最优组合批处理码的单调性质及上下界

陈俊芳^①, 张素梅^②, 张更生^{①③*}

① 河北师范大学数学与信息科学学院, 石家庄 050024;

② 邯郸学院数理学院, 邯郸 056005;

③ 河北省计算数学与应用重点实验室, 石家庄 050024

E-mail: zjfwl@163.com, smzhang66@126.com, gshzhang@hebtu.edu.cn

收稿日期: 2014-04-09; 接受日期: 2014-11-19; * 通信作者

国家自然科学基金 (批准号: 11371121) 和河北省自然科学基金 (批准号: A2013205073) 资助项目

摘要 具有参数 n, k 和 m 的组合批处理码可以看作一个 n 元集以及它的 m 个子集 $B_1, B_2, \dots, B_m$ 组成的集合系统, 满足对于任意 k 个元素都能通过从每个子集中至多取一 (可以一般化为 t) 个元素来取得. 一个优化问题是, 确定 m 个子集中元素总数 $|B_1| + |B_2| + \dots + |B_m|$ 的最小值 $N(n, k, m)$. 这种问题不仅具有理论意义, 而且有着重要的应用价值.

本文研究 $N(n, k, m)$ 的变化规律, 给出 $N(n, k, m)$ 的一个上下界, 当 $2 \leq k < m \leq n - 3$ 时, 如果 $m + 1 - k \geq \lceil \sqrt{k+1} \rceil$,

$$(n - m)k + m \geq N(n, k, m) \geq 2n - m + k - 6 + \lceil 2\sqrt{k+1} \rceil;$$

如果 $m + 1 - k < \lceil \sqrt{k+1} \rceil$,

$$(n - m)k + m \geq N(n, k, m) \geq 2n - 6 + \left\lceil 1 + \frac{k+1}{m-k+1} \right\rceil.$$

然后确定 $N(m+3, 4, m) = m+9$ (当 $m \geq 6$ 时), $N(8, 4, 5) = 15$, 得到的结果部分解决了 Paterson 等人提出的未解决问题.

关键词 组合批处理码 最优 CBC 单调性 上下界

MSC (2010) 主题分类 05B30, 94B25

1 引言

2004 年, 文献 [1] 首次提出了批处理码 (batch codes) 的概念. 在这里, 我们用一个具体的问题来说明它, 假设有一个含 n 项数据的大数据库, 现在要把这 n 项数据分配到 m 个装置中 (不同的装置可以放有相同数据, 这个过程可看作是编码), 此处的装置可以是一种物理装置, 如服务器或磁盘, 也完全可以虚拟的. 数据分配完毕之后, 当用户需要这 n 项数据中的任意 k 项时, 希望可以通过读取装置中存储的数据来恢复它们 (这个过程可看作是译码). 显然, 把这 n 项数据通过编码分配到 m 个装置中的方式有很多种, 但是为了合理起见, 分配方式应该满足如下条件: 首先使得这 m 个装置的最大载荷都能达到最小 (装置的载荷大小是根据从它里面读取的数据个数来测量的); 其次为了节约装置的存

英文引用格式: Chen J F, Zhang S M, Zhang G S. Optimal combinatorial batch code: Monotonicity, lower and upper bounds (in Chinese). Sci Sin Math, 2015, 45: 311–320, doi: 10.1360/N012014-00061

存储空间, 应使得这 m 个装置中的总存储量 N 尽可能的小. 基于上述条件的限制, 就应该合理的分配这 n 项数据. 于是, 寻找一种既能使每个装置的载荷尽可能达到平衡同时又能够缩小总存储量 N 的分配方式成为了一个问题. 把上述问题抽象化就形成了批处理码.

定义 1.1^[1] 设码 C 定义在字符集 Σ 上, 如果它满足下列条件:

- (i) 码 C 能把字符串 $x \in \Sigma^n$ 编码成 m 组字符串 $y_1, y_2, \dots, y_m \in \Sigma^*$ (也称作是服务器), 它的总长度为 N ;
- (ii) 对任意 k 个不同的索引指标 $i_1, i_2, \dots, i_k \in \{1, 2, \dots, n\}$ 来说, x 中对应的 $x_{i_1}, x_{i_2}, \dots, x_{i_k}$, 这 k 个位置上的信息都可以通过从每个服务器中至多读取 t 个信息来译出, 则称码 C 为一个 (n, N, k, m, t) 批处理码.

这个定义中的 m 组字符串 $y_1, y_2, \dots, y_m$ 相当于上述问题中的 m 个装置, 它们的总长度 N 相当于装置的总存储量, 参数 t 相当于每个装置的最大负荷. 研究批处理码的目的就是, 通过对数据进行合理的存储, 既能使得各个计算机的载荷平衡, 又能缩小计算机占用的总存储空间. 文献 [1] 介绍了几种构造批处理码的具体方法, 同时介绍了它在计算机方面的两类应用情形.

2009 年, Paterson 等人^[2] 在上述批处理码的基础上把每个装置中储存的数据简化成总数据的一个子集, 译码简化成从存储装置中直接读出数据. 于是, 构造这样的批处理码就成为一个纯组合的问题, 此时的批处理码称为组合批处理码 (combinatorial batch codes, CBC). 在给出 CBC 的定义之前, 首先介绍集合系统的定义.

定义 1.2^[2] 设 $(X, \mathcal{B})$ 是一个二元组, 如果它满足下列条件:

- (i) X 是一个有限集合;
- (ii) $\mathcal{B}$ 是 X 的一个子集族, 则称 $(X, \mathcal{B})$ 是一个集合系统, 其中 X 中的元素叫作点, $\mathcal{B}$ 中的元素叫作区组.

有了集合系统的定义之后, 下面给出 CBC 的定义.

定义 1.3^[2] 如果一个集合系统 $(X, \mathcal{B})$ 满足下列条件:

- (i) $|X| = n$;
- (ii) $|\mathcal{B}| = m$;
- (iii) $N = \sum_{B \in \mathcal{B}} |B|$;
- (iv) 对 X 的每个 k 元子集 $\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\} \subseteq X$, 都存在 B_i 的一个子集 $C_i \subseteq B_i$, 其中 $|C_i| \leq t$, $i = 1, \dots, m$, 使得

$$\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\} = \bigcup_{i=1}^m C_i,$$

则称它为一个 (n, N, k, m, t) -CBC.

在上述定义中, 如果 $t = 1$, 则说明在这样的 CBC 中只允许从每个服务器里至多读取一个元素, 我们把这样的 CBC 简记为 (n, N, k, m) -CBC. 对于 $t > 1$ 的情形, 文献 [3] 中有所研究. 本文与大多数文献一样, 只讨论 $t = 1$ 时的情形. 显然, 当 $t = 1$ 时, 我们可以假设 $n \geq m \geq k$.

由于在 (n, N, k, m) -CBC 中, N 表示的是所有区组中含有点的总数, 所以, 在给定 n, k 和 m 之后, 我们当然希望 N 的值越小越好. 对应到运用计算机存储数据这一实际应用中, 则是希望缩小计算机占用的存储空间, 以达到节约资源的目的. 于是就有了最优 CBC 的概念.

定义 1.4^[2] 对给定的 n, k 和 m 而言, 如果对所有的 (n, N', k, m) -CBC, 存在一个 (n, N, k, m) -CBC, 使得 $N \leq N'$, 则称此 (n, N, k, m) -CBC 是最优的, 并用 $N(n, k, m)$ 来表示这个最优值.

关于确定给定参数 n, k 和 m 的 CBC 的最优值的研究无疑是 CBC 研究的重要课题之一, 到目前为止, 这方面的研究有了许多进展. 表 1 给出主要结果: 表中第 1 和 2 行的结果来自文献 [1, 2], 第 3 和 5 行来自文献 [2], 第 4 行来自文献 [4, 5], 最后一行来自文献 [3, 6, 7]. 另外, 文献 [8–12] 给出了一些零散结果.

从表 1 中可以看出, 当 $n < \binom{m}{k-2}$ 时, $N(n, k, m)$ 的紧界问题还没有得到全面的解决. 文献 [6] 对于 $1 \leq n \leq (k-1)\binom{m}{k-1}$ 给出了下面的界. 设 $1 \leq c \leq k-1$ 是一个满足下面条件的最小整数,

$$n \leq \frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}},$$

则有

$$N(n, k, m) \geq nc - \left\lfloor \frac{(k-c)\left(\frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}} - n\right)}{m-k+1} \right\rfloor. \quad (1.7)$$

本文针对最优 CBC 的有关问题进行研究, 得到了如下的主要结论:

(i) 研究了当参数 n, k 和 m 变化时 $N(n, k, m)$ 的变化规律, 即所谓的单调性质,

$$N(n, k, m+1) \leq N(n, k, m) - 1,$$

$$N(n+1, k, m) \geq N(n, k, m) + 2,$$

$$N(n, k+1, m) \geq N(n, k, m),$$

其中 $2 \leq k \leq m < n$.

(ii) 得到了 $N(n, k, m)$ 的一个上下界.

当 $2 \leq k < m \leq n-3$ 时,

$$(n-m)k + m \geq N(n, k, m) \geq \begin{cases} 2n - m + k - 6 + \lceil 2\sqrt{k+1} \rceil, & \text{如果 } m+1-k \geq \lceil \sqrt{k+1} \rceil, \\ 2n - 6 + \left\lceil 1 + \frac{k+1}{m-k+1} \right\rceil, & \text{如果 } m+1-k < \lceil \sqrt{k+1} \rceil. \end{cases} \quad (1.8)$$

对于 $n-m=0, 1, 2$, $m=k$ 和 $k=1$ 的情形, $N(n, k, m)$ 的值都已确定, 故上述结论实际上覆盖了余下的所有情形. 在我们验证的几种情形中, (1.8) 的结果优于 (1.7).

(iii) 利用上述结果确定了 $N(m+3, 4, m)$ 的值,

表 1 $N(n, k, m)$ 的主要已有结果

n	k	m	$N(n, k, m)$	标号
n	k	$m = n$	n	(1.1)
n	k	$m = k$	$kn - k(k-1)$	(1.2)
$n = m+1$	k	m	$m+k$	(1.3)
$n = m+2$	k	m	$\begin{cases} m+k-2 + \lceil 2\sqrt{k+1} \rceil, & \text{如果 } m+1-k \geq \lceil \sqrt{k+1} \rceil, \\ 2m-2 + \left\lceil 1 + \frac{k+1}{m-k+1} \right\rceil, & \text{如果 } m+1-k < \lceil \sqrt{k+1} \rceil \end{cases}$	(1.4)
$n \geq (k-1)\binom{m}{k-1}$	k	m	$kn - (k-1)\binom{m}{k-1}$	(1.5)
$\binom{m}{k-2} \leq n \leq (k-1)\binom{m}{k-1}$	k	m	$n(k-1) - \left\lfloor \frac{(k-1)\binom{m}{k-1} - n}{m-k+1} \right\rfloor$	(1.6)

$$N(m+3, 4, m) = \begin{cases} 15, & m = 5, \\ m+9, & m \geq 6. \end{cases}$$

用同样的方法也可以确定 $N(m+3, 5, m)$ 和 $N(m+3, 6, m)$ 等值 (我们将另文讨论). 显然, 当 $k \geq 4$ 时, 对于 $m > k$, 恒有 $m+3 < \binom{m}{k-2}$ 成立. 所以, 本文的结果部分地解决了文献 [2] 中提出的未解决问题.

2 $N(n, k, m)$ 的单调性质

首先给出两个重要引理.

引理 2.1 设集合系统 $(X, \mathcal{B})$ 是一个 (n, N, k, m) -CBC, 当 $2 \leq k \leq m < n$ 时, 则在 X 中必存在一点 x 至少属于 $\mathcal{B}$ 中的两个区组.

证明 假设不存在这样的点 x , 即 X 中所有的点都分别只属于一个区组. 由于 $n > m$, 所以必存在在一个区组, 不妨设为 B_1 , B_1 中至少含有两个点. 设 $B_1 = \{x_1, x_2, \dots, x_q\}$ ($q \geq 2$). 现在, 我们取 B_1 中的 x_1, x_2 两点和 X 中其余 $n-2$ 个点中的任意 $k-2$ 个点. 对于这 k 个点, 因为 $x_1, x_2 \notin \bigcup_{i=2}^m B_i$, 所以, 这 k 个点便不能按照 CBC 定义中的规则被取出. 这与 $(X, \mathcal{B})$ 是一个 CBC 相矛盾, 说明假设不成立. 引理得证. $\square$

进一步地, 对于最优 CBC, 我们有下面的引理.

引理 2.2 设集合系统 $(X, \mathcal{B})$ 是一个最优的 (n, N, k, m) -CBC, 当 $2 \leq k \leq m < n$ 时, 则在 X 中有一点 x 至少属于两个以上的区组, 而且这些包含 x 的区组中都含有异于 x 的点.

证明 由于最优 CBC 一定是一个 CBC, 所以, 由引理 2.1 知, 这个 CBC 中必存在 X 中一点 x 至少属于两个区组. 我们不妨设它属于区组 $B_1, B_2, \dots, B_{m_1}$ ($m_1 \geq 2$). 下证区组 $B_1, B_2, \dots, B_{m_1}$ 中都至少含有两个点.

假设 $B_1, B_2, \dots, B_{m_1}$ 中存在一个区组只含点 x , 不妨设 $B_1 = \{x\}$, 下分两种情形来讨论:

(i) 若 $B_2, B_3, \dots, B_{m_1}$ 中存在一个区组, 这个区组中至少含有两个点, 不妨设为 B_2 . 首先可令 $B_2' = B_2 \setminus \{x\}$, 进而令 $\mathcal{B}' = \{B_1, B_2', \dots, B_{m_1}, \dots, B_m\}$, 此时 $(X, \mathcal{B}')$ 为一个新的集合系统, 其中 $|X| = n$, $|\mathcal{B}'| = m$, $\mathcal{B}'$ 中所有区组含有的点的总数为 $N-1$. 下面说明当 k 不变时这个集合系统仍是一个 CBC.

如果从 X 中取的一个 k 元子集含点 x , 则点 x 可以从 B_1 中取出, 又因为 $(X, \mathcal{B})$ 是一个 (n, N, k, m) -CBC, 所以, 其余的 $k-1$ 个点可以从余下的区组 $B_2, \dots, B_{m_1}, \dots, B_m$ 中取出, 而 $B_2' = B_2 \setminus \{x\}$, 则其余的 $k-1$ 个点可以从 $B_2', B_3, \dots, B_{m_1}, \dots, B_m$ 中取出. 如果这个 k 元子集不含点 x , 显然, 这 k 个点可以从 $B_1, B_2', B_3, \dots, B_{m_1}, \dots, B_m$ 这 m 个区组中取出. 因此, $(X, \mathcal{B}')$ 仍为一个 CBC, 即存在一个 $(n, N-1, k, m)$ -CBC, 这与 (n, N, k, m) -CBC 是一个最优 CBC 相矛盾.

(ii) 若 $B_2, B_3, \dots, B_{m_1}$ 中都只含点 x , 即 $B_2 = B_3 = \dots = B_{m_1} = \{x\}$, 则 X 中其余的 $n-1$ 个点必须分配到其余的 $m-m_1$ 个区组中去, 不妨设这 $m-m_1$ 个区组分别为 $B_{m_1+1}, B_{m_1+2}, \dots, B_m$. 由于 $n-1 > m-m_1$, 所以, $B_{m_1+1}, B_{m_1+2}, \dots, B_m$ 中必存在一个区组, 这个区组中至少含有两个点, 不妨设为 B_{m_1+1} . 假设 $y \in B_{m_1+1}$, 可令 $B_{m_1+1}' = B_{m_1+1} \setminus \{y\}$, $B_1' = \{y\}$, 进而令 $\mathcal{B}' = \{B_1', B_2, \dots, B_{m_1}, B_{m_1+1}', B_{m_1+2}, \dots, B_m\}$, 此时 $(X, \mathcal{B}')$ 为一个新的集合系统, 其中 $|X| = n$, $|\mathcal{B}'| = m$, $\mathcal{B}'$ 中所有区组含有的点的总数为 $N-1$. 因为 $(X, \mathcal{B})$ 是一个 (n, N, k, m) -CBC, 所以, 当 k 不变时, 显然, $(X, \mathcal{B}')$ 仍为一个 CBC, 即存在一个 $(n, N-1, k, m)$ -CBC, 这与 (n, N, k, m) -CBC 是一个最优 CBC 相矛盾.

综合 (i) 和 (ii) 两种情形可知, 原假设不成立, 所以, x 所在的 m_1 个区组中都至少含有两个点. $\square$

应用引理 2.2, 我们可以得到当 $2 \leq k \leq m < n$ 时, 最优 CBC 分别关于 m 和 n 的两个单调性质.

定理 2.3 $N(n, k, m) \geq N(n, k, m+1) + 1$.

证明 设 $N(n, k, m) = t$, $N(n, k, m+1) = p$, 则存在一个 (n, t, k, m) -CBC, 设它是一个 $(X, \mathcal{B})$ 集合系统, 则由引理 2.2 知, 这个 CBC 中必存在 X 中一点 x 属于 m_1 个区组 $B_1, B_2, \dots, B_{m_1}$ ($m_1 \geq 2$), 且这 m_1 个区组中都至少含有两个点. 当 n 和 k 均不变, 区组个数变为 $m+1$ 时, 我们可构造这样一个 CBC:

设 $\mathcal{B} = \{B_1, B_2, \dots, B_{m_1}, B_{m_1+1}, B_{m_1+2}, \dots, B_m\}$, 令 $B_i' = B_i \setminus \{x\}$ ($i = 1, 2, \dots, m_1$), $B_{m_1+1} = \{x\}$, $\mathcal{B}' = \{B_1', B_2', \dots, B_{m_1}', B_{m_1+1}, B_{m_1+2}, \dots, B_m, B_{m+1}\}$. 于是, $(X, \mathcal{B}')$ 为一个新的集合系统, 其中 $|X| = n$, $|\mathcal{B}'| = m+1$, $\mathcal{B}'$ 中所有区组含有的点的总数为 $t - m_1 + 1$. 显然, $(X, \mathcal{B}')$ 仍为一个 CBC, 即存在一个 $(n, t - m_1 + 1, k, m+1)$ -CBC. 因为 $N(n, k, m+1) = p$, 所以, $p \leq t - m_1 + 1$, 即 $t \geq p + m_1 - 1$. 又知, $m_1 \geq 2$, 因此, $t \geq p + m_1 - 1 \geq p + 2 - 1 = p + 1$, 即 $N(n, k, m) \geq N(n, k, m+1) + 1$. $\square$

定理 2.4 $N(n+1, k, m) \geq N(n, k, m) + 2$.

证明 设 $N(n, k, m) = t$, $N(n+1, k, m) = t'$, 则存在一个 $(n+1, t', k, m)$ -CBC. 设它是一个 $(X, \mathcal{B})$ 集合系统, 则由引理 2.2 知, 这个 CBC 中必存在 X 中一点 x 属于 m_1 个区组 $B_1, B_2, \dots, B_{m_1}$ ($m_1 \geq 2$), 且这 m_1 个区组中都至少含有两个点. 当 k 和 m 均不变, 点的个数变为 n 时, 我们可构造这样一个 CBC:

设 $\mathcal{B} = \{B_1, B_2, \dots, B_{m_1}, B_{m_1+1}, \dots, B_m\}$. 首先可令 $X' = X \setminus \{x\}$, $B_i' = B_i \setminus \{x\}$ ($i = 1, 2, \dots, m_1$), 进而令 $\mathcal{B}' = \{B_1', B_2', \dots, B_{m_1}', B_{m_1+1}, \dots, B_m\}$, 此时, $(X', \mathcal{B}')$ 为一个新的集合系统, 其中 $|X'| = n$, $|\mathcal{B}'| = m$, $\mathcal{B}'$ 中所有区组含有的点的总数为 $t' - m_1$. 显然, $(X', \mathcal{B}')$ 仍为一个 CBC, 即存在一个 $(n, t' - m_1, k, m)$ -CBC. 因为 $N(n, k, m) = t$, 所以, $t \leq t' - m_1$, 即 $t' \geq t + m_1$. 又知, $m_1 \geq 2$, 因此, $t' \geq t + m_1 \geq t + 2$, 即 $N(n+1, k, m) \geq N(n, k, m) + 2$. $\square$

文献 [5, 推论 2] 给出了一个当 m 和 n 同时变化时最优 CBC 的单调性质.

定理 2.5 [5] $N(n+1, k, m+1) \leq N(n, k, m) + 1$, 其中 $1 \leq k \leq m \leq n$.

最后, 我们给出一个最优 CBC 关于 k 的单调性质.

定理 2.6 $N(n, k, m) \leq N(n, k+1, m)$, 其中 $1 \leq k \leq m \leq n$.

证明 设 $N(n, k+1, m) = t$, 则存在一个 $(n, t, k+1, m)$ -CBC, 设它是一个 $(X, \mathcal{B})$ 集合系统, 其中 $X = \{x_1, x_2, \dots, x_n\}$, $\mathcal{B} = \{B_1, B_2, \dots, B_m\}$. 如果取 X 的任意一个 $k+1$ 元子集 $\{x_{i_1}, x_{i_2}, \dots, x_{i_{k+1}}\}$, 则存在 B_i 的一个子集 $C_i \subseteq B_i$, 其中 $|C_i| \leq 1$, $i = 1, 2, \dots, m$, 使得 $\{x_{i_1}, x_{i_2}, \dots, x_{i_{k+1}}\} = \bigcup_{i=1}^m C_i$. 那么, 当取 X 的任意一个 k 元子集 $\{x_{j_1}, x_{j_2}, \dots, x_{j_k}\}$ 时, 因为 $k < k+1$, 所以, 必能找到 $C_i' \subseteq B_i$, 其中 $|C_i'| \leq 1$, $i = 1, 2, \dots, m$, 使得

$$\{x_{j_1}, x_{j_2}, \dots, x_{j_k}\} = \bigcup_{i=1}^m C_i'.$$

也就是说, 原先的 $(n, t, k+1, m)$ -CBC 也是一个 (n, t, k, m) -CBC, 即存在一个 (n, t, k, m) -CBC, 所以, $N(n, k, m) \leq t$, 即 $N(n, k, m) \leq N(n, k+1, m)$. $\square$

3 $N(n, k, m)$ 的界

利用第 2 节的结论, 我们可以得到下面的界.

定理 3.1 当 $2 \leq k < m \leq n - 3$ 时,

$$(n-m)k + m \geq N(n, k, m) \geq \begin{cases} 2n - m + k - 6 + \lceil 2\sqrt{k+1} \rceil, & \text{如果 } m+1-k \geq \lceil \sqrt{k+1} \rceil, \\ 2n - 6 + \left\lceil 1 + \frac{k+1}{m-k+1} \right\rceil, & \text{如果 } m+1-k < \lceil \sqrt{k+1} \rceil. \end{cases}$$

证明 一方面根据定理 2.4 知,

$$\begin{aligned} N(n, k, m) &\geq N(n-1, k, m) + 2 \\ &\geq N(n-2, k, m) + 4 \\ &\vdots \\ &\geq N(m+2, k, m) + 2(n - (m+2)). \end{aligned}$$

再利用 (1.4) 即得到 $N(n, k, m)$ 的下界.

另一方面根据定理 2.5 可得

$$\begin{aligned} N(n, k, m) &\leq N(n-1, k, m-1) + 1 \\ &\leq N(n-2, k, m-2) + 2 \\ &\vdots \\ &\leq N(n-(m-k), k, k) + (m-k). \end{aligned}$$

由 (1.2) 得 $N(n-(m-k), k, k) = k(n-(m-k)) - k(k-1)$. 于是得到 $N(n, k, m)$ 的上界. □

利用上面的定理可以给出一些具体最优 CBC 的上下界,

$$\begin{aligned} m+9 &\leq N(m+3, 4, m) \leq m+12, \quad m \geq 5, \\ m+10 &\leq N(m+3, 5, m) \leq m+15, \quad m \geq 6, \\ m+12 &\leq N(m+3, 6, m) \leq m+18, \quad m \geq 7. \end{aligned}$$

我们没有全面地比较 (1.7) 与 (1.8) 的优劣, 只给出一个例子. 对 $N(9, 4, 6)$ 而言, 用 (1.7) 可以得到其下界为 14, 用 (1.8) 可以得到其下界为 15. 实际上, 在第 4 节, 我们证明了 15 是其紧界. 不仅如此, 我们还证明了 $m+9$ 是 $N(m+3, 4, m)$ ($m \geq 6$) 的紧界.

4 确定 $N(m+3, 4, m)$ ($m \geq 5$) 的值

对于 $N(m+3, k, m)$ 来说, 当 $k=1, 2, 3$ 时可以由已有的结论推出. 为了清楚起见, 我们把它列示如下:

- (i) $k=1$ 时, 是平凡的, $N(m+3, 1, m) = m+3$;
- (ii) $k=2$ 时, 由 (1.5) 可知, $N(m+3, 2, m) = m+6$;
- (iii) $k=3$ 时, 由 (1.5) 和 (1.6) 可知,

$$N(m+3, 3, m) = \begin{cases} 12, & m=4, \\ m+7, & m \geq 5. \end{cases}$$

本节的主要任务就是利用上面得到的结论确定 $N(m+3, 4, m)$ 的值, 这个过程我们分为 3 步.

第 1 步 确定 $N(9, 4, 6)$ 的值.

在这个过程中, 我们需要文献 [2] 中给出的一个利用所谓的对偶集合系统判定一个 (n, N, k, m) -CBC 的存在的非常有用的方法 (这个定理在文献 [1] 中也被用不同的术语所描述).

定理 4.1 [2] 如果一个集合系统 $(Y, \mathcal{A})$ 满足下列条件:

- (1) $|Y| = m$;
- (2) $|\mathcal{A}| = n$;
- (3) $\sum_{A \in \mathcal{A}} |A| = N$;
- (4) 对于 $1 \leq i \leq k$, 任意 i 个区组的并集至少包含 i 个点,

则 $(Y, \mathcal{A})$ 是一个 (n, N, k, m) -CBC 的对偶集合系统.

我们先来看一下 $N(9, 4, 6)$ 的取值范围. 由定理 2.4 可知,

$$N(9, 4, 6) \geq N(8, 4, 6) + 2.$$

而由 (1.4) 可得 $N(8, 4, 6) = 6 + 4 - 2 + \lceil 2\sqrt{5} \rceil = 13$, 所以, $N(9, 4, 6) \geq 15$.

定理 4.2 我们构造一个如下的集合系统 $(Y, \mathcal{A})$, 其中 $Y = \{1, 2, 3, 4, 5, 6\}$,

$$\mathcal{A} = \{\{1\}, \{2\}, \{3\}, \{1, 4\}, \{2, 5\}, \{3, 6\}, \{4, 6\}, \{4, 5\}, \{5, 6\}\},$$

则 $(Y, \mathcal{A})$ 是一个 $(9, 15, 4, 6)$ -CBC 的对偶集合系统.

证明 不妨把 $\mathcal{A}$ 中的区组按照给定的排列顺序分别命名为 $B_1, B_2, \dots, B_9$. 根据定理 4.1, 只需验证 $(Y, \mathcal{A})$ 满足以下条件即可,

- (1) $|Y| = 6$;
- (2) $|\mathcal{A}| = 9$;
- (3) $\sum_{i=1}^9 |B_i| = 15$;
- (4) 对于 $1 \leq i \leq 4$, 每 i 个区组的并集至少包含 i 个点.

(1)–(3) 显然成立, 关键是证明 (4). 设 $\{a, b, c\} = \{1, 2, 3\}$, $\{x, y, z\} = \{4, 5, 6\}$. 可以看出, $\mathcal{A}$ 由下面的三类区组构成.

第一类是集合 $\{1, 2, 3\}$ 的三个 1-子集: $\mathcal{B}_1 = \{\{a\}, \{b\}, \{c\}\}$; 第二类是集合 $\{4, 5, 6\}$ 的三个 2-子集: $\mathcal{B}_2 = \{\{x, y\}, \{y, z\}, \{x, z\}\}$; 第三类是如下的三个 2-子集: $\mathcal{B}_3 = \{\{a, x\} \mid a \in \{1, 2, 3\}, x = a + 3\}$.

(i) 当 $i = 1$ 时, 显然成立.

(ii) 对于 $i = 2$ 时, 假设 $\mathcal{A}$ 中存在两个区组的并至多包含一个点. 但是显然, 一个点不可能构成 $\mathcal{A}$ 中的两个区组, 这就说明了 $\mathcal{A}$ 中任意两个区组的并不可能至多包含一个点. 也就是说 $i = 2$ 时结论成立.

(iii) 对于 $i = 3$ 时, 假设 $\mathcal{A}$ 中存在三个区组的并至多包含两个点. 那么, 这两个点的构成无非是下列三种情形: 一是 a 和 b , 这种情形下它们至多构成 $\mathcal{A}$ 中的两个区组; 二是 x 和 y , 这种情形下它们至多构成 $\mathcal{A}$ 中的一个区组; 三是 a 和 x , 这种情形下它们也是至多构成 $\mathcal{A}$ 中的一个区组. 这就说明了 $\mathcal{A}$ 中任意三个区组的并不可能至多包含两个点. 也就是说 $i = 3$ 时结论成立.

(iv) 对于 $i = 4$ 时, 我们假设 $\mathcal{A}$ 中存在 4 个区组的并至多包含三个点. 那么, 这三个点的构成有如下 4 种情形:

(a) 三个点分别为 a, b, c . 此时它们至多构成 $\mathcal{A}$ 中的三个区组 $\{a\}, \{b\}, \{c\}$;

(b) 三个点分别为 a, b, x . 此时它们至多构成 $\mathcal{A}$ 中的三个区组 $\{a\}, \{b\}, \{a, x\}$ 或者 $\{a\}, \{b\}, \{b, x\}$;

(c) 三个点分别为 a, x, y . 此时它们至多构成 $\mathcal{A}$ 中的三个区组 $\{a\}, \{a, x\}, \{x, y\}$ 或者 $\{a\}, \{a, y\}, \{x, y\}$;

(d) 三个点分别为 x, y, z . 此时它们至多构成 $\mathcal{A}$ 中的三个区组 $\{x, y\}, \{y, z\}, \{x, z\}$.

综合上面的 4 种情形说明 $\mathcal{A}$ 中任意 4 个区组的并不可能至多包含三个点. 也就是说当 $i = 4$ 时结论成立.

因此, $(Y, \mathcal{A})$ 满足 (4), 所以, 它是一个 $(9, 15, 4, 6)$ -CBC 的对偶集合系统. $\square$

于是, $N(9, 4, 6) = 15$.

第 2 步 确定 $N(m+3, 4, m)$ ($m \geq 7$) 的值.

一方面根据定理 2.4 知,

$$N(m+3, 4, m) \geq N(m+2, 4, m) + 2.$$

又当 $m \geq 7, k = 4$ 时,

$$m+1-k = m-3 \geq \lceil \sqrt{4+1} \rceil.$$

所以, 由 (1.4) 可得

$$N(m+2, 4, m) = m+4-2 + \lceil 2\sqrt{4+1} \rceil = m+7.$$

因此,

$$N(m+3, 4, m) \geq m+9.$$

另一方面, 根据定理 2.5 可得

$$\begin{aligned} N(m+3, 4, m) &\leq N(m+2, 4, m-1) + 1 \\ &\leq N(m+1, 4, m-2) + 2 \\ &\vdots \\ &\leq N(9, 4, 6) + (m-6) \\ &= m+9. \end{aligned}$$

所以,

$$N(m+3, 4, m) = m+9, \quad m \geq 7.$$

第 3 步 确定余下的特殊情形 $N(8, 4, 5)$ 的值.

先来看一下 $N(8, 4, 5)$ 的取值范围. 一方面, 利用定理 2.4 可得 $N(8, 4, 5) \geq N(7, 4, 5) + 2$. 根据 (1.4) 知, $N(7, 4, 5) = 12$, 所以, $N(8, 4, 5) \geq 14$. 另一方面, 根据文献 [2, 定理 2.4] 所提供的构造方法可以构造出一个 $(8, 15, 4, 5)$ -CBC, 所以, $N(8, 4, 5) \leq 15$. 于是有 $14 \leq N(8, 4, 5) \leq 15$. 下面证明 $N(8, 4, 5)$ 不可能是 14.

假设 $N(8, 4, 5) = 14$, 则存在一个 $(8, 14, 4, 5)$ -CBC, 设它是一个 $(X, \mathcal{B})$ 集合系统. 不妨设 $X = \{1, 2, \dots, 8\}$, $\mathcal{B} = \{B_1, B_2, B_3, B_4, B_5\}$. 首先, 这个 CBC 中没有单点构成的区组. 这是因为根据 (1.2) 知, $N(7, 4, 4) = 4 \times 7 - 4 \times (4-1) = 16 > 13$ 的缘故. 其次, 这个 CBC 中至少存在一点 x 属于三个区组. 我们用反证法来证明.

如果 $1, 2, \dots, 8$ 都至多属于两个区组, 因为 $\mathcal{B}$ 中所有区组含有的点的总数为 14, 通过计算可知, 有两个点分别只属于一个区组, 不妨设为 1, 2, 而其他的点 3, 4, 5, 6, 7, 8 都分别属于两个区组.

首先 1, 2 两点不能属于同一个区组, 因为若它们属于同一个区组, 由于其余的 4 个区组中都不含 1, 2 两点, 则当取的 4 个点中含 1, 2 两点时, 这两点便不能同时取出. 所以, 1, 2 两点属于不同的区组, 不妨设 $1 \in B_1, 2 \in B_2$, 下面分情形来讨论:

(1) 假设 $|B_1| = |B_2| = 2$, 如果 $B_1 = \{1, x\}, B_2 = \{2, x\}$, 由于 x 不可能再出现于其他区组, 这样 1, 2, x 不能按规则同时取出. 如果 $B_1 = \{1, x\}, B_2 = \{2, y\}$, 不妨设 $B_1 = \{1, 3\}, B_2 = \{2, 4\}$. 这样 5, 6, 7, 8 点只能分配到 B_3, B_4, B_5 中, 这 4 个点不能同时按规则取出, 故假设不成立.

(2) 假设 $|B_1| = 3, |B_2| = 2$, 首先假如 $B_1 = \{1, x, y\}, B_2 = \{2, x\}$, 显然, 此时 1, 2, x 不能同时取出. 其次, 假如 $B_1 = \{1, x, y\}, B_2 = \{2, z\}$, 不妨设 $B_1 = \{1, 3, 5\}, B_2 = \{2, 4\}$, 则 3, 4, 5 中任意两点都不属于 B_3, B_4, B_5 中的同一个区组, 即它们恰分别属于 B_3, B_4, B_5 . 可令 $3 \in B_3, 4 \in B_4, 5 \in B_5$. 此时, 为了将 $\{1, 3, 5, 6\}, \{1, 3, 5, 7\}, \{1, 3, 5, 8\}$ 顺利取出, 必有 $\{6, 7, 8\} \subset B_4$, 由于它们必须出现在两个区组中, 这样的话, 6, 7, 8 还需在 B_3 或者 B_5 中各出现一次. 于是, 它们中至少有两个要同时属于 B_3 或者 B_5 , 不妨假设 6, 7 同时属于 B_3 或者 B_5 , 但二者都不可行, 如果 6, 7 属于 B_3 , 则 1, 3, 6, 7 不能按规则同时取出, 如果 6, 7 属于 B_5 , 则 1, 5, 6, 7 不能按规则同时取出, 故假设不成立.

(3) 假设 $|B_1| \geq 4, |B_2| = 2$, 易得 $B_1 \cap B_2 = \emptyset$. 令 $B = (B_1 \cup B_2) \setminus \{1, 2\}$, 则 $|B| \geq 4$. 由于余下了三个区组, 把 B 中的点分配到这三个区组中去的话, 则必有两点出现在同一个区组中. 此时这两点和 1, 2 不能按规则同时取出.

综上可知, 这个 CBC 中必满足 X 中一点 x 至少属于三个区组, 而且这些区组都不是单点集. 假设它属于 m 个区组 ($m \geq 3$). 如果令 $X' = X \setminus \{x\}, B_i' = B_i \setminus \{x\}$ ($i = 1, 2, 3, 4, 5$), 进而令 $B' = \{B_1', B_2', B_3', B_4', B_5'\}$, 则当 $k = 4$ 时, 可验证集合系统 (X', B') 是一个 $(7, 14 - m, 4, 5)$ -CBC. 因为 $m \geq 3$, 所以, $14 - m \leq 11$, 进而有 $N(7, 4, 5) \leq 14 - m \leq 11$, 这与 $N(7, 4, 5) = 12$ 矛盾, 所以, $N(8, 4, 5) \neq 14$, 于是, $N(8, 4, 5) = 15$.

综合以上讨论, 有下面的定理.

定理 4.3

$$N(m+3, 4, m) = \begin{cases} 15, & \text{如果 } m = 5, \\ m+9, & \text{如果 } m \geq 6. \end{cases}$$

致谢 审稿人给与了极具价值的修改建议, 在此表示深深的谢意.

参考文献

- 1 Ishai Y, Kushilevitz E, Ostrovsky R, et al. Batch codes and their applications. In: Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing STOC'04. New York: ACM Press, 2004, 36: 262–271
- 2 Patterson M B, Stinson D R, Wei R. Combinatorial batch codes. Adv Math Commun, 2009, 3: 13–27
- 3 Ruj S, Roy B. More on combinatorial batch codes. ArXiv:0809.3357v1, 2008
- 4 Brualdi R A, Kiernan K P, Meyer S A, et al. Combinatorial batch codes and transversal matroids. Adv Math Commun, 2010, 4: 419–431
- 5 Bujtás C, Tuza Z. Optimal combinatorial batch codes derived from dual systems. Miskolc Math Notes, 2011, 12: 11–23
- 6 Bhattacharya S, Ruj S, Roy B. Combinatorial batch codes: A lower bound and optimal constructions. Adv Math Commun, 2012, 6: 165–174
- 7 Bujtás C, Tuza Z. Optimal batch codes: Many items or low retrieval requirement. Adv Math Commun, 2011, 5: 529–541

- 8 Bujtás C, Tuza Z. Combinatorial batch codes: Extremal problems under Hall-type conditions. *Electron Notes Discrete Math*, 2011, 38: 201–206
- 9 Bujtás C, Tuza Z. Relaxations of Halls condition: Optimal batch codes with multiple queries. *Appl Anal Discrete Math*, 2012, 6: 72–81
- 10 Bujtás C, Tuza Z. Turán numbers and batch codes. *ArXiv:1309.6506v1*, 2013
- 11 Balachandran N, Bhattacharya S. On an extremal hypergraph problem related to combinatorial batch codes. *Discrete Appl Math*, 2014, 162: 373–380
- 12 Silberstein N, Gál A. Optimal combinatorial batch codes based on block designs. *Des Codes Cryptogr*, doi: 10.1007/s10623-014-0007-9, 2014

Optimal combinatorial batch code: Monotonicity, lower and upper bounds

CHEN JunFang, ZHANG SuMei & ZHANG GengSheng

Abstract A combinatorial batch code with parameters n, k, m is a system consisting of m subsets $B_1, B_2, \dots, B_m$ of an n -element set such that any k elements can be retrieved by reading at most one (or in general, t) elements from each subset. An optimization problem is to determine $N(n, k, m)$, the minimum of $|B_1| + |B_2| + \dots + |B_m|$. This problem has both theoretical interests and strong practical motivation.

In this article, we study the function $N(n, m, k)$ and give a lower and upper bounds for $N(n, m, k)$: for $2 \leq k < m \leq n - 3$, if $m + 1 - k \geq \lceil \sqrt{k+1} \rceil$ then

$$(n - m)k + m \geq N(n, k, m) \geq 2n - m + k - 6 + \lceil 2\sqrt{k+1} \rceil;$$

if $m + 1 - k < \lceil \sqrt{k+1} \rceil$ then

$$(n - m)k + m \geq N(n, k, m) \geq 2n - 6 + \left\lceil 1 + \frac{k+1}{m-k+1} \right\rceil.$$

We also determined that $N(m+3, 4, m) = m+9$ ($m \geq 6$), $N(8, 4, 8) = 15$. Our results partly settle an open problem of Paterson et al.

Keywords combinatorial batch code, optimal CBC, monotonicity, lower and upper bounds

MSC(2010) 05B30, 94B25

doi: 10.1360/N012014-00061