DOI:10.19789/j.1004-9398.2023.04.001
文献引用:李会芳.关于子序列数量下界的一个改进[J].首都师范大学学报(自然科学版),2023,44(4):1-4. LI H F. An improvement to the lower bound on the number of subsequences[J]. Journal of Capital Normal University(Natural Science Edition),2023,44(4):1-4.

关于子序列数量下界的一个改进

李 会 芳*

(天津大学数学学院, 天津 300354)

摘要:序列删除 t 个数后会产生不同的子序列,由于序列的结构未知,给计算这些子序列的数量带来了一个难题。为了研究子序列的数量,本文在前人的基础上,采用放缩法,研究了当序列前3个块互不相同时子序列的数量,改进了子序列数量的下界。

关键词:子序列;块;下界
中图分类号:O236.2 **文献标识码**:A

An improvement to the lower bound on the number of subsequences

LI Huifang*

(School of Mathematics, Tianjin University, Tianjin 300354)

Abstract: The sequence after deleting t symbols will get different subsequences. Due to the unknown structure of the sequence which is deleted, it poses a problem to calculate the number of these subsequences. So we study the number of subsequences when the first three runs of the sequence are different in this paper. On the basis of previous ones, we adopt the scaling method and give an improvement to the lower bound on the number of subsequences.

Keywords: subsequence; run; lower bound
CLC: O236.2 **DC**: A

0 引 言

删除信道是指信息传输过程中丢失码字符号,造成信息传输错误。为了使信息有效的传输,研究可以纠正这些错误的纠错码具有重要的意义。然而并不是任意的编码都是有效的,Shannon[1]表示纠错码码率不高于信道容量时,才能实现信息的有效传输。为实现有效的编码,这个领域一个传统的问题就是,估计从原始序列中删除固定数量的元素会产生多少个不同的子序列。如,估计序列 X 删除 t 个元素会产生多少个不同的序列,记为 $\left|D_t(X)\right|$。由此可以估计纠错码最佳码本的边界,进而计算纠错码的码率,有助于评估具体结构代码是否能够有效纠错[2-9]。事实上,估计 $|D_t(X)|$ 的范围在DNA存储领域,序列重建等问题中也有应用。目前,DNA存储技术已经成为一种潜在可行的存储技术,与传统的存储介质相比,DNA有显著提高信息密度和持久性的可能,因而受到广泛的关注[10]。另外序列重建问题在信息存储、分子生物学、化学以及传感器网络等领域也有着重要的应用[11-12]。

下面介绍几个与删除相关的包括子序列、块、循环序列等的定义[13-15]。

收稿日期:2022-03-20
* 通信作者:2019233046@tju.edu.cn

令 $\mathscr{F}_q=\{0,1,2,\cdots,q-1\}$,其中 $q\geqslant 2$,设序列 $X=x_1x_2\cdots x_n\in\mathscr{F}_q^n$。

定义1(子序列) 对于序列 $X\in\mathscr{F}_q^n$,若存在 $i_1\leqslant i_2\leqslant\cdots\leqslant i_{n-t}\in[n]$,使得 $Y=x_{i_1}x_{i_2}\cdots x_{i_{n-t}}(0\leqslant t<n)$,即 Y 为从 X 中删除 t 个数得到的序列,则称 Y 是 X 的子序列。将 $D_t(X)$ 定义为 X 的长度为 $n-t$ 的子序列的集合,$|D_t(X)|$ 表示集合 $D_t(X)$ 中元素的个数,$|D_t^{(a)}(X)|$ 表示 X 的以 $a\in\mathscr{F}_q$ 为第1个元素的子序列的集合。

例如 $X=1224$, $t=2$, 则 $Y=14$ 是 X 的子序列,$D_2(X)=\{12,14,22,24\}$, $D_t^{(2)}(X)=\{22,24\}$, $|D_2(X)|=4$。

定义2(块) 序列 $X\in\mathscr{F}_q^n$ 中每一个由连续相同的数组成的最大的集合称为块(run),记 X 的块的个数为 $r(X)$,$1\leqslant r(X)\leqslant n$。

例如 $X=112220$ 有 3 个块,分别为 $11,222,0$。

定义3(循环序列) 设序列 $C_n=c_1\,c_2\cdots c_n$,对 $i\in[n]$, $c_i\in\mathscr{F}_q$,若 $c_i=a$ 满足 $i-1\equiv a\pmod q$,那么称序列 C_n 为循环序列。记 $|D_t(C_n)|=D_q(n,t)$。

注:规定 $D_q(0,0)=1$,当 $t>n$ 或 $t<0$ 时 $D_q(n,t)=0$。

例如当 $q=3$ 时,$n=7$, $C_7=0120120$。

从序列 X 中删除 t 个序列,$|D_t(X)|$ 的大小依赖于 X 的结构。事实上 q 越大,X 的长度越长,删除 t 个数后产生相同的子序列可能越多,$|D_t(X)|$ 便越难计算。Levenshtein[16] 于 1966 年给出当 $q=2$ 时 $|D_t(X)|$ 的一个范围,后在文献[17]中将这个范围推广到 q 的一般情况

$$\binom{r(X)-t+1}{t}\leqslant|D_t(X)|\leqslant\binom{r(X)+t-1}{t}。$$

1969年,Calabi 和 Hartnett[18] 认为当 X 取循环序列时 $|D_t(X)|$ 取得最大值,即对 $q\geqslant 2$, $\forall X\in\mathscr{F}_q^n$,

$$\max|D_t(X)|=|D_t(C_n)|=D_q(n,t)=\sum_{i=1}^{q}D_q(n-i,t-i+1)。$$

随后,2000年,Hirschberg 和 Regnier[13] 经过进一步分析,给出了一个更紧的界为

$$\sum_{i=0}^{t}\binom{r(X)-t}{i}\leqslant\left|D_t(X)\right|\leqslant\sum_{i=0}^{t}\binom{n-t}{i}D_{q-1}(t,t-i),$$

其中上界为 $D_q(n,t)$,即

$$D_q(n,t)=\sum_{i=0}^{t}\binom{n-t}{i}D_{q-1}(t,t-i),$$

当 $q=2$ 时,这个界化为

$$\sum_{i=0}^{t}\binom{r(X)-t}{i}\leqslant|D_t(X)|\leqslant\sum_{i=0}^{t}\binom{n-t}{i}。$$

后来,Liron 和 Langberg[14]证明了 $q=2$ 时更紧的情况。Mercier[19] 给出了 $|D_2(X)|,|D_3(X)|,\cdots,|D_5(X)|$ 的具体表达式,并推广到了一般的结论。

在此基础上,本文继续研究了估计序列 X 删除几个数后产生不同序列的数量问题,拟通过讨论当 X 前3个块的元素互不相同时 $|D_t(X)|$ 的下界,得到更紧的下界。

1 $|D_t(X)|$ 的下界

1.1 预备知识

下面介绍一些已知的命题,包括 $D_2(n,t)$ 的递推公式与具体表达式,向 X 插入一个元素后 $|D_t(X)|$ 值的变化,以及任意序列 X 关于 $|D_t(X)|$ 的下界。

命题1[13] $n\geqslant t\geqslant 2$,

$$D_2(n,t)=D_2(n-1,t)+D_2(n-2,t-1)。\quad(1)$$

证明 将集合 $D_t(C_n)$ 划分为2个部分,一部分以0为第一个元素,另一部分以1为第1个元素,则

$$|D_t(C_n)|=|D_t^{(0)}(C_n)|+|D_t^{(1)}(C_n)|=|D_t(C_{n-1})|+|D_{t-1}(C_{n-2})|,$$

所以等式(1)成立。

命题2[13] 对 $0\leqslant t\leqslant n$,

$$D_2(n,t)=\sum_{i=0}^{t}\binom{n-t}{i}。\quad(2)$$

命题3[13] 对任意 $U,V\in\mathscr{F}_q^m$ 以及任意 $\sigma\in\mathscr{F}_q$,$|D_t(UV)|\leqslant|D_t(U\sigma V)|$。

证明 要证明此命题只需证明对 $\forall s\in D_t(UV)$ 存在单射 $\mathrm{Map}(s)$,使得 $\mathrm{Map}(s)\in D_t(U\sigma V)$。设 $s=uv$,其中 u 是 U 的子序列并且是 s 的最长前缀,定义 $\mathrm{Map}(s)=u\sigma v$。

下面证明 $\mathrm{Map}(s)$ 为单射,即证当 $s=s'$ 时,$\mathrm{Map}(s)=\mathrm{Map}(s')$。

设 $s=uv,s'=u'v'\in D_t(UV)$,u,u' 都是 U 的子序列并且分别为 s,s' 的最长前缀。分以下2种情况考虑:

情况一:当 $u=u'$ 时,$v\neq v'$ 一定成立,则 $u\sigma v\neq u'\sigma v'$。

情况二:当 $u\neq u'$ 时,利用反证法证明。假设 $\mathrm{Map}(s)=\mathrm{Map}(s')$,即 $u\sigma v=u'\sigma v'$。不妨设 $|u|>|u'|$,令 $u=u'x$, $v'=y'v$, x 非空,故 $u'x\sigma v=u'\sigma y'v$,所以

$x\sigma = \sigma y'$。因此可以将 x 表示为 $x = \sigma w$,y' 表示为 $y' = z'\sigma$。从而 $\sigma w\sigma = \sigma z'\sigma$,其中 z' 是非空的,否则 $s = s'$。可见,$u = u'x = u'\sigma w$,$s' = u'v' = u'y'v = u'z'\sigma v = u'w\sigma v$。因为 $u'z' = u'w$ 是 u 的子序列,所以 $u'z'$ 也是 U 的子序列,且 $u'z'$ 是 s' 的前缀,这与 u' 的最大性矛盾。所以 $\mathrm{Map}(s) = \mathrm{Map}(s')$。综上所述 $\mathrm{Map}(s)$ 是单射,又因为 $\mathrm{Map}(s) \in D_t(U\sigma V)$,所以式(2)成立。

推论1[13]　$X \in \mathscr{F}_q^m$,$r(X) = n(n \leqslant m)$,则存在 $Y \in \mathscr{F}_q^n$,$r(Y) = n$,使得

$$|D_t(Y)| \leqslant \left|D_t(X)\right|。$$

证明　序列 X 可通过序列 Y 插入数字得到,重复应用命题3即可得结论正确。例如,X = 11122332,令 Y = 1232,满足 $r(X) = r(Y)$ 且 $|D_t(Y)| \leqslant |D_t(X)|$。

命题4[13]　$Y \in \mathscr{F}_q^n$,$r(Y) = n$,则

$$|D_t(Y)| \geqslant D_2(n,t)。$$

证明　对 n 和 $n - t$ 进行归纳假设,当 $n \leqslant 2$ 或 $n - t \leqslant 2$ 时,引理成立。令 $Y = abZ$,因为 $r(Y) = n$,所以 $a \neq b$,则

$$|D_t(Y)| = |D_t^{(a)}(Y)| + |D_t^{(b)}(Y)| + \sum_{\sigma \neq a,b} |D_t^{(\sigma)}(Y)| \geqslant$$

$$|D_t^{(a)}(Y)| + |D_t^{(b)}(Y)| = |aD_t(bZ)| + |bD_{t-1}(Z)|,$$

利用归纳假设和式(1)可得

$$|D_t(Y)| \geqslant |D_t(bZ)| + |D_{t-1}(Z)| \geqslant D_2(n-1,t) + D_2(n-2,t-1) = D_2(n,t)。$$

由命题4和推论1可推出对任意 X,$|D_t(X)| \geqslant D_2(r(X),t)$,即为文献[13]中的下界。

1.2　主要结论

本文的主要目的是讨论当 X 的前3个块的元素互不相同时 $|D_t(X)|$ 的下界。根据以上命题,得到本文的主要结论:

定理1　$Y \in \mathscr{F}_q^n$,$q \geqslant 3$,$n \geqslant 3$,$r(Y) = n$ 且 Y 前3个块的元素互不相同,则

$$|D_t(Y)| \geqslant D_2(n,t) + D_2(n-3,t-2)。$$

证明　令 $Y = abcZ$,$a \neq b \neq c$,则

$$|D_t(Y)| = |D_t^{(a)}(Y)| + |D_t^{(b)}(Y)| + |D_t^{(c)}(Y)| + \sum_{\sigma \neq a,b,c} |D_t^{(\sigma)}(Y)| \geqslant$$

$$|D_t^{(a)}(Y)| + |D_t^{(b)}(Y)| + |D_t^{(c)}(Y)| = |D_t(bcZ)| + |D_{t-1}(cZ)| + |D_{t-2}(Z)|,$$

根据命题4可知对任意 $Z \in \mathscr{F}_q^{n-3}$ 都有 $|D_t(bcZ)| \geqslant D_2(n-1,t)$,$|D_{t-1}(cZ)| \geqslant D_2(n-2,t-1)$,$|D_{t-2}(Z)| \geqslant D_2(n-3,t-2)$,所以

$$|D_t(Y)| \geqslant D_2(n-1,t) + D_2(n-2,t-1) + D_2(n-3,t-2) = D_2(n,t) + D_2(n-3,t-2)。$$

定理2　$X \in \mathscr{F}_q^n$,当 $q \geqslant 3$,$n \geqslant 3$,且 X 前3个块的元素互不相同时,

$$|D_t(X)| \geqslant \sum_{i=0}^{t} \binom{r(X)-t}{i} + \sum_{i=0}^{t-2} \binom{r(X)-t-1}{i}。\quad (3)$$

证明　通过式(1),推论1和定理1可推得对任意 X 存在长度为 $r(X)$ 且 $r(Y) = r(X)$ 的序列 Y 使得

$$|D_t(X)| \geqslant |D_t(Y)| \geqslant D_2(r(X),t) + D_2(r(X)-3,t-2) = \sum_{i=0}^{t} \binom{r(X)-t}{i} + \sum_{i=0}^{t-2} \binom{r(X)-t-1}{i}。$$

所以不等式(3)成立。

对比定理2中 $|D_t(X)|$ 的下界与 Hirschberg 和 Regnier[13]的下界,因为

$$\sum_{i=0}^{t-2} \binom{r(X)-t-1}{i} \geqslant 0,$$

所以当 $q \geqslant 3$ 时定理2的下界比文献[13]中的下界更紧。

2　结束语

本文研究了估计序列 X 删除 t 个数后产生不同序列的数量问题,得到了当 X 的前3个块的元素互不相同时 $|D_t(X)|$ 的一个下界。事实上,未来在寻找 $|D_t(X)|$ 的简单表达式或更紧的界方面,还需要做出很多研究。

参 考 文 献

[1] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[2] KULKARNI A, KIYAVASH N. Nonasymptotic upper bounds for deletion correcting codes [J]. IEEE Transactions on Information Theory, 2013, 59(8): 5115-5130.

[3] ABROSHAN M. Codes for synchronization in channels and sources with edits[D]. Los Angeles: University of Cambridge, 2019.

[4] SWART T G, FERREIRA H C. A note on double insertion/deletion correcting codes[J]. IEEE Transactions

on Information Theory, 2003, 49(1): 269-273.
[5] MERCIER H, BHARGAVA V K, TAROKH V. A survey of error-correcting codes for channels with symbol synchronization errors [J]. IEEE Communications Surveys & Tutorials, 2010, 12(1): 87-96.
[6] CULLINA D, KIYAVASH N. An improvement to Levenshtein's upper bound on the cardinality of deletion correcting codes[J]. IEEE Transactions on Information Theory, 2014, 60(7): 3862-3870.
[7] LEVENSHTEIN V I. Efficient reconstruction of sequences[J]. IEEE Transactions on Information Theory, 2001, 47(1): 2-22.
[8] SALA F, GABRYS R, SCHOENY C, et al. Exact reconstruction from insertions in synchronization codes [J]. IEEE Transactions on Information Theory, 2017, 63(4): 2428-2445.
[9] CULLINA D, KIYAVASH N, KULKARNI A A. Restricted composition deletion correcting codes [J]. IEEE Transactions on Information Theory, 2016, 62 (9): 4819-4832.
[10] LENZ A, LIU Y, RASHTCHIAN C, et al. Coding for efficient DNA synthesis[J/OL].(2020-06-28)[2022-03-01]. DOI:10.1109/ISIT44484.2020.9174272.
[11] LEVENSHTEIN V I. Efficient reconstruction of sequences from their subsequences or supersequences [J]. Journal of Combinatorial Theory, Series A, 2001, 93(2): 310-332.
[12] GABRYS R, YAAKOBI E. Sequence reconstruction over the deletion channel [J]. IEEE Transactions on Information Theory, 2018, 64(4): 2924-2931.
[13] HIRSCHBERG D S, REGNIER M. Tight bounds on the number of string subsequences[J]. Journal of Discrete Algorithms, 2000, 1(1): 123-132.
[14] LIRON Y, LANBGERG M. A characterization of the number of subsequences obtained via the deletion channel[J]. IEEE Transactions on Information Theory, 2015, 61(5): 2300-2312.
[15] FLAXMAN A, HARROW A W, SORKIN G B. Strings with maximally many distinct subsequences and substrings[J]. The Electronic Journal of Combinatorics, 2004, 11(1): R8.
[16] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions, and reversals [J]. Doklady Akademii Nauk Sssr, 1965, 163: 845-848.
[17] LEVENSHTEIN V I. Perfect codes in the metric of deletions and insertions[J]. Diskretnaya Matematika, 1991, 3(1): 3-20.
[18] CALABI L, HARTNETT W E. Some general results of coding theory with applications to the study of codes for the correction of synchronization errors[J]. Information and Control, 1969, 15(3): 235-249.
[19] MERCIER H, KHABBAZIAN M, BHARGAVA V K. On the number of subsequences when deleting symbols from a string [J]. IEEE Transactions on Information Theory, 2008, 54(7): 3279-3285.
[20] HIRSCHBERG D S, REGNIER M. Bounds on the number of string subsequences [J]. Journal of Discrete Algorithms, 1999, 1(1): 115-122.

(责任编辑:马田田)