Computational Information Gain and Inference

José Hernández-Orallo[1]

Universitat Politècnica de València, Dep. de Sistemes Informàtics i Computació.
Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain. E-mail:jorallo@dsic.upv.es

Abstract. A definition of computational information gain is presented based on Levin descriptional complexity. The measure is applicable to different inference processes, either deductive or inductive, and evaluates the relative value of new inference results.

1 Introduction

Processes which are apparently so different as induction and deduction can be explained in a computational framework as inference processes that *just* generate an output from an input, which must follow some semantical restrictions and/or selection criteria, widely studied in philosophy of science and mathematical logic, respectively. The term information is seen as the result of a computational effort, analogically to the way energy is seen as the result of a physical work. This suggests many questions, especially how to measure this computational effort.

The *answer* was given by Levin in the seventies [10], proving that the weighting $LT(x) = length(x) + logCost(x)$ between space and time was optimal in the sense of universal search problems. In other words, 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. From here, Levin's variant of descriptional (Kolmogorov) Complexity can be defined as follows:

Definition 1. The **Levin's Length-Time Complexity** of an object x given y on a descriptional mechanism β:

$$Kt_\beta(x|y) = min\{LT_\beta(p|y) : \phi_\beta(\langle p, y\rangle) = x\}$$

This is a very practical alternative of Kolmogorov Complexity, because, as well as avoiding intractable descriptions, it is computable. Given two objects x and y, $Kt(y|x)$ represents the effort from x to y.

2 Definition of Computational Information Gain

The Information Gain of an object y wrt. an object x can be 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. More formally,

Definition 2. The **Computational Information Gain** of an object y wrt. an object y in a context β is defined as:

$$G_\beta(y|x) = Kt_\beta(y|x)/Kt_\beta(y)$$

Some properties of this measure can be shown before applying it to inference processes. The proofs of these and other properties can be found in [5].

Theorem 3. *There exists a constant c such that for every x and y,*

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

This gives an interval practically between 0 and 1, which is very appropriate for measuring a relative information gain value.

Secondly, in order to check that the meaning of 'difficulty' that is gathered by G is compatible with computational complexity, it can be shown that if there exists a polynomial time algorithm from a problem y to a solution x, and the problem y is complex, then $G(x|y)$ must be low.

Theorem 4. *Consider a* learning *algorithm $A^* \in \mathcal{P}$ (i.e. polynomial), namely $\exists p \in \mathcal{N} : O(n^{p-1}) \leq O(A^*) \leq O(n^p)$, with A^* being of constant size, i.e. $l(A^*) = c$. This algorithm deterministically transforms y into x, where x is a program for y, with $n = l(y)$. There is a τ such that for all x and y, if $n > \tau$ and there exists a k such that $Kt > k \cdot p \cdot log\ n$, then $G(x|y) \leq 2/k$.*

3 Computational Information Gain and Deduction

Under Carnap Probabilistic Calculus [1], if $P \models Q$ then Q has less information than P. This has popularised the opinion that deduction cannot be informative. However, this view can also be originated from a supposedly omniscient view of logic, where everything that is implicit is immediately and effortlessly made explicit by the rules of the axiomatic systems. This view is not only practically unfeasible but formally erroneous, as it was shown by [3], extending Gödel results of incompleteness to intractability. In practice and theoretically, minimally expressive axiomatic systems are not omniscient. Making explicit what is implicit requires effort. Consequently, deduction is costly and its conclusions are worthy, valuable, informative, and, in some cases, surprising.

By using G, we can clearly establish the difference between informative deduction and non-informative one. More precisely, if y represents the premises and x the conclusion, the following two extreme situations are illustrative:

- Minimum: $G(x|y) = log\ l(x)/(l(x) + log\ l(x)) \approx 0$. The conclusion is evident from the premises. It is easy to describe the conclusion from the data. $Kt(x|y)$ must be low.
- Maximum: $G(x|y) = 1$. We have that $Kt(x|y) = Kt(x)$. The premises are useless (in time-space terms) to describe the conclusion. A great computational effort is necessary to work on the premises y to obtain the conclusion or there is a need for external information.

Between the two extremes, G establishes a generic measure of the gain which is obtained from making explicit something that was implicit, provided that the system is not omniscient and is resource limited. This establishes a clear difference between explicit or surface information, and implicit or depth information, as it was highlighted by Hintikka for first-order logic [8].

4 Computational Information Gain and Induction

In a similar way as for deduction, if x is the theory and y is the data (the evidence), the two extremes given by G are also illustrative:

- Minimum: $G(x|y) = log\ l(x)/(l(x) + log\ l(x)) \approx 0$. The theory is evident from the data. It is very easy to describe the theory from the data. Some examples of this situation can be the fit polynomial obtained from the data, or a theory with a significant proportion of exceptions or extensionalities (part of x is in y), which makes $Kt(x|y)$ low.
- Maximum: $G(x|y) = 1$. We have that $Kt(x|y) = Kt(x)$. The data is useless (in time-space terms) to describe the theory. A great computational effort is necessary to work on the data y to obtain the theory or there is a need for external information.

Between the two extremes, G establishes a generic measure of how informative the hypothesis is wrt. the evidence (in Popper's sense [12]). It is compared with other selection criteria, especially simplicity, in [7]. The MDL principle [13] and the view of learning as compression [14] are useless for most cases, because the vast majority of sequences are incompressible [11].

5 Learning and Inference

Learning has traditionally been seen as inductive inference since Gold introduced the seminal paper on the paradigm of 'identification in the limit' [4].

However, by using G we can demand much more than identification, and we can differentiate between easy inductions and hard (and surprising) inductions. Nonetheless, this fact, as we have seen, is not restricted to induction, and deduction behaves in a similar way wrt. G. More precisely, we can say a concept x (either inductively or deductively obtained) is surprising wrt. y in a context β iff $G_\beta(x|y)$ is high. The notion of discovery is stricter though:

Definition 5. A concept or theory x is a **discovery** wrt. y in a context β iff:

$$G_\beta(x|y) \approx 1 \text{ and } G_\beta(y|x) \approx 0$$

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 other words, a discovering is something that was not known, was difficult to know, but once known, it is almost trivial in the other sense. In the case of induction, discovering must be accompanied by confirmation.

Finally, the view of induction as identification is paradoxical for finite evidence. For finite concepts, an inductive algorithm that gives the extensional theory for the data would have formally learnt. And, the MDL principle (the best theory is the shortest one) gives an *extensional* theory (the evidence itself) for the great majority of data samples (most strings are random). As a response to this situation, we propose a new notion of 'authentic learning'.

Definition 6. The more one **learns** the greater $G(K1|K0)$, where $K0$ is the knowledge before the inference step and $K1$ is the knowledge after it.

This dismisses the notion of learning as a phenomenon exclusively related to non-deductive inference processes.

6 Conclusions

We have introduced a new measure of computational information gain which can be applied to different inference processes in a unified manner. Several other connections with related concepts (explicitness, intensionality, other definitions of gain ratio) have been established in [5].

As a conclusion, the most important result of this work is that deduction and induction can be conciliated in terms of information gain. Interestingness, explicitness and even learning are concepts which are shared both by deduction and induction. This allows more consistent combinations of deductive systems and inductive paradigms for the construction of non-omniscient rational agents.

References

1. Bar-Hillel, Y.; Carnap, R. "Semantic Information". *British Journal for the Philosophy of Science* 4, 1953, 147-157.
2. Blum, L.; Blum, M. "Towards a mathematical theory of inductive inference". *Inform. and Control 28*, pp. 125-155, 1975.
3. Chaitin, Gregory J. "Information-theoretic limitations of formal systems" *Journal of the ACM*, 21, 403-424., 1974.
4. Gold, E. M. "Language Identification in the Limit". *Inform and Control, 10*, pp. 447-474, 1967.
5. Hernandez-Orallo, J. "Computational Measures of Information Gain and Reinforcement in Inference Process" Doctoral Dissertation, September 1999, Dep. of Logic and Philosophy of Science, University of Valencia.
6. Hernandez-Orallo, J.; Garcia-Varea, I. "Distinguishing Abduction and Induction under Intensional Complexity" in Flach, P.; Kakas, A. (eds.) European Conf. on AI (ECAI'98) Ws. on Abduction and Induction in AI, pp. 41-48, Brighton 1998.
7. Hernandez-Orallo, J.; Garcia-Varea, I. "Explanatory and Creative Alternatives to the MDL Principle" *Foundations of Science*, Kluwer, to appear.
8. Hintikka, J.; Suppes, P. "Surface Information and Depth Information" in Hintikka, Jaako; Suppes, Patrick *Information and Inference* Reidel Pub. Company 1970.
9. Kolmogorov, A.N. "Three Approaches to the Quantitative Definition of Information" *Problems Inform. Transmission* , 1(1):1-7, 1965.
10. Levin, L.A. "Universal search problems". *Problems Inform. Transmission, 9*, pp. 265-266, 1973.
11. Li, M.; Vitanyi, P. "An Introduction to Kolmogorov Complexity and its Applications" *2nd Ed. Springer-Verlag*, 1997.
12. Popper, K.R. "Conjectures and Refutations" London 1969.
13. Rissanen, J. "Modelling by the shortest data description" *Automatica-J.IFAC, 14*, pp. 465-471, 1978.
14. Solomonoff, R.J. "A formal theory of inductive inference". *Inf. Control.* Vol. 7, 1-22, Mar., pp. 224-254, June 1964.

This article was processed using the LaTeX macro package with LLNCS style