Числовая оценка Колмогоровской сложности. Определение вероятности через смену событий.

Филатов О.В.

Филатов Олег Владимирович / Filatov Oleg Vladimirovich - инженер-программист НТЦ Модуль, г. Москва, fil_post@rambler.ru

Аннотация: *Определение Колмогоровской сложности через: распределение составных событий в случайной бинарной последовательности, через комбинаторные распределения элементарных событий в словах, и через инверсии в словах; признаки нахождения последовательности в состоянии Колмогоровской сложности; определение понятия «вероятность» через комбинаторное распределение инверсий в словах; классификация случайных слов по вероятностям частот их инверсий.*

Abstract: *The definition of Kolmogorov complexity through the distribution of composite events in a random binary sequence through combinatorial distribution of elementary events in words and by the inversion in the words; signs of finding a sequence able to Kolmogorov complexity; the definition of "probability" through combinatorial distribution of inversions in the words; Classification of random words on probable frequency of inversions.*

Ключевые слова: *Колмогоровская сложность, случайная последовательность, бинарная последовательность, потоковая последовательность, элементарное событие, составное событие.*

Keywords: *Kolmogorov complexity, random sequence, the binary sequence, threading sequence, elementary event, a composite event.*

Сокращения: СБФ - случайный бинарный файл; п-ть – последовательность; ПП - потоковая последовательность.

Введение

Программа сжатия (архивации) данных после каждого этапа сжатия может заново начать сжимать (пытаться сжимать) только что сжатый файл. Что может служить признаком остановки процесса сжатия, для программы сжимающей данные? Пределом сжатия является образование у итогового (сжатого) файла структуры Колмогоровской сложности. То есть, если программа архиватора зафиксировала в полученном сжимаемом файле структуру соответствующую структуре случайного бинарного файла, то работу по дальнейшему сжатию нужно прекращать, так как достигнута Колмогоровская сложность. Так как при достижении данными (сжатым файлом) Колмогоровской сложности самым коротким описанием бинарного кода является сам этот бинарный код. Колмогоровской сложностью (пределом сжимаемости) обладают все достаточно длинные бинарные файлы, полученные путём последовательной генерации и записи случайных бинарных событий. То есть, случайные бинарные файлы не сжимаемы. В статье рассмотрены различные признаки Колмогоровской сложности, обнаружение которых явится сигналом сжимающей программе к прекращению работы.

F - потоковая последовательность (ПП) [1, 2, 3, 4], случайная бинарная последовательность. Число случайных бинарных событий в *F*, на любой момент времени, конечно. ПП достигает бесконечного числа событий за бесконечно большое время (бесконечная случайная бинарная п-ть). Иными словами – ПП постоянно увеличивает число составляющих её элементарных событий (элов). Остановившиеся в своём росте потоковые последовательности являются конечными бинарными последовательностями или бинарными файлами (программная терминология).

Потоковая последовательность *F* образуется элементарными событиями - элами. Для бинарных последовательностей элы отличаются от бит только концептуально – результатом падения монеты являются

элементарные события, а не биты. Но уже при бросании кубика с шестью гранями элы принципиально отличаются от битов. Результаты выпадения кубика можно описывать в элах, но в одноразрядных битах результаты выпадения кубика описывать уже нельзя. Эл, описывающий выпадение сторон кубика может принимать шесть возможных значений, а бит всегда только два значения.

Основная часть

Составные события файла.

Бинарные случайные последовательности образуются из одинаковых случайных элементарных событий – элов. Элы группируются в составные события [1, 2, 3, 4]. Все длинные бинарные случайные последовательности обладают Колмогоровской сложностью (структурой) и состоят из составных событий, численность составных событий рассчитывается по ф.1.

Примеры составных событий ^{n}S. Составные события первой моды [1, 2, 4] $^{n=1}$S: «1», «0». Составные события второй моды $^{n=2}$S: «11», «00». Составные события третьей моды $^{n=3}$S: «111», «000». Составные события четвёртой моды $^{n=4}$S: «1111», «0000». И так далее, для других мод.

Такой принцип образования составных событий позволяет любую случайную последовательность описать этими событиями [1, 2, 3, 4]. Численность составных событий в потоковой (бинарной) последовательности зависит от их длин *n* и числа элементарных событий (бросков монеты) *N*. Ф.1 связывает $^{n}S_N$ - число составных событий n – ой моды с числом бросков монеты *N* (элементарных бинарных событий).

$$^{n}S_N = \frac{N}{2^{n+1}} \qquad \text{Ф. 1}$$

Где: n – длина составного события (номер моды), N – число бросков монеты (элементарных бинарных событий).

Составные события ${}^{n}S_N$ являются упрощающей абстракцией. Их применение избавляет от уточнения, что именно выпало нули или единицы. Составные события сокращают разнообразие анализируемых данных в два раза. Постулаты С. Голомба можно было бы считать логическими предшественниками для составных событий и ф.1 [4].

Имея длину, составные события не имеют значений [1,2,3]. Пример. Составные события: «111», «000» являются одинаковыми трёх позиционными образованиями «***» - обозначаемые как ${}^{n=3}S$. Такая неразличимость очень удобна (вместо двух типов данных - один тип). Это позволило свести строение любой бинарной и потоковой п-ти к ф.1.

Ф.1.1 описывает баланс между длиной бинарной последовательности N и $n \cdot {}^{n}S_N$ - числом элементарных событий (элов) находящихся во всех составных событиях последовательности. Они должны совпасть:

$$\sum_{n=1}^{\infty} n \cdot {}^{n}S_N = \sum_{n=1}^{\infty} \frac{n \cdot N}{2^{n+1}} = N \qquad \text{Ф. 1.1}$$

Отсюда, $\frac{n}{2^{n+1}}$ - вероятность выпадения составного события ${}^{n}S$ в F п-ти. Число составных событий S_N в пос-ти стремится к $N/2$ [3], ф.1.2:

$$S_N = \sum_{n=1}^{\infty} {}^{n}S_N = \sum_{n=1}^{\infty} \frac{N}{2^{n+1}} = \frac{N}{2} \qquad \text{Ф. 1.2}$$

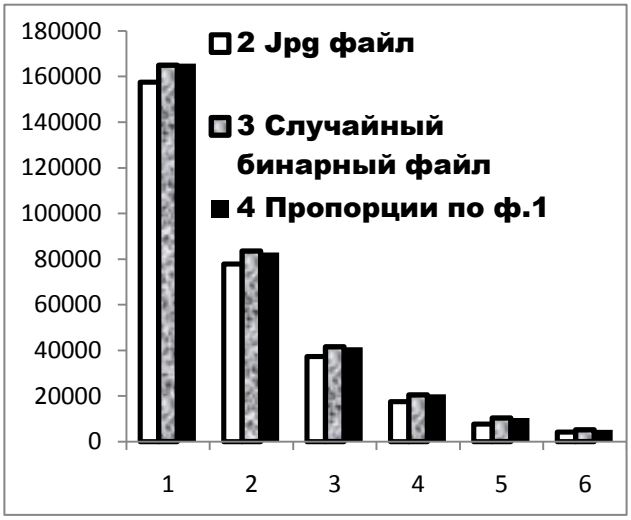
Под структурой пос-ти будем понимать гистограмму (таблица 1, столбец 7), или распределение составных событий (таблица 1, столбцы 2 - 6).

Для обнаружения структуры файла по составным событиям, он просматривается от начала к концу (столбец 3). Все найденные составные события файла сортируют по длинам, в итоге получают гистограмму их

распределения. Далее, по ф.1 производят расчёт числа составных событий для п-ти длины N (столбец 4). Из полученных по ф.1 значений составляют гистограмму распределений мат. ожиданий для составных событий длин n. Гистограммы сравнивают. Если для каждой из длин составных событий отклонение расчётного числа от обнаруженного находится в рамках допуска, то сжатие файла закончено - Колмогоровская сложность, с заданной уровнем точности, достигнута.

Таблица 1.

	Число ${}^{n}S_{(N=662800)}$ в файлах					
1	2 jpg	3 r n d	4 Ф.1	5 txt	6 txt.zip	7
1	**157486**	**164969**	**165700,**	372610	179414	
2	**77788**	**83513**	**82850,**	113283	87227	
3	**37340**	**41547**	**41425,**	17448	43655	
4	**17451**	**20456**	**20712,**	2576	22253	
5	**7688**	**10338**	**10356,**	179	10712	
6	**4219**	**5157**	**5178,**	11	5662	
7	**5918**	**2580**	**2589,**	2	2776	
8	**1027**	**1246**	**1294,**		1438	
9	**558**	**681**	**647,**		666	
10	**298**	**359**	**323,**		348	
11	**152**	**163**	**161,**		165	
12	**83**	**75**	**80,**		87	
13	**59**	**43**	**40,**		31	
14	**57**	**30**	**20,**		13	
15	**37**	**10**	**10,**		11	
16	**21**	**3**	**5,**		5	
17	**16**	**1**	**2,**		4	
18	**6**	**3**	**1,**		1	
S_N	310204	331174	331398,	506109	354476	



С возможностью сравнения структур файлов (столбцы 1,2,5,6) с рассчитанными по ф.1 мат. ожиданием (столбец 4) стало возможно говорить о числовых оценках структуры Колмогоровской сложности. Поэтому, таблица 1 содержит данные для числовой оценки структуры Колмогоровской сложности (столбцы: 2 - 6).

Числовой оценкой структуры Колмогоровской сложности является расхождение между числом составных событий длины *n* (столбцы 2,3,5,6) с составными событиями длины *n* из столбца 4 (рассчитанными по ф.1 для N=662800). Такое *N* взято потому, что именно в это количество бит была сжата по jpg формату картинка, в столбце 2 приведены результаты анализа её файла. По ф.1 получено эталонное распределение, для случайного бинарного файла с N=662800 (столбец 4) , обладающее Колмогоровской сложностью. Анализ чисел событий (столбцы: 2, 4), показывает, что в области малых длин jpg формата числовые оценки соответствуют Колмогоровской сложности. Полное число составных событий в jpg файле равно 310406.

В столбце 5 дано распределение составных событий в txt файле (размер 662800 бита). Как видно, его составные события сильно отклонены от распределения по ф.1 (столбец 4). Это значит - файл сжимаем дальше. Последующее сжатие txt файла архиватором zip, уменьшило его размер в 7,5 раз, до 88235 бит. Структура сжатого txt файла (столбец 6) приняла пропорции соответствующие получаемым по ф.1(столбец 4). После сжатия txt число составных событий в файле сократилось с 506109 до 354476 (учтены события с n>18, не вошедшие в таблицу 1).

Для сравнения распределения составных событий соответствующих Колмогоровской сложности (столбец 4) был получен на генераторе случайных чисел бинарный файл (раскладка в столбце 3), длиной в 662800 эла (бита). Величины составных событий, рассчитанные по ф.1 (Колмогоровская сложность), хорошо соответствуют значениям в столбце 3, этого случайного файла.

Выше была рассмотрена идея построчного сравнения составных событий в исследуемом файле с идеальным распределением (рассчитанным по ф.1). Такое построчное сравнение является простейшей числовой оценкой структуры Колмогоровской сложности. Из-за ограниченного формата статьи развитие этой темы прекращается.

В столбце 7 размещена гистограмма столбцов: 2, 3, 4.

Константа 0,5 – уровень Колмогоровской сложности

Иногда по одному числу (а не по рядам, как в таблице 1) желательно судить о степени случайности последовательности. В ниже приведённом подразделе вводится константа, связанная с распределениями составных событий по ф.1, степень отклонения от которой является показателем приближения к уровню Колмогоровской сложности. Отклонение от этого числа говорит о возможности дальнейшего сжатия файла.

В файле, с Колмогоровским уровнем сложности, число составных событий S_N стремится к $\frac{N}{2}$, ф.1.2. Это является следствием теоремы «Об амплитудно-частотной характеристике идеальной бинарной случайной последовательности» [3, 4]. Число составных событий в случайной последовательности $S_N \rightarrow \frac{N}{2}$ является показателем приближения структуры файла (последовательности) к Колмогоровской сложности [2, 3]. Чем ближе S_N к $\frac{N}{2}$, тем меньше этот файл можно сжать.

Для примера оценим по критерию $S_N \rightarrow \frac{N}{2}$ два файла из таблицы 1.

В случайном бинарном файле (столбец 3) N=662800, отсюда: $\frac{N}{2}$ = 331400 ≈ 331174 = S_N. Можно сделать вывод, что файл больше несжимаем.

В текстовом файле (столбец 5). N=662800, отсюда: $\frac{N}{2}$ = 331400 ≠ 506109 =S_N. Из неравенства $S_N \neq \frac{N}{2}$ следует, что файл сжимаем.

Избавимся в оценке сжимаемости файла ($S_N \rightarrow \frac{N}{2}$) от зависимости от N. Разделив S_N на N, получим ф.2:

$$\frac{S_N}{N} = \frac{N}{2} \cdot \frac{1}{N} = \frac{1}{2} \qquad \text{Ф. 2}$$

Малое отклонение от константы 0.5 явится признаком достижения Колмогоровской сложности в сжимаемом файле, ф.2.1:

$$\Delta = \left|\frac{S_N}{N} - \frac{1}{2}\right| \to 0; \quad \Delta \leq k_f \qquad \text{Ф. 2.1}$$

В ф.2.1 величина коэффициента k_f выбирают таким образом, чтобы последовательностям с Δ меньшим k_f можно было присвоить Колмогоровский уровень сложности.

Пример. В таблице 1 в столбце 3 дано распределение по длинам составных событий в случайном бинарном файле (СБФ) длиной: *N = 662800* (эл, бит). Сумма всех составных событий в СБФ: S_N =*331174*. Делим по ф.2 составные события S_N на число эл: *331174 / 662800 = 0,4997* . Делим составные события S_N jpg файла на число эл: *310406 / 662800 = 0,4683*. Оценим по модулю отклонения: $\Delta_{\text{СБФ}} = |0{,}5 - 0{,}4997| = 0{,}0003$; $\Delta_{\text{jpg}} = |0{,}5 - 0{,}4997| = 0{,}0317$. Отклонение интегрального сквозного показателя $\Delta_{\text{СБФ}}$ от 0, оказалось меньше чем Δ_{jpg} для jpg файла. У не сжатого текстового файла (txt), столбец 5, Δ_{txt} = 0,2524. Чем больше Δ (отклонение от 0), тем эффективнее может быть сжат файл.

Фрагменты с элами.

Порой аппаратура имеет малую разрядность, что ограничивает размер исследуемых файлов. Поэтому файл исследуют по частям, то есть файл делят на фрагменты. В ниже приведённом подразделе описано определение Колмогоровской сложности через анализ содержимого фрагментов файла (последовательности).

Последовательность (файл) разбивают на фрагменты. В каждом фрагменте считают количество только нулей (или единиц), набирая статистику по всем фрагментам. По набранной статистике строят распределение нулей (единиц) и сравнивают с теоретически рассчитанным

(комбинаторным) их распределением. По результату сравнения Δ_e судят о достижении Колмогоровской сложности (с заданной точностью k_f). Если значения $\Delta_e \leq k_f$, ф.4.1, то файл, с заданной степенью точности, обладает Колмогоровской сложностью и более не сжимаем.

Таблица 2.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
№№	${}^{\mathrm{m}}_{\mathrm{n}}e$ (*m elementary events from n*)						**СБФ**	**МО**	
m, i	jpg«0»	jpg «1»	**СБФ**«0»	**СБФ**«1»	${}^{m}_{n}M(N)$	C_n^m	${}^{n=8}I[i]$	${}^{n=8}I(i)$	$I(C_i^n)$
0	334	5590	302	357	323,	1	659	647,	2
1	1742	6561	2598	2592	2589,	8	4500	4530,	14
2	7318	8090	9021	9125	9061,	28	13584	13592,	42
3	16455	15955	18303	17870	18123,	56	22609	22654,	70
4	20805	20805	22682	22682	22654,	70	22933	22654,	70
5	15955	16455	17870	18303	18123,	56	13532	13592,	42
6	8090	7318	9125	9021	9061,	28	4405	4530,	14
7	6561	1742	2592	2598	2589,	8	628	647,	2
8	5590	334	357	302	323,	1	нет	нет	нет
Sum	**82850**	**82850**	**82850**	**82850**	**82850**	**256 = 2^8**	**82850**	**82850**	**256 = 2^8**
N = 662800; n=8; Sum = $\frac{N}{n}$ S в СБФ = 331175; S(теор.) в СБФ = 331400							$\frac{N}{n \cdot 2^n} = 323{,}6328125$		

В исследуемых фрагментах, длиной *n* считают выпавшие элементарные события *m* (элы: нули или единицы). По аналогии с составными событиями ${}^{n}S$ (ф.1) введём обозначение множества элов - ${}^{m}_{n}e$ (Elementary events), символ *m* – обозначает число выпавших одинаковых элементарных событий в фрагменте длиной из *n* элементарных событий. Пример: ${}^{m}_{8}e$ - *m* элементарных событий (эл) из восьми; ${}^{3}_{8}e$ - три эл из 8 («10010001», «00011100», «00101111»); ${}^{0}_{8}e$ - ноль эл (либо «0» либо «1» отсутствует в фрагменте) из 8 («00000000», «11111111»).

Полярная симметрия (таблица 2). Видно, что ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}$ – распределение описывает: и распределение нулей ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0$ среди единиц (столбец 2) и распределение единиц ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$ среди нулей (столбец 3), так как ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0 = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$

(столбцы 2, 3). Поэтому, если не требуется точно указывать ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0$ и ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$, то достаточно ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}$. Свойство: ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0 = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$ назовем «Полярной симметрией».

Центральная симметрия (таблица 2). Видно, что в ${}^{\mathrm{m}}_{\mathrm{n}}e$ – эл распределениях, в каждом столбце: 2 – 7, присутствует симметрия относительно центра столбца ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$ (не путать с полярной симметрией в соседних столбцах: ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0 = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$). Симметрия в столбцах 2, 3 описывается по ф.3.1; в 4, 5 – по ф.3.2; в 6, 7 – по ф.3.3:

(Ф. 3.1): ${}^{\mathrm{m}}_{\mathrm{n}}e \approx {}^{n-m}_{\mathrm{n}}e$; (Ф. 3.2): ${}^{\mathrm{m}}_{\mathrm{n}}e \to {}^{n-m}_{\mathrm{n}}e$; (Ф. 3.3): ${}^{\mathrm{m}}_{\mathrm{n}}e = {}^{n-m}_{\mathrm{n}}e$

Примеры на центральную симметрию (относительно центра столбца) из таблицы 2. Столбец 2: $({}^{1}_{8}e = 1742) \neq ({}^{8-1}_{8}e = 6561)$; ст-ц 3: $({}^{1}_{8}e = 6561) \approx ({}^{8-1}_{8}e = 1742)$. Ст-ц 4: $({}^{1}_{8}e = 2598) \to ({}^{8-1}_{8}e = 2592)$; ст-ц 5: $({}^{1}_{8}e = 2592) \to ({}^{8-1}_{8}e = 2598)$. Ст-ц 6: $({}^{1}_{8}e = 2589) = ({}^{8-1}_{8}e = 2589)$.

О сжимаемости файла можно судить по наличию (отсутствию) оси симметрии по центру столбца в ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}$ - распределении (таблица 2): у чётных фрагментов – одна вершина в центре, у не чётных фрагментов – две вершины в центре. Если это не так (${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} \approx {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$; ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} \neq {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$), то файл сжимаем. В случаях: ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} \to {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$; ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$ - файл близок или обладает Колмогоровской сложностью (мало сжимаем либо не сжимаем).

В столбце 6, таблицы 2 рассчитаны по ф.4 - мат. ожидания ${}^{n}_{m}M(N)$ последовательности из N = 662800 событий:

$$ {}^{m}_{n}M(N) = \frac{C_n^m}{2^n} \cdot \frac{N}{n} = \frac{n!}{m!\,(n-m)!} \cdot \frac{1}{2^n} \cdot \frac{N}{n} \qquad \text{Ф. 4} $$

Мат. ожидание для выпадения нулей равно мат. ожиданию для выпадения единиц: ${}^{m}_{n}M0(N) = {}^{m}_{n}M1(N) = {}^{m}_{n}M(N)$ – полярная симметрия ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e}0 = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}1$, центральная симметрия ${}^{\mathrm{m}}_{\mathrm{n}}\mathrm{e} = {}^{\mathrm{n}-\mathrm{m}}_{\mathrm{n}}\mathrm{e}$. Поэтому будем писать в ф.4 обозначение ${}^{m}_{n}M(N)$ без уточнения, что выпало («0», «1»).

В столбцах 2,3 таблицы 2 представлено распределение ${}^{m}_{\mathrm{n}}e$ для некоего jpg файла. Оно, как и все реальные распределения ${}^{m}_{\mathrm{n}}e$ не совпадает с мат. ожиданием (ф.4): ${}^{m}_{\mathrm{n}}e_N \neq {}^{m}_{n}M(N)$. Сравнение ${}^{m}_{\mathrm{n}}e_N$ и ${}^{m}_{n}M(N)$ распределений по ф.4.1 выявляет коэффициент рассогласования формы k_f:

$$\Delta_e = \sum_{m=0}^{m=n} \left(|{}^{m}_{\mathrm{n}}e_N - {}^{m}_{n}M(N)| : \frac{N}{n} \right) = \frac{n}{N} \cdot \sum_{m=0}^{m=n} |{}^{m}_{\mathrm{n}}e_N - {}^{m}_{n}M(N)| \qquad \text{Ф. 4.1}$$

$$\Delta_e \leq k_f$$

Величина коэффициента k_f выбирается таким образом, чтобы последовательностям с Δ_e меньшим k_f можно было присвоить Колмогоровский уровень сложности.

Распределения в таблице 2, столбцы 2 – 6, связаны между собой равенством: $\sum_{m=0}^{m=n} {}^{m}_{\mathrm{n}}e_N = \sum_{m=0}^{m=n} {}^{m}_{n}M(N) = \frac{N}{n}$. Распределение в столбце 7 описывается равенством: $\sum_{m=0}^{m=n} C_n^m = 2^n$. Связь между C_n^m и ${}^{m}_{n}M(N)$ в столбцах 6, 7: ${}^{m}_{n}M(N) = C_n^m \cdot \frac{N}{n \cdot 2^n}$. Для последовательности, стремящейся к Колмогоровской сложности: ${}^{m}_{\mathrm{n}}e_N \to {}^{m}_{n}M(N)$, отсюда: ${}^{m}_{\mathrm{n}}e_N \to C_n^m \cdot \frac{N}{n \cdot 2^n}$.

По величине Δ определяют возможность дальнейшего сжатия файла. Пример. Рассчитать коэффициент рассогласования формы Δ, ф.4.1, для jpg, СБФ файлов и мат. ожидания (таблица 2, столбцы 2, 4, 6).

$\Delta_{\mathrm{jpg}} = \frac{8}{662800} \cdot \sum_{\substack{m=0, \\ n=8}}^{m=8} |{}^{m}_{\mathrm{n}}e_N - {}^{m}_{n}M(N)| = \frac{8}{662800} \cdot 18496 = 0{,}223$. Отклонение от идеальной формы более 22%. Файл не достаточно соответствует Колмогоровской сложности. Возможно уменьшение его размера на 20%.

$\Delta_{\mathrm{СБФ}} = \frac{8}{662800} \cdot 641 = 0{,}008$. Для $\Delta_{\mathrm{СБФ}}$ отклонение от идеальной формы меньше 1%. Файл в высокой степени соответствует Колмогоровской сложности.

$\Delta_M = \frac{8}{662800} \cdot 0 = 0$. Для Δ_M отклонение от идеальной формы равно нулю.

Фрагменты с инверсиями.

В таблице 2, столбцы: 4, 5 дано распределение нулей и единиц в 8 разрядном слове. Но, постоянно указывать какое внедрение («0» в «1» или «1» в «0») производится внутрь несущей основы слов неудобно. Например, один ноль можно восемью различными способами разместить среди единиц $C_{m=1}^{n=8}$: «011..1», «101..1», «1101..1», .. «111..10» (таблица 2), но и семь единиц $C_{m=7}^{n=8}$ можно разместить восемью способами на подложке из нулей (от подложки останется видимым только один ноль). То есть $C_{m=1}^{n=8} = C_{m=7}^{n=8}$. Поэтому в столбцах 4, 5 наблюдается полярная (перекрёстная) симметрия: значение в строке m=0 столбца 4 равно значению в строке m=8 столбца 5, значение в строке m=1 столбца 4 равно значению для m=7 столбца 5, и т.д.

Избавиться от употребления цифр «0», «1» можно используя переходы от нулей к единицам («01») и от единиц к нулям («10») – инверсии. Переходы «01» и «10» образуют «Инверсные события». Как и предыдущие логические абстракции: составные события, элы [1,2,4] – инверсные события могут обозначаться с полярностью перехода, но в данной статье раскрываются, наоборот, преимущества скрытия излишней информации. Поэтому описывается применение только неполярных «Инверсных событий».

В слове длиной n можно комбинаторным путём разместить $I(C_i^n)$ инверсий ***i***. Формула для расчета числа инверсионных комбинаций ф.5:

$$I(C_i^n) = \frac{2 \cdot (n-1)!}{i! \cdot (n-1-i)!}\ ;\ \textit{где}\ n = 1,2,3,\dots;\ i \le n-1 \qquad \text{Ф. 5}$$

$I(C_i^n)$ – число комбинаций содержащих внутри себя по i инверсий; n – длина слова; i – число инверсий внутри слова.

Поскольку число отрезков с элами $\sum C_n^m$ и с инверсиями $\sum I(C_i^n)$ одно и тоже, то: $\sum C_n^m = \sum I(C_i^n) = 2^n$.

Ф.5.1 – формула расчёта мат. ожидания инверсий ${}^{n}I(i)$ для слов длиной n, содержащих $i \leq n-1$ инверсий, в файле из N эл (событий), получается путём умножения комбинаторного распределения $I(C_i^n)$ на коэффициент связи с числом событий в файле: $\frac{N}{n \cdot 2^n}$.

$$ {}^{n}I(i) = I(C_i^n) \cdot \frac{N}{n \cdot 2^n} \qquad \text{Ф. 5.1} $$

$$ \textit{где } n = 1,2,3,\ldots;\ i \leq n-1 $$

Переход от найденных фрагментов ${}^{n}I[i]$ с i инверсиями в файле Колмогоровской сложности (столбец 8 таблицы 2) к комбинаторному распределению $I(C_i^n)$ (столбец 10, ф. 5) осуществляется делением числа ${}^{n}I[i]$ на $\frac{N}{n \cdot 2^n}$: $\quad I(C_i^n) \cong {}^{n}I[i] : \frac{N}{n \cdot 2^n} = \frac{{}^{n}I[i] \cdot n \cdot 2^n}{N}$.

Ф.5.2 - вероятность выпадения фрагментов длины n с i переходами в них:

$$ p(C_i^n) = \frac{(n-1)!}{i! \cdot (n-1-i)!} \cdot \frac{1}{2^n} \qquad \text{Ф. 5.2} $$

$$ \textit{где } n = 1,2,3,\ldots;\ i \leq n-1 $$

Сумма вероятностей $p(i,n)$ в ф.5.2 равна 1: $\sum_{i=0}^{i=n-1} \frac{2 \cdot (n-1)!}{2^n \cdot i! \cdot (n-1-i)!} = 1$, и сумма вероятностей по 3.1 равна 1. Поэтому верно равенство: $\sum_{i=0}^{i=n-1} \frac{(n-1)!}{2^{n-1} \cdot i! \cdot (n-1-i)!} = \sum_{m=0}^{m=n} \frac{n!}{m!(n-m)!} \cdot \frac{1}{2^n} = 1$, это равенство перепишем в виде ф.5.3 (пример: таблица 2, столбцы 7, 10):

$$ \sum_{i=0}^{i=n-1} \frac{2 \cdot (n-1)!}{i! \cdot (n-1-i)!} = \sum_{m=0}^{m=n} \frac{n!}{m!\,(n-m)!} = 2^n \qquad \text{Ф. 5.3} $$

Выпишем левую часть ф.5.3: $\sum_{i=0}^{i=n-1} \frac{(n-1)!}{i! \cdot (n-1-i)!} = 2^{n-1}$. Заменив n – 1= k, получим равенство: $\sum_{i=0}^{i=k} \frac{k!}{i! \cdot (k-i)!} = 2^k$; где: $k = n-1;\ 0 \leq i \leq k$. Таким образом, можно переписать ф.5.3 в виде ф.5.4:

$$\sum_{i=0}^{i=k} \frac{2 \cdot k!}{i! \cdot (k-i)!} = \sum_{m=0}^{m=n} \frac{n!}{m!\,(n-m)!} = 2^n \qquad \text{Ф. 5.4}$$

где: $k = n-1;\ \ 0 \le i \le k;\ \ m \le n;\ \ n > 0$

Комбинаторные формулы для инверсий и элов (комбинаций монет) имеют один вид и одну сумму комбинаций 2^n, ф.5.4.

Естественно, что число фрагментов файла $\frac{N}{n}$ не зависит от способа учёта их содержимого (инверсий: $\sum_i^{(i=n-1)} {}_i^n I(N)$, элов: $\sum_{m=0}^{m=n} {}_m^n M(N)$). Сумма всех фрагментов содержащих внутри себя *i* инверсий ($0 \le i \le n-1$): $\sum_i^{(i=n-1)} {}_i^n I(N) = \sum_i^{(i=n-1)} \frac{2\cdot(n-1)!}{i!\cdot(n-1-i)!} \cdot \frac{N}{n\cdot 2^n} = \frac{N}{n}$. Отсюда: $\sum_{m=0}^{m=n} {}_m^n M(N) = \sum {}_n^m M0(N) = \sum {}_n^m M1(N) = \sum_i^{(i=n-1)} {}_i^n I(N) = \frac{N}{n}$.

Коэффициент рассогласования формы для инверсий. По аналогии с ф.4.1 введём формулу ф.5.5 - расчёта отклонения Δ_I от ожидаемой формы (Колмогоровской сложности) инверсий внутри фрагментов длины *n*, фала из *N* эл (элементарных событий). Обозначим через ${}_i^n I_N$ – распределение инверсий в неком файле, а через ${}_i^n I(N)$ – мат. ожидание выпадения отрезков с *i* инверсиями в них. Тогда модуль разности $\Delta_i = |{}_i^n I_N - {}_i^n I(N)|$ - будет расхождением между числом найденных отрезков с *i* инверсиями в них и ожидаемым числом этих отрезков в файле с Колмогоровской сложностью. Сумма по всем Δ_i делённая на число отрезков в файле даст отклонение от Колмогоровской сложности Δ_I :

$$\Delta_I = \frac{n}{N} \cdot \sum_{i=0}^{i=n-1} |{}_i^n I_N - {}_i^n I(N)| \qquad \text{Ф. 5.5}$$

где: $0 \le i \le n-1$

Пример на соответствие Колмогоровской сложности Δ_I, ф.5.5, файла (таблица 2, столбец 8). Файл создан путём компьютерной генерации

случайных бинарных значений N = 662800 (эл). Разделим файл на равные отрезки длиной: $n = 8$. Обозначим: $k = \frac{n}{N} = \frac{8}{662800}$ и рассчитаем сумму значений Δ_i , где $0 \leq i \leq n-1$, вычитая из соответствующих значений столбца 8, ${}_i^n I_N$, таблицы 2, значения столбца 9, ${}_i^n I(N)$, с последующим суммированием: $\Delta_i = k \cdot \Sigma\Delta_i = k \cdot 578 = 0{,}007$. Из полученного результата Δ_i = 0,007 видно, что отклонение от идеальной формы меньше 1%. Файл в высокой степени соответствует Колмогоровской сложности. Отметим, что ранее рассчитанный по распределению элов коэффициент отклонения от Колмогоровской формы $\Delta_{\text{СБФ}} = 0{,}008$.

Обсуждение.

Для выявления степени соответствия структуры файла Колмогоровской сложности удобнее использовать фрагментацию файла, а не выявление его составных событий по ф.1. Для описания фрагментированного случайного бинарного файла (СБФ) в терминах «0» / «1» (столбцы 4, 5, таблица 2) требуется указывать, что нули «0» распространяются среди единиц «1» или наоборот. Хотя, для многих задач это неважно, так как полярная (перекрёстная) симметрия в распределениях (столбцы 4, 5) обеспечивает равенство сумм по столбцам 4, 5. Распределение, оперирующее инверсиями (столбец 9, таблица 2), а не нулями и единицами, лишено путаницы в описании того, что в чём распределено (нули в единицах или наоборот). Причём итоговая сумма по столбцу 9 равна итоговым суммам по столбцам 2 - 6. Это равенство итоговых сумм приводит к мысли, что описания Колмогоровской сложности по распределению элементарных событий в фрагментах последовательности (столбцы 2 - 6), может быть заменено на распределение инверсий в фрагментах последовательности (столбцы 8, 9). То есть, предлагается использовать для расчёта вероятностей выпадения монеты не комбинации из цифр / гербов («0» / «1»), а число инверсий в фрагментах - серии выпадений монеты из n событий (n > 1).

Построение вероятности выпадений на учёте инверсий, $p(C_i^n)$, ф.5.2, кроме логических преимуществ (нет путаницы нулей и единиц), даёт ещё преимущество в большей информативной ёмкости. А именно, при чётном количестве выпавших инверсий, в последнем событии серии (фрагмента), монета выпадает в том же состоянии, что и в первом событии серии. При нечётном количестве инверсий, в последнем событии серии, монета выпала в ином состоянии, чем при первом броске. То есть, при использовании инверсионных вероятностей, мы в дополнение к самой величине вероятности, получаем знание о связи (состояниях) первого и последнего события в описываемом фрагменте. Во всём остальном информационность описания при помощи инверсионных вероятностей $p(C_i^n)$, ф.5.2, равна информационности описания вероятности в терминах выпадения «0» / «1».

Надо отметить, что в случае, когда надо набрать статистику выпадений именно герба (цифры), то инверсные вероятности (описания) не эффективны. А если требуется найти соответствие Колмогоровской сложности, то наоборот, единичный учёт выпадения орлов (цифр) будет полностью не эффективен, так как он ведёт арифметический учёт выпадений, а потоковая последовательность (состоящая из составных событий), даже порезанная на фрагменты, учитывает значения соседних событий элементарных событий. Спереди и сзади, до и после выпавшего эла (элементарного события). Именно этот учёт соседних элементарных событий делает систему вероятностей, построенную на инверсиях $p(C_i^n)$, ф.5.2, более информативно ёмкой, по сравнению с простым учётом выпадения элементарных событий. Рассмотрим столбцы 4 (учёт элов) и 8 (учёт инверсий), в таблице 2. По обоим столбцам можно узнать о степени случайности файла (Колмогоровскую сложность). Но при инверсной организации описания выпадения монеты (столбец 8) для решения этой задачи требуется на одно значение (строку таблицы) меньше, нет строки 8, что является проявлением большей информационной ёмкости инверсного типа описания.

В инверсном представлении вероятностей $p(C_i^n)$, ф.5.2, оба элементарных события образующих инверсный перепад: «01» , «10» (не образующих инверсию: «00» , «11») одинаково равноправны и не зависят друг от друга. Но образуемое ими инверсное событие зависит от них обоих. Поэтому перейдя в пространство инверсных событий (вероятностей) всегда можно знать: равны ли между собой первое и последние события рассматриваемого фрагмента (последовательности).

На рисунке 1 представлены графики вероятностей для фрагментов длиной 8 элементарных событий ($n=8$). График 1 - вероятности выпадений m нулей «0» («1») внутри фрагмента длиной $n=8$ эл (элементарных событий). Вероятность выпадения фрагмента содержащего внутри себя m нулей или единиц ${}_n^m p$ рассчитывается по ф.6:

$$ {}_n^m p = \frac{C_n^m}{2^n} = \frac{n!}{m!\,(n-m)!} \cdot \frac{1}{2^n} \qquad \text{Ф. 6} $$

На графике 2, рисунок 1, дано распределение вероятностей $p(C_i^n)$, ф.5.2, для нахождения i инверсий в нутрии фрагмента длиной n эл (элементарных событий).

Базовые графики вероятностей в системах: «элы: 0,1» и «инверсия»

m, ***i***	график 1 (элы: "0", "1")	график 2 (инверсии)
0	0,00390625	0,0078125
1	0,03125	0,0546875
2	0,109375	0,1640625
3	0,21875	0,2734375
4	0,2734375	0,2734375
5	0,21875	0,1640625
6	0,109375	0,0546875
7	0,03125	0,0078125
8	0,00390625	

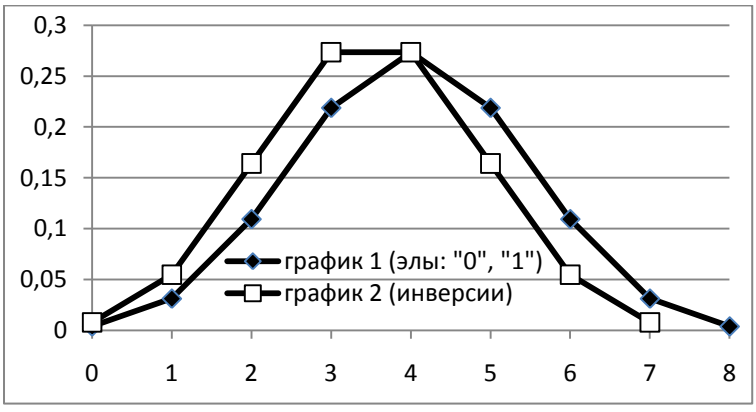


Рисунок 1

Графики 1, 2 совпадают в точке x=4. Это означает, что вероятность выпадения четырёх инверсий и четырёх нулей (единиц), в восьми разрядном слове, одинакова в обоих системах вероятностей (классической и инверсной).

В столбце «график 2» рассчитаны вероятности выпадений фрагментов (слов) содержащих внутри себя *i* инверсий (*i* =0, … , 8). Как видно распределение инверсионных вероятностей имеет центральную симметрию.

Графики вероятностей выпадения инверсий обладают центральной (вертикальной) симметрией, как и графики традиционной вероятности ${}_n^m p$, ф.6, для выпадений нулей (единиц) в фрагментах. Поэтому, восприятие вероятностных распределений для обоих вероятностных систем (основанной на орлах – решках ${}_n^m p$, ф.6, основанной на инверсиях элементарных событий $p(C_i^n)$, ф.5.2,) одинаково.

Классификация инверсионных фрагментов по вероятностным группам (Колмогоровской сложности).

В таблице 2, в столбцах 6, 9, показаны мат. ожидания (частоты встреч) ${}_n^m M(N)$ и ${}^{n=8} I(i)$, в достаточно длинных последовательностях, фрагментов с m элами в фрагменте, и с i инверсиями в фрагменте. Избавляясь от N зависимости, переходим к распределениям: C_n^m (столбец 7), $I(C_i^n)$ (столбец 9).

Так группа $I(C_{i=0}^n)=2$ содержит всего два слова, причём длина фрагментов n любая. Эти слова не содержат внутри себя ни одной инверсии, легко сжимаются и имеют наименьшую Колмогоровскую сложность. Перечень всех фрагментов группы $I(C_{i=0}^{n=8})$: «00000000», «11111111». Эти фрагменты предлагается считать принадлежащей группе с наименьшей вероятностью выпадения и с наименьшей, нулевой, Колмогоровскуой сложностью.

Симметричная ей по величине вероятности, группа $I(C_{i=n-1}^n)=2$, так же содержит всего два слова при любой длине фрагмента n. То же имеет низкую

Колмогоровскую сложность. Пример всех фрагментов группы $I(C_{i=7}^{n=8})$: «01010101», «10101010». Эти фрагменты предлагается считать принадлежащей группе с наименьшей вероятностью выпадения, но с первой Колмогоровской сложностью – так как по мнению автора сложность слов: «00000000» и «01010101» различна.

И так далее.

Например, число фрагментов в группе $I(C_{i=3}^{n=8})=56$, рассчитывается по ф. 5. Колмогоровская сложность каждого из пятидесяти шести фрагментов входящих в эту группу с тремя переходами выше, чем у фрагментов групп $I(C_{i=0}^{n})$ и $I(C_{i=n-1}^{n})$. Пример одного из 56 фрагментов группы с тремя инверсиями внутри $I(C_{i=3}^{n=8})$: «01000111».

Можно заметить, чем меньше вероятность выпадения группы $p(C_i^n)$, тем меньше Колмогоровская сложность входящих в неё фрагментов. Поэтому, для фрагментов длины *n* можно предложить в качестве меры Колмогоровской сложности использовать биномиальное распределение инверсий $I(C_i^n)$. Наибольшая Колмогоровская сложность, по этой мере, будет принадлежать фрагментам содержащие числа переходов равные и близкие к $i \approx n / 2$. А наименьшая Колмогоровская сложность будет принадлежать фрагментам содержащим числа переходов i равные и близкие к крайним значениям диапазона: *0, 1, 2, …, n-2, n-1, n*. Все фрагменты сортируются по вероятностным группам. Число таких групп зависит от чётности / не чётности длины фрагмента *n*.

Для нечётных длин фрагментов *n* число вероятностных групп $\frac{n+1}{2}$. В каждую группу входят две подгруппы со своей Колмогоровской сложностью, за исключением группы с наивысшей вероятностью выпадения фрагментов – в ней одна подгруппа с Колмогоровской сложностью.

Для чётных длин фрагментов *n* число вероятностных групп $\frac{n}{2}$. В каждую группу входят две подгруппы со своей Колмогоровской сложностью.

Образование стохастических подпоследовательностей из СБ n-ей.

Поскольку в работе широко использовались случайные бинарные последовательности и СБФ файлы, а в математических кругах идёт дискуссия об операциях над ними, то допустимо в конце статьи привести взгляд практика на эту тему (операции над СБ пос-ми).

Эксперименты по образованию стохастических подпоследовательностей (СПП) и кортежей из случайных последовательностей показали, что из достаточно длинной исходной п-ти можно брать по совершенно любому закону или без всякого закона любой её эл (элементарное событие) значение которого не должно быть известно до его взятия. Но этот эл может использоваться для построения СПП (кортежа) только один раз.

Выводы

В достаточно длинных случайных бинарных последовательностях Колмогоровская сложность описывается небольшим набором чисел.

Колмогоровскую сложность можно определять разными способами (анализ составных событий в непрерывном файле, анализ фрагментов при делении файла на фрагменты) и поэтому, в каждом из процессов определения, Колмогоровская сложность описывается своей собственной формулой. Так при сквозном обнаружении составных событий в случайной бинарной потоковой последовательности Колмогоровская сложность описывается формулами: ф.1, ф.2.1. При сборе данных о структуре

случайной бинарной последовательности путём анализа фрагментов заданной длины Колмогоровская сложность описывается формулами: ф.4, ф.4.1. При переходе в пространство инверсных событий Колмогоровская сложность характеризуется формулами: ф.5.1, ф.5.5.

Таким образом, используя выше описанные методы исследования бинарных последовательностей, структуру любой последовательности можно оценить на степень приближения к структуре последовательности обладающей Колмогоровской сложностью.

Известно несколько концептуальных определений понятия «вероятность». В работе было обращено внимание на удобство использования, в качестве вероятности, событий образованных инверсиями - сменами значений состояний двух последовательных элементарных событий. Таким образом, на основе механизма инверсий, вводится ещё одна трактовка понятия «вероятность», по крайней мере, для фрагментов бинарных случайных последовательностей.

Формулы математических ожиданий выпадений числа фрагментов с определённым количеством параметров: ф.4, ф.5.1 – описывают комбинаторную структуру (амплитудно-частотное распределение) случайных бинарных последовательностей. Эти формулы позволяют говорить, что последовательности Колмогоровской сложности порождаются комбинаторным распределительным законом. Сами комбинаторные распределения, рассчитанные по формулам ф.4, ф.5.1, можно сгруппировать по вероятностным группам. А вероятностным группам присвоить вероятности выпадения, и уровни Колмогоровской сложности.

Библиографический список

1. Филатов О. В., Филатов И.О., Макеева Л.Л. и др. «Потоковая теория: из сайта в книгу». Москва, «Век информации», 2014, с. 200.

2. Филатов О. В., Филатов И.О., статья «О закономерностях структуры бинарной последовательности», «Журнал научных публикаций аспирантов и докторантов», №5, 2014.
3. Филатов О. В., статья «Теорема «О амплитудно-частотной характеристике идеальной бинарной случайной последовательности», «Проблемы современной науки и образования», № 1 (31), 2015 г.
4. Филатов О. В., Филатов И.О. «Закономерность в выпадении монет – закон потоковой последовательности». Германия, Издательский Дом: LAP LAMBERT Academic Publishing, 2015, с. 268.