The Information Estimation System for Data Processing Results

Nataly Zhukova, Ildar Baimuratov, Nguyen Than
ITMO University
St. Peterburg, Russia
nazhukova@mail.ru,
{baimutov.i,nguyenngocthan92}@gmail.com

Nikolay Mustafin
Saint-Peterburg Electrotechnical University "LETI"
St. Peterburg, Russia
ngmustafin@etu.ru

***Abstract*— Information measures can be used to evaluate data processing results. At the moment there are many different unrelated information concepts. The purpose of this paper is to propose a theoretical classification for these concepts. Existing information concepts are overviewed and analysed. As a result, three basic information definitions are identified. Next, these basic definitions are derived from a common basis - the function concept. The resulting classification can be used for a comprehensive evaluation of data processing results, as demonstrated by the example of logical formulas.**

I. Introduction

The results of data processing can be evaluated in various ways. One of such methods is the information evaluation. Information evaluation is most relevant in relation to tasks, in which there is no idea of the final result. An example of such tasks is machine learning without a teacher, for which, unlike training with a teacher, it is impossible to apply accuracy criteria.

At the moment there are many different, unrelated approaches to determining the amount of information. Some researchers propose to consider these approaches as different, non-exclusive interpretations. In this paper it is proposed to consider the set of information definitions as an integrated system, in which each definition corresponds to a certain formal property of the information bearer. In other words, the purpose of this paper is to construct a theoretically grounded classification of information measures.

To achieve this goal, first of all, a brief overview of the various existing definitions of the amount of information is presented. Then a preliminary analysis of them is made in order to determine which of the existing concepts are fundamental and which are derivatives. Further, a formal foundation for fundamental concepts classification is given and the classification itself is constructed. Finally, the application of resulting evaluation system is demonstrated, using logical formulas as examples.

II. Review of existing information definitions

The article "Information" [1] of the Stanford Encyclopedia of Philosophy lists the main historical information concepts. We present this list to demonstrate their heterogeneity:

- *Fisher Information* [2]: the amount of information $I_F(\theta)$, which random variable X contains about the dependent value θ. Let $f(x, \theta)$ be some likelihood function of θ. If f has sharp jumps, this means that the values of X, corresponding to the jumps, contain a large amount of information about the value of θ. This dependence is reflected in the Fisher information $I_F(\theta)$, which is the variance of the derived likelihood function $f(x, \theta)$:

$$I_F(\theta) = \int \left(\frac{\delta}{\delta\theta} log f(x,\theta) \right)^2 f(x,\theta) dx \quad (1)$$

- *Shannon information* [3]: the amount of information $I_S(x)$, which contains some value x of a random variable X. The smaller the probability $P(x)$ of some value x, i.e. the more "unexpected" it is, the more information it contains:

$$I_F(x) = -logP(x) \quad (2)$$

  The amount of information defined for each value of the random variable X is also a random variable, then its expectation, or information entropy $H(X)$, determines the amount of information contained in the random value X as a whole:

$$H(X) = -\sum_i P(x_i) log P(x_i) \quad (3)$$

- *Kolmogorov complexity* [4] [5] [6]: characterizes the number of computational resources required to reproduce the object x based on the description d. It is defined as the length of the shortest program p whose input is d, and the result of execution $-$ y:

$$K(x) = min(l(p): p(d) = y) \quad (4)$$

- *Quantum information*: a generalization of the Shannon concept of information for quantum states. The unit of information contained in a certain quantum state is the Qubit, which, in addition to the discrete values 0 and 1, can also take values in the interval [0,1]. The amount of information contained in the entire quantum system described by the density matrix ρ is measured by the von Neumann entropy [7]:

$$S(\rho) = -Tr(\rho \ ln \ \rho) \quad (5)$$

- *Information as a state of an agent*: a logical formalization of such concepts as knowledge and belief. For example, in epistemic logic [8], the statement that some *i-th* agent knows the fact *a* is true if and only if *a* is true on all worlds reachable for the *i-th* agent *w'*:

$$w \vDash K_i a \Leftrightarrow \forall w'(R_i(w,w') \rightarrow w' \vDash a) \quad (6)$$

- *Semantic information*: information *inf (s)* is defined as the content of the statement *s* and is measured based on its logical probability *q(s)* [9]:

$$inf(s) = log\ 1/q(s) \quad (7)$$

We supplement this list with the following definitions of informativity:

- *Hartley information* [10]: usually considered within the framework of the Shannon's theory of information, however Kolmogorov [5] singled it out as a separate approach. For the combinatorial approach independence from any probabilistic assumptions is essential. The combinatorial informativity allows us to determine the number of characters required for encoding a message *x* of length *l* in the *m*-valued code with an alphabet, consisting of *N* characters:

$$I_H(x) = l[log_m N] \quad (8)$$

- *The Kullback-Leibler distance* [11]: (relative entropy) of discrete random variables *X* and *Y* is interpreted as the magnitude of the information gain in the transition from the distribution *X* to the distribution *Y*:

$$D_{KL}(Y|X) = \sum_i P(y_i) log \frac{P(yi)}{P(xi)} \quad (9)$$

- *Akaike information criterion* [11]: estimates the informativeness of a statistical model with *k* parameters and likelihood function *L*, as the difference between the complexity and accuracy of the model:

$$AIC = 2k - 2ln(L) \quad (10)$$

This review allows us to conclude that there are different, unrelated definitions of information. Although there are studies of the relationship between these concepts [13], the final classification was never found. Many researchers, for example, the authors of [14], explain this fact by the idea that information is a formal concept that can have different, non-exclusive interpretations. Thus, the modern theory of information contains a large set of various unrelated measures. There is a problem of their valid formal systematization.

III. Analysis of existing information definitions

Despite various interpretations, some information concepts use the same formalism. Therefore, various information concepts can be partially systematized according to the common formalism. An example of such an approach can be found in A. N. Kolmogorov's article "Three approaches to the definition of the concept «quantity of information»" [5].

Consider Shannon information (2) and semantic information (7). From a formal point of view, these definitions are reduced to the expression

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

where P (x) is some probability distribution. The only difference is that in the case of Shannon information the distribution of P (x) is interpreted as statistical, and in the case of semantic information - as logical. Thus, these definitions can be identified from a formal point of view.

Consider also Shannon entropy (3) and von Neumann entropy (7). Formally, these expressions have the form

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

where *P(x)* is also some probability distribution. In the case of Shannon entropy, *P(x)* is also a statistical probability, in the case of von Neumann entropy, *P(x)* is the density matrix. Thus, these definitions can also be identified

In addition, measures of information can be divided into basic and derivative. For example, consider the following information measures used in information theory [15]:

- *Self-information* of some value x_i of a discrete random variable *X* with the probability distribution $P(x_i)$:

$$I(x_i) = -logP(x_i) \quad (11)$$

- *Entropy* of a discrete random variable *X*:

$$H(X) = -\sum_i P(x_i) logP(x_i) \quad (12)$$

- *Conditional entropy* of a discrete random variables *X* and *Y* with the probability distribution $P(x_i)$ and $P(y_j)$ and the joint probability $P(x_i|\ y_j)$:

$$H(Y|X) = -\sum_i \sum_j P(x_i)P(y_j|x_i) logP(y_j|x_i) \quad (13)$$

These measures are derived from a single basic definition, in this case, the Shannon definition (2), using various methods of probability theory, such as calculating the average value and using conditional probability, respectively. Thus, it becomes possible to systematize some concepts of information by the methods of probability theory used. The results of this analysis are shown in the Table 3. Some of the cells in the table are not filled in, since the corresponding measures of the information are not given in the original survey, however, they exist.

Thus, most of the information measures are systematized and come down to three basic definitions: Shannon information (2), Kolmogorov complexity (4) and Hartley information (8). The exceptions are the information concept as a state of the agent and the Akaike information criterion. The

first is qualitative concept, not a quantitative one, the second is a combination of two basic definitions: Shannon's and Hartley's. However, to build a complete and sound theoretical classification, it is still necessary to bring these basic definitions under a single basis.

TABLE I. RESULT OF THE INFORMATION CONCEPTS ANALYSIS

P(x)	P(y|x)	M(P(x))	M(P(y|x))	$\frac{P(y\|x)P(x)}{P(y)}$	D(x)
I(x)	-	H(x)	H(Y|X)	$D_{KL}(Y\|X)$	$I_F(x)$
K(x)	-	-	-	-	-
$I_H(x)$	-	-	-	-	-

IV. CLASSIFICATION OF INFORMATION MEASURES

In this paper, it is proposed to derive a classification of information measures from the formal properties of an information bearer. As information bearers, it is proposed to consider various data processings, starting with mathematical formulas and ending with machine learning algorithms, including the data itself as "zero" processing. In a very general way, these processings are mappings. Thus, the proposed solution consists in deriving the classification of information measures from the properties of mappings.

Let there be some arbitrary mapping

$$X_1 \times ... \times X_n \rightarrow f_1 ... \rightarrow f_l \rightarrow Y \quad (14)$$

where X_i, Y are some sets, $x_i \in X_i$, $y \in Y$ and f_j is some j-th mapping.

Each mapping from the set-theoretic point of view can be considered as a set of tuples T:

$$T = \{x_1 ... x_n y : x_i \in X_i, y \in Y\} \quad (15)$$

in which, in turn, one can select a subset Y_i:

$$y_i = \{x_1 ... x_n y : y = y_i\}. \quad (16)$$

For a set of tuples T, it is possible to distinguish the following structures:

- Domain $X = X_1 \times ... \times X_n$;
- Codomain $Y = Y_1 \cup ... \cup Y_k$;
- The set of mappings $F = Y^X$

Each of these structures also has certain elements:

- The domain X consists of the parameters $X_1, .., X_n$;
- The codomain Y consists of the subsets $Y_1, ..., Y_k$.
- The set of mappings F consists of subsets $F_1 \cup ... \cup F_m$ of compositions of a certain length.

Finally, a measure can be specified for each atomic element listed above:

- for a parameter X_i: the number of values $|X_i|$;
- for a subset of values Y_i: volume $|Y_i|$;
- for a subset of mappings F_i: volume $|F_i|$;

Thus, we assume that the basic information definitions are derived from the listed measures of mappings:

- Kolmogorov complexity is derived from the number of parameter values $|X_i|$,
- Shannon information - from the volumes of codomain subsets $|Y_i|$,
- Hartley information - from the volumes of subsets $|F_i|$.

i. Domain informativeness

In this subsection it is shown that the Kolmogorov complexity (4) is derived from the notion of function (14). Let us define that the description $d(y_i)$ of an object $y_i \in Y$ is some tuple from the set Y_i:

$$d(y_i) = x_1 ... x_n y_i \in Y_i \quad (17)$$

For each parameter X_j in Y_i, the number of unique values $|X_j|$ is specified. Then we determine the length $l(d(X_j))$ of the minimal description of the parameter X_j:

$$l(d(X_j)) = log|X_j| \quad (18)$$

This description is minimal, because

$$l(d(X_j)) = 0 \Leftrightarrow |X_j| = 1 \quad (19)$$

In other words, if the value of the parameter X_j in the description of y_i is constant, then this parameter is not considered in the minimum description.

Finally, we define the minimum description length $l(d(y_i))$ as the sum of the lengths of the descriptions of the parameters X_j:

$$l(d(y_i)) = \sum_j l(d(X_j)) \quad (20)$$

Thus, the complexity $K(y_i)$ is derived from the volume of the parameters X_j:

$$K(y_i) = l(d(y_i)) \quad (21)$$

and characterizes the informativeness of the function domain.

ii. Codomain informativeness

Now we show that the Shannon information (2) is also derived from the concept of function (14). The ratio of the volumes of the set of tuples T and its subset Y_i for some y_i can be considered as the probability $P(y_i)$:

$$P(y_i) = \frac{|Yi|}{|T|} \quad (22)$$

Since this value characterizes the probability that the function value is y_i. If every tuple in T ends with y_i, then y_i can be considered as a reliable event:

$$P(y_i) = 1 \Leftrightarrow |Y_i| = |T| \quad (23)$$

Conversely, if there is a y_i such that no tuple contains it, then y_i is an impossible event:

$$P(y_i) = 0 \Leftrightarrow |Y_i| = 0 \quad (24)$$

The volume of a set cannot be less than zero, therefore, $\frac{|Yi|}{|T|}$ is always positive.

Thus, Shannon information is the negative logarithm of the ratio of the volumes Y_i and T:

$$I(y_i) = -\log\frac{|Yi|}{|T|} \quad (25)$$

and characterizes the informativeness of the function codomain.

iii. Mappings informativeness

Finally, let us show that the Hartley information (8) is derived from the notion of function (14). Let there be arbitrary sets X and Y, then there is a set $F = Y^X$ of possible mappings from X to Y. Moreover, for every $x \in X$ and $y \in Y$ there exists a mapping $f(x) = y \in F$. If for some reason only some subset $F' \subset F$ is available, then for some x and y there may not be mapping $f \in F'$ Such that $f(x) = y$. However, if the subset F' is functionally complete, then there exists some finite number l, such that

$$F = F' \times_1 \ldots \times_l F' \quad (26)$$

Consequently,

$$l = [\log_{F'}F] \quad (27)$$

Which coincides with the Hartley information, if we consider F' as an alphabet, F as an encoded message, and l as the maximum code length.

Thus, the Hartley information is the maximum length of a composition of mappings for some set $F' \subseteq$ F, that is necessary for mapping each element of X to Y:

$$I_H(F) = [\log_{F'}F] \quad (28)$$

i.e. it characterizes the informativeness of the mapping itself.

Thus, we built a complete analytical systematization of information measures. Three basic definitions identified as a result of a preliminary analysis of existing information measures correspond to the properties of mappings: Kolmogorov complexity corresponds to the domain informativeness, Shannon information – the codomain informativeness, Hartley information – the mapping informativeness.

V. Comprehensive information estimation

The resulting classification of information measures can be used for a comprehensive estimation of the data processing results. As an example, consider the task of finding dependencies in data. Suppose, as a result of the application of a certain algorithm, a dependence was found between the attributes a and b, which can be modeled as two alternative formulas: f_1 and f_2. The formulas and their truth tables are the Table 2 and Table 3, respectively. Both formulas are mappings of the form $2 \times 2 \rightarrow 2$, therefore, $|T| = 4$, $|F| = 16$. Let us estimate the characteristics of the object *1* from each formula. The set of tuples Y_1 corresponds to the formula *f1*, the set Y_2 – to f_2, see the Table 4 and Table 5. The information measures calculation is presented in the Table 6.

TABLE II. *f1*

(a	∨	b)	∧	a
0	0	0	0	0
0	1	1	0	0
1	1	0	1	1
1	1	1	1	1

TABLE III. f2

a	∨	b
0	0	0
0	1	1
1	1	0
1	1	1

TABLE IV. Y_1

a	b
1	0
1	1

TABLE V. Y_2

a	b
0	1
1	0
1	1

TABLE VI. CALCULATION OF INFORMATION MEASURES

	$K(1)$	$I(1)$	$I_H(F)$
Y_1	1	1	4
Y_2	2	0,4	n/a

Thus, the truth of the formula f_1 is less complex, which corresponds to reality, as

$$(a \vee b) \wedge a \Leftrightarrow a \quad (29)$$

while the formula $a \vee b$ is irreducible. At the same time, f_1 truth is more informative according to Shannon, since the truth of f_1 is less likely than the truth of f_2. Finally, since only one mapping is used in f_2, its Hartley information cannot be compared.

VI. CONCLUSION

We reviewed existing approaches to defining information to demonstrate that existing definitions are not related to each other and that information theory lacks a theoretical basis for their systematization. We also made a preliminary analysis of existing measures of information, as a result of which three basic definitions of information were identified. Secondly, it was established that some concepts of information are formally identical and differ only in interpretation. As a result, existing measures of information have been systematized, but such systematization still lacks theoretical foundation.

As a solution, we proposed to derive information measures from the properties of information bearer, and in the general case to consider mappings as such a bearer. As a result, the basic measures of information were reduced to the mapping properties: Kolmogorov complexity — to the domain informativeness, Shannon information — to the codomain informativeness, Hartley information — to the informativeness of the mapping itself. As a result, we obtained a complete and theoretically grounded classification of information measures. The application of the resulting classification for a comprehensive estimation of information was demonstrated on the example of logical formulas.

REFERENCES

[1] Stanford Encyclopedia of Philosophy, Information, Web: https://plato.stanford.edu/entries/information/.

[2] R.A. Fisher, “Theory of statistical estimation”, *Proceedings Cambridge Philosophical Society*, 22(5), 1925, pp. 700–725.

[3] C.E. Shannon, “A Mathematical Theory of Communication”, *Bell System Technical Journal*, vol.27, 1948, pp. 379-423.

[4] R.J. Solomonoff, “A Formal Theory of Inductive Inference”, *Information and Control*, 7(1), 1964, pp. 1–22; 7(2), 1964, pp. 224–254.

[5] A.N. Kolmogorov, “Three approaches to the definition of the concept of the amount of information”, *Prob. Transfer Inform*, vol.1, 1965, p. 3-11.

[6] G.J Chaitin, “On the Length of Programs for Computing Finite Binary Sequences”, *Journal of Association for Computing Machinery*, 13(4), 1966, pp. 547–569.

[7] J. Von Neumann, *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer, 1995.

[8] J. Hintikka, *Knowledge and Belief*. Ithaca: Cornel University Press, 1962.

[9] Y. Bar-Hillel, R. Carnap, “An Outline of a Theory of Semantic Information”, *Research Laboratory of Electronics,* Technical Report, No. 247, October 27, 1952.

[10] R.V.L Hartley, “Transmission of Information”, *Bell System Technical Journal*, 7(3), 1928, pp. 535–563.

[11] S. Kullback, R.A. Leibler, “On information and sufficiency”, *The Annals of Mathematical Statistics*, 2(1), 1951, pp. 79-86.

[12] H. Akaike , “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control*, vol.19, 1974, pp. 716—723.

[13] S. Galatolo, M. Hoyrup, C. Rojas, “Effective symbolic dynamics, random points, statistical behavior, complexity and entropy”, *Information and Computation*, 208, 2010, pp. 23–41.

[14] S. Fortin, O. Lombardi, L. Vanni, “A pluralist view about information”, *Philosophy of Science*, 82(5), 2015, pp. 1248-1259.

[15] E. M. Gabidulin, N. I. Pilipchuk. Lectures on information theory, Moscow: MEPI, 2007.