第11卷 第1期
1992 年 2月

红外与毫米波学报
J. Infrared Millim. Waves

Vol. 11, No. 1
February, 1992

差值"基础比特+溢出比特"编码方法

钱神恩　李树钦　王汝勤

(中科院长春光学精密机械研究所应用光学国家重点实验室，长春，130022)

戴逸松

(吉林工业大学电子工程系，长春，130025)

摘要: 提出了差值"基础比特+溢出比特"编码方法. 与霍夫曼方法相比，具有无需知道信源的统计特性，算法简便，易于实时处理，编码效率高 (可达 90%以上) 等特点，并在成象光谱仪机上实时数据压缩中取得了良好的应用效果.

关键词: 编码，实时处理，数据压缩，成象光谱.

引言

在对空间高分辨率成象光谱仪(HIRIS)遥感信息实时数据压缩系统研究中[1]，为了用最少的比特表示经二真值线性预测法[2]压缩后的数据，最终获得尽可能高的比特压缩比，需对压缩后的数据进行高效、实时的编码. 由于作为最佳源编码方法的霍夫曼编码需要知道信源的统计特性，使用很不方便，而且算法烦琐，不易实时处理[3]. 为此，我们研究提出了一种不需要知道信源统计特性，算法简便，适合在线处理的编码方法——差值"基础比特+溢出比特"法.

1 差值"基础比特+溢出比特"编码

原始数据经二真值线性预则法 (简称二真值法) 压缩后，去掉了大量的冗余信息，但若对压缩后结果不经编码直接表示，则需用大量的比特，不利于提高最终的系统压缩比. 实际上，二真值法压缩后的数据还存在较大的冗余度，若进行有效的编码进一步去除冗余信息，可使系统总压缩比进一步提高. 下面我们分两步来讨论这种编码方法.

1.1 邻差化缩小被编码数据的数值范围

一般来说，被编码数据的平均值 μ 大，其数值范围就大，编码表示这些数据所需的平均码长 $\bar{L}$ 就长；反之，平均值 μ 小，被编码数据的数值就小，平均码长 $\bar{L}$ 就短. 因此，为了减小平均码长 $\bar{L}$，首先应该缩小被编码数据的数值变化范围.

本文 1990 年 10 月 29 日收到，修改稿 1991 年 5 月 24 日收到.

表1列出了植物类玉米光谱数据经二真值法压缩后48组待编码数据. 由表1可见, $y1(i)$、$y2(i)$的数值都较大，平均值分别为 $\mu_1=72.75$，$\mu_2=73.06$. $RL(i)$数值相对小些，平均值$\overline{RL}=3.35$. 无疑，表示48个$y1(i)$或$y2(i)$所需的比特数肯定比$RL(i)$要多，平均码长也长. 实际上，二真值法压缩后的待编码数据$y1(i)$、$y2(i)$是相邻的两个采样值，$y1(i)$与$y2(i-1)$是相邻两条预测线上最近的两个采样值，它们之间存在着相关性，它们的差值为:

$$\left.\begin{aligned}\Delta y1(i)&=y1(i)-y2(i-1),\ i=2,3,\cdots,N; &(1)\\ \Delta y2(i)&=y2(i)-y1(i),\ i=1,2,\cdots,N; &(2)\end{aligned}\right\}$$

$$\Delta y1(1)=y1(1); \qquad (3)$$

其数值必定小于$y1(i)$、$y2(i)$的数值，平均值也小于原数据的平均值.

表1 玉米光谱压缩后待编码数据 ($N=48$)

Table 1 Coding data (corn spectrum compressed) ($N=48$)

i	$y1(i)$	$y2(i)$	$RL(i)$	i	$y1(i)$	$y2(i)$	$RL(i)$	i	$y1(i)$	$y2(i)$	$RL(i)$
1	9	8	1	18	163	164	4	35	51	54	10
2	10	11	3	19	165	162	5	36	81	78	0
3	11	11	4	20	150	149	3	37	78	83	0
4	13	15	3	21	149	149	6	38	83	82	9
5	20	19	10	22	147	145	0	39	70	68	2
6	12	12	2	23	145	139	1	40	67	67	3
7	14	17	0	24	134	132	1	41	64	52	1
8	17	24	1	25	127	116	1	42	41	33	1
9	33	66	0	26	102	90	0	43	26	22	4
10	66	79	0	27	90	66	0	44	9	8	3
11	79	111	0	28	66	55	0	45	7	8	17
12	111	131	0	29	55	51	0	46	28	30	2
13	131	143	0	30	51	37	0	47	31	32	3
14	143	147	2	31	37	32	1	48	32	31	27
15	152	153	8	32	30	28	1				
16	158	158	12	33	26	29	7				
17	160	162	2	34	48	48	1	μ	72.75	73.06	3.35

表2列出了玉米光谱压缩后数据的差值$\Delta y1(i)$、$\Delta y2(i)$，其数值范围大大缩小，$\mu_1'=4.35$，$\mu_2'=5.77$. 若用差值$\Delta y1(i)$，$\Delta y2(i)$分别替代$y1(i)$、$y2(i)$，就可减少表示被编码数据的比特数，降低平均码长$\overline{L}$. 恢复时原待编码数据可按下式求得:

$$y1(i)=\Delta y1(i); \qquad (4)$$

$$\left.\begin{aligned}y2(i)&=\Delta y2(i)+y1(i),\ i=1,\ 2,\ \cdots N; &(5)\\ y1(i)&=\Delta y1(i)+y2(i-1),\ i=2,\ 3,\ \cdots N; &(6)\end{aligned}\right\}$$

这种用被编码数据相邻值的差代替原编码数据的过程我们称为邻差化，这种邻差化具有普遍性，对任何被编码数据均可进行.

1.2 基础比特+溢出比特编码

一般说，被编码数据总存在一个数值$R(=2^n)$，其中大多数被编码数据值小于R. 例

如前面提到的玉米光谱压缩后数据，48个游长值$RL(i)$中有35个数值小于$R=4=2^2$ $(n=2)$，占总数的73%；表2所列的邻差化后的编码数据，48个$\Delta y1(i)$中有37个数值小于$R=8=2^3$ $(n=3)$，占总数的77%；48个$\Delta y2(i)$中有36个数值小于$R=8$ $(n=3)$，占总数的75%. 对于这种被编码数据大部分数值小于某个临界值情况的编码，我们采用一种叫做"基础比特+溢出比特"的编码方法 (简称"基+溢"法). 它将编码结果分成"基础比特"与"溢出比特"两部分，基础比特为固定字长部分，字长一般由R值决定，用于表示小于R的数值；溢出比特为随机字长部分，每一位溢出比特"1"代表数值$R=(2^n)$，溢出比特长度由大于R的数值决定，大于多少个R就有多少个溢出比特位；为标志非固定字长的溢出比特位的结束，用"0"作为溢出比特位结束逗号. 这种方法编码时由于大部分数据小于R，溢出比特长度为零，只需$n+1$比特即可，其中1比特是逗号位，对有符号的数据需增加1比特符号位，共需$n+2$比特；对少部分大于R的数据，编码结果为基础比特加溢出比特两部分.

表2　邻差化后待编码数据

Table 2　Coding data after difference

i	$\Delta y1(i)$	$\Delta y2(i)$	i	$\Delta y1(i)$	$\Delta y2(i)$
1	9	−1	40	−1	0
2	2	1	41	−3	−12
3	0	0	42	−11	−8
4	2	2	43	−7	−4
5	5	−1	44	−13	−1
6	−7	0	45	−1	1
7	2	3	46	20	2
8	0	7	47	1	1
9	9	33	48	0	−1
10	0	13			
⋮	⋮	⋮	μ	4.35	5.77

这种编码方法直观、简便，无需知道信源的统计特性，可用最少的比特表示被编码的数据. 对邻差化后二真值法压缩后数据的编码，若设$\Delta y1(i)$、$\Delta y2(i)$和$RL(i)$的基础比特长度分别为k、n、m，并考虑到$RL(i)$无符号位，对大部分数据编码后只需$5+k+n+m$比特即可. 例如表1所列的压缩后数据，$k=n=3$，$m=2$，只需13比特便可表示，比不编码直接表示时的24比特(3字节)大大减少.

2　基础比特长度对编码效果的影响

与霍夫曼码相比，"基+溢"法编码时唯一需要设置的参数是基础比特长度L_b. 图1是"基+溢"编码方法的基础比特长度L_b与编码效率η的关系曲线图. 横座标为被编码数据平均值μ之倒数$(1/\mu)$. 由图可见，若能求得被编码数据的平均值μ，则可从图中找到最大编码效率时对应曲线的基础比特长度L_b. 被编码数据的平均值μ越大，$1/\mu$越小，基础比特长度L_b就越大. 如表1中48个$\Delta y2(i)$的平均值为$\mu_2=73.06$，倒数为0.0137，由图可得$L_b=6$时编码效率最高；而邻相化后$\Delta y2(i)$的平均值$\mu_2'=5.77$，倒数为0.17，$L_b=2$时编码效率最高，当然邻差化后产

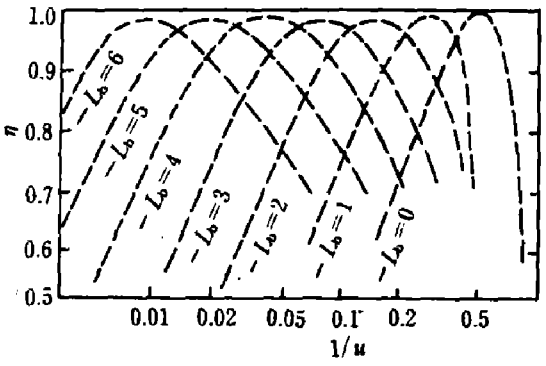

图1　基础比特长度L_b与编码效率η关系

Fig.1　curves of L_b *vs.* η

生了负值，还需增加1位符号位，实际为3比特. 可见，邻差化后每个$\Delta y2(i)$编码平均可节省3比特.

实际编码时可根据被编码数据的情况，大致估计其平均值，然后选择相应的基础比特长度. 图1的横座标采用对数座标，目的是压缩横座标的比例尺. 若取线性座标可更明显地看出各L_b曲线对应的μ差别并不大，这就使得基础比特长度L_b的选择变得相对容易，即使被编码数据的平均值估计偏差较大，也不致于使编码效果相差太大.

以植物类玉米光谱压缩后数据为例来说明L_b偏差对编码效果的影响. 邻差化后$\Delta y2(i)$的平均值为$\mu_2 = 5.77$，倒数为0.17，由图1可知，当$L_b = n = 2$时编码效率最高. 现分别取$L_b = 1, 2, 3$三个值，对$\{\Delta y2(i)\}$编码，求得三种不同L_b时编出的平均码长分别为$\overline{L}_{\Delta y2} = 5.61, 5.23, 5.54$，以$n = 2$时平均码长最短. 48条预测线的平均游长$\overline{RL} = 3.35$，倒数为0.3，由图1可得$L_b = m = 1$时编码效率最高. 分别取$L_b = m = 0, 1, 2$，三个值对$\{RL(i)\}$编码，求得三种不同$L_b$情况的平均码长分别为$\overline{L}_{RL} = 3.69, 3.50, 3.63$，$m = 1$时平均码长最短. 从上面对$\{\Delta y2(i)\}$、$\{RL(i)\}$编码过程$L_b$的选取情况看，即使基础比特长度选择有偏差，编码结果相差也不太大. L_b取其平均值对应值的相邻值时，$\{\Delta y2(i)\}$平均码长偏差分别为7.2%、5%；$\{RL(i)\}$平均码长的偏差分别为5.4%，3.7%，可见偏差都不大，说明L_b的选择具有较宽的适应区域.

由上面讨论可知，“基+溢”法编码时选择适当的基础比特长度L_b对提高编码效率是有益的.

3 实验结果与分析

我们对二真值法压缩后HIRIS中每个GIFOV的光谱数据先邻差化，转换成$\{\Delta y1(i), \Delta y2(i), RL(i)\}$ $(i = 1, 2, \cdots, N)$，然后再用“基+溢”法编码. 表3列出了21种典型地物光谱压缩后数据的编码结果. 为便于分析和评价编码效果，表3中同时列出了每种压缩后待编码数据$\{y1(i), y2(i), RL(i)\}$的熵$H_o(x)$和邻差化后被编码数据的熵$H_d(x)$，以及对应这两种熵的编码效率η_1、η_2. 表3是在L_b取被编码数据平均值μ之倒数所对应的最大编码效率时取得的结果. 由香农(Shonon)无失真编码定理可知，任何编码方法编出的平均码长$\overline{L}$都不能小于被编码数据的熵$H(x)$. 而表3中21种典型地物光谱压缩后数据用差值“基+溢”法编码，平均码长$\overline{L}$几乎都不到5比特，均小于待编码数据的熵$H_o(x)$，编码效率

$$\eta_1 = \frac{H_o(x)}{\overline{L}} \geqslant 100\%.$$

这是因为N组待编码源数据的熵$H_o(x)$是按各数值出现的概率求得的，邻差化过程去掉了待编码数据间的某些相关性，而使其熵变小为$H_d(x)$. “基+溢”编码过程实际上是对邻差化后的数据$\{\Delta y1(i), \Delta y2(i), RL(i)\}$ $(i = 1, 2, \cdots, N)$进行编码，按$H_d(x)$与平均码长$\overline{L}$之比求得的“基+溢”法编码效率

$$\eta_2 = \frac{H_d(x)}{\overline{L}} < 100\%,$$

满足香农无失真编码定理. 由表3可见，η_2都大于90%，这结果与霍夫曼码结果类似.

表 3　21 种典型地物光谱压缩后数据编码结果

Table 3　Cooling results of 21 typical earth resources spectra compressed

序号	地物的光谱名称	源数据特性		编码结果		
		$H_o(x)$	$H_d(x)$	$\bar{L}$	η_1(%)	η_2(%)
1	有机物为主土壤	4.59	3.85	4.01	114.3	95.9
2	含铁土壤	4.94	4.37	4.58	107.9	95.4
3	方解石碳酸岩	5.16	4.11	4.40	117.2	93.5
4	高岭土	5.51	4.63	4.87	113.1	95.0
5	豆　类	5.64	4.33	4.46	126.7	97.1
6	玉　米	5.64	4.73	5.03	113.0	94.0
7	红　松	5.48	4.53	4.96	110.6	91.4
8	桦　树	5.52	4.42	4.59	120.3	96.4
9	水	4.88	4.03	4.33	112.6	92.9
⋮						
21	雪	5.51	4.64	4.86	113.2	95.4

差值"基+溢"编码方法是一种不需要知道信源统计特性的编码方法，算法简便，实时性好．它首先通过邻差化将待编码数据的数值范围缩小，然后再用"基+溢"法进行编码．这种方法编码时唯一需要设置的参数是基础比特长度L_b，由被编码数据的平均值μ决定．适宜的L_b可获得最短的平均码长．L_b的选择具有较宽的适应区域，相邻L_b引起的平均码长偏差一般为 5%左右．编码效率可达 90%以上，可获得类似霍夫曼码的效果．经对 21 种典型地物光谱数据实验，在 8 比特数字系统中编码的平均码长一般不超过 5 比特，提高编码效率近 1 倍．

参 考 文 献

1　钱神恩．博士学位论文，吉林工业大学电子工程系，1990

2　钱神恩．**光学学报**，1990，10(3)：260～266

3　周炯槃．**信息理论基础**，北京：人民邮电出版社，1983，358～378

4　Qian Shen-en. *Proc. SPIE*, 1990, 1244: 331～342

DIFFERENCE BASE-BIT PLUS OVERFLOW-BIT CODING

Qian Shenen, Li Shuqiu, Wang Ruqin

(*National Laboratory of Applied Optics, Changchun Institute of Optics and Fine Mechanics, Changchun, Jilin* 130022, *China*)

Dai Yisong

(*Electronics Engineering Department, Jilin University of Technology, Changchun, Jilin* 130025, *China*)

Abstract: A new coding method—Difference Base-bit Plus Overflow-bit Coding (DBOC) is proposed. Compared with the Huffman coding, this method does not require the statistical properties of coded data, and it can be realized in real-time. Experiments show that the coding efficiency can reach over 90%. With this coding method an excellent result has been obtained in an on-board data compression system for imaging spectrometer.

Key words: coding, real-time processing, data compression, imaging spectrometer.