CRC 编码检错性能考察

赵启满

College of Electronic and Information Engineering (CEIE), Tongji University
同济大学 电子与信息工程学院

2024 年 5 月 16 日

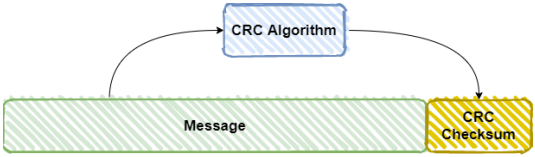
需关注

- 信息传输、数据存储过程中，使用 CRC 校验码保障完整性，出错而未被检测到的概率 P_{ue}。
- CRC 的极限性能，即多少位的错误可以保证 100% 的检出。
- 误码率、数据长的最差情况下，CRC 的未检出率 P_{ue} 是否有下限值，即最低检测性能。

背景知识

- Cyclical Redundancy Check (CRC) 校验码是一种广泛应用的错误检测码。
- 数据 i 与校验码 r 相连，生成 codeword 码字 c。c = [i, r].
- 数据 i、校验码 r 视作多项式，采取模 2 运算。

$$r(x) = i(x) \cdot x^m \mod g(x)$$
$$i(x) = i_0 + i_1 x + i_2 x^2 + \cdots + i_{k-1} x^{k-1}$$
$$r(x) = r_0 + r_1 x + r_2 x^2 + \cdots + r_{m-1} x^{m-1}$$
$$g(x) = x^m + g_{m-1} x^{m-1} + \cdots + g_1 x + g_0$$



重要性质

$c_1 \oplus c_2 \in C$，故原码 c 异或错误图案 $c \oplus c_e \in C$
$c = [i, r], c_e = [e, r_e]$ $c' = [i \oplus e, r \oplus r_e]$
给定数据的错误图样，前后校验码的汉明距离与数据内容无关。
例: $c_e = [e, r_e] = [0110, 1011]$

data	checksum	flipped by 0110	checksum'	codeword dist
0100	0101	0010	1110	5
0110	1011	0000	0000	5
1100	1111	1010	0100	5

表: CRC-4 校验码对比

给定三个原始数据，按 0110 翻转。**三者前后的检验码都满足 1011 翻转。**

检错指标

指定码字长度 n，CRC 多项式 $g(x)$，数据长度 k，校验码长度 m，此 (n,k) 码共有 2^k 种码字，可用如下指标衡量此 (n,k) 码的性能：

最小汉明距离 $d_{\min}$

数据 i 与其校验码 r 组成码字 c，对于任意的码字对 c_1 和 c_2，都满足汉明距离大于最小值 $d_{\min}$。
(n,k) 码具有最小汉明距离 $d_{\min}$，可以检测出所有不大于 $d_{\min}-1$ 位的错误。

未检出率 P_{ue}

未检出率 P_{ue} 是指在传输过程中，出错而未被检测到的概率，即错码同样为 (n,k) 的有效码字。计算公式如下，其中 A_j 是汉明距离为 j 的码字数量（码重），B_j 为对偶码码重。$P_{ue}(C,p) = \sum_{j=d_{\min}}^{n} A_j p^j (1-p)^{n-j} = 2^{-m} \sum_{j=0}^{n} B_j (1-2p)^j - (1-p)^n$

Depends on

生成多项式 $g(x)$

生成多项式的选择最重要，也最可控。$g(x)$ 是 CRC 算法的核心，决定了 CRC 算法的性能。

- 生成多项式的次数 m 决定了校验码的长度。
- 同次的不同生成多项式，其间性能差异巨大。

码字长度 n

给定生成多项式 $g(x)$，数据越长，则最小汉明距离 $d_{\min}$ 越小。
可人为控制码字长度，保证 CRC 算法在较好性能的长度区间内。

信道参数

误码率、二元对称...

To Compute d_{min}

d_{min} 及汉明距离分布是衡量 CRC 码性能的重要指标，其计算没有通用的方法，只能穷举所有出错可能。
k 位数据中发生 t 位错误，可能的图样有 C_k^t 种，是复杂度约 $O(k^t)$ 的问题。

示例：4 字节数据，发生 4 bits 翻转错误，出错后重新计算 CRC-32 校验码。考虑所有可能错误，比较与原校验码的汉明距离，统计结果如下：

```
9dc937f5e94b664a 25ac7bdd3ebad8e8 e89e756e5c7ed8b9
[0, 0, 0, 0, 0, 0, 30, 32, 106, 166, 562, 1044, 1896, 2921, 4003, 4784, 4907, 4756, 3853, 2869, 2013, 1102, 505, 268, 101, 33, 8, 1, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 30, 32, 106, 166, 562, 1044, 1896, 2921, 4003, 4784, 4907, 4756, 3853, 2869, 2013, 1102, 505, 268, 101, 33, 8, 1, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 30, 32, 106, 166, 562, 1044, 1896, 2921, 4003, 4784, 4907, 4756, 3853, 2869, 2013, 1102, 505, 268, 101, 33, 8, 1, 0, 0, 0, 0, 0]
```

可以看出，4 位错误时校验码最小距离为 6，码字最小距离为 10。若寻找 d_{min}，无需计算 9 位以上错误。

To Compute P_{ue}

$$P_{ue}(C,p) = \sum_{j=d_{\min}}^{n} A_j p^j (1-p)^{n-j} = 2^{-m} \sum_{j=0}^{n} B_j (1-2p)^j - (1-p)^n$$

A_j 的计算获得与 $d_{\min}$ 原理相同，获得精确的 P_{ue} 需要考虑所有的 A_j，复杂度为 $O(2^k)$, 计算极其困难。

CRC 适用误码率 p 较小的信道，通常 $n >> d_{\min}$，纳入所有项意义不大且计算困难，数据较短且 p 很小时可只计算 $d_{\min}$ 及后几项近似。

或者借由构建对偶码 $C^{\perp}$，对偶码字空间仅有 2^{n-k}，利用对偶码的码重 B_j 计算 P_{ue}。

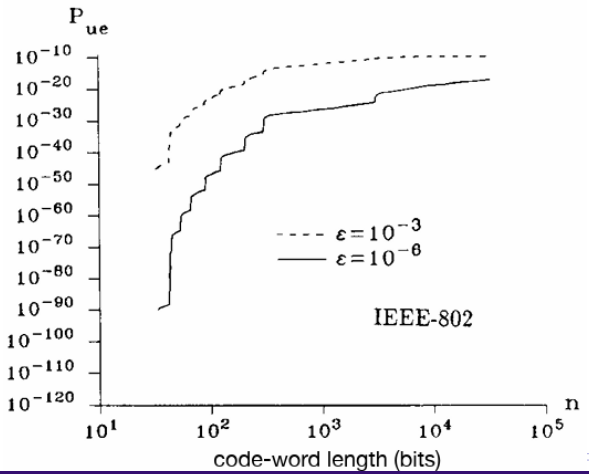
IEEE 802.3 CRC-32

以太网标准 IEEE 802.3 CRC-32 不同码字长度下的的最小汉明距离。

code length n	$d_{min}(n)$
3007,...,12144	4
301,...,3006	5
204,...,300	6
124,...,203	7
90,...,123	8
67,...,89	9
54,...,66	10
45,...,53	11
43,...,44	12
33,...,42	15

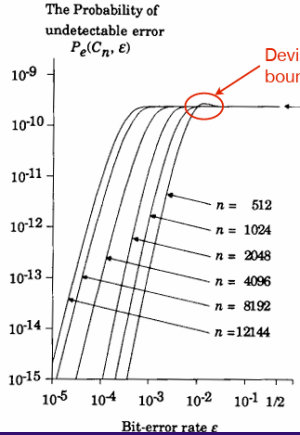
不同误码率下的 P_{ue}

以 IEEE 802.3 CRC-32 为例，对于误码率 $p = 10^{-3}$ 和 $p = 10^{-6}$ 的二元对称信道



P_{ue} 最差极限

以 IEEE 802.3 CRC-32 为例，不论误码率、不论码字长度，P_{ue} 最终都有一个最差极限值，约为 2^{-32}。



$P_{ue} \approx 2^{-32}$ for any n and $p > 0.5 \cdot 10^{-3}$

总结

对 CRC 多项式的性能评估在计算上非常困难，没有特别便捷的数学工具。

CRC 校验保证 100% 的检出错误位数 d_{min} - 1 的计算是可行。

错误未检出率 P_{ue} 可以使用对偶码的码重分布进行精确计算。

错误未检出率 P_{ue} 存在**最差极限**，可以在恶劣条件下保证检出率的下限值，m 位校验码可以实现的极限约为 2^{-m}，即增加校验码位数可以极大提高检出率。

参考文献

- G. Castagnoli, S. Bräuer and M. Herrmann. “Optimization of Cyclic Redundancy-Check Codes with 24 and 32 Parity Bits”. IEEE Transactions on Communications, vol. 41, no. 6, pp. 883-892, June 1993.
- T. Fujiwara, T. Kasami and S. Lin. “Error Detecting Capabilities of the Shortened Hamming Codes Adopted for Error Detection in IEEE Standard 802.3”. IEEE Transactions on Communications, vol. 37, no. 9, pp. 986-989, September 1989.
- P. Koopman. “32-Bit Cyclic Redundancy Codes for Internet Applications”. International Conference on Dependable Systems and Networks (DSN), pp. 459-468, July 2002.