УДК 621.327:681.5

В.В. Баранник[1], А.В. Хаханова[2]

[1]*Харьковский университет Воздушных Сил им. И. Кожедуба*
[2]*Харьковский национальный университет радиоэлектроники*

ОЦЕНКА ИНФОРМАТИВНОСТИ ДВОИЧНЫХ МАССИВОВ НА ОСНОВЕ КОМБИНАТОРНОГО ПОДХОДА

Излагаются основные этапы методики оценки информативности двоичных данных на основе построения комбинаторной информационной модели двоичных массивов. Данная методика учитывает ограничения на число серий единиц в двоичных последовательностях плавающих размеров. Это позволило обосновать, что учет закономерностей по числу серий единиц позволяет сократить избыточность в условиях не стационарности статистических свойств исходных данных. Показано, что данный подход к оценки информативности двоичных данных обеспечивает сокращение избыточности в условиях: равновероятных символов алфавита; для наихудших условий выявленных структурных закономерностей.

Ключевые слова: *число серий единиц, информационная модель, сжатие данных.*

Введение

Постановка проблемы и анализ литературы. Существующие подходов относительно компактного представления двоичных данных, не адекватны современным требованиям процессов функционирования информационных систем [1, 2]. Увеличиваются объемы данных, которые предназначены для хранения и передачи в информационно-регистрирующих системах. Основной причиной является организация сжатия двоичных данных на основе устранения вероятностных видов избыточности. В тоже время количество такой избыточности зависит от типа законов распределения вероятностей появления символов в сообщении. Это приводит к снижению степени сжатия данных

Поэтому **актуальной научной задачей** является сокращение объемов двоичных данных без внесения погрешности.

Целью статьи является обоснование выбора подхода для построения технологии компрессии двоичных данных.

Разработка методики оценки информативности двоичных массивов

Для дополнительного увеличения коэффициента сжатия, обрабатываемых двоичных данных предлагается дополнительно проводить сокращение избыточности в двоичных матрицах G, $G=\{g_{k\ell}\}$, $k=\overline{1,n}$, $\ell=\overline{1,n}$, $g_{k\ell}\in\{0;1\}$, где $g_{k\ell}$ – $(k;\ell)$-й элемент матрицы знаков. Двоичные матрицы имеют низкую избыточность $R_{ст}$, определяемую на основе учета статистических закономерностей $R_{ст}=(H_0-H_{ст})/H_0\to 0$, где H_0 – количество информации, в среднем приходящееся на один элемент матрицы, для случая, когда закономерности не выявлены, $H_0=1$ бит; $H_{ст}$ – среднее количество информации, содержащееся в одном элементе с учетом ограничений на закон распределения вероятностей $p_{k\ell}$ появления элементов $g_{k\ell}$: $H_{ст}=p(0)\ell og_2\,p(0)+p(1)\ell og_2\,p(1)\to H_0$, где $p(0)$, $p(1)$ – вероятности появления соответственно нулевого и единичного элементов. Это снижает потенциальные возможности относительно обеспечения необходимой степени компрессии двоичных матриц. Следовательно, сжатие двоичных матриц достигается за счет сокращения комбинаторной избыточности, независящей от статистических свойств матриц G.

Анализируемая характеристика двоичных матриц должна учитывать взаимное расположение нулевых и единичных элементов и их количество. Причем количество нулевых элементов может равняться количеству единичных. Количество информации H в двоичных последовательностях с учетом анализируемых ограничений должно удовлетворять следующим требованиям:

1) количество информации на один элемент должно быть меньшим единицы для источника информации с равномерным появлением единиц и нулей

$$H <<< 1 \text{ для } p(0)\approx p(1); \qquad (1)$$

2) величина H для произвольных значений анализируемой характеристики η должна быть меньше единицы

$$\max_{\eta_{min}\le\eta\le\eta_{max}}\{H\}<1, \qquad (2)$$

где η_{min}, η_{max} – соответственно минимальное и максимальное значения величины η.

Предлагается рассматривать выявления закономерностей на основе характеристики количества серий η. Признак серии задается системой выражений:

$$\begin{cases} g_{k-1,\,\ell}\neq g_{k\ell}\to k\le n;\\ g_{k,\,\ell}\neq g_{1,\,\ell+1}\to k>n. \end{cases}$$

Вычисление количества серий единиц η для ν элементов матрицы G:

– на нулевом шаге $g_{0,\ell}=0$; $\eta_\ell=0$;

– на k-м шаге значения числа серий увеличивается на 1, $\eta_{k\ell}=\eta_{k-1,\ell}+1$, если $g_{k-1,\ell}<g_{k\ell}$;

– в противном случае $\eta_{k\ell}=\eta_{k-1,\ell}+0$, если $g_{k-1,\ell}\geq g_{k\ell}$;

– для конечного шага при $k=\nu$ получаем искомое значение количества серий единиц $\eta=\eta_\nu$ для j-го столбца.

Количество $V_{\nu,\eta}$ двоичных последовательностей G_ν, содержащих ν элементов и η серийных перепадов равно [3, 4]:

$$V_{\nu,\eta}=(\nu+1)\,!/(2\eta)!\;(\nu+1-2\eta)\,!. \qquad (3)$$

В соответствии с (3) количество информации H в среднем приходящееся на один элемент двоичной последовательности с η сериями единиц равно

$$H=\left(\sum_{\xi=1}^{\nu+1}\ell og_2\,\xi\;-\sum_{\xi=1}^{2\eta}\ell og_2\,\xi-\sum_{\xi=1}^{\nu+1-2\eta}\ell og_2\,\xi\right)/\nu. \qquad (4)$$

Рассмотрим соответствие H условиям (1) и (2):

1) для фиксированного ν максимальное значение $V_{\nu,\eta}$ двоичных последовательностей достигается для $\eta=\eta_{cp}=[(\nu+1)/4]$. Это вытекает из того, что величина $V_{\nu,\eta}$ для $0\leq\eta\leq[(\nu+1)/2]$ имеет один максимум в области среднего значения η. Поэтому для проверки условия (2) необходимо определить значение величины $V_{\nu,\eta_{cp}}$:

$$V_{\nu,\eta_{cp}}=\begin{cases}(\nu+1)\,!/(\nu/2)!\;(\nu/2+1)\,!,\\ \to \nu\bmod(2)=0;\\ (\nu+1)\,!/((\nu+1)/2)!\;((\nu+1)/2)\,!,\\ \to \nu\bmod(2)\geq 1,\end{cases} \qquad (5)$$

где $\nu\bmod(2)=0$, $\nu\bmod(2)\geq 1$ – условия, когда величина ν принимает соответственно четное и нечетные значения.

Значит, для $\eta=\eta_{cp}=[(\nu+1)/4]$ величина H будет принимать максимальное значение, равное:

– если $\nu\bmod(2)=0$, то

$$H_{\nu,\eta_{cp}}=\left(\sum_{\xi=1}^{\nu+1}\ell og_2\,\xi-2\sum_{\xi=1}^{\nu/2}\ell og_2\,\xi-\ell og_2\,(\nu/2+1)\right)/\nu;$$

– если $\nu\bmod(2)\geq 1$, то

$$H_{\nu,\eta_{cp}}=\left(\sum_{\xi=1}^{\nu+1}\ell og_2\,\xi\;-\;2\sum_{\xi=1}^{(\nu+1)/2}\ell og_2\,\xi\right)/\nu, \qquad (6)$$

где $H_{\nu,\eta_{cp}}$ – количество информации, приходящееся на один элемент матрицы в случае когда, число серий единиц равно $\eta=\eta_{cp}=[(\nu+1)/4]$.

Условие (2) выполняется, если

$$H_{\nu,\eta_{cp}}\;<\;1. \qquad (7)$$

Правая часть неравенства (7) соответствует варианту, когда количество допустимых двоичных последовательностей равно 2^ν. Поэтому неравенство (7) будет выполняться для

$$V_{\nu,\eta_{cp}}<2^\nu. \qquad (8)$$

Рассмотрим сумму величин $V_{\nu,\eta}$ по всему диапазону значений числа серий единиц $0\leq\eta\leq[(\nu+1)/2]$:

1. Если ν четное, то $\eta=\overline{0,\nu/2}$. Введем вспомогательные переменные $\alpha=(\nu+1)$ и $\beta=2\eta$, то

$$\sum_{\beta=0}^{(\alpha-1)}(\alpha)\,!/((\beta)!\;(\alpha-\beta)\,!). \qquad (9)$$

Переменная β принимает все значения в диапазоне $\beta=\overline{0,(\alpha-1)}$. В то время как величина 2η принимает в этом диапазоне только положительные четные значения. Поэтому

$$\sum_{\eta=0}^{\nu/2}V_{\nu,\eta}=\sum_{\beta=0}^{(\alpha-1)}h_{\alpha,\beta}; \qquad (10)$$

$$h_{\alpha,\beta}=\begin{cases}(\alpha)\,!/((\beta)!\;(\alpha-\beta)\,!), & \to(-1)^\beta=1;\\ 0, & \to(-1)^\beta=-1.\end{cases}$$

Тогда с учетом свойства сочетаний без повторений

$$\sum_{\beta=0}^{\alpha}((-1)^\beta(\alpha)\,!)/((\beta)!\;(\alpha-\beta)\,!)=0 \qquad (11)$$

получим $\sum_{\eta=0}^{\nu/2}V_{\nu,\eta}=2^{\alpha-1}=2^\nu$. В тоже время $V_{\nu,\eta_{cp}}<\sum_{\eta=0}^{\nu/2}V_{\nu,\eta}$. Отсюда следует, что для четного ν неравенство (8) выполняется.

2. Если ν нечетное, то $\eta=\overline{0,(\nu+1)/2}$, а

$$\sum_{\eta=0}^{(\nu+1)/2}V_{\nu,\eta}=\sum_{\eta=0}^{(\nu+1)/2}\frac{(\nu+1)\,!}{(2\eta)!\;(\nu+1-2\eta)\,!}.$$

С учетом свойств, заданных соотношениями (10) и (11), получим

$$\sum_{\eta=0}^{\nu/2}V_{\nu,\eta}=2^{\alpha-1}=2^\nu.$$

Следовательно, неравенство (8) выполняется для нечетного ν. Таким образом, условие (2) для предлагаемой количественной характеристики «число серий единиц» выполняется.

Рассмотрим условие (1) применительно к характеристики «число серий единиц». Количество $V(s)_{\nu,\eta}$ двоичных последовательностей для числа единиц, равного s, распределенных по η сериям равно

$$V(s)_{\nu,\eta} = \frac{(s-1)!}{(\eta-1)!\,(s-\eta)!} \times \frac{(\nu-s+1)!}{(\eta)!\,(\nu-s+1-\eta)!},$$

$$\eta = \overline{0,\ \min(s;\ \nu/2)}. \quad (12)$$

В случае равных вероятностей $p(0) \approx p(1)$ имеем одинаковое количество нулевых и единичных элементов, равное $\nu/2$. Для такого условия выражение (12) будет иметь вид:

1. Если ν – четное ($\eta = \overline{0,\ s}$), то

$$V(\nu/2)_{\nu,\eta} = \frac{((\nu/2)-1)!}{(\eta-1)!((\nu/2)-\eta)!} \times \frac{((\nu/2)+1)!}{(\eta)!((\nu/2)+1-\eta)!}. \quad (13)$$

Поскольку сомножители в правой части выражения (13) представляют собой сочетания без повторений, то они достигают максимума при $\eta/2$. Значит максимальное значение величина $V(s)_{\nu,\eta}$ при фиксированных значениях ν и s достигает для $\eta = s/2 = \nu/4$:

$$V(\nu/2)_{\nu,\nu/4} = \frac{((\nu/2)-1)!}{((\nu/4)-1)!\ ((\nu/4))!} \times \frac{((\nu/2)+1)!}{((\nu/4))!((\nu/4)+1)!}.$$

В соответствии с данным выражением количество информации $\overline{H}(\nu/2)_{\nu,\nu/4}$, приходящееся на один элемент двоичной последовательности, содержащей $\eta = \nu/4$ серий единиц и $\nu/2$ единичных элементов

$$\overline{H}(\nu/2)_{\nu,\nu/4} = \ell og_2\left(\frac{((\nu/2)-1)!}{((\nu/4)-1)!((\nu/4))!} \times \frac{((\nu/2)+1)!}{((\nu/4))!((\nu/4)+1)!}\right)/\nu.$$

Поскольку количество η серий единиц дополнительно ограничено величиной s, то выполняется

$$\sum_{\eta=1}^{s} V(s)_{\nu,\eta} = \sum_{\eta=1}^{\nu/2} V(\nu/2)_{\nu,\eta} < \sum_{\eta=0}^{\nu/2} V_{\nu,\eta}. \quad (14)$$

Правая часть неравенства (14) имеет ограничение $\sum_{\eta=0}^{\nu/2} V_{\nu,\eta} = 2^{\nu}$. Значит $\sum_{\eta=1}^{s} V(s)_{\nu,\eta} < 2^{\nu}$, т.е.

$$2\sum_{\xi=1}^{((\nu/2)-1)} \ell og_2\,\xi - \sum_{\xi=(\nu/2)}^{((\nu/2)+1)} \ell og_2\,\xi - 2\sum_{\xi=1}^{((\nu/4)-1)} \ell og_2\,\xi - \sum_{\xi=(\nu/4)}^{((\nu/4)+1)} \ell og_2\,\xi - 2\sum_{\xi=1}^{(\nu/4)} \ell og_2\,\xi < \nu.$$

Следовательно, максимальное значение количества информации $\overline{H}(\nu/2)_{\nu,\nu/4}$ будет меньшим единицы. Условие (1) выполняется. По аналогии доказывается, что условие (1) выполняется для нечетного ν. Таким образом, двоичные матрицы данных имеют комбинаторную избыточность, обусловленную ограничениями на число серий единиц.

Выводы

Построена комбинаторная модель оценки информативности двоичных матриц данных. Обосновано, что учет закономерностей по числу серий единиц позволяет сократить избыточность в условиях не стационарности статистических свойств исходных фрагментов данных.

Научная новизна состоит в том, что получила дальнейшее развитие методика оценки информативности двоичных данных на основе построения комбинаторной информационной модели двоичных массивов. В отличии от других моделей учитывается ограничений на число серий единиц в двоичных последовательностей плавающих размеров. Это позволило обосновать, что учет закономерностей по числу серий единиц позволяет сократить избыточность в условиях не стационарности статистических свойств исходных данных.

Список литературы

1. Прэтт У. Цифровая обработка изображений: Том 1, 2. – М.: Мир, 1985. – 736 с.

2. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / В. Ватолин, А. Ратушняк и др. – М.: ДИАЛОГ – МИФИ, 2002. – 384 с.

3. Королев А.В., Баранник В.В. Оценка количества информации изображения по числу серий одинаковых элементов // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ, 2002. – Вип. 2 (18). – С. 43-46.

4. Баранник В.В., Юдин А.К. Рекуррентное двухпризнаковое двоичное полиадическое кодирование // Открытые информационные и компьютерные интегрированные технологии. – Х.: НАКУ «ХАИ», 2006. – Вып. 33. – С. 22-28.

Поступила в редколлегию 12.08.2008

Рецензент: д-р техн. наук, проф. Ю.В. Стасев, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

ОЦІНКА ІНФОРМАТИВНОСТІ ДВІЙКОВИХ МАСИВІВ НА ОСНОВІ КОМБІНАТОРНОГО ПІДХОДУ

В.В. Баранник, Г.В. Хаханова

Висловлюються основні етапи методики оцінки інформативності двійкових даних на основі побудови комбінаторної інформаційної моделі двійкових масивів. Дана методика враховує обмеження на число серій одиниць в двійкових послідовностях плаваючих розмірів. Це дозволило обґрунтувати, що облік закономірностей за числом серій одиниць дозволяє скоротити надмірність в умовах не стаціонарності статистичних властивостей початкових даних. Показано, що даний підхід до оцінки інформативності двійкових даних забезпечує скорочення надмірності в умовах: рівноімовірних символів алфавіту; для якнайгірших умов виявлених структурних закономірностей.

Ключові слова: *число серій одиниць, інформаційна модель, стиснення даних.*

ESTIMATION OF INFORMING OF BINARY ARRAYS ON THE BASIS OF COMBINATORY APPROACH

V.V. Barannik, A.V. Hahanova

The basic stages of method of estimation of informing of binary information are expounded on the basis of construction of combinatory informative model of binary arrays. This method takes into account limit on the number of cerouss of units in binary sequences of floating sizes. It allowed to ground, that the account of conformities to the law on the number of cerouss of units allows to shorten surplus in the conditions of not stationarity of statistical properties of basic data. It is rotined that this fitting to for the estimations of informing of binary information are provided by reduction of surplus in terms: equal probabilities characters of alphabet; for the worst terms of the exposed structural conformities to the law.

Keywords: *number of series of units, informative model, compression of data.*