ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.7, July 2009, pp.1967−1976
doi: 10.3724/SP.J.1001.2009.03277

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于熵的随机性检测相关性研究*

范丽敏[1,2+], 冯登国[1], 陈 华[1]

[1](中国科学院 软件研究所 信息安全国家重点实验室,北京 100190)

[2](中国科学院 研究生院,北京 100049)

Study on the Correlation Between Randomness Tests Based on Entropy

FAN Li-Min[1,2+], FENG Deng-Guo[1], CHEN Hua[1]

[1](State Key Laboratory of Information Security, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: lm_fan@163.com

Fan LM, Feng DG, Chen H. Study on the correlation between randomness tests based on entropy. *Journal of Software*, 2009,20(7):1967−1976. http://www.jos.org.cn/1000-9825/3277.htm

Abstract: There exist a lot of randomness test methods, and most of them have parameters. As it is not practical to do all randomness tests in practice, it is important to study the relations among these methods. In this paper, four kinds of relations between randomness tests and a conception of "correlation degree of randomness tests" are defined firstly based on the statistics theory firstly. And the correlation degree is measured by means of the entropy method. Then, the relevancies between the four relations and the correlation degree are proved. And an algorithm of calculating correlation degree and a selection policy are provided as well. The work of this paper is helpful for selecting reasonable and scientific randomness tests and parameters. In addition, the randomness test methods adopted by NIST (National Institute of Standards and Technology) in AES (advanced encryption standard) are explored by using the correlation degree and some dependence relations are found.

Key words: randomness test; correlation degree; entropy; *P*-Value; parameter selection

摘 要: 目前,存在众多的随机性检测项目,并且许多项目都带有参数.选择所有的项目进行检测不现实,因此需要研究检测项目之间的关系.从统计学角度出发,对检测项目的相关性进行研究,首先定义了检测项目之间存在的 4 种关系,提出了检测项目相关度的概念,然后利用熵值法对检测项目相关度进行度量,并证明了这 4 种关系与相关度的联系,同时给出了一种计算相关度的算法和一个基于相关度的检测项目选择策略.所研究的结果为随机性检测项目及其参数选择提供了理论依据.与此同时,利用相关度对 NIST 在评选 AES 中所采用的检测项目进行研究,发现了其中一些检测项目之间存在着依赖关系.

关键词: 随机性检测;相关度;熵;*P*-Value;参数选择

中图法分类号: TP309 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60503014, 60603013 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.2007AA01Z470, 2008AA01Z417 (国家高技术研究发展计划(863)); the Beijing Municipal Natural Science Foundation of China under Grant No.4072026 (北京市自然科学基金)

Received 2007-10-18; Accepted 2008-02-04

密码算法是构建安全信息系统的核心要素之一,是保障信息与数据机密性、完整性和真实性的重要技术.密码算法检测评估是密码算法研究的重要组成部分,它为密码算法的设计、分析提供客观的量化指标和技术参数,对密码算法的应用具有重要的指导意义.在密码算法的设计和评测过程中,需要从多个方面对其进行检测和分析.Shannon 利用统计特性对密码的无条件安全进行了定义[1]:“如果密文块和明文块是统计独立的,则该密码提供完全保密”.在当前存在的密码算法中,只有一次一密才能达到这种统计独立性,其他已知密码算法只能最大程度地与统计独立进行逼近.因此,分析密码算法的统计性能是密码算法安全性研究的重要内容.并且,用统计检测的方法对密码算法进行评测可以为理论分析提供大量参考数据,从而减少理论分析者的工作量,同时可以暴露出用现有的分析方法无法发现的安全漏洞[2].

统计检测通过统计密码算法输出序列的随机特性来实现.目前,已有多种用于检测序列随机性的方法.典型的有[3−7]:单比特频数、扑克、重叠子序列、票券收集器、置换、游程、碰撞、生日间隔、序列相关性、矩阵秩、比特流、停车场、3D、压缩、重叠模板匹配、非重叠模板匹配、通用统计、随机游动、随机游动变量、二元推导、变换点、序列复杂度、线性复杂度等等.同时,许多检测项目带有参数,如果将不同的参数视作不同的项目,那么参数的不同选择又衍生出更多的检测项目.例如,参数为 9 的非重叠模板有 148 个,随机游动的检测结果有 8 个,随机游动变量的检测结果有 18 个[8].

显然,在实际应用中选择所有的项目和参数不现实,因此有必要对检测项目及其参数可能存在的关系进行研究,从而选择合适的检测项目子集,提高检测效率和检测实用性.但是各个检测项目提出的背景和数学基础不尽相同,所以检测项目及参数的相关性研究也是比较困难的问题.

目前,关于检测项目及参数的相关性研究成果并不多见.对参数的研究主要集中在选择正确的参数上,文献[9]对碰撞检测的参数进行研究,通过大量的实验确定了一个较为合理的参数取值.在检测项目的相关性研究方面,文献[10]提出通用统计检测,并指出该检测项目在序列无限长时可以代替一些其他检测项目;文献[11]通过大量的统计实验发现了 3 个不易通过的检测项目,从统计角度说明,如果通过了这 3 个检测项目的序列一定程度上也可以通过其他项目的检测;NIST 利用主成分分析(principal component analysis)方法研究了其在评选 AES 过程中用到的部分检测项目之间的依赖关系[5],研究表明,这些检测项目之间的依赖关系很小.目前国内外的相关研究结果或者是基于具体检测项目的特点,如文献[9,10],或者从统计学角度给出依赖关系的笼统结论,但是未给出项目之间关系的定量描述,如文献[5,11].总之,当前缺乏对检测项目之间关系的详细刻画和相关性的定量描述.本文对检测项目和参数进行了统一的处理,将不同参数看作不同的项目,从统计学角度刻画检测项目之间可能存在的 4 种关系,利用熵值法对各种关系进行量化,给出了一种计算相关度的算法.因此,本文的研究工作为随机性检测中项目及参数的选择提供了理论依据和一种通用的方法.

本文第 1 节介绍研究所需的背景知识.第 2 节对检测项目之间可能存在的关系加以定义,提出检测项目相关度的概念并利用熵值法对相关度进行度量.同时,以定理的形式给出检测项目之间关系与相关度的联系.第 3 节给出一种计算相关度的算法和一个基于相关度的检测参数选择策略.同时,利用相关度对 NIST 采用的 16 个检测项目进行研究.第 4 节是结论.

1 背景知识

1.1 随机性检测

随机性检测利用概率统计的方法对随机数发生器或者密码算法产生序列的随机性进行描述.不同的检测项目从不同的角度刻画待检测序列与真随机序列之间的差距.

随机性检测通常采用假设检验[12]的方法.假设检验就是在总体分布未知或者只知其形式但不知其参数的情况下,为了推断总体的某些性质而提出某些关于总体的假设,然后根据样本对提出的假设做出判断.随机性假设检验,就是已知真随机序列的某一方面符合一个特定的分布,那么假设待检测序列是随机的,则该待检测序列在这方面也应该符合这个特定的分布.以随机序列的某种统计值 V 符合自由度为 n 的卡方分布为例:

原假设(零假设)H_0:序列是随机的,待测序列的统计值 V 服从 $\chi^2(n)$ 分布;

备择假设 H_α:序列不是随机的,待测序列的统计值 V 不服从$\chi^2(n)$分布.

通过判断一个待测序列的统计值 V 是否服从$\chi^2(n)$分布来确定是否接受原假设,从而判断该序列是否通过了该项随机性检测.

在随机性检测中判断是否接受原假设通常采用 P-Value 方法[13].P-Value 是一个序列比真随机序列的随机性要好的概率.利用统计值 V 求出 P-Value,并将 P-Value 与显著性水平α比较.如果 P-Value≥α,则接受原假设,判断该待测序列通过了该项随机性检测.

1.2 熵的基本概念和性质

熵的概念最初源于热力学,后来由 Shannon 引入信息论[14].熵的定义如下:

系统 U 由一系列的事件构成,$U=\{u_i, i=1,2,...,n\}$,u_i 出现的概率 $P(u_i)\geq 0$,且满足 $\sum_{i=1}^{n} p(u_i)=1$,则系统 U 的熵定义为

$$H(U)=-\sum_{i=1}^{n} p(u_i)\times \log_2 p(u_i) \tag{1}$$

熵有如下性质[15]:

可加性:熵具有概率性质,系统的熵等于各个状态的熵之和.

非负性:系统的熵是非负的.

对称性:系统的熵与其状态出现概率的排列次序无关.

极值性:当系统的状态概率为等概率时,熵值达到最大.

加法性:两个独立的系统,复合起来的熵等于两个系统的熵值之和.

2 随机性检测相关性研究

2.1 符 号

下面列出本文用到的一些符号及其含义:

T_A:检测项目 A;

$R(A,B)$:检测项目 A 和 B 之间的相关度;

$\tilde{R}(A,B)$:检测项目 A 和 B 之间的不相关度;

P-Value(A):T_A 对一个序列检测得到的 *P-Value* 值;

P_{A-B}:对同一个序列进行检测, *P-Value*(A)> *P-Value*(B) 的概率,即 $P_{A-B}=P\{$ *P-Value*(A)> *P-Value*(B)$\}$;

$P_{A=B}$:对同一个序列进行检测, *P-Value*(A)= *P-Value*(B) 的概率,即 $P_{A=B}=P\{$ *P-Value*(A)= *P-Value*(B)$\}$;

$P_{A-B}(i)$: $-1\leq$ *P-Value*(A) $-$ *P-Value*(B) ≤ 1,将 (0,1] 分为 n 个区间,则 *P-Value*(A) $-$ *P-Value*(B) 落入第 $i(1\leq i\leq n)$个区间内的概率记作 $P_{A-B}(i)$.

2.2 定 义

要研究检测项目的相关性,首先需要定义检测项目之间可能存在的关系.本文利用检测项目 P-Value 的关系定义检测项目之间可能存在的 4 种关系:包含关系、等价关系、无关关系和相关关系.各种关系分别定义如下:

定义 1. 两个检测项目的包含关系(⊂).

两个不同的检测项目 T_A 和 T_B,如果对于任何序列都有 P_{A-B} 等于 1,则称 T_A 包含 T_B,记为 $T_B\subset T_A$.

定义 2. 两个检测项目的等价关系(=).

两个不同的检测项目 T_A 和 T_B,如果对于任何序列都有 $P_{A=B}$ 等于 1,则称 T_A 等价 T_B,记为 $T_B=T_A$.

定义 3. 两个检测项目的无关(独立)关系(×).

两个不同的检测项目 T_A 和 T_B,对于相同的序列,如果二者 *P-Value* 的概率分布是独立的,则称 T_A 与 T_B 无关,也可称为独立,记为 $T_B\times T_A$.

定义 4. 两个检测项目的相关关系(∩).

两个不同的检测项目 T_A 和 T_B,如果不存在包含关系、等价关系和无关关系,则称二者存在相关关系,记为 $T_B \cap T_A$.

基于上述定义,检测项目之间的包含、等价和无关这 3 种关系比较明确,需要研究的是如何对检测项目之间的相关关系及其相关性进行定量描述.本文利用两个检测项目 A,B 的 P-Value 之差的分布来对 T_A 和 T_B 的相关度进行度量.P-Value 是一个[0,1]的实数,那么 $P\text{-}Value(A)-P\text{-}Value(B)$就是[−1,1]之间的实数.设 $f(z)$为 $P\text{-}Value(A)-P\text{-}Value(B)$的概率密度函数.$f(z)$落入某个区间$[z_1,z_2]$的概率是 $\int_{z_1}^{z_2} f(z)\mathrm{d}z$.如果将(0,1)分为 $n(n\geq 1)$个区间,在每个区间内构建一个系统,则该系统由两个事件组成,第 1 个事件发生的概率是 $P\text{-}Value(A)<P\text{-}Value(B)$的概率所占的比例,第2个事件发生的概率是$P\text{-}Value(A)>P\text{-}Value(B)$的概率所占的比例.利用熵值的基本思想定义在这个区间内两个检测项目的不相关度,然后累加求得整体的不相关度,进而求得两个项目之间的相关度.

定义 5. 两个检测项目的不相关度$\tilde{R}(A,B)$:

$$\tilde{R}(A,B)=-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}+\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right)$$

.其中,将(0,1)分成 $n(n\geq 1)$个区间,$P_{A-B}(i)$为 $P\text{-}Value(A)-P\text{-}Value(B)$落入第 $i(1\leq i\leq n)$个区间中的概率,$P_{B-A}(i)$为 $P\text{-}Value(B)-P\text{-}Value(A)$ 落入第 $i(1\leq i\leq n)$个区间中的概率.这里约定,若在某一个区间中 $P_{B-A}(i)=P_{A-B}(i)=0$,则该区间的不相关度为 0.

自然地,我们定义相关度如下:

定义 6. 两个检测项目的相关度 $R(A,B)$:

$$R(A,B)=1-\tilde{R}(A,B).$$

2.3 性质和定理

第 2.2 节对检测项目之间可能存在的关系进行了定义,并且利用熵值法对检测项目的相关度进行了刻画.本节给出根据上述定义得到的性质和定理,并分别对其进行了证明.首先介绍两个等价关系和包含关系的性质.

性质 1. 对于两个不同的检测项目 A 和 B,假如二者之间存在等价关系,即 $T_A=T_B$,那么对于任何序列,如果通过了 T_A 的检测,那么该序列一定能够通过 T_B 的检测.

证明:根据两个检测项目等价的定义可知 $P_{A=B}=1$,则对于任何序列都有 $P\text{-}Value(A)=P\text{-}Value(B)$.显然,通过了 T_A 的检测,就有 $P\text{-}Value(A)\geq\alpha$,那么一定有 $P\text{-}Value(B)\geq\alpha$,所以该序列一定通过 T_B 的检测. □

性质 2. 对于两个不同的检测项目 A 和 B,假如二者之间存在包含关系,即 $T_B\subset T_A$,那么对于任何序列,① 如果通过了 T_B 的检测,那么它一定能够通过 T_A 的检测;② 通过 T_A 的检测,却不一定能够通过 T_B 的检测.

证明:根据检测项目包含的定义可知,若 $T_B\subset T_A$ 则有 $P_{A-B}=1$.对于任何序列,总有 $P\text{-}Value(A)>P\text{-}Value(B)$,因此,如果 $P\text{-}Value(B)\geq\alpha$,总有 $P\text{-}Value(A)\geq\alpha$,那么通过了 T_B 的检测就一定能够通过 T_A 的检测.①得证.

总能找到一条序列,使得 $PValue(A)=\alpha$,这条序列通过了 T_A 的检测,但是 $P\text{-}Value(A)>P\text{-}Value(B)$,因此,$P\text{-}Value(B)<\alpha$,该序列未能通过 T_B 的检测.②得证. □

下面以定理的形式给出检测项目之间的关系与相关度的联系.首先引入一个引理:

引理 1. 两个检测项目 A 和 B,如果 $T_A\times T_B$,那么 $P_{A-B}=P_{B-A}$;如果将(0,1]分为 n 个区间,在每个区间中均有 $P_{A-B}(i)=P_{B-A}(i)$.

证明:A 检测项目的 P-Value 均匀分布,其分布设为 X,其密度函数为

$$f(x)=\begin{cases}0, & x<0\\ 1, & 0\leq x\leq 1\\ 0, & x<0\end{cases} \tag{2}$$

B 检测项目的 P-Value 均匀分布,其分布设为 Y,其密度函数为

$$f(y)=\begin{cases}0, & y<0\\ 1, & 0\le y\le 1\\ 0, & y>1\end{cases} \tag{3}$$

则 $W=-Y$ 的概率密度为

$$f(w)=\begin{cases}0, & w>0\\ 1, & -1\le w\le 0\\ 0, & w<-1\end{cases} \tag{4}$$

在此定义:

$$Z=X-Y=X+W \tag{5}$$

由于有 $T_A\times T_B$,所以二者的 P-Value 分布是独立的,即 X 与 Y 独立,X 与 W 也独立.

对于独立的两个随机变量 $Z=X+W$ 的概率密度为[12]

$$f(z)=f_X\times f_W=\int_{-\infty}^{\infty}f_X(x)f_W(z-x)\mathrm{d}x \tag{6}$$

带入式(3)、式(4)可知:

$$f(z)=\begin{cases}0, & z<-1\\ z+1, & -1\le z\le 0\\ -z+1, & 0\le z\le 1\\ 0, & z>1\end{cases} \tag{7}$$

$f(z)$的概率密度曲线如图 1 所示.

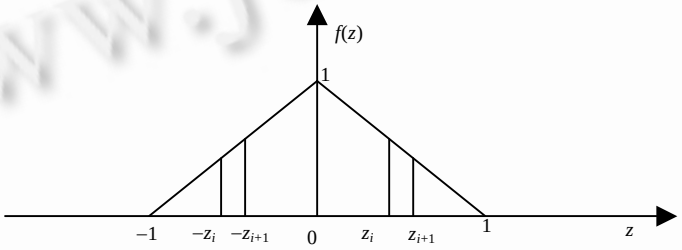

Fig.1 Curve of the probability density function $f(z)$ of $Z=X+W$

图 1 $Z=X+W$ 的概率密函数 $f(z)$的曲线

这里,P_{A-B} 是 P-*Value*(A)>P-*Value*(B)的概率,也就是 $f(z)>0$ 的概率;与之对应,P_{B-A} 是 $f(z)<0$ 的概率.因为

$$P_{A-B}=\int_0^{\infty}f(z)\mathrm{d}z=\int_0^1(-z+1)\mathrm{d}z=0.5\,,\ P_{A-B}=\int_{-\infty}^{0}f(z)\mathrm{d}z=\int_{-1}^{0}(z+1)\mathrm{d}z=0.5\,.$$

因此,$P_{A-B}=P_{B-A}$ 得证.

若将(0,1)分为 n 个区间,设第 i 个区间为 $(z_i,z_{i+1}]$.$P_{A-B}(i)$代表在该区间中,检测项目 A 的 P-Value 大于检测项目 B 的 P-Value 的概率,即 $P_{A-B}(i)=\int_{z_i}^{z_{i+1}}f(z)\mathrm{d}z$;$P_{B-A}(i)$代表在该区间内检测项目 B 的 P-Value 大于检测项目 A 的 P-Value 的概率,即 $P_{B-A}(i)=\int_{-z_{i+1}}^{-z_i}f(z)\mathrm{d}z$;由式(7)和图 1 可以计算出,$P_{A-B}(i)=P_{B-A}(i)$.引理得证. □

定理 1. 两个无关检测项目的相关度为 0.即两个检测项目 A 和 B,如果 $T_A\times T_B$,则 $R(A,B)=0$.

证明:根据引理 1 可知,如果 $T_A\times T_B$,则将(0,1)分为 n 个区间,每个区间中均有 $P_{A-B}(i)=P_{B-A}(i)$.

那么在第 $i(1\le i\le n)$区间中:

$$-\left(\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}+\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right)=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1\,.$$

所以, $\tilde{R}(A,B)=-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}+\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right)=1.$

因此,$R(A,B)=1-\tilde{R}(A,B)=0$ 得证. □

定理 2. 两个等价或包含的检测项目相关度为 1.即两个检测项目 A 和 B,如果 $T_B\subset T_A$ 或者 $T_B=T_A$,则 $R(A,B)=1$.

证明:如果 $T_B=T_A$,根据等价关系的定义可知,$P_{A=B}$ 恒等 1,$P_{A-B}=P_{B-A}=0$,对于 $i(1\le i\le n)$,有 $P_{A-B}(i)=P_{B-A}(i)=0$,因此,$\tilde{R}(A,B)=0$,从而有 $R(A,B)=1$.

如果 $T_B\subset T_A$,根据等价的定义有 $P_{A-B}=1,P_{B-A}=0$.因此,对于 $i(1\le i\le n)$,有

$$\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}=1,\ \frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}=0,$$

也有 $\tilde{R}(A,B)=0$,因而 $R(A,B)=1$.定理 2 得证. □

引理 1 表明,两个无关的检测项目在每个区间中有 $|P_{A-B}(i)-P_{B-A}(i)|=0$.定理 3 给出相关的检测项目在各个区间中概率的差异与相关度大小的关系.

定理 3. 两个检测项目 A 和 B,如果 $T_B\cap T_A$,则 $|P_{A-B}(i)-P_{B-A}(i)|$ 越大,$R(A,B)$越大.

证明:设一个系统 U 有两个事件 $U=\{u_1,u_2\}$,在这个系统中,u_1 定义的概率记为 P_1,u_2 的概率记为 P_2,并且有 $P_1+P_2=1$.那么根据熵的定义,该系统的熵 $H(U)=-(P_1\log_2 P_1+P_2\log_2 P_2)$.

当两个事件的概率相等,即 $P_1=P_2=1/2$ 时,熵值 $H(U)=1$,达到最大值.因为有 $P_1+P_2=1$,那么 $0\le|P_1-P_2|\le1$ 等价于 $0\le|2P_1-1|\le1$.$|P_1-P_2|$逐渐变大的过程是 P_1 由 1/2 向 0 或者向 1 逐渐变化的过程,在这个过程中,熵值 $H(U)$逐渐变小.

因为有 $T_B\cap T_A$,即两个项目不是等价、包含和无关的.因此,可以有区间满足 $P_{A-B}(i)+P_{B-A}(i)\ne0$,这个结论很显然,可以通过反证法来证明.

对于满足 $P_{A-B}(i)+P_{B-A}(i)\ne0$ 的区间,系统 U 中 P_1,P_2 分别以 $\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}$ 和 $\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}$ 代替,则 $-\left(\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}+\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\log_2\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right)$ 与 $H(X)$具有相同的性质.因为有 $\left|\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}-\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right|=\left|\frac{P_{A-B}(i)-P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right|$ 且 $P_{A-B}(i)+P_{B-A}(i)\ne0$,所以 $\left|\frac{P_{A-B}(i)}{P_{A-B}(i)+P_{B-A}(i)}-\frac{P_{B-A}(i)}{P_{A-B}(i)+P_{B-A}(i)}\right|$ 逐渐变大的过程就是$|P_{A-B}(i)-P_{B-A}(i)|$逐渐变大的过程.由熵值的非负性和加法性可知,若$|P_{A-B}(i)-P_{B-A}(i)|$逐渐变大,则不相关度 $\tilde{R}(A,B)$逐渐变小.即$|P_{A-B}(i)-P_{B-A}(i)|$越大,$R(A,B)$越大.定理 3 得证. □

3 相关度计算

目前对检测项目和参数的选择通常都是基于检测者的经验,因此,若能求得检测项目之间的相关度,就可以为随机性检测项目及其参数的选择提供理论上的依据.本节给出一种通过统计实验近似求得相关度的算法,并通过大量的实验对 NIST 采用的检测项目相关度进行计算.

需要注意的是,计算相关度的一个基本假设是所取的样本(序列)是在整个样本空间(所有的随机序列)中随机选取的.即进行统计实验的样本应该是随机的样本.本文采用 BBS(Blum-Blum-Shub generator)[16]随机数发生器来产生模拟真随机的样本数据.

3.1 求解相关度的算法

下面给出实验求解相关度 $R(A,B)$的算法 *Correlation_Degree*(*A*,*B*).

算法 1. *Correlation_Degree*(*A*,*B*).

Step 1. 将[−1,1]均匀分为 20 个子区间[−1,−0.9),[−0.9,−0.8),...,[−0.1,0),(0,0.1],(0.1,0.2],...,(0.9,1],记各子区间为 SA_i,1≤i≤20,置各子区间的初始个数 N_i=0;

Step 2. 利用 BBS 随机数发生器产生 S 条随机序列(本文实验中 S=500),每条序列长度为 n 比特(本文实验中 n=10^6 比特);

Step 3. 对第 1 条至第 S 条序列重复步骤 Step 3.1~Step 3.4:

Step 3.1. 进行 T_A 检测,P-Value 结果记作 P-$Value$(A);

Step 3.2. 进行 T_B 检测,P-Value 结果记作 P-$Value$(B);

Step 3.3. 计算差值 Dis=P-$Value$(A)−P-$Value$(B);

Step 3.4. 如果 $Dis\in SA_i$,则 N_i++;

Step 4. 计算 $\tilde{R}(A,B)=-\frac{1}{10}\sum_{i=1}^{10}\left(\frac{N_i}{N_i+N_{21-i}}\log_2\frac{N_i}{N_i+N_{21-i}}+\frac{N_{21-i}}{N_i+N_{21-i}}\log_2\frac{N_{21-i}}{N_i+N_{21-i}}\right)$;

Step 5. 计算 R(A,B)=1−$\tilde{R}$(A,B);

Step 6. 返回结果 R(A,B).

3.2 实验结果

利用第 3.1 节中的算法 1 对 NIST 在评选 AES 中用到的 16 个检测项目的相关度进行计算.实际上,NIST 用到的 16 个检测项目根据不同的参数可以分为 189 个子项目.本文的研究综合考虑这 189 个子项目,对它们之间的相关度进行计算.检测项目及编号见表 1,计算出的相关度结果列入表 2 和表 3.

Table 1 Breakdown of the randomness tests applied during experimentation

表 1 实验中各随机性检测项目列表

Randomness test	Test ID	Number of P-Value	Randomness test	Test ID	Number of P-Value
Monobit	1	1	Serial (16)	10~11	2
Block frequency (100)	2	1	Aperiodic templates (9)	12~159	148
Cusum	3~4	2	Periodic template (9)	160	1
Runs	5	1	Random excursions	161~168	8
Rank	6	1	Random excursions variant	169~186	18
Long runs of ones	7	1	Spectral DFT	187	1
Universal statistical	8	1	Lempel-Ziv compression	188	1
Approximation entropy (10)	9	1	Linear complexity	189	1

Table 2 Partial experimental result of correlation degree between randomness tests

表 2 随机检测检测项目相关度实验结果(部分)

Correlation degree	1	2	3	5	6	7	8	9	10	11	12	160	161	169	187
1	1.0	0.116	**0.566**	0.133	0.025	0.011	0.047	0.112	0.057	0.010	0.081	0.026	0.044	0.110	0.135
2	0.116	1.0	0.021	0.021	0.038	0.106	0.117	0.011	0.116	0.036	0.099	0.033	0 .046	0.036	0.108
3	**0.566**	0.021	1.0	0.042	0.058	0.037	0.046	0.113	0.119	0.044	0.092	0.045	0.014	0.018	0.123
5	0.133	0.021	0.042	1.0	0.153	0.027	0.033	0.016	0.016	0.023	0.106	0.021	0.113	0.030	0.052
6	0.025	0.038	0.058	0.153	1.0	0.020	0.142	0.118	0.016	0.168	0.090	0.030	0.019	0.005	0.012
7	0.011	0.106	0.037	0.027	0.020	1.0	0.107	0.131	0.026	0.020	0.109	0.134	0.016	0.014	0.018
8	0.047	0.117	0.046	0.033	0.142	0.107	1.0	0.121	0.038	0.035	0.113	0.038	0.012	0.041	0.119
9	0.112	0.011	0.113	0.016	0.118	0.131	0.121	1.0	0.031	0.036	0.109	0.119	0.113	0.011	0.115
10	0.057	0.116	0.119	0.016	0.016	0.026	0.038	0.031	1.0	0.412	0.093	0.030	0.111	0.161	0.020
11	0.010	0.036	0.044	0.023	0.168	0.020	0.035	0.036	0.412	1.0	0.104	0.014	0.036	0.015	0.036
12	0.081	0.099	0.092	0.106	0.090	0.109	0.113	0.109	0.093	0.104	1.0	0.101	0.097	0.093	0.132
160	0.026	0.033	0.045	0.021	0.030	0.134	0.038	0.119	0.030	0.014	0.101	1.0	0.117	0.125	0.133
161	0.044	0.046	0.014	0.113	0.019	0.016	0.012	0.113	0.111	0.036	0.097	0.117	1.0	0.122	0.110
169	0.110	0.036	0.018	0.030	0.005	0.014	0.041	0.011	0.161	0.015	0.093	0.125	0.122	1.0	0.140
187	0.135	0.108	0.123	0.052	0.012	0.018	0.119	0.115	0.020	0.036	0.132	0.133	0.110	0.140	1.0

Table 3 Breakdown of randomness test pairs whose correlation degree is larger than 0.5

表 3 相关度超过 0.5 的随机性检测项目对

ID	Result of correlation degree	Remark
1	$R(1,3)$=0.566	Monobit and cusum
2	$R(169,170)$=0.611, $R(170,171)$=0.524, $R(171,172)$=0.531 $R(172,173)$=0.410, $R(173,174)$=0.501, $R(182,183)$=0.506 $R(183,184)$=0.513, $R(184,185)$=0.602, $R(185,186)$=0.604	Random excursions variant
3	$R(12,86)$=0.913,$R(85,159)$=0.908	Aperiodic templates (9)

3.3 实验结果分析

受篇幅所限,本文只列出了部分结果,见表 2.同时总结了相关度超过 0.5 的项目列表,见表 3.通过实验结果也可看到,$R(A,A)=1$,$R(A,B)=R(B,A)$.大部分的检测项目之间的相关度均小于 0.5,并且绝大部分小于 0.5 的相关度均小于 0.2.NIST 曾利用主成分分析研究了这些检测项目的依赖关系,其研究结论是这些项目之间的依赖关系很小,但其研究并没有给出量化的结果,并且其研究也没有发现本文表 3 所列的一些检测项目之间存在的依赖关系.

从表 3 可以看出,频数与累加和之间、随机游动变量的结果之间、不同的非重叠模板之间的 *P*-Value 值并不是完全独立的,而是存在一定的关系.下面进行具体分析.

由第 3.3 节的公式(7)可以计算得出,如果有 500 条随机的序列并将[–1,1]均匀分为 20 个相等的区间,则两个完全无关检测项目的 *P*-Value 之差分散于各区间的序列个数应为$\{SA_i\}$={2.5,7.5,12.5,17.5,22.5,27.5,32.5,37.5,42.5,47.5,47.5,42.5,37.5,32.5,27.5,22.5,17.5,12.5,7.5,2.5}.本文将$\{SA_i\}$作为下一步继续分析中的“期望值”.

首先讨论检测项目相关度小于 0.2 的情况,以单比特频数与矩阵秩为例,$R(1,6)$=0.025,二者的 *P*-Value 之差的分布与期望值对比情况如图 2 所示.从图 2 中我们可以看出,二者 *P*-Value 之差的分布与期望值的分布相差不大.

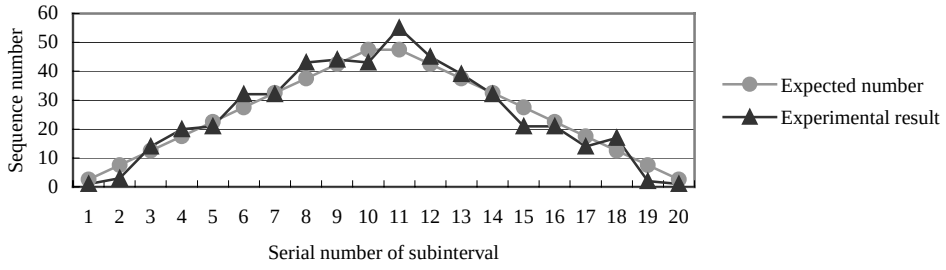

Fig.2 Contrast between distribution and expected for *P*-Value dispersion of monobit and rank

图 2 单比特频数与矩阵秩 *P*-Value 之差的分布与期望分布的对比

单比特频数与累加和的相关度为 0.566,实验所得的二者 *P*-Value 之差在各区间分布的序列条数与期望值的对比情况如图 3 所示,可以看出,单比特频数与累加和的 *P*-Value 差值的分布与无关项目的期望分布差异明显,二者的 *P*-Value 之间存在着很强的依赖关系.由图 2 与图 3 的对比也可以看出,图 3 中两条曲线的差异要小于图 2.单比特频数与矩阵秩的相关程度比单比特频数与累加和的相关程度要小,更接近于无关.

对于非重叠模板的分析如下,序号 12 是非重叠模板为 000000001 的检测,序号 86 是非重叠模板为 100000000 的检测.$R(12,86)$=0.913,说明二者的相关度非常大.实际上,也可以直观地看出这两个非重叠模板本身就具有很大的关联性.同样,序号为 113 的检测是非重叠模板为 011111111 的检测,序号为 159 的检测是非重叠模板为 111111110 的检测,二者的相关性也很大.因此,就统计计算的情况来看,序号 12 与序号 86 的检测项目、序号为 113 与序号 159 的检测项目之间存在着冗余.

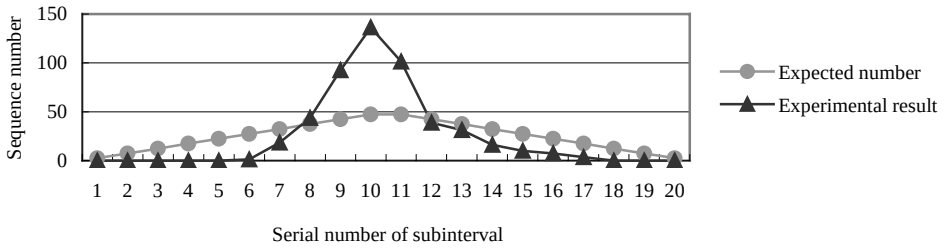

Fig.3 Contrast between distribution and expected for *P*-Value dispersion of monobit and cusum

图 3 单比特频数与累加和的 *P*-Value 之差的分布与期望值的对比

对随机游动变量的相关度检测结果分析与非重叠模板的分析类似,在随机游动所得到的 18 个检测结果中,部分结果存在着依赖关系.

3.4 相关度的一个具体应用

检测项目相关度可以为检测项目和参数的选择提供科学依据.本节给出利用相关度进行检测参数选择的一个具体应用.

目前已有了 *N* 个检测项目,现在需要从另外一个角度来检测序列的随机性,那么需要新增加一个检测项目,并且这个新的检测项目带有参数.假设已有检测项目为:单比特频数、块内频数、累加和、游程分布、块内最大“1”游程、矩阵秩、通用统计、重叠模板、非重叠模板、随机游动、随机变量,新增检测项目为近似熵,近似熵检测带有参数.那么选择参数 2 还是参数 10 的近似熵,就需要考虑新增加的近似熵与已有检测项目之间的相关程度.经过计算,参数为 2 的近似熵和参数为 10 的近似熵与已存在项目之间的相关度对比见表 4.

Table 4 Contrast the correlation degree between entropy with parameter of 2 and entropy with parameter of 10

表 4 参数为 2 的近似熵与参数为 10 的近似熵的相关性比较

Test ID	1	2	3	5	6	7	8	10	11	12	160	161	169	187
9 (Parameter=2)	**0.422**	0.016	**0.305**	**0.233**	0.020	0.040	0.029	0.108	0.021	0.130	0.024	0.047	0.043	0.043
9 (Paramter=10)	0.112	0.011	0.113	0.016	0.118	0.131	0.121	0.031	0.036	0.109	0.119	0.113	0.011	0.115

由表 4 可以看出,参数为 10 的近似熵与现存项目的相关度很小,均未超过 0.2.而参数为 2 的近似熵与序号为 1,3 和 5 的检测项目相关度分别为 0.422,0.305 和 0.233.因此,通过相关度的比较,建议优先选择参数 10.

对于其他检测项目及其参数,同样可以利用这种方法和策略来进行选择.相关度为项目和参数的选择提供了一种可操作的通用方法.

4 结束语

本文从统计学的角度对检测项目之间的关系进行了量化研究.定义了检测项目之间可能存在的 4 种关系:包含、等价、相关和无关,并提出了检测项目相关度的概念.主要研究结果可以总结如下:1) 提出了一种利用熵值对相关度进行度量的方法,量化的相关度是[0,1]之间的实数;2) 证明了检测项目之间的关系与相关度的联系;3) 给出了计算相关度的算法和基于相关度的参数选择策略;4) 利用本文提出的相关度对 NIST 采用的检测项目进行研究,发现其中一些检测项目之间存在着依赖关系.

本文提出的相关度概念及求解算法可以广泛应用于检测项目及其参数的选择中,为检测项目及其参数的选择提供了理论依据和可操作的方法.

References:

[1] Shannon CE. Communication theory of secrecy system. Bell System Technical Journal, 1949,28(10):656−715.

[2] Chen H. Security test on cryptographic algorithms and design of key cryptographic components [Ph.D. Thesis]. Beijing: Institute of Software, the Chinese Academy of Sciences, 2004 (in Chinese with English abstract).

[3] Knuth DE. The Art of Computer Programming, Volume 2: Seminumerical Algorithms. 3rd ed., Addison-Wesley, 1981. 59−73.

[4] Preneel B, Biryukov A, Oswald E, Van Rompay B, Granboulan L, Dottax E, Murphy S, Dent A, White J, Dichtl M, Pyka S, Schafheutle M, Serf P, Biham E, Barkan E, Dunkelman O, Quisquater JJ, Ciet M, Sica F, Knudsen L, Parker M, Raddum H. NESSIE security report. Techenical Report, D20, Belgium: Information Society Technologies (IST) Programme of the European Commission, 2003. 24−25. https://www.cosic.esat.kuleuven.be/nessie/deliverables/D20-v2.pdf

[5] Rukhin A, Soto J, Nechvatal J, Smid M, Barker E, Leigh S, Levenson M, Vangel M, Banks D, Heckert A, Dray J, Vo S. A statistical test suite for random and pseudorandom number generators for cryptographic applications. Technical Report, SP 800-22, Washington: National Institute of Standard and Technology, 2001.

[6] Filiol E. A new statistical testing for symmetric ciphers and hash functions. In: Deng R, Qing S, Bao F, Zhou J, eds. Information and Communications Security: The 4th Int'l Conf. LNCS 2513, Berlin, Heidelberg: Springer-Verlag, 2002. 342−353.

[7] Ryabko YB, Pestunov AI. "Book stack" as a new statistical test for random numbers. Probability Information Transmission, 2004,40(1):66−71.

[8] Soto J, Bassham L. Randomness testing of the advanced encryption standard finalist candidates. Technical Report, IR 6483, Washington: National Institute of Standards and Technology, 2000.

[9] Tsang WW, Hui LCK, Chow KP, Chong CF, Tso CW. Tuning the collision test for power. In: Estivill-Castro V, ed. Proc. of the 27th Australasian Conf. on Computer science-Volume 26 Dunedin. New Zealand: Australian Computer Society, Inc., 2004. 23−30.

[10] Aurer U. A universal statistical test for random bit generators. Journal of Cryptology, 1992,5(2):89−105.

[11] Marsaglia G, Tsang WW. Some difficult-to-pass tests of randomness. Journal of Statistical Software, 2002,7(3):1−9. http://www.jstatsoft.org/v07/i03/paper

[12] Sheng Z, Xie SQ, Pan CY. Probability and Statistics. 2nd ed., Beijing: High Education Press, 1989. 135−137, 189 (in Chinese).

[13] Soto J. Statistical testing of random number generators. In: Proc. of the 22nd National Information Systems Security Conf. Crystal City, 1999. http://csrc.ncsl.nist.gov/nissc/1999/proceeding/papers/p24.pdf

[14] Shannon CE. A mathematical theory of communication. Bell System Technical Journal, 1948,27(4):623−656.

[15] Zhang WQ, Zhang SY, Jiang LQ. A decision assessment model based on entropy and its application. Journal of Systems Engineering, 1995,10(3):69−74 (in Chinese with English abstract).

[16] Junod P. Cryptographic secure pseudo-random bits generation: The Blum-Blum-Shub generator. 1999. http://crypto.junod.info/bbs.pdf

附中文参考文献:

[2] 陈华.密码算法的安全性检测及关键组件的设计[博士学位论文].北京:中国科学院软件研究所,2004.

[12] 盛骤,谢式千,潘承毅.概率论与数理统计.第 2 版,北京:高等教育出版社,1989.135−137, 189.

[15] 张文泉,张世英,江立勤.基于熵的决策评价模型及应用.系统工程学报,1995,10(3):69−74.

范丽敏(1978－),女,内蒙古赤峰人,博士生,助理研究员,主要研究领域为密码学,信息安全.

陈华(1976－),女,博士,副研究员,主要研究领域为密码学,信息安全.

冯登国(1965－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为密码学,信息安全.