

ЭНТРОПИЯ И ИНФОРМАЦИЯ

Г. М. КОШКИН

Томский государственный университет

ENTROPY AND INFORMATION

G. M. KOSHKIN

The basic properties of entropy and information for discrete random objects are described. The problem of communications coding is discussed; the Shannon-Fano coding principle for a binary alphabet is explained, taking a Russian text as an example.

Рассмотрены основные свойства энтропии и информации для дискретных случайных объектов. Обсуждена задача кодирования сообщений и изложен принцип кодирования Шеннона–Фано в двоичном алфавите на примере русского текста.

ВВЕДЕНИЕ

Новые направления, возникшие в математике в XX веке, обычно оперируют со сложными понятиями и представлениями, которые с трудом поддаются популяризации. На этом фоне весьма значительна заслуга выдающегося американского математика и инженера Клода Шеннона, который в 1947–1948 годах исходя из элементарных соображений открыл новую область математики — теорию информации [1]. Толчком к созданию новой науки были чисто технические проблемы передачи информации по телеграфным и телефонным проводам, однако к настоящему времени благодаря общему характеру теория Шеннона находит применение в исследованиях, относящихся к передаче и сохранению любой информации в природе и технике.

В данной статье рассмотрены базовые понятия энтропии и информации по Шеннону, показана их важная роль при решении задач оптимального кодирования при передаче информации по линиям связи.

ЭНТРОПИЯ КАК МЕРА
НЕОПРЕДЕЛЕННОСТИ ВЫБОРА
В КОНЕЧНЫХ СХЕМАХ

Понятие энтропии как меры необратимого рассеяния энергии впервые было введено в термодинамике в 1865 году немецким физиком Р. Клаузиусом, который показал, что каждому состоянию термодинамической системы соответствует определенное значение энтропии. В статистической физике энтропия S , согласно концепции Л. Больцмана (1872 год), связывается с термодинамической вероятностью W макроскопического состояния соотношением

$$S = k \ln W,$$

где k — постоянная Больцмана. В свою очередь, в теории информации энтропия по К. Шеннону определяется как мера неопределенности опыта с разными исходами. Последние две трактовки энтропии имеют весьма глубокую связь: на базе информационной энтропии, например, выводятся канонические гиббсовские распределения статистической физики.

Смысл понятия энтропии как меры неопределенности раскроем на простых примерах случайного выбора в конечных схемах. Случайные события, как известно, характеризуются тем, что у нас нет полной уверенности в их наступлении, то есть мы имеем некую неопределенность при изучении опытов, связанных с такими событиями. Понятно, что степень этой неопределенности в различных случаях может быть разной. Например, если опыт состоит в выяснении того, окажется ли первый встреченный на улице Томска человек космонавтом или нет, то можно с полной уверенностью считать, что это будет не космонавт. Но уже труднее предсказать, будет ли первый встреченный человек мужчиной или женщиной, и практически невозможно предсказать выигрышную комбинацию цифр в спортлото.

Для практики необходимо уметь численно выражать степень неопределенности различных опытов, чтобы сравнивать их посредством такой характеристики между собой. В качестве меры неопределенности случайного объекта (системы) с конечным множеством возможных состояний A_1, A_2, \dots, A_n с соответствующими вероятностями p_1, p_2, \dots, p_n или дискретной случайной величины X , принимающей значения X_1, X_2, \dots, X_n с теми же вероятностями, Клод Шеннон предложил использовать функционал [1]

$$H(A) = H(p_1, p_2, \dots, p_n) = -\sum_{k=1}^n p_k \log p_k, \quad (1)$$

названный им энтропией. Понятия непрерывной и дискретной случайных величин, их распределений, математических ожиданий (средних), которые нам понадобятся в минимальной степени, можно найти в [2].

Отметим, что в (1) логарифмы берутся при произвольном основании, но оказалось, что в технике удобнее всего использовать логарифмы при основании 2. В этом случае за единицу измерения степени неопределенности можно принять неопределенность, содержащуюся в опыте с двумя равновероятными исходами (например, в опыте с подбрасыванием монеты при выяснении того, что выпало: цифра или герб). Такая единица измерения неопределенности называется двоичной единицей, или битом. Если использовать десятичные логарифмы, то единицей степени неопределенности будет служить неопределенность опыта с 10 равновероятными исходами (таким является, например, опыт, состоящий в извлечении шара из урны с десятью перенумерованными шарами). Такая единица степени неопределенности называется десятичной единицей, или дитом, и она примерно в 3,32 раза больше двоичной единицы, так как $\log_2 10 = 3,32$.

Таким образом, согласно Шеннону, состоянию A_i объекта A следует приписать неопределенность, рав-

ную $-\log p_i$, а в качестве меры неопределенности самого объекта принимается среднее значение неопределенности отдельных состояний, то есть среднее значение дискретной случайной величины, принимающей значения, приведенные в табл. 1. Отсюда следует, что если набору $A_i, i = 1, 2, \dots, n$, состояний случайного объекта поставить в соответствие дискретную случайную величину X , задаваемую табл. 2, то энтропия такой случайной величины совпадет с энтропией объекта A . Последнее объясняется тем, что мера Шеннона не может претендовать на полный учет всех факторов, вызывающих неопределенность опыта. Она, например, не зависит от самих состояний A_i случайного объекта A или значений X_i дискретной случайной величины X . Тем не менее ее удобно использовать при решении некоторых вопросов теории передачи сообщений по линиям связи. Так, для определения времени, необходимого для передачи некоторого сообщения, конкретное содержание сообщения несущественно; это проявляется в независимости энтропии $H(A)$ от состояний A_1, A_2, \dots, A_n . К тому же понятно, что вероятности отдельных сообщений, вообще говоря, незначительны для теории связи.

Как следует из вводной части данного раздела, нужно различать термодинамический и информационный подходы к пониманию энтропии. В следующем разделе рассмотрены основные свойства информационной энтропии.

Таблица 1

X_i	$-\log p_1$	$-\log p_2$...	$-\log p_n$
p_i	p_1	p_2	...	p_n

Таблица 2

X_1	X_2	...	X_n
p_1	p_2	...	p_n

СВОЙСТВА ИНФОРМАЦИОННОЙ ЭНТРОПИИ

Энтропия обладает интересными свойствами, которые подтверждают, что она является разумной количественной мерой неопределенности.

1. Энтропия $H(p_1, p_2, \dots, p_n) = 0$ тогда и только тогда, когда все вероятности p_i , кроме одной, равны нулю, а эта единственная вероятность равна единице. Во всех других случаях энтропия положительна.

2. При заданном n величина H максимальна и равна $\log n$, когда все p_i равны между собой, то есть

$$p_1 = p_2 = \dots = p_n = \frac{1}{n}.$$

3. Если A и B — два независимых случайных объекта с числом состояний n и m соответственно, то

$$H(AB) = H(A) + H(B). \quad (2)$$

Доказательство. Так как вероятность r_{kl} состояния $A_k B_l$ в случае независимости A и B равна произведению $p_k q_l$, то

$$\begin{aligned} H(AB) &= -\sum_{k,l} r_{kl} \log r_{kl} = \\ &= -\sum_k p_k \log p_k \sum_l q_l - \sum_l q_l \log q_l \sum_k p_k, \end{aligned}$$

откуда, учитывая условие нормировки для вероятностей состояний объектов $\sum_l q_l = \sum_k p_k = 1$, получаем (2).

Понятно, что в силу свойства 1 значение H равно нулю только в случае полной определенности исхода опыта, то есть когда отсутствует всякая неопределенность. Доказательство свойства 2, а также некоторые другие свойства энтропии можно найти в [1, 3, 4].

Как следует из свойства 3, информационная энтропия, как и термодинамическая, обладает свойством аддитивности. Принцип аддитивности в применении к игральной кости гласит, что энтропия n бросаний кости в n раз больше, чем энтропия одного бросания. Другим важным применением этого принципа может служить следующий пример: энтропия нескольких сообщений равна сумме энтропий отдельных сообщений.

Заметим, что уже этих трех простых свойств энтропии достаточно для изложения дальнейшего материала.

ИНФОРМАЦИЯ

Определив энтропию как меру неопределенности состояния случайного объекта, мы видим, что в результате получения сведений неопределенность такого объекта может быть разве что уменьшена. Поэтому естественно количество информации измерять уменьшением энтропии того объекта или системы, для уточнения состояния которого предназначены сведения.

Рассмотрим некоторый случайный объект A и оценим информацию, получаемую в результате того, что состояние системы A становится полностью известным. До получения сведений, то есть априори, энтропия системы равнялась $H(A)$, а после получения сведений, то есть апостериори, состояние объекта определилось полностью, и энтропия стала равной нулю. Обозначим через I_A информацию, получаемую в результате выяснения состояния объекта A . Ясно, что она равна уменьшению энтропии:

$$I_A = H(A) - 0$$

или

$$I_A = H(A), \quad (3)$$

то есть количество информации, приобретаемой при полном выяснении состояния некоторого объекта, равно энтропии этого объекта.

Представим формулу (3) в виде

$$I_A = -\sum_{i=1}^n p_i \log p_i, \quad (4)$$

где $p_i = P(A = A_i)$. Формула (4) есть средняя информация частных информаций, получаемых от отдельных сообщений:

$$I_{A_i} = -\log p_i, \quad i = 1, 2, \dots, n,$$

и может быть записана в виде

$$I_A = \sum_{i=1}^n p_i I_{A_i}.$$

Так как все числа p_i не больше единицы, то как частная информация I_{A_i} , так и средняя информация I_A не могут быть отрицательными. Если все возможные состояния объекта априори одинаково вероятны, то есть $p_1 = p_2 = \dots = p_n = \frac{1}{n}$, то, естественно, частная информация I_{A_i} от каждого отдельного сообщения

$$I_{A_i} = -\log \frac{1}{n} = \log n \quad (5)$$

равна средней информации:

$$I_A = -n \frac{1}{n} \log \frac{1}{n} = \log n. \quad (6)$$

В случае, когда состояния объекта обладают различными вероятностями, информации от разных сообщений неодинаковы: наибольшую информацию несут сообщения о тех состояниях, которые априори были наименее вероятны. Например, сообщение о том, что в Томске 1 января выпал снег, несет гораздо меньше информации, чем сообщение, что в этом городе снег выпал 1 июля.

Приведем три иллюстративных примера, с помощью которых облегчается понимание средней и частной информаций.

Пример 1. На шахматную доску произвольным образом ставится ферзь. Определить информацию, получаемую от сообщения, в какой именно клетке находится фигура.

Решение. Так как все состояния данной системы равновероятны, то частная информация типа “ферзь находится в квадрате $e7$ ”

$$I_{A_i} = -\log \frac{1}{64} = \log 64 = 6 \text{ бит.}$$

Средняя информация о системе в силу формул (5) и (6) также равна 6 битам.

Пример 2. Определить частную информацию, полученную в сообщении наугад выбранного прохожего: “сегодня мой день рождения”.

Решение. Так как вероятность сообщения $p = \frac{1}{365}$, то частная информация от данного сообщения

$$I = -\log \frac{1}{365} = \log 365 \approx 8,51 \text{ бит.}$$

Пример 3. Определить среднюю информацию, полученную в сообщении наугад выбранного прохожего: “сегодня мой день рождения или сегодня не мой день рождения”.

Решение. Система имеет два возможных состояния: A_1 — день рождения с вероятностью $p_1 = \frac{1}{365}$ и A_2 — не день рождения с вероятностью $p_2 = \frac{364}{365}$. Средняя информация

$$I_A = H(A) = -\frac{1}{365} \log \frac{1}{365} - \frac{364}{365} \log \frac{364}{365} \approx 0,063 \text{ бит.}$$

Если информация выражена в битах, то ей можно дать наглядное истолкование. Рассмотрим систему с двумя состояниями: A_1, A_2 , вероятности которых равны p_1, p_2 . Чтобы выяснить состояние данной системы, достаточно задать один вопрос: находится ли система в состоянии A_1 ? Ответ да или нет на этот вопрос доставляет максимальную информацию в 1 бит, когда оба состояния априори равновероятны, то есть $p_1 = p_2 = \frac{1}{2}$.

Если информация от какого-то сообщения равна n битам, то она равносильна информации, даваемой n ответами да или нет на равновероятные вопросы. Так, в примере 1 необходимо задать шесть таких вопросов, а в примере 2 — девять вопросов.

ЗАДАЧИ КОДИРОВАНИЯ СООБЩЕНИЙ

При передаче сообщений по линиям связи часто приходится пользоваться некоторыми кодами, например азбукой Морзе. С помощью азбуки Морзе любое сообщение можно представить в виде комбинации элементарных сигналов или символов: точка, тире, пауза (пробел между буквами), длинная пауза (пробел между словами).

Кодированием в общем смысле назовем отображение состояния одной физической системы с помощью состояний некоторой другой системы. Например, при телефонном разговоре звуковые сигналы кодируются с помощью электромагнитных колебаний, которые затем декодируются на другом конце линии снова в звуковые сигналы.

Мы ограничимся наиболее простым случаем кодирования, когда обе системы A и B имеют конечное число возможных состояний. Пусть имеется некоторая система A (например, буква русского алфавита), которая случайным образом принимает одно из состояний A_1, A_2, \dots, A_n , и пусть мы ее хотим закодировать с помощью другой системы B , возможные состояния которой B_1, B_2, \dots, B_m , причем $m < n$. Ясно, что в случаях, когда $m < n$, одно состояние системы A приходится отображать с помощью определенной комбинации состояний системы B (кодирование букв какого-то алфавита азбукой Морзе).

Кодированием в узком смысле назовем выбор таких комбинаций и установление соответствия между передаваемым сообщением и этими комбинациями.

Коды различаются по числу элементарных символов или, что то же самое, по числу возможных состояний системы B . В азбуке Морзе таких элементарных символов, как было указано выше, четыре. Код с двумя элементарными символами называется двоичным, и обычно его изображают нулем и единицей (0 и 1).

КОД ШЕННОНА–ФАНО

Так как одно и то же сообщение можно закодировать различными способами, то естественно возникает вопрос об оптимальных или наиболее выгодных в каком-то смысле способах кодирования. Будем считать оптимальным такой код, при котором на передачу сообщений затрачивается минимальное время. Если на передачу каждого элементарного символа (например, 0 или 1) тратится одно и то же время, то оптимальный код на передачу сообщения заданной длины потребует минимального количества элементарных символов.

Поставим следующую задачу: закодировать двоичным кодом буквы русской азбуки так, чтобы каждой букве соответствовала определенная комбинация элементарных символов 0 и 1 и чтобы среднее число этих символов на букву текста было минимальным.

Рассмотрим 31 букву русской азбуки вместе с неразличимыми в телеграфии ъ и ь плюс промежуток между словами, обозначаемый знаком тире. Самый простой способ кодирования состоит в приписывании всем символам подряд номеров от 0 до 31, а затем в переводе всех номеров в двоичную систему исчисления. Напомним, что в двоичной системе единицы разных

разрядов являются разными степенями двойки. Например,

$$11 = 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0,$$

и в двоичной системе десятичное число 11 запишется как 1011.

Так как каждое из чисел 0, 1, 2, ..., 31 изображается пятизначным двоичным кодом, то наш код принимает следующий вид:

а	00000
б	00001
в	00010
г	00011
д	00100
.....	
я	11110
—	11111

В этом коде на представление каждой буквы тратятся пять элементарных символов. Понятно, что это не оптимальный код, так как было бы разумнее, чтобы часто встречающиеся буквы были закодированы меньшим числом символов, а реже встречающиеся — большим.

Чтобы составить такой код, необходимо знание частот букв в русском тексте, которые можно принять в качестве оценок вероятностей появления букв в тексте. Эти частоты (в порядке их убывания) сведены в табл. 3, заимствованную из книги [5].

Перейдем к составлению наиболее экономичного кода на базе табл. 3 из соображений теории информации. Очевидно, код будет оптимальным, если каждый элементарный символ будет передавать максимальную информацию. А это, согласно свойству 2 энтропии, имеет место только в случае равновероятности состояний, поэтому в основе оптимального кодирования лежит требование, чтобы элементарные символы в закодированном тексте встречались в среднем с одинаковой частотой.

Теперь изложим способ построения кода, удовлетворяющего поставленному выше условию, который известен под названием кода Шеннона—Фано. Согласно этому способу, кодируемые символы разделяются на две приблизительно равновероятные подгруппы: для первой группы символов на первом месте комбинации ставится 0, а для второй группы символов — 1. Далее каждая группа снова делится на две приблизительно равновероятные подгруппы; для символов первой подгруппы на втором месте ставится 0, а для второй подгруппы — 1 и т.д.

Покажем принцип построения кода Шеннона—Фано на примере русского алфавита. Взяв из табл. 3 первые шесть букв (от — до т) и просуммировав их вероятности (частоты), получим 0,498; на все остальные

Таблица 3

Буква	Частота	Буква	Частота	Буква	Частота	Буква	Частота
—	0,145	р	0,041	я	0,019	х	0,009
о	0,095	в	0,039	ы	0,016	ж	0,008
е	0,074	л	0,036	з	0,015	ю	0,007
а	0,064	к	0,029	ъ, ь	0,015	ш	0,006
и	0,064	м	0,026	б	0,015	ц	0,004
т	0,056	д	0,026	г	0,014	щ	0,003
н	0,056	п	0,024	ч	0,013	э	0,003
с	0,047	у	0,021	й	0,010	ф	0,002

буквы (от и до ф) приходится приблизительно такая же вероятность 0,502. Таким образом, первые шесть букв будут иметь на первом месте знак 0, а остальные буквы — 1. Далее снова разделим первую группу на две приблизительно равновероятные подгруппы, и для всех букв первой подгруппы на втором месте поставим 0, а для второй подгруппы — 1. Процесс продолжается до тех пор, пока в каждом подразделении не останется ровно одна буква, которая будет закодирована своим двоичным знаком. Полученный код приводится в табл. 4, согласно которой механизм построения кода становится совсем понятным.

С помощью этого кода можно закодировать и декодировать любое сообщение. Например, фраза “Сороковский журнал” кодируется так:

1001001101000011001001101011001101110001111010000
111110011010010100100010110110

Отметим, что при этом нет необходимости отделять буквы друг от друга специальным знаком, так как декодирование здесь выполняется однозначно. Однако при таком коде любая ошибка кодирования делает невозможным декодирование всего следующего за ошибкой текста.

Таблица 4

Буква	Двоичное число	Буква	Двоичное число	Буква	Двоичное число
—	000	к	10111	ч	111100
о	001	м	11000	й	1111010
е	0100	д	110010	х	1111011
а	0101	п	110011	ж	1111100
и	0110	у	110100	ю	1111101
т	0111	я	110110	ш	11111100
н	1000	ы	110111	ц	11111101
с	1001	з	111000	щ	11111110
р	10100	ъ, ь	111001	э	111111110
в	10101	б	111010	ф	111111111
л	10110	г	111011		

Убедимся, что составленный нами код при отсутствии ошибок в самом деле является оптимальным. Для этого найдем сначала среднюю информацию, содержащуюся в одной букве текста, то есть энтропию на одну букву:

$$H = -\sum_{i=1}^{32} p_i \log p_i = -0,145 \log 0,145 - \dots - 0,002 \log 0,002 \approx \\ \approx 4,42 \text{ бит.}$$

Далее определяем среднее число элементарных символов на букву:

$$k = -\sum_{i=1}^{32} k_i p_i = 3 \cdot 0,145 + 3 \cdot 0,095 + \dots + 9 \cdot 0,002 \approx 4,45.$$

Деля энтропию H на k , получаем информацию на один элементарный символ

$$I_s = \frac{4,42}{4,45} \approx 0,994 \text{ бит.}$$

Таким образом, информация на один символ близка к своему верхнему пределу 1, и, следовательно, выбранный нами код весьма близок к оптимальному.

В случае же простейшего кода мы имели изображение каждой буквы пятью двоичными знаками, и информация на один символ

$$I_s = \frac{4,42}{5} \approx 0,884 \text{ бит,}$$

что заметно меньше, чем при оптимальном кодировании.

В заключение отметим, что кодирование по буквам по большому счету не является оптимальным, так как между соседними буквами осмысленного текста всегда

имеется зависимость. В связи с этим обстоятельством более экономный код можно построить, если кодировать также и целые блоки из букв. В частности, в русском тексте можно кодировать такие часто встречающиеся комбинации букв, как *тся*, *ает*, *ние* и т.п., причем принцип двоичного кодирования сохраняется, поэтому кодируемые блоки следует располагать в порядке убывания частот, как и буквы в табл. 3.

ЛИТЕРАТУРА

1. Шеннон К. Математическая теория связи // Работы по теории связи и кибернетике. М.: Изд-во иностр. лит., 1963. С. 243–332.
2. Родкина А.Е. О некоторых понятиях и проблемах финансовой математики // Соросовский Образовательный Журнал. 1998. № 6. С. 122–127.
3. Тарасенко Ф.П. Введение в курс теории информации. Томск: Изд-во Том. ун-та, 1963. 240 с.
4. Яглом А.М., Яглом И.М. Вероятность и информация. М.: Наука, 1973. 512 с.
5. Вентцель Е.С. Теория вероятностей. М.: Наука, 1964. 576 с.

Рецензент статьи Ю.П. Соловьев

* * *

Геннадий Михайлович Кошкин, доктор физико-математических наук, профессор кафедры теоретической кибернетики Томского государственного университета. Область научных интересов – теория статистического оценивания в условиях неопределенности с приложениями к проблемам идентификации и управления в сложных системах. Соавтор двух монографий, автор и соавтор около 60 научных статей в отечественных и зарубежных журналах и двух учебных пособий для студентов.