Unified Information Gain Measures for Inference Processes

José Hernández-Orallo
Universitat de València[1]
Dep. of Logic and Philosophy of Science
Apdo. 22109, 46071 València, Spain
E-mail: jorallo@dsic.upv.es

Joint Conference of
the 5th Barcelona Logic Meeting
and the 6th Kurt Gödel Colloquium
Wednesday, June 16th, to Saturday, June 19th, 1999,
Barcelona.

[1] Also at Universitat Politècnica de València, Departament de Sistemes Informàtics i Computació, Camí de Vera s/n, E-46022, València, Spain.

Introduction

Inference Processes like (computational) deduction and induction can be regarded in a computational framework as processes that generate an output O from an input I.

In the case of deduction:

I = *Premises P* + *Axiomatic System S*

O = *Conclusion C.*

Some semantical restrictions should be preserved.

In the case of induction:

I = *Evidence E* + *Background Theory B.*

O = *the hypothesis H*

Some evaluation criteria should be ensured.

Let us consider the term *information* seen as the result of a computational effort, analogously to the way energy is seen as the result of a physical work.

⇓

Both *C* and *H* provide information.

How this computational effort is to be measured?

Descriptional Complexity

Descriptional Complexity (Kolmogorov Complexity or Algorithmic Information Theory):

DEFINITION 1. KOLMOGOROV COMPLEXITY

The *Kolmogorov Complexity* (KC) of a string x on a bias β:

$$K_\beta(x|y) = \min \{ l_\beta(p_x(y)) \}$$

where p_x denotes any "prefix-free" β-program for x using input y and $l_\beta(p_x)$ denotes the length of p_x in β.

$K_\beta(x) = K_\beta(x|\varepsilon)$ where ε denotes the empty string.

Kolmogorov Complexity is an absolute and objective criterion of simplicity. It is independent (up to a constant term) of the descriptional mechanism used.

DEFINITION 2. ABSOLUTE INFORMATION GAIN

The *absolute information gain* of an object x wrt. an object y:

$$V(x \mid y) = K(x \mid y) \,/\, K(x)$$

The information which is needed to describe x from y wrt. the information which is needed to describe x alone.

Kt as Effort

DEFINITION 3. LEVIN'S LENGTH-TIME COMPLEXITY
The *Levin Complexity* of a string x on a bias β:

$$Kt_\beta(x|y) = \min \{ LT_\beta(p_x(y)) \}$$

where $LT_\beta(p_x) = l(p_x) + log_2 \text{Cost}_\beta(p_x)$

Why *Kt* instead of *K*?

- *Kt* is computable.
- Levin showed that the weighting $LT(x) = \text{length}(x) + \log \text{Cost}(x)$ between space and time is *optimal* in the sense of universal search problems.

Intuitively, given any problem, either an amount of time is needed to obtain the answer to the problem or either an amount of the data (space) of the solution is needed.

Given two objects, the effort from x to y is then measured as $Kt(y \mid x)$.

Computational Information Gain

The Information Gain of object y wrt. object x is then defined as the quotient between the effort which is necessary to describe y from x and the effort which is necessary to describe y alone.

DEFINITION 4. COMPUTATIONAL INFORMATION GAIN
The *Computational Information Gain* of an object x wrt. an object y is:

$$G(x \mid y) = Kt(\mathrm{x} \mid \mathrm{y}) \,/\, Kt(\mathrm{x})$$

THEOREM 1. LIMITS OF $G(x \mid y)$
There exists a constant c such that for every x and y,

$$\log l(x)/(l(x) + \log l(x) + c) < G(x \mid y) \leq 1$$

THEOREM 2. ROBUSTNESS TO POLYNOMIALITY
Consider a *learning* algorithm A^* in $\mathcal{P}$ (i.e. polynomial), namely $\exists p \in \mathbb{N}^+ : O(n^{p-1}) \leq O(A^*) \leq O(n^p)$, being A^* of constant size, i.e., $l(A^*) = c$. This algorithm deterministically transforms y into x, where x is a program for y, being $n = l(y)$. There is a τ such that for all x and y, if $n > \tau$ and $Kt(x) > k \cdot p \cdot \log n$, then $G(x \mid y) \leq 2 \,/\, k$.

Proof of Theorem 1

PROOF OF THEOREM 1. The second inequality $G(x \mid y) \leq 1$ is obtained by considering that y must only be read if it is necessary for obtaining x, so $\forall x,y\ Kt(x \mid y) \leq Kt(x)$. The limit 1 is obvious by choosing $y = \varepsilon$ and the definition of $Kt(x)$ as $Kt(x \mid \varepsilon)$. The first inequality is justified by the fact that the numerator follows

$$Kt(x \mid y) \geq \log l(x)$$

because x must be printed and this takes at least $l(x) + c_2$ units of time. In fact this limit can be come close if $x = y$ because the program "print y" has cost approximately $2 \cdot l(x)$. The denominator must follow this disequality.

$$Kt(x) < l(x) + log\ l(x) + c$$

because in the worst case, when x is random, we need $l(x) + c_1$ bits of information for the program "print x" and at most $l(x) + c_2$ units of time to be printed. Both constants can be represented by a negligible c. By (1) and (2) we have that $log\ l(x)/(l(x) + log\ l(x) + c) < G(x \mid y)$. □

Proof of Theorem 2

PROOF OF THEOREM 2. For every string of data y, let us construct x in the following way: x = *"apply A* to y"*. Since we can construct x from $<A^*, y>$ in an easy way p= "apply 1st argument to 2nd argument" $Kt(x \mid <A^*, y>) \leq LT(p) = l(p) + log\ cost\ (p) < c' + \log n^p$). It is obvious that $Kt(x \mid y) < Kt(x \mid <A^*, y>)$. So we have that $Kt(x \mid y) < c' + \log n^p = c' + p \log n$.

If, as supposed, $Kt(x) > k \cdot p \cdot log\ n$, then the quotient $G(x \mid y) = Kt(x \mid y) / Kt(x) \leq ((c' / (p \cdot log\ n)) + 1) / k$. Since $p > 0$, just choose $\tau = n$ such that $c' / (p \cdot log\ n) < 1$. From here, $G(x \mid y) \leq 2 / k$. □

Information and Deduction

Carnap Probabilistic Calculus:

$P \models Q$ means that Q has less information than P.

Is deduction non-informative?

In omniscient systems:

- Everything implicit is immediately and effortlessly made explicit.

In non-omniscient (real) systems:

- Making explicit what was implicit requires effort and, consequently,
- Deduction is costly and the conclusions are worthy, valuable, i.e. *informative*, and, in some cases, surprising.

Hintikka (1970) introduced the theory of semantic information to distinguish between:

- 'cash' information → Surface Information
- 'potential' information → Depth Information

Based on the constituents of first-order logic.

Information Gain and Deduction

If y represents the premises and x the conclusion:

Minimum: $G(x \mid y) = \log l(x) \,/\, (l(x) + \log(l(x)) \approx 0$

The conclusion is evident from the premises. It is easy to describe the conclusion from the data. $Kt(x \mid y)$ ↓↓

Maximum: $G(x \mid y) = 1$

We have that $Kt(\mathrm{x} \mid \mathrm{y}) = Kt(\mathrm{x})$. The premises are useless (in time-space terms) to describe the conclusion. It is necessary a great computational work on the premises y to obtain the conclusion or there is a need for external information.

G($x \mid y$)	**V($x \mid y$)**	**Meaning**
≈ 0	0	x is explicit from y.
≈ 0	1	Impossible
1	0	x is deeply implicit in y.
1	1	x is independent wrt. y.

V represents the limit (wrt. to efficiency) of G (*in the same way as depth information is the limit of surface information*).

Representational Optimality

Given a deductive system with axioms A and its set of consequences (or theorems) S. The *representational optimality* is defined as:

$$Opt(S) = \text{argmin}_T \{ \tau \cdot L(T) + \sum_{s \in S} Kt(s|T) \text{ s.t. } \forall s \in S : V(s|T)=0 \}$$

Intermediate Information is useful to maintain explicitly:
A theorem P such that $A \models P$ should be left explicitly in T if $\sum_{s \in S} Kt(s|<A,P>) < \sum_{s \in S} Kt(s|A)$ and T does not get too long.

This formalises the well-known necessity of theorems (or lemmata) and the use of extensional properties for mathematical practice, in order to avoid difficult derivations that were already done (while still maintaining under control the whole size of the theory).

There are many more views of 'optimality' that can be expressed or combined in this framework.

Information and Induction

Carnap Probabilistic Calculus:

$P \models Q$ means that P has more information than Q.

Is induction always informative?

Popper recognised that not every inductive inference is informative.

Example: Given data x, the theory "print x for ever" is little informative.

Popper advocated for the "informativeness" of hypotheses.

Information Gain and Induction

If x is the theory and y is the data (the evidence):

Minimum: $G(x \mid y) = \log l(x) \; / \; (l(x) + \log(l(x)) \approx 0$

The theory is evident from the data. It is very easy to describe the theory from the data. $Kt(x \mid y)$ ↓↓

Examples: → the polynomial obtained from the data.
→ Exceptions ($Kt(x \mid y)$ ↓↓)
→ Extensionalities (part of x is in y)

Maximum: $G(x \mid y) = 1$

We have that $Kt(x \mid y) = Kt(x)$. The data is useless (in time-space terms) to describe the theory. It is necessary a great computational work on the data y to obtain the theory or there is a need for external information.

What is to Discover?

A concept x is *surprising* wrt. y in a context β iff:

$$G_\beta(x \mid y) \uparrow\uparrow$$

A concept or theory x is a *discovery* wrt. y in a context β iff:

$$G_\beta(x \mid y) \uparrow\uparrow \text{ and } G_\beta(y \mid x) \approx 0$$

i.e, x is surprising for y and y is explicit from x (e.g. x is an efficient theory or explanation for y).

In a proper way, discovering must be accompanied by confirmation, however, x is valuable *per se*.

Learning, Identification and Gain

Gold (1967): Learning as "*identification in the limit*":

For finite concepts, however:

- An inductive algorithm that gives the extensional theory for the data has formally learnt!!!

The MDL principle (the best theory is the shortest one) gives that "extensional" theory for the great majority of data samples (most strings are random).

Authentic Learning:

The more that one learns the greater $G(h \mid d)$.

However, G is not a plausibility criterion.

Investment and Selection Criteria

There are infinite theories for some data.

Which ones are useful to remember?

Only the best one according to a plausibility selection criterion? If this best one is refuted later, all the work should be made again...

Is it valuable to store the computational effort which has been invested for more than one hypothesis?

Memory is not unlimited...

Oblivion Criterion:

$$OC(h \mid d) = G(h \mid d) \cdot PC(h \mid d)$$

For instance, if the plausibility criterion is the MDL principle we have:

$$OC(h \mid d) = G(h \mid d) \cdot 2^{-l(h)}$$

Related Concepts

- Kirsh's theory of explicitness is perfectly formalised by G.

- Quinlan's Gain Ratio: (induction of binary trees) If C is set of class labels and X is a feature for splitting the tree then a descriptional variant of Gain Ratio is equal to $1 - V(X \mid C)$.

- Intensionality: Extensional concepts have $G = 0$. Comprehensive theories usually have more information gain.

- Compression is positively related to G. If the compression ratio between the data d and the hypothesis is greater than $l(d) / \log l(d)$ then $G(h \mid d) = 1$. (*the data must be read*)

Combination of Ind/Ded and Plausibility/Informativeness Criteria

Axioms are given by experience (evidence).

Which things should be left explicit in a theory?
- Those properties which are useful to cover the evidence.
- Those properties which justify other properties.

Which should be forgotten?
- Evidence which is well covered by the theory.
- Consequences which are easy to recover from axioms (low gain).
- Properties which are not used for the consequences or for other properties.

This leads to a theory of *reinforcement*[2]. A rule or property is more reinforced if it is more used by the evidence or by other rules / properties.

---

[2] (Hernández-Orallo 1999) *Intl. J. of Intelligent Systems*, to appear.

Conclusions

- Deduction and induction can be conciliated in terms of information gain.
- This allows more consistent combinations of deductive systems and inductive paradigms for the construction of non-omniscient (resource-bounded) rational agents.
- Interestingness and Explicitness are concepts which are shared both by deduction and induction.

Drawbacks:

- V is not computable and G must be approximated....