关于线性时间复损码的研究

慕建君,孙韶辉,王新梅
(西安电子科技大学综合业务网国家重点实验室,陕西西安 710071)

摘 要: 本文对基于随机二部图的复损码进行了深入的研究.提出了给定度分布对的复损码成功译码时可接受最大损失 δ 的一上界,通过对此上界的详细分析提出了求解复损码度分布对的一种算法.这就从理论上说明了具有如上算法选取度分布对的复损码,应该优于文[2]所给度分布对的复损码.而且证明了具有某一确定度分布对的复损码能以线性时间可编码和可成功地译码.

关键词: 复损码; 随机二部图; 编译码复杂度; 度分布对算法
中图分类号: TN919.3 **文献标识码**: A **文章编号**: 0372-2112 (2002) 01-0122-04

Study on Linear Time Loss-Resilient Codes

MU Jian-jun, SUN Shao-hui, WANG Xin-mei
(National Key Lab. of ISN, Xidian University, Xi'an, Shaanxi 710071, China)

Abstract: A detailed study of loss-resilient codes based on random bipartite graphs are made in this paper. We propose the upper bound on the maximum tolerable loss fraction δ for which the decoding of the loss-resilient code with a given degree distribution pair is successfull. The algorithm to find the degree distribution pair of loss-resilient codes is presented by making a detailed analysis of this upper bound. This result shows that codes constructed from degree distribution pair obtained by the algorithm above should perform better than the degree distribution pair given in [2]. Moreover, we prove loss-resilient code with the definite degree distribution pair can be both encoded and decoded successfully in linear time.

Key words: loss-resilient code; random bipartite graph; encoding and decoding complexity; degree distribution pair algorithm

1 引言

近年来,ARQ(Automatic Repeat reQuest)技术和分层恢复等新技术可有效地提高数据传输的可靠性和避免网络拥塞的出现,但可能导致大的时延.克服这一缺点的办法是利用编码的方法[1],即把要传输的 k 比特的原数据编码为 $n(n>k)$ 比特的数据后发送出去,若接收方接收到足够量的数据,则运用适当的译码方法就可恢复 k 比特的原数据,称这种码为前向纠错(FEC, Forward Error Correcting)码[1],或称复损码(Loss-Resilient Code)[2].通常在接收到的编码数据流中数据的位置是已知的,所采用的信道模型是删除信道[3],此信道中每个编码符号丢失的概率均为 p,且在传输中编码符号的丢失是相互独立的.若取 MDS 码为一 (n, k) 复损码,则利用接收到的任意 k 比特的数据均可重构原数据.标准的 MDS 码类由 RS 码来给定的,RS 码的编码和译码的时间复杂度分别为 $O(n\log n)$ 和 $O(n\log^2 n\log\log n)$[4],但是这类码不能与删除信道的信道容量很接近.Elias 已证明删除信道的信道容量为 $1-p$,而且随机线性码可在删除信道下以任意的速率 $R(R<1-p)$ 传输,其编码和译码的时间复杂度分别为 $O(n^2)$ 和 $O(n^3)$[3].于是一方面得到基于 RS 码的 MDS 码具有较低的编译码复杂度,但是不能与删除信道的信道容量任意接近;另一方面有随机线性码可与删除信道的信道容量任意接近,但具有高的编译码复杂度.

为了解决这个问题,文[2]利用级联码的方式给出了一种能以任意接近删除信道信道容量的速率进行传输的线性时间复损码,但没有证明编译码的线性时间性.本文证明了给定度分布对的复损码成功译码时可接受最大损失 δ 的一上界,通过对此上界的详细分析提出了求解复损码度分布对的一种算法.而且证明了具有某一确定度分布对的复损码能以线性时间可编码和可成功地译码.从而得到了一既具有线性时间编译码复杂度又能以任意接近删除信道信道容量的速率进行传输的复损码.

2 复损码的构造及其编译码的思想和复杂度

定义一有 n 个信息比特和 βn 个校验比特的码 $C(B)$,使得它们分别与给定二部图 B(左边集有 n 个结点,右边集有 βn 个结点)的两结点集中的结点相对应.

收稿日期:2000-09-19;修回日期:2001-03-03
基金项目:国家自然科学基金(No. 69972035)

$C(B)$型码的编码方式为每个校验比特等于二部图 B 中此校验比特的所有邻接比特的和(如图 1). 若称二部图 B 中左边的结点为变量结点,右边的结点为校验结点(如图 2),则 $C(B)$型码的删除错误译码算法为[2]:(1) [初始化]初始化校验结点为 0,把非删除错误变量结点的值加到校验结点的当前值上,同时去掉这些变量结点和由这些变量结点引出的所有边,并令 $i=1$,称所得的 B 的子图为 G_i;(2) [恢复变量结点的值]在以上所得子图 G_i 中寻找一度数为 1 的校验结点,将此校验结点的值传给它唯一的邻接变量结点 ℓ,于是恢复了结点 ℓ 的值,然后将结点 ℓ 的值加到其所有邻接校验结点的当前值上,同时去掉从变量结点 ℓ 引出的所有边,称所得的 G_i 的子图为 G_{i+1};(3) [终止条件]令 $i:=i+1$ 对(2)所得的子图重复进行(2),直到其所得的子图中找不到度为 1 的校验结点,或所有的变量结点均已恢复. 若所有的变量结点均已恢复,则称此删除错误译码算法可成功译码.

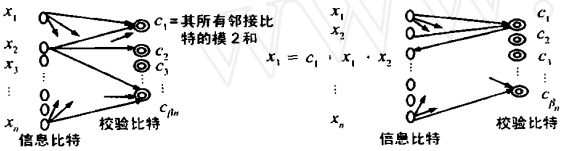

图 1　用二部图 B 定义从信息比特特到校验比特的一个映射

图 2　码 $C(B)$ 中用比特 x_1, x_2 和 c_1 来求解 x_3

若级联 $C(B)$型码 $C(B_0)$, $C(B_1)$, …, $C(B_m)$和传统复损码 $C(C(B_i)$有 βn 个信息比特和 $\beta^{+1}n$ 个校验比特, $C(B_i)$的校验比特即 $C(B_{i+1})$型码的信息比特),与码 $C(B_i)$对应的二部图 B_i 的左边和右边分别有 βn 个和 $\beta^{+1}n$ 个结点($i=0,1,\cdots,m$),并取 m 使得 $\beta^{m+1}n\approx\sqrt{n}$,最后一级码 C 是一有 $\beta^{m+1}n$ 个信息比特和码率为 $1-\beta$ 的传统纠删码(如柯西码[5]). 于是,得到一有 n 个信息比特和

$$\sum_{i=1}^{m+1}\beta n+\beta^{m+2}n/(1-\beta)=n\beta/(1-\beta)$$

个校验比特的码率为 $1-\beta$ 的复损码 $C(B_0,B_1,\cdots,B_m,C)$.

从复损码 $C(B_0,B_1,\cdots,B_m,C)$ 的构造知其编码思想由 $C(B)$型码的编码方式易得,其译码思想为:若码 C 的译码算法可恢复其所有损失,当然译出 $C(B_m)$ 的所有校验比特,由此就能恢复 $C(B_m)$ 的所有信息比特,如此逐步递推译码直到用 $C(B_0)$ 的校验比特来恢复原来的 n 个信息比特.

Luby[2]等人用微分方程为工具,将译码过程模型化后给出了给定度分布对可接受最大损失的一个分析性条件. 具体地,定义二部图 B 的一条边的左边(右边)度数为图 B 中此边左边(右边)邻接结点的度数,并用 λ_i 和 ρ_i 分别表示左边和右边度数为 i 的边的比率. 设二部图 B 左边和右边结点的最大度数分别为 d_v 和 d_c,令 $\lambda=(\lambda_2,\lambda_3,\cdots,\lambda_{d_v})$, $\rho=(\rho_2,\rho_3,\cdots,\rho_{d_c})$,称偶对 (λ,ρ) 为码 $C(B)$ 的一度分布对,若设码 $C(B)$ 度分布对的生成函数表示为 $\lambda(x)=\sum_{i=2}^{d_v}\lambda_i x^{i-1}$ 和 $\rho(x)=\sum_{i=2}^{d_c}\rho_i x^{i-1}$,则有:

引理 1[2]　对于度分布对为 (λ,ρ) 的随机二部图 B,码 $C(B)$ 的所有校验位的值已知且其删除错误的产生是随机的,若对于所有的 $x\in(0,1]$ 有 $\rho(1-\delta\lambda(x))>1-x$($\delta\in(0,1)$),则码 $C(B)$ 的删除错误译码算法能以大概率恢复所有信息比特的 δ 部分损失.

文[2]对二部图 B 提出了一度分布对 (λ,ρ),即选定正整数 d 后取 $\lambda(x)=\frac{1}{H(d)}\sum_{i=1}^{d}\frac{x^i}{i}$ 和 $e(x)=e^{\alpha(x-1)}$,其中 $H(d)=\sum_{i=1}^{d}\frac{1}{i}$,选取 α 使得 $\alpha e^{\alpha}/(e^{\alpha}-1)=a_R$,而二部图 B 左边和右边结点的平均度数 a_L 和 a_R 满足 $R=1-a_L/a_R$(R 为码 $C(B)$ 的码率),并给出如下结论:

引理 2[2]　设 $C(B)$型码中二部图的度分布 $\lambda(x)$ 和 $\rho(x)$ 如上选取,若 $\delta\leq\beta\left(1-\frac{1}{d}\right)$,则对 $\forall x\in(0,1]$ 有 $\rho(1-\delta\lambda(x))>1-x$.

引理 3[5]　某一有限域上的柯西码为 MDS 码,而且具有平方时间的编译码复杂度.

因此有如下定理:

定理 1　设复损码 $C(B_0,B_1,\cdots,B_m,C)$ 为由级联如上选取度分布对的 $m+1$ 个码 $C(B_j)_{(j=0,1,\cdots,m)}$ 和传统复损码 C(一般取柯西码)所得的复损码,则对 $\forall\varepsilon>0$,复损码能 $C(B_0,B_1,\cdots,B_m,C)$ 以线性时间可编码,同时具有能以大概率从(所有比特的) $\beta(1-\varepsilon)$ 部分的随机损失中恢复其所有信息比特的线性时间译码算法.

证明　(1)证明此复损码恢复错误的能力. 对 $\forall\varepsilon>0$,若取 C 为柯西码,由引理 3 知 C 为 MDS 码,从而 C 能从(所有比特的) $\beta(1-\varepsilon)$ 部分的随机损失中恢复其所有信息比特. 由引理 2 和引理 1 知对于如上选择的度序列,在所有校验比特已知时,其对应的 $C(B)$型码的译码算法能以大概率从(信息比特的) $\beta(1-\varepsilon)$ 部分的随机损失中恢复其所有信息比特. 于是逐步递推译码即知此复损码能以大概率从(所有比特的) $\beta(1-\varepsilon)$ 部分的随机损失中恢复其所有信息比特.

(2)证明此复损码的编译码复杂度. 由复损码 $C(B_0,B_1,\cdots,B_m,C)$ 的编码方式和译码算法知码 $C(B_j)$ 的编码时间与二部图 B_j 的边数成正比,其译码时间至多与 B_j 的边数成正比,下证复损码 $C(B_0,B_1,\cdots,B_m,C)$ 能以线性时间进行编码

对码 $C(B_j)$ $(j=0,1,\cdots,m)$,由二部图 B_j 左边结点的平均度数 $a_L=H(d)(d+1)/d$ 知 B_i 的边数为

$$E_i=\beta n a_L=\beta nH(d)(d+1)/d$$

因此, $C(B_0)$, $C(B_1)$, …, $C(B_m)$级联码的编码时间与下式成正比

$$\frac{nH(d)(d+1)}{d}+\frac{\beta nH(d)(d+1)}{d}+\cdots+\frac{\beta^n nH(d)(d+1)}{d}$$

$$=(1+\beta+\beta^2+\cdots+\beta^n)\frac{nH(d)(d+1)}{d}=\frac{n}{1-\beta}(1-\beta^{m+1})\frac{H(d)(d+1)}{d}$$

柯西码 C 的码长为 $n\beta^{m+1}+\frac{\beta^{m+2}n}{1-\beta}=\frac{\beta^{m+1}n}{1-\beta}$,由引理 3 知

柯西码 C 能以平方时间可编码. 而且注意到 $\left(\frac{\beta^{m+1}}{1-\beta}\right)^2 \approx \frac{(\sqrt{n})^2}{(1-\beta)^2} = \frac{n}{1-\beta}\cdot\frac{1}{1-\beta}$,因此,由复损码 $C(B_0, B_1, \cdots, B_m, C)$ 的构造知它能以线性时间可编码. 译码情况的证明类似.

由定理 1 可见:通过如上级联得到的复损码能以线性时间进行编码和成功译码,而传统的复损码的编译码时间至少为平方时间数量级的[5,6],于是,如此选择度分布对的复损码的编译码时间可大大降低. 这就是要采用级联的方式构造新的复损码的原因.

3 复损码可接受最大损失的上界和其度分布对的求解算法模型

对给定的度分布对 (λ, ρ) (它所对应的随机二部图为 B),Luby[7]等人利用新的概率分析工具,对与译码过程相关联的"与或(And-Or)树"进行了分析后,得出与引理 1 的条件等价的条件,即对 $\forall x \in (0, \delta]$ 有 $\delta\cdot\lambda(1-\rho(1-x)) < x$,以此条件为工具可得码 $C(B)$ 的删除错误译码算法成功译码时可接受最大损失 δ 的一上界.

引理 4 设 B 是度分布为 $\lambda(x)$ 和 $\rho(x)$ 的随机二部图,则 B 的左边和右边结点的平均度数 a_L 和 a_R 分别为

$$a_L = 1/\int_0^1 \lambda(x)\,dx,\ a_R = 1/\int_0^1 \rho(x)\,dx.$$

证明 设二部图 B 有 E 条边,且 B 的左边有 n 个结点,则有

$$a_L = \frac{E}{n} = \frac{E}{\sum_{i\geq 2}\lambda_i E/i} = \frac{1}{\sum_{i\geq 2}\lambda_i/i} = \frac{1}{\int_0^1 \lambda(x)\,dx}$$

因此 $a_L = 1/\int_0^1 \lambda(x)\,dx$. 同理可证 $a_R = 1/\int_0^1 \rho(x)\,dx$.

对于正整数 u 和实数 $x \geq 0$,令 $g(x) = u^x$,$g''(x) = u^x(\ln u)^2 \geq 0$,即 $g(x)$ 为 $x \in [0, \infty)$ 上的下凸函数,故对于满足 $\sum_{i=1}^m a_i = 1$ 的任意非负实数 a_i 及 $\forall \beta_i \geq 0 (i = 1, 2, \cdots, m; m$ 为正整数$)$ 有 $g\left(\sum_i a_i\beta_i\right) \leq \sum_i a_i g(\beta_i)$,取 $\beta_i = i$ 得如下引理:

引理 5 对于非负实数 $a_i (i = 1, 2, \cdots, m)$ 以及实数 $u > 0$,若 $\sum_{i=1}^m a_i = 1$,则 $\sum_{i=1}^m a_i u^i \geq u^{\sum_{i=1}^m i a_i}$.

由引理 4 和引理 5 可得如下上界.

定理 2 设 $\lambda(x)$ 和 $\rho(x)$ 为一随机二部图 B 的度分布,a_L 和 a_R 为二部图 B 左边和右边结点的平均度数,若对实数 $\delta \in (0,1)$ 及 $\forall x \in (0, \delta]$ 有 $\delta\cdot\lambda(1-\rho(1-x)) \leq x$,则 $\delta \leq \frac{a_L}{a_R}(1-(1-\delta)^{a_R})$.

证明 由 $\lambda(x)$ 为 $[0,1]$ 上的严格单调递增函数知它的逆函数 $\lambda^{-1}(x)$ 唯一存在且为严格单调递增函数,于是对 $\forall x \in (0, \delta]$,条件 $\delta\cdot\lambda(1-\rho(1-x)) \leq x$ 等价于条件 $1-\rho(1-x) \leq \lambda^{-1}(x/\delta)$,对上式两边从 0 到 δ 求积分并利用 $\int_0^1 \lambda^{-1}(x)\,dx = 1 - \int_0^1 \lambda(x)\,dx$ 后有

$$\int_0^1 \rho(x)\,dx - \int_0^\delta \rho(x)\,dx) \geq \delta\int_0^1 \lambda(t)\,dt$$

由引理 4 即得

$$\delta \leq \frac{a_L}{a_R} - a_L\int_0^\delta \rho(x)\,dx = \frac{a_L}{a_R}\left(1 - \int_0^\delta \rho(x)\,dx/\int_0^1 \rho(x)\,dx\right)$$

令 $B(x) = \int_0^x \rho(t)\,dt/\int_0^1 \rho(t)\,dt$,在引理 5 中取 a_i 为 $B(x)$ 中 x^i 的系数和 $u = 1-\delta$,由 $B(x)$ 的定义和引理 4 可知 a_i 为二部图 B 中度为 i 的右边结点的比率,从而有 $a_R = \sum_i i a_i$,由引理 5 得 $B(1-\delta) \geq (1-\delta)^{a_R}$,因此 $\delta \leq \frac{a_L}{a_R}(1-(1-\delta)^{a_R})$.

定理 2 中为了得到满足条件 $\delta\cdot\lambda(1-\rho(1-x)) < x (\forall x \in (0, \delta])$ 的 δ 的最优上界必须要求引理 5 中等式成立,而引理 5 中等式成立当且仅当除了某一 $a_i = 1$ 之外,其它 $a_j = 0 (j = 1, 2, \cdots m, j \neq i)$,这就意味着 $\rho(x)$ 只有唯一的非零系数,即二部图 B 的右边结点只有唯一的一个度数(即图 B 为右边正则的二部图),这启发我们在寻找最优度分布对 (λ, ρ) 时先选定右边正则的二部图,然后设计与其相匹配的最优的左边度分布. 这就从理论上说明了具有如上方法选取度分布对的复损码应该优于文[2]所给度分布对的复损码,因为对于给定的码率这样的复损码允许较大的信息损失. 于是,对于码率为 R 的码 $C(B)$ 可得二部图 B 的右边度分布 $\rho(x) = x^{a-1}$ (a 为图中右边结点的度数) 已知时,求解满足条件 $\delta\cdot\lambda(1-\rho(1-x)) < x (\forall x \in (0, \delta])$ 的左边度分布 $\lambda(x) = \sum_{i=2}^{d_v}\lambda_i x^{i-1}$ 的算法(而且使可接受的损失率 δ 尽量大一些),此算法把 $(0, \delta]$ 区间离散化,对充分大的正整数 N,取 $x_i = i\delta/N (i = 1, 2, \cdots, N)$ 有

问题模型:

$$\begin{cases} \min & \sum_{i=1}^N [x_i - \delta\cdot\lambda(1-(1-x_i)^{a-1})] \\ \text{s.t.} & \delta\cdot\lambda(1-(1-x_1)^{a-1}) < x_1 \\ & \delta\cdot\lambda(1-(1-x_1)^{a-2}) < x_2 \\ & \cdots \\ & \delta\cdot\lambda(1-(1-x_N)^{a-1}) < x_N \\ & \lambda_2 + \lambda_3 + \cdots + \lambda_{d_v} = 1 \\ & \sum_{i=2}^{d_v}\lambda_i/i = a(1-R)^{-1} \\ & \lambda_i \geq 0, i = 2, 3, \cdots, N, a \geq 2 \end{cases} \quad (1)$$

定理 3 适当选择 δ(可取小一些的 $\delta, \delta \in (0,1)$) 时线性规划 (1) 的解一定存在.

证明 由 $0 \leq \lambda(x) \leq 1$ 知总可以选取小一些的 δ 使得对所有 $x \in (0, \delta]$ 有 $\delta\cdot\lambda(1-\rho(1-x)) < x$,当然也可取到小的 δ 使线性规划 (1) 的约束条件满足,在满足 (1) 的约束条件的 δ 中取最大的 δ 便可使线性规划 (1) 的目标函数最小.

已知二部图 G 的右边度分布 $\rho(x) = x^{a-1}$ 时,求左边度分布 $\lambda(x) = \sum_{i=2}^{d_v}\lambda_i x^{i-1}$ 的算法(同时使可接受损失 δ 尽量大) 为:

(1) 令 $low := 0$, $high := 1$, $\delta = (low + high)/2$;

(2) 判断线性规划 (1) 的约束条件是否满足;

(3) 若线性规划 (1) 的约束条件满足,则令 $low := \delta$, $\delta = (low + high)/2$,转向 (2);否则令 $high := \delta$, $\delta = (low + high)/2$,转向 (2);

(4)如此进行,直到找到某一个大的 δ 使线性规划(1)有解为止.

4 结束语

本文研究了基于随机二部图的复损码的编译码算法思想.证明了若以某种方式选定随机二部图边的度分布对,则此复损码能以线性时间可编码,且具有能以大概率从(所有比特的) $\beta(1-\varepsilon)$ 部分的随机损失中恢复其所有信息比特的线性时间译码算法.同时对给定度分布对证明了此复损码成功译码时可接受最大损失 δ 的一上界,最后,通过对此上界的详细分析提出了求解复损码度分布对的一种算法,而且证明了此算法的可行性.这对复损码度分布对的设计有着重要的理论指导意义.关于复损码的研究目前主要有如下几个方面:

(1)利用最优化理论的知识和上面给出的度分布对设计算法,研究构造获得具体的最优度分布对 (λ,ρ) 的方法,从而构造良好性能的复损码;

(2)进一步从理论上分析并设计复损码的迭代译码算法,同时寻找此译码算法稳定收敛的条件.

参考文献:

[1] J W Byers,et al. A digital fountain approach to reliable distribution of bulk data [BD]. Available at http://www.icsi.berkeley.edu/~luby/,1998.

[2] M Luby,et al. Practical loss-resilient codes [A]. Proc. of the 29th ACM Symposium on the Theory of Computing [C],1997:150-159.

[3] P Elias Coding for two noisy channels [A]. Information Theory,Third London Symposium [C],1955:61-67.

[4] R E Blahunt. Theory and practice of error control codes [M]. Addison Wesley,Reading,MA,1983.

[5] J Blomer,M Mitzenmacher,A Shokrollahi. An XOR-based erasure-resilient coding scheme [BD]. Available at http://www.icsi.berkeley.edu/~luby/1995.

[6] L Rizzo. Effective erasure codes for reliable computer communication protocols. ACM Computer Communication [J]. Review,1997,27(2):24-36.

[7] M Luby,et al. Analysis of random processes via and-or tree evaluation [A]. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms [C],Francisco California,1998:364-373.

作者简介:

慕建君　男.1965 年 3 月出生于陕西吴堡县.1965 年出生.1997 年获西安电子科技大学应用数学专业硕士学位并留校任教.现为该校通信与电子系统专业在职博士研究生.目前的研究兴趣为编码、信息论与应用数学.

孙韶辉　1972 年出生.分别于 1994 年和 1999 年获西安电子科技大学学士学位和硕士学位.现为该校通信与电子系统专业博士研究生.目前的研究兴趣包括信息论、编码/调制理论和通信网络与交换.

王新梅　男.1937 年出生.西安电子科技大学教授,博士生导师.中国电子学会会士.长期从事信息论、编码和密码学的教学与研究.