**УДК** 006.72

И.Р. Баймуратов, Н.А. Жукова

# СИСТЕМА КОМПЛЕКСНОЙ ОЦЕНКИ ИНФОРМАТИВНОСТИ

*Баймуратов И.Р., Жукова Н.А.* **Система комплексной оценки информативности.**

**Аннотация.** На данный момент в теории информации существует множество различных подходов к определению количества информации. В данной работе кратко представлены основные существующие меры информативности. Представленный обзор позволяет сделать вывод о существование различных, несвязанных друг с другом определений информации. Большинство исследователей сходятся на том, что информация — это формальный концепт, который может иметь различные, невзаимоисключающие интерпретации. Таким образом, имеет место проблема обоснованной формальной систематизации этих методов. Цель данной работы — предложить подход, в рамках которого станет возможно, опираясь на основополагающие концепции теории информации, определить для различных существующих мер информативности общий формальный контекст и интегрировать их в единую систему.

В рамках предлагаемого подхода информативность определяется как количественная характеристика структурных свойств процессов обработки данных. При этом под процессами обработки данных подразумеваются любые возможные манипуляции с данными, начиная с математических формул и заканчивая алгоритмами машинного обучения. Предлагаемое решение заключается в определении общей модели процесса обработки данных и разработке общего метода оценки информативности. Общая модель процессов обработки данных основывается на теоретико-множественном анализе: процессы рассматриваются как отображения из множеств входных значений во множество выходных. Общий метод оценки информативности выводится из теоретико-игровых концепций: оценка информативности рассматривается как игра с природой, информативность — как функция выигрыша. Применение метода к различным элементам модели в совокупности образует систему комплексной оценки информативности.

Метод демонстрируется на примере комплексной оценки информативности логической формулы. Среди прочего, предложенный метод позволяет определить оптимальность данных с точки зрения практической ценности, а также степень абстрактности и конкретности данных.

**Ключевые слова:** Теория информации, системный подход, теория игр, логика

**1. Введение.** На данный момент в теории информации существует множество различных подходов к определению количества информации и множество мер информативности. Цель данной работы — предложить подход, в рамках которого станет возможно, опираясь на основополагающие концепции теории информации, определить единый формальный контекст и построить в этом контексте систему комплексного оценивания информативности.

Для достижения поставленной цели, во-первых, будут кратко представлены различные существующие на данный момент в теории информации определения количества информации. Затем мы предложим общий подход, позволяющий рассматривать различные существующие

определения количества информации в едином контексте, построить классификацию различных существующих мер информативности и систему комплексной оценки информативности данных. Мы проанализируем рассмотренные ранее определения количества информации в рамках предложенной классификации и продемонстрируем применение предложенной системы оценки информативности на данных логического типа.

**2. Обзор существующих определений информации.** В статье "Информация" [1] Стэнфордской энциклопедии философии перечисляются основные исторически сложившиеся концепции информации:

– *Информация Фишера* [2]: количество информации $I(\theta)$, которое случайная величина $X$ содержит о зависимой величине $\theta$. Пусть $f(x, \theta)$ — некоторая функция правдоподобия величины $\theta$. Если $f$ имеет резкие скачки, это значит, что соответствующие скачкам значения $X$ содержат большое количество информации о величине $\theta$. Эту зависимость и отражает информация Фишера $I(\theta)$, которая представляет собой дисперсию производной функции правдоподобия $f(x, \theta)$:

$$I(\theta) = \int (\tfrac{\delta}{\delta\theta} \log f(x, \theta))^2 f(x, \theta) dx.$$

– *Информация Шэннона* [3]: количество информации $I(x)$, которое содержит некоторое значение $x$ случайной величины $X$. Чем меньше вероятность $P(x)$ некоторого значения $x$, т.е. чем более оно "неожиданное", тем больше информации оно содержит:

$$I(x) = -\log P(x).$$

Количество информации, определенное для каждого значения случайной величины $X$, также является случайной величиной, тогда ее математическое ожидание, или информационная энтропия $H(X)$, определяет количество информации, содержащейся в случаной величине $X$ в целом:

$$H(X) = -\sum_i P(x_i) \log P(x_i).$$

– *Колмогоровская сложность* [4] [5] [6]: характеризует количество вычислительных ресурсов, необходимых для воспроизведения объекта $y$ на основе описания $x$. Определяется как длина кратчайшей программы $p$, входными данными которой является $x$, а результатом выполнения — $y$:

$$K(y|x) = min(l(p) : p(x) = y).$$

– *Квантовая информация*: обобщение шэнноновской концепции информации для квантовых состояний. Единицей информации, содержащейся в некотором квантовом состоянии, является кубит, который помимо дискретных значений 0 и 1 также может принимать значения из интервала $[0, 1]$. Количество информации, содержащейся во всей квантовой системе, описанной матрицей плотности $\rho$, измеряется энтропией фон Нейманна [7]:

$$S(\rho) = -Tr(\rho \ln \rho).$$

– *Информация как состояние агента*: логическая формализация таких понятий как знание и убеждение. Например, в эпистемической логике [8] утверждение, что некоторый $i$-ый агент знает факт $a$ истинно тогда и только тогда, когда $a$ истинно во всех достижимых для $i$-го агента мирах $w'$:

$$w \models K_i a \Leftrightarrow \forall w'(R_i(w, w') \rightarrow w' \models a).$$

– *Семантическая информация*: информация $inf(s)$ определяется как содержание высказывания $s$ и измеряется на основе его логической вероятности $q(s)$ [9]:

$$inf(s) = \log \tfrac{1}{q(s)}.$$

.

Мы дополняем этот список следующими определениями информативности:

– *Комбинаторная информация*: обычно рассматривается в рамках теории информации Шэннона, однако Колмогоров [5] выделил ее как отдельный подход. Для комбинаторного подхода существенна независимость от каких-либо вероятностных допущений. Комбинаторная информативность позволяет определить количество знаков, необходимое для кодирования в $m$-значном коде сообщения $x$ длиной $l$ в алфавите, состоящем из $N$ символов:

$$H(x) = l[\log_m N].$$

– *Информационный критерий Акаике* [10]: оценивает информативность статистической модели, имеющей $k$ параметров и функцию правдоподобия $L$, как разность сложности и точности модели:

$$AIC = 2k - 2ln(L).$$

В современных российских исследованиях также выделяется несколько способов определения колоичества информации. Например, в [11], где рассматривается проблема предварительного анализа информативности признаков, используемых в системах поддержки принятия решений, выделяется пять подходов к отбору информативных признаков:

– на основе дискретных методов поиска в обучающей выборке информативной зоны;

– на основе методов кластеризации;

– на основе предположения о нормальности распределений объектов в кластерах;

– на основе теоретико-информационного понятия энтропии;

– на основе непараметрических оценок плотности.

Выбор метода предлагается осуществлять на основе его эффективности на контрольной выборке.

Представленный обзор позволяет сделать вывод о существование различных, несвязанных друг с другом определений информации. Многие исследователи, например, авторы [12], объясняют этот факт тем, что информация — это формальный концепт, который может иметь различные, невзаимоисключающие интерпретации. Таким образом, современная теория информации содержит большой набор различных несвязанных методов. Имеет место проблема обоснованной формальной систематизации этих методов.

**3. Комплексный подход к определению информации.** Прежде всего определим, что под информацией мы подразумеваем количественную характеристику различных свойств процессов обработки данных. При этом под процессами обработки данных мы имеем в виду любые возможные манипуляции с данными, начиная с математических формул и заканчивая алгоритмами машинного обучения, включая сами данные как тождественное отображение в самих себя. Предлагаемое нами решение заключается в определении общей модели процесса обработки данных и разработке системы комплексной оценки информативности на основе этой модели. В рамках этого комплексного подхода различные меры информативности сопоставляются со свойствами структуры различных элементов процесса обработки данных. Таким образом, в первую очередь необходимо определить общую модель процесса обработки данных.

**3.1. Общая модель процесса обработки данных.** Общая модель процесса обработки данных — отдельное направление наших исследований. Подробее с ним можно ознакомиться в [13]. В рамках данной работы процесс обработки данных В общем виде предлагается представлять как множество вида

$$\{v_1 ... v_n y : V_1, ..., V_n \rightarrow_{f_1^1} ... \rightarrow_{f_l^m} Y\},$$

где $V_i, Y$ — некоторые множества, $v_i \in V_i$, $y \in Y$ и $f_j^i$ — некоторое $i$-ое отображение с порядковым номером $j$.

В этом множестве представляется возможным выделить следующие структурные элементы:

– признаковое пространство $V = V_1 \times ... \times V_n\}$;

– композиция отображений $F = f_1^1 ... f_l^m$;

– распределение выходных значений $Y_1, ..., Y_k$, таких что $\bigcup_i Y_i = Y$ и $Y_i = \{v_1 ... v_n y_i\}$.

Каждый из этих элементов также в свою очередь обладает некоторой структурой:

– для признакового пространства $V = V_1 \times ... \times V_n\}$: количество измерений $n$ и "размер" каждого измерения $|V_i|$;

– для композиции отображений $F = f_1^1 ... f_l^m$: длина композиции $l$ и количество различных отображений $m$;

– для распределения выходных значений $Y_1, ..., Y_k$: количество подмножеств $k$ и объем каждого подмножества $|Y_i|$.

Анализируя получившуюся модель, можно прийти к выводу, что структуру каждого из ее элементов в свою очередь можно рассматривать как трехуровневую иерархию, состоящую из следующих подчиненных элементов:

1. элемент $x$;

2. множество $X = \{x_1, ..., x_n\}$.

3. множество всех комбинаторно возможных альтернатив $\mathcal{X} = \{X_1, ..., X_m\}$.

При этом элементы первого уровня, будем называть их атомарными объектами, обладают некоторой величиной $|x|$: измерение $V_i$ обладает объемом $|V_i|$, композиция отображений $f_1^1 ... f_l^m$ обладает длиной $l$, подмножество выходных значений $Y_i$ также обладает объемом $|Y_i|$.

Таким образом, необходимо определить критерий оценки для каждого уровня этой иерархии. Полная структура приведена в Таблице 1.

Таблица 1. Общая структура процесса обработки данных

| | Элемент | Множество | Множество альтернатив |
|---|---|---|---|
| Признаковое пространство | Измерение $V_i$ | Пространство $V = \{V_1, ..., V_n\}$ | Множество возможных пространств $\mathcal{V}$ для различного числа $n$ |
| Композиции отображений | Композиция $f_1^1 ... f_l^m : X_1, ..., X_n \to Y$ | Набор композиций $F = \{f_1^1 ... f_l^m : X_1 \times ... \times X_n \to Y^Y\}$ | Множество наборов композиций $\mathcal{F}$ для различного $m$ |
| Выходные значения | Подмножество $Y$ | Распределение $Y_1, ..., Y_k$ | Множество возможных распределений $\mathcal{Y}$ для различного $k$ |

**3.2. Теоретико-игровое обоснование метода оценки информативности.** Для построения системы комплексной оценки информативности необходимо определить меры информативности элементов и множеств. Прежде всего, мы хотим дать теоретическое обоснование метода, который мы будем использовать для оценки информативности множеств.

Пусть заданы некоторое множество $X$ и некоторая функция информативности $I(x)$ элемента $x \in X$. Мы предлагаем рассматривать проблему оценки информативности элемента $x \in X$ как ситуацию выбора $x$, где функция информативности $I(x)$ играет роль функции выигрыша. При этом выигрыш $I(x)$ зависит от некоторой объективной величины $|x|$, определяемой в зависимости от контекста, поэтому проблему оценки информативности элемента $x$ можно рассматривать как игру с природой.

Определим эту игру как парную игру рационального игрока $X$ и природы П. Игрок $X$ обладает $n$ возможными стратегиями $\chi_1, ..., \chi_n$. Природа П также обладает множеством состояний $\pi_1, ..., \pi_m$. Результат выбора каждой из стратегий $\chi_i$ зависит от состояния природы $\pi_j$, поэтому определим результат выбора игрока $X$ некоторой стратегии $\chi_i$ при состоянии природы $\pi_j$ как исход $\chi_i\pi_j$. Игрок $X$ оценивает каждый исход $\chi_i\pi_j$ с помощью функции выигрыша $I(\chi_i\pi_j)$. Таким образом, игра может быть задана матрицей выигрыша, представленной в Таблице 1:

Таблица 2. Матрица выигрыша

| | $\pi_1$ | ... | $\pi_m$ |
|---|---|---|---|
| $\chi_1$ | $I(\chi_1\pi_1)$ | ... | $I(\chi_1\pi_m)$ |
| ... | ... | ... | ... |
| $\chi_n$ | $I(\chi_n\pi_1)$ | ... | $I(\chi_n\pi_m)$ |

В играх с природой при поиске оптимальной стратегии используются различные критерии. Использование того или иного критерия зависит от различных условий, таких как знание вероятностного распределения $P$ состояний природы П, заданной степени оптимизима $\gamma$, заданной степени доверия $\nu$ к распределению $P$ и т.д. Рассмотрим следующие критерии [14]:

– критерий Байеса. Если известно вероятностное распределение $P$ состояний природы П, то эффективность $H_b(\chi_i)$ стратегии $\chi_i$, согласно критерию Байеса, можно определить как средневзвешенный выигрыш $I(\chi_i\pi_j)$ с весами $P_1, ..., P_m$:

$$H_b(\chi_i) = \sum_j P_j I(\chi_i\pi_j).$$

Тогда ценностью $H_b(X)$ всей игры для игрока $X$ будет стратегия с максимальной эффективностью:

$$H_b(X) = \max_i H(\chi_i) = \max_i \sum_j P_j I(\pi_j \chi_i).$$

– критерий Лапласа. Если вероятностное распределение $P$ не известно, то можно предположить, что состояния природы $\pi_j$ равновероятны, тогда эффективность $H_l(\chi_i)$ стратегии $\chi_i$, согласно критерию Лапласа, определяется формулой

$$H_l(\chi_i) = \frac{1}{m} \sum_j I(\chi_i \pi_j),$$

а ценностью $H_l(X)$ игры будет стратегия с максимальной эффективностью:

$$H_l(X) = \max_i H(\chi_i) = \max_i \sum_j \frac{1}{m} I(\chi_i \pi_j).$$

– критерий Вальда. Если вероятностное распределение $P$ не известно, то игрок может стремиться минимизировать свои риски. Тогда эффективность $H_v(\chi_i)$ стратегии $\chi_i$, согласно критерию Вальда, определяется формулой

$$H_v(\chi_i) = \min_j I(\chi_i \pi_j),$$

а ценностью $H_v(X)$ игры будет стратегия с максимальной эффективностью:

$$H_v(X) = \max_i H(\chi_i) = \max_i \min_j I(\chi_i \pi_j).$$

Таким образом, информативность $H(X)$ множества $X$ можно оценить различными способами:

– $H_b(X) = \max_i \sum_j P_j I(\chi_i \pi_j)$, если известно вероятностное распределение $P$;

– $H_l(X) = \max_i \sum_j \frac{1}{m} I(\chi_i \pi_j)$, если состояния природы $\pi_j$ равновероятны;

– $H_v(X) = \max_i \min_j I(\chi_i \pi_j)$, если вероятностное распределение $P$ не известно.

В дальнейшем мы будем предполагать, что вероятностное распределение $P$ состояний природы $\Pi$ известно, следовательно, мы будем использовать критерий Байеса $H_b(X)$ как определение информативности $H(X)$ множества элементов $X$.

**3.3. Метод оценки информативности.** Теперь мы можем определить метод оценки информативности структурных элементов процесса обработки данных. Согласно общей структуре процесса обработки данных, необходимо определить меры информативности

– некоторого произвольного элемента $x \in X$;

– самого множества $X$;

– и множества всех комбинаторно возможных альтернатив $\mathcal{X}$.

А, согласно теоретически-игровому обоснованию, оценку информативности можно рассматривать как игру игрока $X$ с природой $\Pi$, где

– информативность $I(x)$ элемента $x$ — это функция выигрыша $I(\chi_i\pi_j)$, с помощью которой игрок $X$ оценивает исход $\chi_i\pi_j$;

– информативность $H(X)$ множества $X$ — это средневзвешенный выигрыш $H(X) = \sum_j P_j I(\chi_i\pi_j)$.

Таким образом, остается только задать функцию выигрыша $I(\chi_i\pi_j)$.

Пусть полезность стратегии $\chi$ при состоянии природы $\pi_j$ для игрока $X$ определяется некоторой величиной $|x|$, которая нормирована на суммарную величину элементов $|X|$ с помощью логарифмической функции, тогда информативность $I(x)$ элемента $x$, соответствующего стратегии $\chi$ определяется формулой

$$I(x) = \log_{|X|} |x|.$$

Рассмотрим пару комплементарных понятий: информативность $I(x)$ и сложность $C(x)$. Сложность $C(x)$ равна:

$$C(x) = 1 - \log_{|X|} |x| = -\log_{|X|} \tfrac{|x|}{|X|},$$

где $\frac{|x|}{|X|}$ можно рассматривать как частоту $P(x)$, тогда определение сложности $C(x)$ совпадает с нормированным определением собственной информации

$$C(x) = -\log_{|X|} P(x).$$

Далее для избежания усложнения мы будем рассматривать только информативность.

Далее, оценку информативности множества $X = \{x_1, ..., x_n\}$ можно рассматривать как игру, в которой игрок $X$ имеет всего одну стратегию $x$, которую можно интерпретировать как выбор произвольного элемента $x$ из множества $X$, а природа $\Pi$ имеет $n$ состояний, таких что $\chi\pi_i = |x_i|$, тогда, согласно критерию Байеса, информативность множеста $X$ равна средневзвешенной информативности $I(x_i)$:

$$H(X) = \sum_i P_i I(x_i),$$

что совпадает с определением информационной энтропии.

Мы предлагаем также рассматривать множество элементов $X$ как один из вариантов среди множества альтернатив $\mathcal{X}$, обладающего вероятностным распределением $P$, определяемым в зависимости от контекста. В сравнении с этим множеством альтернатив $\mathcal{X}$ становится возможным оценить само множество $X$.

Информативность $H(\mathcal{X})$ множества возможных альтернатив $\mathcal{X} = \{X_1, ..., X_n\}$ мы предлагаем определить аналогичным образом — как

выбор некоторого набора стратегий $X \in \mathcal{X}$ при $n$ состояний природы, таких что $X\pi_i = |X_i|$, т.е. как средневзвешенную информативность $H(X_i)$:

$$H(\mathcal{X}) = \sum_i P_i H(X_i).$$

Таким образом, мы можем определить информативность каждого структурного элемента процесса обработки данных.

**4. Система комплексной оценки информативности.** В данном разделе мы конкретизируем метод оценки информативности для каждого элемента общей модели процессов обработки данных.

**4.1. Информативность признакового пространства.** Согласно общей структуре процесса обработки данных, структура признакового пространства состоит из

- измерения $V_i$;
- пространства $V_1, ..., V_n$;
- множества возможных пространств $\mathcal{V}$.

В результате подстановки этих элементов в определения мер информативности получаем

- информативность измерения $I(V_i) = \log_{|V_1 ... V_n|} |V|$;
- информативность пространства $H(V) = \sum_i P_i I(V_i)$;
- информативность множества возможных пространств $H((V)) = \sum_i P_i H(V)$.

Таким образом, остается только определить размер $|V_i|$ измерения $V_i$, размер $|V|$ признакового пространства $V$ и закон распределения $P(V_i)$.

Определим $|V_i|$ измерения $V_i$ как количество его уникальных компонент $v_1, ..., v_m \in V_i$ и размер $|V|$ пространства $V$ как произведение $|V_1|...|V_n|$. Тогда информативность $I(V_i)$ измерения $V_i$ равна:

$$I(V_i) = \log_{|V|} |V_i|,$$

а информативность $H(V)$ всего пространства $V$:

$$H(V) = \sum_i P_i \log_{|V|} |V_i|,$$

где $P_i = \frac{1}{n}$, т.е.

$$H(V) = \sum_i \frac{1}{n} \log_{|V|} |V_i| = \frac{1}{n}$$

для $|V_i| > 1$.

Это определение основано на интуитивном представлении, что размерность, состоящая всего из одной компоненты, в некотором смысле вырождена, неинформативна и ее можно заменить на константу, сократив тем самым размерность пространства. В самом деле:

$$I(V_i) = 0 \Leftrightarrow |V_i| = 1.$$

И наоборот, если некоторое измерение содержит в себе все компоненты пространства, это измерение содержит всю информацию:

$$I(V_i) = 1 \Leftrightarrow |V_i| = |V|.$$

Определим, что множество альтернативных пространств $(V)$ для некоторого пространства $V$ — это множество всех возможных $V'$, таких что $|V'| = |v|$, в том числе и само $V$:

$$(V) = \{V' : \forall V' |V'| = |V|.$$

Поскольку $|V| = |V_1|...|V_n|$, множество альтернативных пространств $(V)$ можно рассматривать как множество $Fact(|V|)$ всех возможных факторизаций $fact_i(|V|)$ числа $|V|$. Для краткости будем говорить о факторизациях $Fact(V)$ пространства $V$. Тогда информативность $H(\mathcal{V})$ определяется как математическое ожидание $H(fact_i(V))$ по каждой факторизации $fact_i(V) \in Fact(V)$:

$$H(\mathcal{V}) = \sum_i P_i H(fact_i(V)),$$

где $P_i = \frac{1}{|Fact(V)|}$, а $|Fact(V)|$ — объем множества $Fact(|V|)$. Либо представляется возможным определить подмножество $Fact_i(V) \subseteq Fact(V)$ факторизаций пространства $V$, различающихся лишь порядком элементов, так как их порядок не влияет на значение $H(V)$. Объем $|Fact_i(V)|$ множества $Fact_i(V)$ вычисляется как количество перестановок с повторениями:

$$|Fact_i(V)| = \frac{n!}{n_1!...n_m!},$$

где $n$ — количество элементов факторизации $Fact_i(V) = |V_1|...|V_n|$ и $n_i$ — количество совпадающих элементов некоторой величины. Тогда

$$P_i = \frac{|Fact_i(V)|}{|Fact(V)|}$$

и

$$H(\mathcal{V}) = \sum_i \frac{|Fact_i(V)|}{|Fact(V)|} H(Fact_i(V_i)).$$

**4.2. Информативность композиции отображений.** Согласно общей структуре, структура композиции отображений состоит из

- композиции $f = f_1^1...f_l^m : X_1, ..., X_n \to Y$;
- набора композиций $F = \{f : X_1 \times ... \times X_n \to Y^Y\}$;
- множества наборов композиций $\mathcal{F}$.

В результате подстановки в общие определения получаем

- $I(f) = \log_{|F|} |f|$;
- $H(F) = \sum_i P_i I(f_i)$;
- $H((F)) = \sum_i P_i H(F)$.

Таким образом, достаточно определить размер $|f|$ композиции $f$, размер $|F|$ набора композиций $F$ и закон распределения $P(f)$.

Прежде всего уточним, что мы имеем в виду под композицией отображений и набором отображений. Пусть имеется некоторое отображение $f : X_1 \times ... \times X_n \to Y$, тогда существует некоторая минимальная композиция $f_1^1 ... f_l^m$ элементарных отображений $f^1, ..., f^m$, эквивалентная отображению $f$ в следующем смысле:

$$\forall x_1, ..., x_n, y f(x_1, ..., x_n) = y \Rightarrow f_1^1 ... f_l^m(x_1, ..., x_n) = y.$$

Далее, под набором композиций $F$ мы имеем в виду множество всех возможных отображений из заданных множеств $X_1, ..., X_n$ в заданное множество $Y$ при заданном наборе элементарных отображений $f^1, ..., f^m$, т.е. множество отображений $f$ из множеств $X_1, ..., X_n$ в множество отображений $Y$ в само себя:

$$F = \{f : X_1 \times ... \times X_n \to Y^Y\}.$$

Таким образом, определим $|f|$ как длину $l$ композиции $f_1^1 ... f_l^m$ и $|F|$ как объем набора $F$. Количество композиций $|F|$ определяется формулой

$$|F| = |Y|^{|X_1| ... |X_n|}.$$

Следовательно, для заданного числа элементарных отображений $m$ длина $l$ композиции $f_1^1 ... f_l^m$ имеет место неравенство:

$$l \leqslant \log_m |F|.$$

В противном случае композиция $f_1^1 ... f_l^m$ избыточна.

Информативность $I(f)$ композиции $f$ равна:

$$I(f) = \log_{|F|} |f|,$$

а информативность $H(F)$ набора композиций $F$:

$$H(F) = \sum_i P_i \log_{|F|} |f_i|,$$

где $P_i = \frac{1}{|F|}$, т.е.

$$H(F) = \sum_i \frac{1}{|F|} \log_{|F|} |f_i|.$$

Также можно рассматривать подмножество композиций $F_i \subseteq F$ одинаковой длины $l$:

$$F_i = \{f' : |f'| = |f_i|\},$$

тогда $P_i = \frac{|F_i|}{|F|}$ и

$$H(F) = \sum_i \frac{|F_i|}{|F|} \log_{|F|} |f_i|.$$

Это определение основано на интуитивном восприятии композиции отображений как более информативной, чем эквивалентное ей элементарное отображение, так как оно явным образом раскрывает процесс обработки данных. Информативность элементарного отображения минимальна:

$$I(f) = 0 \Leftrightarrow |f| = 1,$$

а информативность композиции $f_1^1 ... f_l^m$, наоборот, максимальна, если она максимально полно раскрывает процесс обработки:

$$I(f) = 1 \Leftrightarrow |f| = |F|.$$

Определим, что множество наборов композиций $(F)$ для некоторых множеств $X_1, ..., X_n$ и $Y$ — это множество наборов $F'$ композиций $f_1^1 ... f_l^m$ с различными наборами элементарных отображений $f^1, ..., f^m$:

$$(F) = \{F' : \forall f^1, ..., f^m | F'| = |F|\}.$$

Тогда информативность $H(\mathcal{F})$, согласно предложенному методу оценки информативности, определяется как математическое ожидание $H(F)$ для каждого набора $F \in \mathcal{F}$:

$$H(\mathcal{F}) = \sum_i P_i H(F),$$

где $P_i = \frac{1}{|\mathcal{F}|}$.

Поскольку состав и порядок наборов элементарных отображений $f^1, ..., f^m$ не влияет на величину $H(F)$, возможно рассматривать сочетания $comb(F)$ элементарных отображений $f^1, ..., f^m \in F$. Количество сочетаний $|comb(F)|$ для заданного набора элементарных отображений $f^1, ..., f^m$ определяется формулой

$$comb(F) = \frac{n!}{m!(n-m)!},$$

где $n = |F|$, а множество наборов композиций $\mathcal{F}$ становится возможно определить как объединение сочетаний $Comb(F) = \bigcup_i comb_i(F)$ и объем $|\mathcal{F}|$ — как сумму объемов сочетаний:

$$|Comb(F)| = \sum_i |comb_i(F)|.$$

Таким образом,

$$P_i = \frac{|comb(F)|}{|Comb(F)|}$$

и

$$H(\mathcal{F}) = \sum_i \frac{|comb_i(F)|}{|Comb(F)|} H(F).$$

**4.3. Информативность распределения выходных значений.** Структура распределения выходных значений состоит из

- подмножества $Y_i$;
- распределения $Y = \{Y_1, ..., V_k\}$;
- множества возможных распределений $\mathcal{Y}$.

В результате подстановки полуаем

- $I(Y_i) = \log_{|Y|} |Y_i|$;
- $H(Y) = \sum_i P_i I(Y_i)$;
- $H((Y)) = \sum_i P_i H(Y)$.

Следовательно, для определения инфомративности распределения выходных значений достаточно определить размер $|Y_i|$ подмножества $Y_i$, размер $|Y|$ распределения $Y$ и закон распределения $P(Y_i)$.

Напомним, что $Y_i = \{x_1...x_n y_i\}$ для некоторого $y_i$, тогда определим объем $|Y_i|$ как объем множества $\{x_1...x_n y_i\}$, а объем $|Y|$ — как объем объединения $\bigcup_i Y_i$, т.е. как объем множества $\{x_1...x_n y\}$ для каждого $y \in Y$. Тогда информативность $I(Y_i)$ подмножества $Y_i$ определяется формулой

$$I(Y_i) = \log_{|Y|} |Y_i|,$$

информативность $H(Y)$ распределения $Y$ — формулой

$$H(Y) = \sum_i P_i \log_{|Y|} |Y_i|,$$

где $P_i = \frac{|Y_i|}{|Y|}$, т.е.

$$H(Y) = \sum_i \frac{|Y_i|}{|Y|} \log_{|Y|} |Y_i|.$$

Таким образом, подмножество, состоящая всего из одного элемента неинформативно:

$$I(Y_i) = 0 \Leftrightarrow |Y_i| = 1,$$

а подмножество, которое содержит в себе все элементы множества, содержит в себе всю информацию:

$$I(Y_i) = 1 \Leftrightarrow |Y_i| = |Y|.$$

Определим множество всех возможных распределений $\mathcal{Y}$ как множество всех возможных отображений множества $Y$ в самого себя:

$$\mathcal{Y} = Y^Y,$$

следовательно,

$$|\mathcal{Y}| = |Y|^{|Y|}$$

и

$$H(\mathcal{Y}) = \sum_i P_i H(Y),$$

где $P_i = \frac{1}{|\mathcal{Y}|}$.

Поскольку величина $H(Y)$ не зависит от порядка либо состава элементов в подмножествах $Y_i \subseteq Y$, имеет смысл рассматривать $\mathcal{Y}$ как множество всех возможных разбиений $Part(Y) = \{part(Y)\}$, где

$$part(Y) = \{Y' : \forall Y_i \subseteq Y \forall Y_i' \subseteq Y' |Y_i| = |Y_i'|\}.$$

Объем $|part(Y)|$ разбиения $part(Y)$ вычисляется по формуле

$$|part(Y)| = \frac{n!}{n_1!...n_k!} \frac{n!}{k!(n-k)!} \frac{k!}{k_1!...k_m!},$$

где $n = |Y|, n_i = |Y_i|, k$ — количество подмножеств и $k_j$ — количество подмножеств с некоторым одинаковым количеством элементов $n_i$. Таким образом,

$$P(part(Y)) = \frac{|part(Y)|}{|\mathcal{Y}|}$$

и

$$H(\mathcal{Y}) = \sum_i \frac{|part_i(Y)|}{|Part(Y)|} H(part_i(Y)).$$

**5. Комплексная оценка информативности логических формул.** Рассмотрим практическое применение предложенной системы комплексной оценки информативности. Напомним, что в общем виде процесс обработки данных можно представить в виде отображения

$$\{v_1...v_n y : V_1, ..., V_n \rightarrow_{f_1^1} ... \rightarrow_{f_l^m} Y\}.$$

Поскольку информативность элементов процесса обработки данных зависит только от объемов множеств $V_1, ..., V_n$ и $Y$, имеет смысл разбить области практического применения системы оценки информативности в зависимости от типа обрабатываемых данных. Мы предлагаем рассматривать следующие типы:

– логический;
– целый;
– вещественный.

Для демонстрации предложенной системы мы рассмотрим ее применение к данным логического типа.

Пусть имеется следующая логическая формула:

$$a \wedge b \rightarrow c.$$

Ее таблица истинности приведена в Таблице 3.

Таблица 3. Таблица истинности

| $a$ | $b$ | $c$ | $a \wedge b$ | $a \wedge b \rightarrow c$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 1 |

Оценим информативность этой формулы.

**Информативность признакового пространства.** Для оценки информативности признакового пространства этой формулы необходимо определить объем каждого из признаков. Поскольку каждый из признаков является пропозициональной переменной, а пропозициональная переменная может принимать значения 0 и 1,

$$|a| = 2, |b| = 2 \text{ и } |c| = 2.$$

Таким образом, объем всего признакового пространства

$$|a \times b \times c| = |a||b||c| = 2 * 2 * 2 = 8.$$

Информативность каждого признака

$$|a| = \tfrac{1}{3}, |b| = \tfrac{1}{3} \text{ и } |c| = \tfrac{1}{3}.$$

Наконец, информативность признакового пространства

$$I(a \times b \times c) = \tfrac{1}{3}.$$

Таблица 4. Возможные признаковые пространства

| $Fact(V)$ | $H(Fact(V))$ | $P(Fact(V))$ | P(Fact(V))H(Fact(V)) |
|---|---|---|---|
| 2*2*2 | 1/3 | 0,25 | 1/12 |
| 4*2 | 0,5 | 0,5 | 0,25 |
| 8 | 1 | 0,25 | 0,25 |

Информативность возможных признаковых пространств приведена в Таблице 4.

Таким образом, информативность множества признаковых пространств равна 0,58(3). Пространство $2 * 2 * 2$ обладает наименьшей взвешенной информативностью $P(Fact(V))H(Fact(V))$ и его информативность $H(Fact(V))$ меньше информативности множества возможных пространств, т.е. меньше средневзвешенного значения.

**Информативность композиции отображений.** Для оценки информативности композиции отображений этой формулы необходимо определить количество используемых элементарных отображений $m$ и объем набора отображений $F$. В формуле представлены две пропозициональные операции, $\wedge$ и $\rightarrow$, следовательно, $m = 2$, а $|F| = |a \wedge b \rightarrow c|^{|a||b||c|} = 2^{2*2*2} = 256$. Таким образом,

$$I(\wedge \rightarrow) = 0,125.$$

Расчет информативности набора композиций $F$ приведен в Таблице 5. Информативность $H(F) \approx 0,32$.

Таблица 5. Набор композиций отображений

| $l$ | $I(f_i)$ | $P(F_i)$ | $P(F_i)I(f_i)$ |
|---|---|---|---|
| 1 | 0 | 1/128 | 0 |
| 2 | 0,125 | 1/64 | 0,002 |
| 3 | 0,2 | 1/32 | 0,00625 |
| 4 | 0,25 | 1/16 | 0,015625 |
| 5 | 0,29 | 1/8 | 0,03625 |
| 6 | 0,323 | 1/4 | 0,08075 |
| 7 | 0,35 | 1/2 | 0,175 |
| 8 | 0,375 | 1/128 | 0,003 |

Мы не будем приводить полный расчет информативности множества наборов комбозиций, т.к. он состоит из 256 строк, фрагмент расчета приведен в Таблице 6.

Таблица 6. Множество наборов композиций

| $m$ | $H(F)$ | $P(F)$ | P(F)H(F) |
|---|---|---|---|
| 1 | 0,82 | 2,2108591e-75 | 1,8179095e-75 |
| 2 | 0,32 | 2,8188454e-73 | 8,998289e-74 |
| ... | ... | ... | ... |
| 255 | 0,0005 | 2,2108591e-75 | 1,079521e-78 |
| 256 | 0 | 8.6361685e-78 | 0 |

Таким образом, информативность множества наборов отображений равна 0,0625. Информативность набора элементарных отображений $\wedge, \rightarrow$ значительно превышает средневзвешенное значение, хотя взвешенная информативность набора близка к минимальной.

**Информативность распределения значений.** Для оценки информативности распределения значений этой формулы необходимо определить объем каждого подмножества векторов значений таблицы истинности формулы. Формула истинна в 7 случаях и ложна в 1, т.е. $|\top| = 7$ и $|\bot| = 1$. Дальнейший расчет информативности подмножеств приведен в Таблице 7.

Таблица 7. Подмножества

| $Y_i$ | $I(Y_i)$ | $P(Y_i)$ | $P(F_i)I(f_i)$ |
|---|---|---|---|
| $\top$ | 0,9358 | 0,875 | 0,82 |
| $\bot$ | 0 | 0,125 | 0 |

Информативность $H(Y) \approx 0,82$.

Фрагмент расчета информативности множества распределений значений приведен в Таблице 8.

Таблица 8. Множество распределений значений

| $Part(Y)$ | $H(Part(Y))$ | $P(Part(Y))$ | P(Part(Y))H(Part(Y)) |
|---|---|---|---|
| $[1, 1, 1, 1, 1, 1, 1, 1]$ | 0 | 0.0024 | 0 |
| $[2, 1, 1, 1, 1, 1, 1]$ | 0,08(3) | 0.0673 | 0,0056 |
| ... | ... | ... | ... |
| $[7, 1]$ | 0,82 | 2.670288e-05 | 2,1864635e-05 |
| $[8]$ | 1 | 4.7683716e-07 | 4.7683716e-07 |

Таким образом, информативность множества распределений значений равна 0,2514. Взвешенная информативность соответствующего

формуле разбиения $[7, 1]$ близка к минимальной, однако не взвешенная информативность разбиения значительно выше средневзвешенного значения.

Подводя итог, формула $a \wedge b \to c$ обладает информативными композицией отображений и распределением значений, неинформативным признаковым пространством, все взвешенные показатели информативности близки к минимальным. Мы интерпретируем это следующим образом:

– высокое взвешенное значение говорит о балансе между конкретной и абстрактной информацией, поэтому соответствующие структуры представляют наибольшую практическую ценность;

– низкое взвешенное значение при высокой информативности говорит об излишней абстрактности информации, однако такие структуры полезны с теоретической точки зрения;

– низкое взвешенное значение при низкой информативности говорит об излишней конкретности.

Следовательно, данная формула излишне абстракта, что согласуется с интуитивным восприятием. Также можно ознакомиться с примером применения мер оценки информативности распределения значений в рамках задачи кластеризации в [15].

**6. Заключение.** Таким образом, мы провели обзор существующих подходов к определению информативности, чтобы продемонстрировать, что существующие определения не связаны друг с другом и что в теории информации недостает теоретической основы для их систематизации. В качестве решения мы предложили системный подход к определению информации в рамках которого определили общую модель процесса обработки данных и структурные элементы этой модели, информативность которых необходимо оценивать. Далее, мы предложили общий метод оценки информативности, обосновали его с помощью теории игр и показали, что существующие методы оценки информации, такие как собственная информация и информационная энтропия, выводимы из него. Наконец, мы конкретизировали применение предложенного метода для оценки информативности различных структурных элементов процесса обработки данных и продемонстрировали работу метода на примере данных логического типа. Полученные результаты согласуются с интуитивным восприятием информативности.

## Литература


1. *Adriaans P.* Information // Stanford Encyclopedia of Philosophy, 2012, https://plato.stanford.edu/entries/information/
2. *Fisher R.A.* Theory of statistical estimation // Proceedings Cambridge Philosophical Society, 1925, 22(5), pp. 700–725.

3. *Shannon C.E.* A Mathematical Theory of Communication // Bell System Technical Journal, 1948. Vol. 27, pp. 379-423.
4. *Solomonoff R.J.* A Formal Theory of Inductive Inference // Information and Control, 1964, v. 7, No. 1, pp. 1–22; No.2, pp. 224–254
5. *Колмогоров А.Н.* Три подхода к определению понятия количество информации. Пробл. передачи информ., 1965, том 1, выпуск 1, С. 3-11
6. *Chaitin G.J.* On the Length of Programs for Computing Finite Binary Sequences // Journal of Association for Computing Machinery, 1966, v. 13, No. 4, pp. 547–569
7. *Von Neumann J.* Mathematische Grundlagen der Quantenmechanik, 1955, Berlin: Springer.
8. *Hintikka J.* Knowledge and Belief, 1962, Cornel University Press, Ithaca.
9. *Bar-Hillel Y., Carnap R.* An Outline of a Theory of Semantic Information // Technical Report No. 247, October 27, 1952, Research Laboratory of Electronics. – 49.
10. *Akaike H.* A new look at the statistical model identification // IEEE Transactions on Automatic Control, 1974, Vol. 19, Pp. 716—723.
11. *С.И. Колесникова* Методы анализа информативности разнотипных признаков // Вестник Томского Государственного Университета, 2009, №1(6), сс. 69-80.
12. *Fortin S., Lombardi O., Vanni L.* A pluralist view about information // Philosophy of Science, 2015, 82(5), pp. 1248-1259.
13. *Baymuratov I., Zhukova N.* Logical and Mathematical Models of Data Fusion // International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), 2017, pp. 121-126.
14. *Блягоз З.У., Попова А.Ю.* Принятие решений в условиях риска и неопределенности // Вестник Адыгейского государственного университета. 2006. №4.
15. *Baymuratov I., Zukova N.* An Approach to Clustering Models Estimation // 22nd Conference of Open Innovations Association FRUCT, 2018, pp. 19-24.

**Баймуратов Ильдар Раисович** — аспирант кафедры информатики и прикладной математики Университета ИТМО. Область научных интересов: математическая логика, теория информации, интеллектуальный анализ данных. baimuratov.i@gmail.com; Университет ИТМО, Кронверкский пр-кт, д. 49, г. Санкт-Петербург, 197101.

**Жукова Наталия Александровна** — к.т.н., ст. науч. сотр. СПИИРАН, доцент кафедры информатики и прикладной математики Университета ИТМО. Область научных интересов: интеллектуальный анализ данных, машинное обучение, интеллектуальные системы и технологии. nazhukova@mail.ru; Университет ИТМО, Кронверкский пр-кт, д. 49, г. Санкт-Петербург, 197101.

I.R. Baimuratov, N.A. Zhukova

# COMPLEX INFORMATION ESTIMATION SYSTEM

*Baimuratov I.R., Zhukova N.A.* **Complex information estimation system.**

**Abstract.** There are numerous approaches to information estimation present at the moment in information theory. Some of them are briefly reviewed in the text. With this overview it is suggested that these numerous definitions of information are incoherent. Most of researches conclude that information is a formal concept which has different compatible interpretations. It means that there is a lack of formal foundation for these concepts. Therefore, we aim to develop an approach, which would allow to define a general framework for different information measures and integrate them into one system, using fundamental mathematical concepts.

In this framework information is considered as quantitative characterization for structural properties of data processing. And by data processing we understand any possible manipulations with data, from mathematical functions to machine learning. The proposal is to define the general data processing model and to develop the general method of information estimation. The general data processing model is produced by set-theoretical analysis. Data processing is considered as projection from input sets to output set. The general information estimation method is inferred from game-theoretical concepts. Information estimation is considered as game with nature and information — as utility function. The system of complex information estimation is formed by applying the method to different elements of the model.

The system workflow is examplified with complex information estimation of logic formula. Among other results, the method allows to define is data optimal from practical point of view or is it abstract or detailed.

**Keywords:** Information Theory, Systems theory, Game theory, Logic

**Baimuratov Ildar Raisovich** — postgraduate, Informatics and Applied mathematics Department, ITMO University. Research interests: mathematical logic, information theory, data analysis. baimuratov.i@gmail.com; ITMO University, 49 Kronverksky Pr., St. Petersburg, 197101, Russia.

**Zhukova Nataly Alexandrovna** — Candidate of Engineering Sciences, senior researcher, SPIIRAS, associated professor, Informatics and Applied mathematics Department, ITMO University. Research interests: data analysis, machine learning, intelligent systems. nazhukova@mail.ru; ITMO University, 49 Kronverksky Pr., St. Petersburg, 197101, Russia.

## References

1. *Adriaans P.* Information // Stanford Encyclopedia of Philosophy, 2012, https://plato.stanford.edu/entries/information/
2. *Fisher R.A.* Theory of statistical estimation // Proceedings Cambridge Philosophical Society, 1925, 22(5), pp. 700–725.
3. *Shannon C.E.* A Mathematical Theory of Communication // Bell System Technical Journal, 1948. Vol. 27, pp. 379-423.
4. *Solomonoff R.J.* A Formal Theory of Inductive Inference // Information and Control, 1964, v. 7, No. 1, pp. 1–22; No.2, pp. 224–254
5. Kolmogorov A.N. [Three approaches to the definition of the concept “quantity of information”]. *Probl. Peredachi Inf.*, 1965. no., Vol.1, pp. 3—11. (In Russ.).
6. *Chaitin G.J.* On the Length of Programs for Computing Finite Binary Sequences // Journal of Association for Computing Machinery, 1966, v. 13, No. 4, pp. 547–569
7. *Von Neumann J.* Mathematische Grundlagen der Quantenmechanik, 1955, Berlin: Springer.

8. *Hintikka J.* Knowledge and Belief, 1962, Cornel University Press, Ithaca.
9. *Bar-Hillel Y., Carnap R.* An Outline of a Theory of Semantic Information // Technical Report No. 247, October 27, 1952, Research Laboratory of Electronics. – 49.
10. *Akaike H.* A new look at the statistical model identification // IEEE Transactions on Automatic Control, 1974, Vol. 19, Pp. 716—723.
11. *С.И. Колесникова* Методы анализа информативности разнотипных признаков // Вестник Томского Государственного Университета, 2009, №1(6), сс. 69-80.
12. *Fortin S., Lombardi O., Vanni L.* A pluralist view about information // Philosophy of Science, 2015, 82(5), pp. 1248-1259.
13. *Baymuratov I., Zhukova N.* Logical and Mathematical Models of Data Fusion // International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), 2017, pp. 121-126.
14. *Блягоз З.У., Попова А.Ю.* Принятие решений в условиях риска и неопределенности // Вестник Адыгейского государственного университета. 2006. №4.
15. *Baymuratov I., Zukova N.* An Approach to Clustering Models Estimation // 22nd Conference of Open Innovations Association FRUCT, 2018, pp. 19-24.