GB

NATIONAL STANDARD OF THE

PEOPLE'S REPUBLIC OF CHINA

ICS 35.040
L 80

GB/T 32915-2016

Information security technology - Randomness test methods for binary sequence

信息安全技术

二元序列随机性检测方法

Issued on: August 29, 2016 **Implemented on: March 01, 2017**

Issued by: General Administration of Quality Supervision Inspection and Quarantine of PRC;

Standardization Administration of PRC.

Table of Contents

Information security technology - Randomness test methods for binary sequence

1 Scope

This standard specifies the randomness test indicators and test methods in commercial password applications.

This standard applies to the randomness test of binary sequences generated by random number generators.

2 Terms and definitions

The following terms and definitions apply to this document.

2.1

Binary sequence

A bit string consisting of “0” and “1”.

2.2

Random number generator

A device or program that produces a random binary sequence.

2.3

Randomness hypothesis

When performing randomness test on a binary sequence, first assume that the sequence is random. This assumption is called the original hypothesis or null hypothesis and is recorded as H_0. The hypothesis opposite to the null hypothesis, that this sequence is not random, is called the alternative hypothesis, which is denoted as H_α.

2.4

Randomness test

A function or process used for binary sequence test to determine whether to accept the randomness null hypothesis.

2.5

Significance level

The probability of erroneously determining a random sequence as a non-random sequence in randomness test, which is represented by α.

2.6

Sample

A binary sequence for randomness test, which is called a sample.

2.7

Sample length

The number of bits in a sample.

2.8

Sample size

The number of samples that are randomly tested.

2.9

Test parameter

Parameters that are required to be set for randomness test.

2.10

Run

A self-sequence consisting of consecutive “0” or “1” in a sequence, and the preamble and successor elements of the subsequence are different from their own elements.

3 Symbol

The following symbols apply to this document.

α: Significance level

H_0: Original hypothesis (null hypothesis)

H_α: Alternative hypothesis

ε: Sequence to be tested

n: Bit length of the sequence to be tested

ε_i: A bit in the sequence to be tested, ε_i = (0, 1)

ε': A new sequence generated in accordance with certain rules on the basis of ε

Xi: $2\varepsilon_i - 1$

m: Bit length of the subsequence

Σ: AND symbols

*: Multiplication, sometimes omitted

ln(x): Natural logarithm of x

$Log_2(x)$: Logarithm of x with base 2

$\lfloor x \rfloor$: The largest integer not greater than x

max: Taking the maximum value from several elements

Φ(x): Standard normal distribution function

V: Statistic value

P_value: Complementary error function

erfc: A measure of the quality of the sample randomness.

igamc: Incomplete gamma function

π: The ratio of 1 in the sequence to be tested

$V_{n(}$obs): Total number of runs in the sequence to be tested

ApEn(m): Approximate entropy of sequence to be tested

K: Number of L-bit subsequences in the sequence to be tested in universal statistical test

L: Length of subsequence in general statistics

L_i: Linear complexity of subsequences in linear complexity test

M: Number of matrixes in matrix rank test

$1.5 < T_i \leq 2.5$, v_5 plus 1;

$T_i > 2.5$, v_6 plus 1.

Step 6: Calculate the statistical value $V = \sum_{i=0}^{6} \frac{(v_i - N\pi_i)^2}{N\pi_i}$. Where π_i values are: $\pi_0 = 0.010417$, $\pi_1 = 0.03125$, $\pi_2 = 0.12500$, $\pi_3 = 0.5000$, $\pi_4 = 0.25000$, $\pi_5 = 0.06250$, $\pi_6 = 0.020833$.

Step 7: Calculate $P\_value = igamc\left(3, \frac{V}{2}\right)$.

4.13.3 Result determination

The P_value result calculated in 4.13.2 is compared to the significance level α. If P_value ≥ α, the sequence to be tested is deemed to pass the linear complexity test.

4.14 Maurer universal statistical test method

4.14.1 Overview

The Maurer universal statistical test is used to test whether the sequence to be tested can be losslessly compressed. Since a random sequence cannot be significantly compressed, if the sequence to be tested can be significantly compressed, the sequence is deemed not to be random.

4.14.2 Test procedures

The Maurer universal statistical test procedures are as follows:

Step 1: Divide the sequence ε to be tested into two parts: the initial sequence and the test sequence. The initial sequence consists of Q L-bit non-overlapping subsequences. The test sequence consists of K L-bit non-overlapping subsequences, discard the extra bits (not enough to form a complete L-bit subsequence), $K = \left\lfloor \frac{n}{L} \right\rfloor - Q$.

Step 2: For the initial sequence, create a table with the L-bit value as the index value in the table, T_j ($1 \leq j \leq 2^L$) represents the value of the j^{th} element in the table, calculate $T_j = i (1 \leq i \leq Q)$, where j is the decimal representation of the i^{th} L-bit subsequence in the initial sequence.

Appendix A

(Informative)

Random test principle

A.1 Single bit frequency test

Single-bit frequency test is the most basic test used to test whether the number of 0 and 1 in a binary sequence are similar. That is, if a binary sequence of length n is known, it tests whether the sequence has a good 0, 1 balance. Let n_0, n_1 denote the number of 0 and 1 in the sequence, respectively. For a random sequence, when its length is sufficiently large, its statistical value V shall obey the standard normal distribution:

$$V = 2\sqrt{n}\left(\frac{n_1}{n} - \frac{1}{2}\right)$$

A.2 Intra-block frequency test

The intra-block frequency test is used to test whether the number of 1 in the m-bit subsequence of the sequence to be tested is close to m/2. For a random sequence, the number of 1 in an m-bit subsequence of any length shall be close to m/2.

The intra-block frequency test divides the sequence to be tested into N subsequences, each of which has a length of m and n = N × m. Of course, if n cannot be divisible by m, there will be extra bits, and the extra bits will be discarded at this time. Calculate the ratio of 1 in each subsequence, set it to $\pi_i = \frac{\sum_{j=1}^{m} \varepsilon_{(i-1)m+j}}{m}$, 1 ≤ i ≤ N . The cumulative sum of the ratios of 1 of all N subsequences is taken as a statistical value.

$$V = 4m\sum_{i=1}^{N}\left(\pi_i - \frac{1}{2}\right)^2$$

This statistic shall obey the χ^2 distribution with a degree of freedom of N.

A.3 Poker test

For any positive integer m, the binary sequence of length m has 2^m types. The sequence to be tested is divided into $N = \left\lfloor \frac{n}{m} \right\rfloor$ non-superimposed

the sequence to be tested obeys the randomness requirement.

Let $V_n(obs)$ denote the total number of runs of the sequence to be tested, π denotes the proportion of 1 in the sequence, and Φ(z) is the standard normal distribution, then:

$V=\frac{V_n(obs)-2n\pi(1-\pi)}{2\sqrt{n}\pi(1-\pi)}$ obeys the standard normal distribution.

A.6 Run distribution test

A run of consecutive 1 (or 0) is called a block (or a discontinuity). If the binary sequence to be tested is random, the number of runs of the same length is nearly identical. The expected value of the number of runs of length i in a random binary sequence of length n is $e_i=\frac{n-i+3}{2^{i+2}}$. Let k be the largest integer that satisfies $e_i \geq 5$. Let b_i, g_i denote the number of "1" runs and "0" runs of length i in a binary sequence, respectively, for each i, $1 \leq i \leq k$. The statistical value V approximately obeys the χ^2 distribution with a degree of freedom of 2k - 2:

$$V=\sum_{i=1}^{k}\frac{(b_i-e_i)^2}{e_i}+\sum_{i=1}^{k}\frac{(g_i-e_i)^2}{e_i}$$

A.7 Maximum "1" run test in the block

The sequence to be tested is divided into N equal-length subsequences, and the randomness of the sequence to be tested is evaluated in accordance with the distribution of the largest 1 run in each subsequence.

The sequence to be tested is divided into N subsequences of length m, where n = N × m. In accordance with the size of m, corresponding to (K + 1) sets (related to the size of m), then calculate the length of the largest "1" run of each subsequence and classify it into the corresponding set. Let the number of elements in the (K + 1) sets be v_0, v_1, v_2, ..., v_k ($v_0 + v_1 + v_2 + ... + v_k = N$), and the statistical value V shall obey the χ^2 distribution with a degree of freedom K:

$$V=\sum_{i=0}^{K}\frac{(v_i-N\pi_i)^2}{N\pi_i}$$

The values of K and π_i are related to m. Table A.1, Table A.2 and Table A.3 respectively show the corresponding K value, v_i definition and π_i value when m takes 8, 128 and 10000 respectively.

π_I represents the relative frequency of the mode I = (i_1, i_2, ..., i_m) appearing in the sequence to be tested;

$-\varphi^{(m)}$ represents the entropy of the relative frequency distribution of all 2^m m-bit subsequence modes.

The approximate entropy ApEn(m) is defined as: $ApEn(m) = \varphi^{(m)} - \varphi^{(m+1)}$, where $ApEn(0) = -\varphi^{(1)}$.

The approximate entropy gives the difference between the frequency between the m-bit overlappable subsequence mode and the (m + 1)-bit overlappable subsequence mode when the subsequence length m is increased by one. Therefore, a small ApEn(m) value indicates that the sequence to be tested is regular and continuous; and a large ApEn(m) value indicates that the sequence to be tested is irregular and discontinuous.

For any m, the approximate entropy of the random sequence (irregular sequence) ApEn(m) shall be approximately equal to ln2. Therefore, the statistical value $V = 2n[\ln 2 - ApEn(m)]$ shall obey the χ^2 distribution with a degree of freedom of 2^m.

A.13 Linear complexity test

The sequence to be tested is divided into N subsequences of length M, where n = N × M, then the linear complexity L_i of each subsequence is calculated by the Berlekamp-Massey algorithm, and $T_i = (-1)^M (L_i - \mu) + 2/9$, where $\mu = M/2 + [9 + (-1)^{M+1}]/36 - 1/2^M (M/3 + 2/9)$.

Select (K + 1) disjoint independent sets, then classify the T_M of each subsequence in accordance with the set, count the number of T_M appearing in each set, which are respectively recorded as v_0, v_1, ..., v_K, obviously $v_0 + v_1 + ... + v_K = N$.

Statistical value $V = \sum_{i=0}^{K} \frac{(v_i - N\pi_i)^2}{N\pi_i}$ shall obey the χ^2 distribution with a degree of freedom of K.

This standard selects K = 6, set 7 positive integers v_0, v_1, ..., v_6, set the initial values of these 7 positive integers to 0, for all $i \in [1, N]$:

If: $T_i \le -2.5$, v_0 plus 1;

$-2.5 < T_i \le -1.5$, v_1 plus 1;

$-1.5 < T_i \le -0.5$, v_2 plus 1;

$-0.5 < T_i \le 0.5$, v_3 plus 1;

$0.5 < T_i \leq 1.5$, v_4 plus 1;

$1.5 < T_i \leq 2.5$, v_5 plus 1;

$T > 2.5$, v_6 plus 1.

Wherein, the corresponding π_i values are: $\pi_0 = 0.010417$, $\pi_1 = 0.03125$, $\pi_2 = 0.12500$, $\pi_3 = 0.5000$, $\pi_4 = 0.25000$, $\pi_5 = 0.06250$, $\pi_6 = 0.020833$.

A.14 Maurer universal statistical test

Maurer general statistics (referred to as general statistics) test mainly tests whether the sequence to be tested can be losslessly compressed. If the sequence to be tested can be significantly compressed, the sequence is considered to be non-random because the random sequence cannot be significantly compressed.

Universal statistical testing can be used to test various aspects of the sequence to be tested, but this does not mean that the universal statistical test is the assembly of the previous tests, but the universal statistical test completely adopts a different method from other tests. A sequence can be tested by universal statistical test if and only if the sequence is incompressible. The purpose of universal statistical testing is to test any statistical defects in the sequence to be tested.

The universal statistical test requires a large amount of data, which divides the sequence into subsequences of length L, then divides the sequence to be tested into two parts: initial sequence and test sequence. The initial sequence includes Q subsequences, Q shall be greater than or equal to 10×2^L; the test sequence includes K subsequences, K shall be greater than or equal to 1000×2^L. Therefore, the sequence length n shall be $10 \times 2^L + 1000 \times 2^L$, and L shall be in the range of $1 \leq L \leq 16$, it is recommended that L take a value not less than 6. Obviously, when L = 6, there is n = 387840. When the sequence length n is constant, it should select $K = \left\lfloor \frac{n}{L} \right\rfloor - Q$, Q shall ensure that all 2^L modes of the L-bit subsequence appear at least once in the initial sequence.

First, traverse the initial sequence (in the unit of blocks) from the beginning, find the position (block number) where each L-bit mode last appears in the initial sequence. If an L-bit mode does not appear in the initial sequence, set its position to 0; thereafter, traversing the test sequence from the beginning, each time will get an L-bit subsequence, calculate the position of the subsequence and the position difference of the last occurrence of the subsequence, that is, the block number is subtracted, and the result of the subtraction is the distance. Then calculate the base 2 logarithm of the distance; finally, add all the results of the logarithm. In this way, it can get the statistical value:

This is an excerpt of the PDF (Some pages are marked off intentionally)

------ The End ------