Энтропийные методы анализа информации*

И. М. Макаров, А. А. Ахрем, В. З. Рахманкулов,
Т. Л. Вашевник, Р. Ю. Филюков

Аннотация. В статье изучаются основные количественные характеристики алгебраической теории информации — энтропия и мера элементарной информации. Мера информации определяется относительно пользователя, т. е. с учетом уже имеющейся у него исходной информации. Приведены примеры вычисления (оценки) энтропии и меры информации для различных видов элементарной информации.
Ключевые слова: *решетки данных и знаний, энтропия, мера информации решеток.*

Введение

Настоящая статья является непосредственным продолжением работ авторов [1, 2], посвященных изучению базовых понятий алгебраической теории информационных систем. В этих работах было введено и исследовано основное понятие алгебраической теории информации — элементарная информация об одной точке x_0 опорного множества сведений X. В данной статье рассматриваются количественные методы сравнения элементарных информаций о точке x_0. В основу исследований положен разработанный в математической теории связи энтропийный подход к анализу информации [3–6]. Количественная мера (энтропия) анализируемой информации определяется относительно пользователя, т. е. с учетом уже имеющейся начальной информации у пользователя. Приводятся примеры вычисления (оценки) энтропии для различных классов элементарной информации о точке x_0 из множества сведений X.

1. Количественный анализ элементарной информации

Пусть заданы: X — произвольное множество элементов (сведений), B — некоторая решетка понятий (подмножеств) для X, BS — атомарная (дизъюнктная) шкала для B: $BS = b_k, k \in K\}$, где K — конечное или счетное множество символов [7–9].

* Работа выполнена при финансовой поддержке РФФИ (проект № 07-01-00572) и программы Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация» (проект № 209).

Предположим дополнительно, что для каждой точки $x_0 \in X$ определено дискретное распределение достоверностей, заданное на шкале BS, т. е. определена неотрицательная числовая функция $p(b_k, x_0)$, $p(b_k, x_0) \geq 0, b_k \in BS$, с условием нормировки

$$\sum_{b_k \in BS} p(b_k, x_0) = 1.$$

Определение 1 [7–9].

Дискретное распределение достоверностей называется начальным распределением достоверностей для точки x_0 из X относительно решетки понятий B. Величина

$$\sum_{b_k \in \sigma} p(b_k, x_0) = p,$$

определенная для любого подмножества $\sigma \in B$, называется достоверностью (вероятностью) факта принадлежности $x_0 \in \sigma$.

Утверждение 1.

Достоверность удовлетворяет следующим соотношениям: $0 \leq p \leq 1$.

В дальнейшем максимальную достоверность $p = 1$ будем трактовать как истинность факта $x_0 \in \sigma$. Минимальную достоверность $p = 0$ будем трактовать как ложность факта $x_0 \in \sigma$, т. е. истинность факта $x_0 \in \sigma' = X \setminus \sigma$. В случае, когда решетка представляет собой несчетное множество, начальное распределение достоверностей для точки $x_0 \in X$

будем задавать интегрируемой функцией плотности $f(x,x_0)\ge 0$ с условием нормировки

$$\int_x f(x,x_0)\,dx=1\,.$$

В этом случае предполагается, что решетка подмножеств B состоит из измеримых подмножеств. Величина

$$\int_\sigma f(x,x_0)\,dx=p\,,$$

определенная для любого $\sigma\in B$, трактуется как достоверность факта принадлежности $x_0\in\sigma$.

Конкретизация начального распределения $p(b_k,x_0)$ или $f(x,x_0)$ для $x_0\in X$ зависит от пользователя и характеризует имеющуюся у него исходную информацию о точке . При отсутствии каких-либо начальных сведений о точке будем задавать либо равномерное распределение, либо нормальное (гауссовское) распределение и т. д.

Следуя [7–9], введем следующее определение.

Определение 2.

Пусть дано начальное распределение достоверностей $p(b_k,x_0)$ для точки $x_0\in X, b_k\in BS\subset B,$ где B — решетка понятий для X с конечной или счетной атомарной шкалой BS. Для любого сведения $\sigma(x_0)$ о точке x_0, при условии, что $\sigma\in B$, определим вторичное распределение достоверностей $p_\sigma(b_k,x_0)$ для x_0 по формуле Байеса [9–11]:

$$p_\sigma(b_k,x_0)=\begin{cases}0, & b_k\not\subset\sigma,\\ p(b_k,x_0)\cdot\left(\sum\limits_{b_k\subset\sigma_k}p(b_k,x_0)\right)^{-1}, & b_k\subset\sigma.\end{cases}$$

Величину изменения энтропии для начального и вторичного распределений назовем количеством информации, принесенной сведением $\sigma(x_0)$ и обозначим

$$I(\sigma(x_0))=H(X(x_0))-H(\sigma(x_0))\,. \qquad (1)$$

В (1) энтропия задается формулами

$$H(X(x_0))=-\sum_{b_k\subset X}p(b_k,x_0)\cdot\log_2 p(b_k,x_0),$$

$$H(\sigma(x_0))=-\sum_{b_k\subset\sigma}p_\sigma(b_k,x_0)\cdot\log_2 p_\sigma(b_k,x_0).$$

Количество информации измеряется в битах.

Пример 1 [7, 8]

Пусть даны отрезок [0,1] числовой прямой и решетки B для него с атомарной шкалой $BS=$ $=\{b_k,k=0,1,\ldots,9\}$, где b_k — дизъюнктные подмножества $b_0=[0;0,1],b_1=(0,1;0,2],\ldots,b_9=(0,9;1]$. Допустим, что про точку $x_0\in[0;1]$ известно сведение $\sigma(x_0)=[0;0,3]$ (отметим, что $\sigma\in B$). Определим количество информации, принесенной сведением σ. Предположим вначале, что исходное распределение $p(b_k,x_0)$ является равномерным [10, 11]:

$$p(b_k,x_0)=0{,}1\,,\ k=0,1,\ldots,9\,.$$

Тогда вторичное распределение задается соотношениями

$$p_\sigma(b_k,x_0)=\begin{cases}\frac{1}{3}, & k=0,1,2,\\ 0, & k\ge 3.\end{cases}$$

В этом случае получаем, что

$$I(\sigma(x_0))=\log\frac{10}{3}\ \text{(где }\log a=\log_2 a,\ a>0\text{)}$$

При начальном распределении вида

$$p(b_k,x_0)=\begin{cases}\frac{1}{2}, & k=0,1,\\ 0, & k\ge 2\end{cases}$$

мы имеем:

$$x_0\in[0;0,2],\ \text{так как}\ \sum_{k=0}^{1}p(b_k,x_0)=1\,.$$

В этом случае сведение $\sigma(x_0)$ ничего нового о точке x_0 не сообщает, вторичное распределение совпадает с начальным распределением и поэтому $I(\sigma(x_0))=0.$

При начальном распределении вида

$$p(b_k,x_0)=\begin{cases}c_1>0, & k\le 8,\\ c_2>0, & k=9\end{cases}$$

имеем:

$$9c_1+c_2=1.$$

Вторичное распределение в этом случае равно

$$p_\sigma(b_k,x_0)=\begin{cases}\frac{1}{3}, & k=0,1,2,\\ 0, & k\ge 3.\end{cases}$$

Мера (количество) информации $I(\sigma(x_0))$ вычисляется по формуле

$$I(\sigma(x_0))=-9c_1\log c_1-c_2\log c_2-\log 3.$$

При достаточно малых $c_1 \approx 0$ имеем $c_2 \approx 1$, и количество информации $I(\sigma(x_0))$ будет отрицательным. В этом случае начальное распределение находится в конфликте со сведением $\sigma(x_0)$. В самом деле, параметры $c_1 \approx 0$, $c_2 \approx 1$ означают, что $x_0 \in [0;0,9]$ маловероятно и что $x_0 \in (0,9;1]$ высоковероятно. В то же время сведение $\sigma(x_0)$ означает, что $x_0 \in [0;0,3]$.

Для количества информации имеют место следующие утверждения [7–9].

Теорема 1.

1. Если начальное распределение для точки $x_0 \in X$ относительно решетки понятий B равномерно, то для любого сведения $\sigma(x_0)$ о точке x_0 справедливо соотношение

$$I(\sigma(x_0)) = \log \frac{M}{m} \geq 0, \qquad (2)$$

где в случае дискретной решетки M — число атомов решетки, m — число атомов, составляющих подмножество σ, а если B — несчетна, то

$$M = \int_x dx, \quad m = \int_\sigma dx .$$

В соотношении (2) равенство нулю достигается тогда и только тогда, когда сведение является наиболее общим, т. е. $I(X(x_0)) = 0$.

2. Пусть начальное распределение для точки $x_0 \in X$ относительно решетки понятий B равномерное. Тогда для любых двух сведений $\sigma_1(x_0), \sigma_2(x_0)$ при условии $\sigma_1, \sigma_2 \in B$ выполняется соотношение

$$I(\sigma_1(x_0)) \leq I(\sigma_2(x_0)),$$

если $\sigma_1(x_0)$ — более общее сведение, чем $\sigma_2(x_0)$, т. е. $\sigma_1 \supset \sigma_2$.

3. Если начальное распределение для точки $x_0 \in X$ относительно решетки B равномерное, то для любых двух сведений $\sigma_1(x_0), \sigma_2(x_0)$, $\sigma_1, \sigma_2 \in B$, справедливо равенство

$$I(\sigma_1(x_0) \wedge \sigma_2(x_0)) = I(\sigma_1(x_0)) + I\sigma_1(\sigma_2(x_0)),$$

где $I\sigma_1(\sigma_2(x_0))$ — количество информации, принесенной сведением $\sigma_2(x_0)$ при условии, что начальным распределением является вторичное распределение для $\sigma_1(x_0)$.

4. Если в условиях утверждения 3 теоремы имеются $n \geq 3$ сведений $\sigma_1(x_0), \ldots, \sigma_n(x_0)$ об одной точке, причем $\sigma_k \in B, k = 1, \ldots, n$, то

$$I(\sigma_1(x_0) \wedge \ldots \wedge \sigma_n(x_0)) =$$
$$= I(\sigma_1(x_0)) + \sum_{k=2}^{n} I_{\sigma_1 \ldots \sigma_{k-1}}(\sigma_k(x_0)), \qquad (3)$$

где $I_{\sigma_1 \ldots \sigma_{k-1}}(\sigma_k(x_0))$ — количество информации, принесенной сведением $\sigma_k(x_0)$ при условии, что начальное распределение является вторичным для сведения $\sigma_1(x_0) \wedge \ldots \wedge \sigma_{k-1}(x_0)$.

Следуя [7, 8] введем

Определение 3.

Пусть $A(x_0)$ — носитель информации $IN(x_0)$ о точке $x_0 \in X$ и задано начальное распределение для точки x_0 относительно решетки понятий B. Если для любого сведения $\sigma(x_0) \in A(x_0)$ справедливо $\sigma \in B$, то количеством информации $IN(x_0)$, принесенной носителем $A(x_0)$, называется следующая точная верхняя грань по всем конечным конъюнкциям сведений носителя:

$$I(A(x_0)) = \sup I(\wedge \sigma_k(x_0)).$$

Для случая терминального носителя $T(x_0) = \{x_\sigma; \sigma \in A\}(x_0)$ информации $IN(x_0)$ по определению получим

$$I(T(x_0)) = I(A(x_0)).$$

Для количества информации носителей имеют место следующие утверждения [7, 8].

Теорема 2 (о корректности определения количества информации).

Предположим, что начальное распределение для точки $x_0 \in X$ относительно решетки понятий B является равномерным. Пусть $A_1(x_0), A_2(x_0)$ — два эквивалентных носителя информации, любые сведения которых $\sigma_1(x_0) \in A_1(x_0), \sigma_2(x_0) \in A_2(x_0)$ таковы, что $\sigma_1, \sigma_2 \in B$. Тогда

$$I(A_1(x_0)) = I(A_2(x_0)).$$

Теорема 3 (о данных о точке $x_0 \in X$).

Пусть начальное распределение для точки $x_0 \in X$ относительно решетки B является равномерным. Допустим, что решетка B разложена в произведение своих подрешеток

$$B = \prod B_k .$$

Если $A(x_0) = \{\sigma_k(x_0); k \in K\}$ — данные о точке $x_0 \in X$ относительно (3), то

$$I(A(x_0)) = \sum I_k(\sigma_k(x_0)),$$

где $I_k(\sigma_k(x_0))$ — количество информации, принесенной сведением $\sigma_k(x_0)$ при условии, что начальное распределение для x_0 относительно подрешетки B_k равномерное. Следуя [7, 8], приведем примеры вычисления количества информации о точке $x_0 \in X$ на основе результатов теорем 2, 3.

Пример 2.

Предположим, что n-значное натуральное число $x_0 \in N$ задано в десятичной позиционной системе счисления. Десятичную запись числа x_0 рассмотрим с точки зрения данных

$$A(x_0) = \{\varphi_{i_1}^1; \cdots; \varphi_{i_n}^n\}(x_0)$$

относительно разложения $B = B_1 \times \ldots \times B_n$, где подрешетки B_k определяются атомарными шкалами

$$BS_k = \{\sigma_i^k, i = 0,1,\ldots,9\}$$

из подмножеств σ_i^k чисел, которые в десятичной записи имеют в k-м разряде знак i $(i = 0,1,\ldots,9)$. В этом случае количество информации, принесенной такими данными, равно для равномерного распределения

$$I(A(x_0)) = n \cdot \log 10 .$$

Пример 3.

Найдем количество информации домашнего адреса

$$A(x_0) = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}(x_0) \quad (x_0 \in X). \qquad (4)$$

В (4) обозначено:

X — множество возможных местожительств;

x_0 — идентификатор (фамилия, имя, отчество) человека;

$\sigma_k \in B_k, k = 1,2,3,4$, — элементы следующих решеток понятий для X:

B_1 — решетка населенных регионов страны, например республик, краев, областей, городов и т. п. В решетке B_1 используются две шкалы понятий: максимальная и минимальная. В первой шкале BS_1 перечисляются названия всех республик, краев, областей, районов и т. д. Под каждым названием подразумевается подмножество местожительств. Шкала BS_1 является максимальной. Другая шкала $BS_1^{'} = \{\sigma_i^j\}$ — семейство шестизначных почтовых индексов. Здесь σ_i^j — множество мест жительства, соответствующих шестизначным почтовым индексом, у которых в разряде j $(j = 1,2,\ldots,6)$ стоит знак i $(i = 0,1,\ldots,9)$. Имеется вполне определенное соответствие почтовых индексов и множеств мест жительства. Шкала $BS_1^{'}$ является почти минимальной. B_2 — решетка проспектов, улиц, площадей, переулков и т. д. Шкала BS_2 определяется перечислением всех названий проспектов, улиц и т. п. Она является атомарной шкалой, т. е. максимальной. B_3 — решетка номеров домов со шкалой $BS_3 = \{\psi_i^j; i = 0,1,\ldots,9; j = 1,2,3\}$. Здесь ψ_i^j — множество местожительств, соответствующих десятичным номером домов, у которых в разряде j $(j = 1,2,3)$ стоит знак i $(i = 0,1,\ldots,9)$. Шкала BS_3 — почти минимальна в B_3. B_4 — решетка номеров квартир со шкалой $BS_4 = \{x_i^j, i = 0,\ldots 9; j = 1,2,3,4\}$. Здесь x_i^j — множество местожительств, соответствующих десятичным номерам квартир, у которых в разряде j $(j = 1,2,3,4)$ стоит знак i $(i = 0,1,\ldots,9)$. Шкала BS_4 — почти минимальна в B_4.

Приведем пример домашнего адреса: 117036, Москва, ул. Дмитрия Ульянова, дом 24, квартира 168.

Вернемся теперь к вычислению количества информации $I(A(x_0))$. Так как сведение $\sigma_1(x_0)$ о населенном районе однозначно определяется шестизначным числом (почтовым индексом) в десятичной записи, то

$$I_1(\sigma_1(x_0)) = 6 \log 10 . \qquad (5)$$

Сведение $\sigma_2(x_0)$ — это название конкретного проспекта, улицы, площади, переулка, проезда и т. д. Атомарная шкала BS_2 определяется множеством всех проспектов, улиц и т. п. Количество информации, принесенной сведением $\sigma_2(x_0)$, будет равно

$$I_2(\sigma_2(x_0)) = \log n_2 , \qquad (6)$$

где $n_2 = \mathrm{Card} BS_2$ — количество всех возможных проспектов, улиц, площадей, переулков и т. д. Сведение $\sigma_3(x_0)$ — это n_3 (число разрядов в номере дома), а сведение $\sigma_4(x_0)$ есть n_4 (число разрядов в номере квартиры). Тогда имеем:

$$\begin{aligned} I_3(\sigma_3(x_0)) &= n_3 \log 10; \\ I_4(\sigma_4(x_0)) &= n_4 \log 10. \end{aligned} \qquad (7)$$

Из (5)–(7) окончательно находим, что

$$I(A(x_0)) = (6 + n_3 + n_4) \log 10 + \log n_2 .$$

Пример 4.

Пусть задано натуральное число $x_0 = 555$. Если начальное распределение является равномерным, то для носителя $A(x_0) = 555$ будем иметь

$$I(A(x_0)) = 3\log 10 .$$

Если заранее известно, что натуральное число имеет в десятичной записи с такого-то по такое-то место одинаковые знаки, например равные знаки сотен, десятков, единиц (три пятерки), то

$$I(A(x_0)) = I(\sigma_1(x_0)) + I_{\sigma_1}(\sigma_2(x_0)) + \\ + I_{\sigma_1\sigma_2}(\sigma_3(x_0)) = \log 10. \qquad (8)$$

В формуле (8) $I_{\sigma_1}(\sigma_2(x_0)) = I_{\sigma_1\sigma_2}(\sigma_3(x_0)) = 0$, где σ_1 — число сотен, σ_2 — число десятков, σ_3 — число единиц в остатке от деления числа x_0 на 10.

Пример 5.

Рассмотрим десятичную запись дроби $x_0 = \frac{5}{9} = 0,(5)$ (ноль целых и пять в периоде). Если начальное распределение является равномерным, то для носителя $A(x_0)$ имеем бесконечное значение

$$I(A(x_0)) = \infty .$$

Если же заранее известно, что после запятой все знаки равные, то

$$I(A(x_0)) = 2\log 10 .$$

Заключение

Таким образом, в работе исследуется важная числовая характеристика элементарной информации о точке. Количество (мера) одной и той же информации зависит от начального распределения, т. е. от пользователя, но не зависит от носителя этой информации. Показано, что наиболее удобным носителем для вычисления меры информации являются данные о точке относительно разложения решетки понятий в произведение подрешеток с атомарными шкалами. В случае, когда появляются большие или бесконечные значения меры информации, принесенной данными, эти значения можно уменьшить, если учесть имеющиеся начальные (исходные) сведения о точке.

Литература

1. *Макаров И. М., Ахрем А. А., Рахманкулов В. З.* Об основных понятиях теории решетчатых множеств // Труды ИСА РАН «Технология программирования и хранения данных». М.: Изд. дом «Либроком»/URSS, 2009. С. 220–227.
2. *Макаров И. М., Ахрем А. А., Рахманкулов В. З.* Алгебраические методы анализа данных //Труды ИСА РАН «Системные исследования. Методологические проблемы». Вып. 35 (в печати).
3. *Шеннон К.* Работы по теории информации и кибернетики. М.: ИЛ, 1963.
4. *Мазур М.* Качественная теория информации. М.: Мир, 1974.
5. *Стратонович Р. Л.* Теория информации. М.: Советское радио, 1975.
6. *Гоппа В. Д.* Введение в алгебраическую теорию информации. М.: Физмаилит, 1995.
7. *Чечкин А. В.* Математическая информация. М.: Наука, 1991.
8. *Соболева Т. С., Чечкин А. В.* Дискретная математика. М.: Академия, 2006.
9. *Бениаминов Е. М.* Алгебраические методы в теории баз данных и представлений знаний. М.: Научный мир, 2003.
10. *Вентцель Е. С., Овчаров Л. С.* Теория вероятностей. М.: Наука, 1973.
11. *Севастьянов Б. А.* Курс теории вероятностей и математической статистики. М.: Наука, 1982.

Макаров Игорь Михайлович. Советник Президента РАН. Д. т. н., профессор, академик РАН. Закончил МАИ в 1950 г. Количество печатных работ: более 250. Область научных интересов: гибкая автоматизация производства, интеллектуальная робототехника, теория управления, системный анализ, математическое моделирование.

Ахрем Андрей Афанасьевич. С. н. с. ИСА РАН. К. ф.-м. н. Окончил МГУ в 1977 г. Количество печатных работ: более 95. Область научных интересов: математическая теория систем, математическое моделирование сложных компьютерно-интегрированных производств.

Рахманкулов Виль Закирович. Заведующий лабораторией ИСА РАН. Д. т. н., профессор. Окончил МАИ в 1960 г. Количество печатных работ: более 155. Область научных интересов: гибкая автоматизация производства, интеллектуальная робототехника, теория управления, системный анализ, математическое моделирование. E-mail: vilrakh@mail.ru

Вашевник Татьяна Леонидовна. Научный сотрудник ИСА РАН. Окончила МЭСИ в 1980 г. Количество печатных работ: 17. Область научных интересов: математическое моделирование.

Филюков Руслан Юрьевич. Инженер-исследователь ИСА РАН. Окончил МИРЭА в 2002 г. Количество печатных работ: 3. Область научных интересов: математическое моделирование и искусственный интеллект.