Гончаренко Валерий Евстафиевич
Ивановский филиал Российского экономического университета имени Г.В. Плеханова, г. Иваново.

Путешествие по тропе первопроходцев измерения информации
Часть 2

Данная статья является продолжением статьи, в которой автор предлагает читателям заново разработать вероятностно-статистический подход к измерению информации («Потенциал», №7, 2015 г.). Стиль изложения материала делает эту задачу непринуждённой и увлекательной. При желании читатели могут обратиться к более строгому изложению затронутых вопросов в ранее опубликованных статьях, указанных в списке использованных источников.

В предыдущем путешествии по тропе первопроходцев измерения информации мы успешно справились с задачей оценки неопределённости и количества информации для случайных событий с равновероятными исходами. Можно с полной уверенностью предположить, что достаточно часто наблюдаются случайные события с разновероятными исходами. Будут ли и в этом случае правильно работать наши методы и формулы? Заканчиваем привал и двигаемся дальше по тропе.

Только вымолвить успели, а очередной зазывала уже предлагает нам новый аттракцион с разноцветными шарами.

Решение трудной задачи

Неутомимый факир громогласно обращается к уважаемой публике.

В моем барабане по-прежнему 64 практически одинаковых шара, только 63 из них окрашены в белый цвет, а один-единственный – в чёрный. Кто угадает, какого цвета окажется наугад изъятый из барабана шар? Конечно, почти все присутствующие, следуя здравой логике, выбирают шар белого цвета, ведь вероятность того, что выпадет именно белый шар, гораздо больше. Хотя кто его знает, маловероятно, но это может оказаться единственный шар чёрного цвета. Количественные оценки этих вероятностей мы уже умеем рассчитывать. Из общего количества 64 исходов 63 исхода благоприятствуют белому цвету и один исход – чёрному, следовательно,

$$p_б = m_б/N = 63/64 = 0{,}984\ (98{,}4\%),$$

$$p_ч = m_ч/N = 1/64 = 0{,}015(1{,}5\%).$$

Разница видна невооружённым глазом, и рассчитывать на то, что факир достанет шар чёрного цвета, не приходится. А если гипотетически представить огромный барабан, в котором 768 млрд шаров белого цвета и только один чёрного, то вероятность извлечения шара чёрного цвета стремится к нулю. В этом случае присутствующий народ будет задорно кричать факиру: «Не смешите нас, вы, конечно, достанете шар белого цвета, и, возможно, вам придётся вечно повторять свой опыт, пока неожиданно для всех не достанете шар чёрного цвета».

Как мы видим по реакции публики, неопределённость этого случайного события очень мала, а, следовательно, и информация в сообщении о реализации одного из возможных исходов, полностью уничтожающая эту неопределённость, тоже очень мала. Когда факир скажет: «Я достал шар белого цвета», все присутствующие на площади подумают: «Я так и знал». Самое удивительное в нашей схеме, если факир сообщит: «Граждане, я достал шар чёрного цвета». Эта информация уничтожит ту же очень незначительную неопределённость и будет содержать очень мало бит (формула может давать и вещественное число с десятичной точкой, например 0,0000000000001 бит).

Как же так, возмущаются граждане, произошло чрезвычайно редкое событие, а тут утверждают, что практически никакой информации мы не получили?! Будем стойкими и вежливо ответим: «Это результат

принятых соглашений и допущений вероятностно-статистического подхода, мы и впредь будем строго их придерживаться, а все желающие могут разработать на другой тропе свои эксклюзивные подходы к определению и измерению информации».

Итак, в последнем аттракционе случайное событие характеризуется разновероятными исходами, и это создаёт проблему для количественной оценки неопределённости и информации. Для решения этой проблемы придётся придумать более сложные схемы и ряд допущений.

Рассмотрим полную группу исходов условного случайного события, исходы которого будем обозначать множеством их номеров:

$$\{4, 3, 4, 4, 1, 3, 4, 4, 2, 3, 1, 4, 3, 2, 4, 4\}.$$

Если в соответствии с правилом половинного деления поделить его на два подмножества

I $\{\{4, 3, 4, 4, 1, 3, 4, 4\}$

II $\{2, 3, 1, 4, 3, 2, 4, 4\}\}$,

то и на элементарный вопрос «Исход с номером 4 находится в первом подмножестве?» мы не сможем получить элементарного ответа «да» или «нет». Запрашиваемый исход невозможно в такой схеме выделить методом половинного деления.

У рассматриваемого события могут быть 4 вида различных исходов,

а их общее количество $N = 16$. Вероятности исходов:

$$p_1 = m_1/N = 2/16 = 0{,}125,$$
$$p_2 = m_2/N = 2/16 = 0{,}125,$$
$$p_3 = m_3/N = 4/16 = 0{,}250,$$
$$p_4 = m_4/N = 8/16 = 0{,}500.$$

Сумма вероятностей полной группы исходов равна 1.

Принимаем допущение, что исходы равномерно распределяются друг относительно друга. В этом случае для каждого выделенного исхода можно представить свою схему подмножеств во множестве исходов. Для исхода 1 можно отобразить два подмножества

{{1, X, X, X, X, X, X, X}
{1, X, X, X, X, X, X, X}},

где X – любой из остальных возможных исходов в пределах их общего количества.

Если мы задаёмся целью определения положения только исхода 1, то его определение во всём множестве равносильно его определению в одном из подмножеств, и без всяких модификаций работает принцип «на элементарный вопрос можно дать элементарный ответ», или принцип половинного деления, несущий информацию в 1 бит. Количество полученной информации для выделения исхода под номером 1 можно определить по формуле

$$I_1 = \log_2 (8) = 3 \text{ бита.}$$

Мы воспользовались формулой (1), как будто исход 1 находится в системе равновероятных исходов без повторений. Важно подчеркнуть, что такая возможность предоставляется только для исхода 1 в сформированной структуре подмножеств.

По аналогии для остальных исходов получим:

2-й исход: {{2, X, X, X, X, X, X, X}
{2, X, X, X, X, X, X, X}}

$$I_2 = \log_2 (8) = 3 \text{ бита,}$$

3-й исход: {{3, X, X, X}{ 3, X, X, X}
{3, X, X, X}{3, X, X, }}

$$I_3 = \log_2 (4) = 2 \text{ бита,}$$

4-й исход: {{4, X}{4, X}{4, X}{4, X}
{4, X}{4, X}{4, X}{4, X}}

$$I_4 = \log_2 (2) = 1 \text{ бит.}$$

Итак, вместо изначального множества исходов мы использовали четыре вида выделенных подмножеств:

{1,X,X,X,X,X,X,X} – 3 бита;
{2,X,X,X,X,X,X,X} – 3 бита;
{3,X,X,X} – 2 бита;
{4,X} – 1 бит.

Только в структурах этих подмножеств можно было определить методом половинного деления количество информации, необходимое для определения выделенного исхода, принимая допущение о равновероятности исходов в подмножестве. На основании полученных результатов создадим модель системы исходов условного события, в которой выделенные подмножества как бы объединяются между собой в пропорции вероятностей выделенных исходов. Схематично изначальное множество исходов и соответствующую им модель можно представить в следующем виде:

$$\{1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4\} \equiv$$
$$\equiv p_1\{1,X,X,X,X,X,X,X\} +$$
$$+p_2\{2,X,X,X,X,X,X,X\} + p_3\{3,X,X,X\} +$$
$$+ p_4\{4,X\}.$$

В соответствии с полученной моделью количество информации, необходимое для определения одного из исходов, будет

$$I = 0{,}125 \cdot \log_2(16/2) + 0{,}125 \cdot \log_2(16/2) + 0{,}250 \cdot \log_2(16/4) + 0{,}500 \cdot \log_2(16/8) = 0{,}375 + 0{,}375 + 0{,}500 + 0{,}500 = 1{,}75 \text{ бит}. \quad (2)$$

Займёмся теперь преобразованием выражения (2) с целью придания ему обобщённого формального выражения. Допустим, мы рассматриваем случайное событие, у которого всего k различных исходов, номер каждого из них будем обозначать индексом i, количество исходов i-го вида обозначим через m_i, а общее количество исходов во всей совокупности обозначим через N. Тогда неопределённость этого случайного события можно описать выражением

$$H = p_1 \cdot \log_2(N/m_1) + p_2 \cdot \log_2(N/m_2) + p_3 \cdot \log_2(N/m_3) + \ldots + p_k \cdot \log_2(N/m_k). \quad (3)$$

Легко заметить, что аргументами функции двоичного логарифма в этом выражении являются обратные величины вероятностей i-х исходов, т. е. $N/m_i = 1/p_i$. Используем это в выражении (3) и получим:

$$H = p_1 \cdot \log_2(1/p_1) + p_2 \cdot \log_2(1/p_2) + p_3 \cdot \log_2(1/p_3) + \ldots + p_k \cdot \log_2(1/p_k). \quad (4)$$

Выражение $1/p_i$ можно заменить на p_i^{-1}, в итоге получим

$$H = p_1 \cdot \log_2(p_1^{-1}) + p_2 \cdot \log_2(p_2^{-1}) + p_3 \cdot \log_2(p_3^{-1}) + \ldots + p_k \cdot \log_2(p_k^{-1}). \quad (5)$$

В соответствии со свойствами функции логарифма степень аргумента можно вынести множителем перед функцией логарифма. Тогда получим:

$$H = -p_1 \cdot \log(p_1) - p_2 \cdot \log(p_2) - p_3 \cdot \log(p_3) - \ldots - p_k \cdot \log(p_k), \quad (6)$$

или в свёрнутом виде:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) = \sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right). \quad (7)$$

Из реального мира нам подсказывают, что основание логарифма в формуле (7) оказывает влияние только на единицы измерения. Если использовать двоичный логарифм, то мы получим наименьшую единицу измерения неопределённости – бит, а при использовании натурального логарифма – более крупную единицу – **нат**. Может, ещё имеет смысл использовать десятичный логарифм, тогда получим еще более крупную единицу измерения – **дит**.

Учитывая, что количество информации – это полностью снятая неопределённость, выражение (7) можно также использовать для определения количества информации.

Остаётся вопрос по индивидуальному количеству информации

На нашей тропе мы достаточно быстро пришли от выражения (1) к выражению (7), а в реальном мире в 1928 г. Ральф Хартли предложил использовать формулу (1) для определения количества информации для случайного события с равновероятными исходами, и только в 1948 г., т. е. через 20 лет, Клод Шеннон в своей опубликованной работе «Математическая теория связи» [4] предложил и обосновал формулу (7) для количественной оценки неопределённости случайного события с разновероятными исходами, которую также ещё называют менее понятным термином «энтропия множества вероятностей». Между этими формулами нет противоречия, формула (1) является частным случаем формулы (7). Благодаря этим первопроходцам мы достаточно быстро прошли по их тропе. Но и сегодня на ней не всё так просто и однозначно. В частности, в выражениях (6) или (7) составляющие $\log(p_i)$ называют индивидуальными, или частными количествами информации. В трактовке этой величины мнения расходятся. Её трактуют как количество информации в сообщении, когда был реализован i-й исход случайного события. В качестве доказательства различного толкования величин $\log(p_i)$ приведём формулировки и решения двух задач, взятых из других источников.

Популярным примером некорректной формулировки может быть следующая задача и её решение.

Задача 1. *В мешке находятся 20 шаров, из них 15 белых и 5 красных. Какое количество информации несёт сообщение о том, что достали белый шар?*

Решение. *Найдём вероятность того, что достали белый шар:*

$$p_б = 15/20 = 0{,}75.$$

Найдём количество информации в сообщении о вытаскивании белого шара:

$$I_б = \log_2 (1/p_б) = \log_2 (1/0{,}75) =$$
$$= 1{,}15470 \text{ бит}.$$

На аналогичную тему можно встретить другой популярный пример, приведём его дословно.

Задача 2. *Какое количество информации будет содержать зрительное сообщение о цвете вынутого шарика, если в непрозрачном мешочке находятся 50 белых, 25 красных, 25 синих шариков?*

1) Всего шаров

$$50 + 25 + 25 = 100;$$

2) вероятности шаров

$$50/100 = 1/2,\ 25/100 =$$
$$= 1/4,\ 25/100 = 1/4;$$

3) $$I = -\,(1/2 \log_2 1/2 +$$
$$+1/4 \log_2 1/4 + 1/4 \log_2 1/4) =$$
$$= -(1/2(-1) + 1/4(-2) + 1/4(-2)) = 1{,}5 \text{ бит}.$$

Как мы видим, формулировки приведённых примеров задач по сути одинаковые, а их решения существенно различаются. Заслуживает критики предлагаемое решение первой задачи. Из контекста можно догадаться, что речь идёт о случайном событии по извлечению из непрозрачного мешочка шаров одного из двух цветов. Вероятности извлечения шара красного или белого цвета различные. Для решения задачи необходимо использовать формулу Шеннона (7). В приведённом решении не делается различия между количеством информации как меры снятия неопределённости опыта и величиной индивидуального количества информации, которая в вероятностном подходе к измерению информации используется не как самостоятельная величина меры количества информации, а только как отдельная составляющая при расчёте интегрального показателя неопределённости (энтропии вероятностей).

Более детально с материалами затронутой темы можно ознакомиться в ранее опубликованных статьях [1, 2, 3].

Литература

1. *Гончаренко В.Е.* Проблемы трактовки меры Шеннона в курсе информатики // Информатика и образование. – 2013. – №7.

2. *Гончаренко В.Е.* Сложные задачи на определение информации // Информатика. – 2011. – №10.

3. *Гончаренко В.Е.* Что измеряют по формулам Хартли и Шеннона // Информатика. – 2011. – №6.

4. *Шеннон К.* Математическая теория связи // Работы по теории информации и кибернетике / пер. с англ. ; под ред. Р.Л. Добрушина и О.Б. Лупанова. М., 1963.

Юмор Юмор Юмор Юмор Юмор Юмор

Блиц-ответы

– Что нужно сделать, чтобы не повторять старые ошибки?

– Придумать новые.

***

– Можно ли увидеть край атмосферы?

– Можно, если вооружить глаза.

***

– Как можно исправить изображённый на доске график?

– Стереть его.

***

– Какая опасность грозит водолазу при быстром подъёме из глубин моря на поверхность?

– Он не успеет заметить акулу.

***

– В какой области работал наш знаменитый соотечественник учёный Эдуард Константинович Циолковский?

– В Калужской.