 126a.

[11] LIU Yu, MAN Hong. Network vulnerability assessment using Bayesian networks [J]. **Proceedings of SPIE — The International Society for Optical Engineering**, 2005, **5812**: 61-71.

[12] Ammann P, Wijesekera D, Kaushik S. Scalable, graph-based network vulnerability analysis [C] // **Proceedings of the ACM Conference on Computer and Communications Security**. Washington D C: ACM, 2002: 217-224.

[13] Frigault M, WANG Ling-yu. Measuring network security using Bayesian network-based attack graphs [J]. **Proceedings — International Computer Software and Applications Conference**, 2008: 4591650.

[14] Ishiguro M, Suzuki H, Murase I, *et al*. Internet threat detection system using Bayesian estimation [C] // **16th Annul FIRST Conference on Computer Security Incident Handling**. Hungary: FIRST, 2004.

[15] Sheldon M R. 应用随机过程：概率模型导论[M]. 10 版. 龚光鲁，译. 北京：人民邮电出版社，2011.
Sheldon M R. **Introduction to Probability Models** [M]. 10th ed. GONG Guang-lu, trans. Beijing: Posts&Telecom Press, 2011. (in Chinese)

[16] 何书元. 随机过程[M]. 北京：北京大学出版社，2008.
HE Shu-yuan. **Stochastic Process** [M]. Beijing: Peking University Press, 2008. (in Chinese)

[17] 王志刚. 应用随机过程[M]. 合肥：中国科学技术大学出版社，2009.
WANG Zhi-gang. **Applied Stochastic Process** [M]. Hefei: University of Science and Technology of China Press, 2009. (in Chinese)

[18] 国家互联网应急中心. 网络安全信息与动态周报[R]. 北京：国家互联网应急中心，2013.
CNCERT/CC. Network Security Information and Dynamic Weekly [R]. Beijing: CNCERT/CC, 2013. (in Chinese)

[19] 茆诗松，程依明，濮晓龙. 概率论与数理统计教程[M]. 2 版. 北京：高等教育出版社，2011.
MAO Shi-song, CHENG Yi-ming, PU Xiao-long. **Probability Theory and Mathematical Statistics Course** [M]. 2nd ed. Beijing: Higher Education Press, 2011. (in Chinese)

# Information security incident probability calculation model based on Poisson distribution

SONG Ming-qiu*, HAO Yan

( Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, China )

**Abstract**: From the possibility distribution rules of information security incidents, and according to the basic characteristics of Poisson distribution, the frequency of information security incidents is proved mathematically to obey Poisson distribution . The verification is done by using the statistical data of National Internet Emergency Center (CNCERT/CC). On this basis, an information security incidents probability calculation model with Poisson distribution is established based on the Bayes theorem. Then, taking advantage of the probability mass function, the prior probability distribution of information security incidents is calculated, and a likelihood function is modeled to adjust the prior probability distribution, and the posterior probability distribution of the frequency of information security incidents is got. Finally, the statistical data of CNCERT/CC demonstrate the feasibility and effectiveness of the model.

**Key words**: information security incidents; Poisson distribution; Bayes theorem; posterior probability