УДК 621.398

В.В.Сергеев

ВЫБОР СПОСОБА КОДИРОВАНИЯ ДАННЫХ В МЕТОДАХ АДАПТИВНЫХ ВЫБОРОК

Известным классом методов сжатия данных являются методы адаптивных выборок [I]. Их действие заключается в выборе из предварительно дискретизированного сигнала некоторых "существенных" отсчетов таких, что остальные отсчеты ("избыточные") могут быть восстановлены по "существенным" с максимальной погрешностью, не превышающей некоторое допустимое значение ε. При оценке эффективности таких методов сжатия данных обычно теоретически или экспериментально определяется зависимость от ε коэффициента сжатия по числу отсчетов: $K_0 = N_0/N$, где N_0 - общее число обработанных отсчетов, N - число "существенных" отсчетов. Однако при регистрации (передаче) сжатых данных в цифровой форме эффективность метода сжатия наиболее полно характеризуется коэффициентом сжатия по числу двоичных знаков: $K_{сж} = J_0/J$, где J_0 -объем данных до сжатия, J - объем данных после сжатия. Величина $K_{сж}$ зависит как от значения K_0, так и от выбранного способа кодирования данных. Особенностью методов адаптивных выборок является необходимость включения в сжатые данные информации для указания положения (датирования) "существенных" отсчетов. Выбор способа датирования определяет структуру кода сжатых данных.

В статье рассматриваются некоторые простые способы кодирования сжатых данных. Для каждого из них определяется связь $K_{сж}$ с K_0 и ε. Дается методика выбора по известным ε, $K_0(\varepsilon)$ такого способа, для которого обеспечивается наибольшее значение $K_{сж}$.

В настоящее время в системах обработки данных широко используются ЭВМ третьего поколения. Поскольку для них характерна байтовая структура памяти, данные, подлежащие обработке, как правило, представляются в форме I байт (8 бит) на отсчет [2]. Пусть и в нашем случае сигнал, который подвергается процедуре сжатия данных, в цифровой форме представляет собой последовательность восьмиразрядных двоичных чисел без знака. Не умаляя общности, примем вес младшего разряда за единицу, при этом сигнал имеет шкалу 0-255, его отсчеты, а также погрешность ε принимают целочисленные значения. С точки зрения удобства оперирования сжатыми данными, их целесообразно представлять в подобной же форме, т.е. так, чтобы целому числу отсчетов исходного сигнала всегда соответствовало целое число байтов.

Простейшим способом кодирования, который может удовлетворять этому требованию, является представление информации о существенных отсчетах парами "отсчет-дата" [3]. Пусть b_S и b_T - число двоичных символов, соответственно, для описания значения существенного отсчета и его датирования (указанием числа подряд пропущенных избыточных отсчетов). В соответствии с изложенным, выделим два случая:

а) (первый способ кодирования) $b_S = b_T = 8$, (I)

б) (второй способ кодирования) $b_S + b_T = 8$. (2)

Рассмотрим также третий способ кодирования сжатых данных, при котором пары "отсчет-дата" не формируются. Каждому отсчету сигнала поставим в соответствие двоичный символ для указания принадлежности отсчета к подмножеству существенных или избыточных. Для восьми подряд идущих отсчетов обозначим элементы образовавшейся двоичной "датирующей" последовательности как q_i ($i = \overline{1,8}$). Пусть нулевые значения элементов соответствуют избыточным отсчетам, единичные - существенным. Тогда число существенных отсчетов среди рассматриваемых восьми определится как сумма:

$$V = \sum_{i=1}^{8} q_i .$$

Данный способ кодирования заключается в том, что каждым восьми отсчетам сигнала ставится в соответствие (V + I) байт, первый из них содержит отрезок датирующей последовательности, а следующие - V значений существенных отсчетов.

Некоторые способы кодирования сжатых данных оказывают влияние на зависимость $K_0(\varepsilon)$, во-первых, из-за введения ограничения на максимальную величину интервала между существенными отсчетами и, во-вторых, из-за округления значений отсчетов при их описании малым числом двоичных разрядов. Оценим это влияние.

При анализе систем сжатия данных с адаптивными выборками поток событий, заключающихся в появлении существенных отсчетов, обычно принимается биномиальным [4]. При этом случайная величина

T - число избыточных отсчетов между парой существенных - имеет геометрическое распределение:

$$W(T) = r(1-r)^{T},$$

где r - вероятность отнесения каждого отсчета к подмножеству существенных.

При ограничении T сверху величиной T_M ее новое распределение имеет вид

$$W(T) = \begin{cases} r(1-r)^{T} & \text{при} \quad T = \overline{0,(T_M-1)}, \\ (1-r)^{T_M} & \text{при} \quad T = T_M \\ 0 & \text{при} \quad T = \overline{(T_M+1),\infty}. \end{cases} \tag{3}$$

Коэффициент сжатия по выборкам имеет смысл средней длины интервала между существенными отсчетами при единичном шаге первичной дискретизации, т.е.

$$K_0 = \sum_{T=0}^{\infty} W(T)(T+1). \tag{4}$$

Для геометрического распределения

$$K_0 = \frac{1}{r}. \tag{5}$$

Для распределения выражения (3) после проведения суммирования (4) с учетом равенства (5) получаем:

$$K_0^{*} = K_0\left[1-\left(1-\frac{1}{K_0}\right)^{T_M+1}\right] = K_0(1-\alpha), \tag{6}$$

где K_0^{*} - значение коэффициента сжатия по выборкам с учетом ограничения T;

α - параметр, характеризующий относительное уменьшение коэффициента сжатия.

Если $\alpha <$ 0,05, то из выражения (6) следует приближенное неравенство

$$T_M > \frac{3}{\ln\left(\frac{K_0}{K_0-1}\right)} - 1,$$

которое для $K_0 \gg$ I принимает вид

$$T_M > 3K_0. \tag{7}$$

Это неравенство можно считать условием, при котором можно пренебречь уменьшением коэффициента сжатия из-за ограничения длин интервалов между существенными отсчетами (оно составляет менее 5%). Очевидно, ограничение длин интервалов имеет место для первых двух способов кодирования и определяется величиной b_T :

$$T_M = 2^{b_T} - 1. \qquad (8)$$

Если отсчеты сигнала описываются восьмиразрядными целыми двоичными числами, то при переходе к их описанию более короткими словами в b_S разрядов, возникает погрешность округления ε_0 , равная половине веса младшего разряда нового слова:

$$\varepsilon_0 = 2^{(7-b_s)}. \qquad (9)$$

При этом сжатие данных должно производиться не при допустимой погрешности ε , а при меньшей - ε' . такой, чтобы

$$\varepsilon_0 + \varepsilon' = \varepsilon. \qquad (10)$$

В данном случае вместо значения $K_0(\varepsilon)$ достигается меньшее значение коэффициента сжатия по выборкам, в соответствии с выражением (10) равное $K_0(\varepsilon - \varepsilon_0)$. Связь значений функции $K_0(\varepsilon)$ при различных значениях аргумента приближенно определим, считая эту функцию линейной:

$$K_0(\varepsilon - \varepsilon_0) = 1 + (1 - \frac{\varepsilon_0}{\varepsilon})[K_0(\varepsilon) - 1]. \qquad (11)$$

Округление отсчетов, очевидно, производится во втором способе кодирования сжатых данных.

Итак, определим связь $K_{сж}$ с ε и $K_0(\varepsilon)$ для трех рассмотренных способов кодирования сжатых данных.

Для первого способа кодирования, как следует из (1) и (8), T_M = 255, поэтому в соответствии с (7) при $K_0 < 85$ (т.е. в большинстве практических случаев) влиянием на эффективность ограничения длин интервалов можно пренебречь. Таким образом, учитывая двукратное увеличение объема сжатых данных из-за датирования, здесь имеем:

$$K_{сж}(\varepsilon) = 0{,}5\,K_0(\varepsilon). \qquad (12)$$

Для второго способа искажениями коэффициента сжатия по выборкам пренебречь нельзя. В данном случае $K_{сж}(\varepsilon)=K_0^*(\varepsilon)$, или с учетом подстановки (II) в (6):

$$K_{сж}(\varepsilon)=\left[1-\left(1-\frac{\varepsilon_0}{\varepsilon}\right)(K_0(\varepsilon)-1)\right]\left\{1-\left[\frac{\left(1-\frac{\varepsilon_0}{\varepsilon}\right)(K_0(\varepsilon)-1)}{1+\left(1-\frac{\varepsilon_0}{\varepsilon}\right)(K_0(\varepsilon)-1)}\right]^{T_M+1}\right\}. \quad (I3)$$

В этом выражении, как следует из (2), (8) и (9),

$$T_M=2^{(8-b_S)}-1, \quad \varepsilon_0=2^{(7-b_S)},$$

т.е. эффективность сжатия, кроме прочих факторов, зависит от значения b_S.

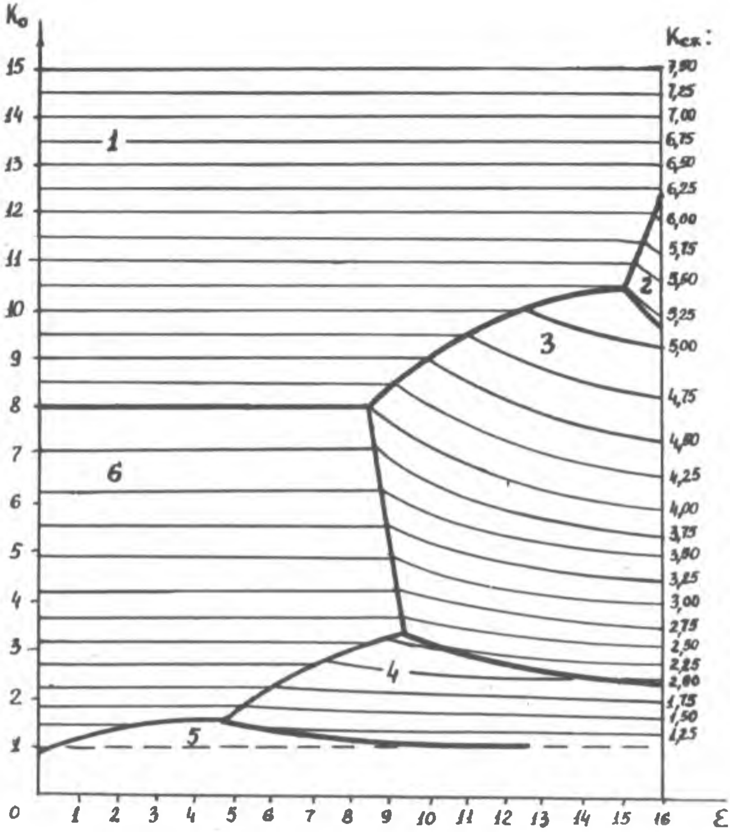
Для третьего способа кодирования нетрудно получить, что

$$K_{сж}(\varepsilon)=\frac{8K_0(\varepsilon)}{8+K_0(\varepsilon)}. \quad (I4)$$

При различных ε и $K_0(\varepsilon)$ наилучшим, в смысле эффективности, может быть тот или иной способ кодирования сжатых данных. Для выявления случаев, когда каждый из рассмотренных способов обеспечивает наибольший эффект сжатия, для $1<K_0\leqslant$ I56, $\varepsilon=$ 0,I6 были проведены расчеты $K_{сж}$ по формулам (I2), (I3) и (I4) и в плоскости $\{\varepsilon, K_0\}$ выделены области, где каждый способ является наилучшим. Результаты расчетов приведены на рис. I. Там же показаны линии равных значений $K_{сж}$. Рисунок дает возможность выбрать для каждой пары $\varepsilon, K_0(\varepsilon)$ наиболее выгодный, с точки зрения эффективности, способ кодирования сжатых данных и определить достигаемое при этом значение коэффициента сжатия по числу двоичных знаков.

Литература

I. Мановцев А.П. Основы теории радиотелеметрии. - М.: Энергия, I973.
2. Евдокимов В.П., Покрас В.М. Методы обработки данных в научных космических экспериментах. - М.: Наука, I977.
3. Бабкин В.Ф., Крюков А.Б., Штарьков Ю.М. Сжатие данных. - В сб.: Аппаратура для космических исследований. - М.: Наука, I972.



Р и с. I. Области применения способов кодирования сжатых данных: I - первый способ; 2-5 - второй способ (2- $V\delta_s$=4, 3- $V\delta_s$ = 5, 4- $V\delta_s$= 6, 5- $V\delta_s$=7); 6 - третий способ

4. С в и р и д е н к о В.А. Анализ систем со сжатием данных. - М.: Связь, 1977.