Partial Information Decomposition for the Analysis of Inductive Inferences with Multiple Premises

Aaron J. Gutknecht[1][0000−0002−2704−6944], Michael Wibral[1][0000−0001−8010−5862], and Abdullah Makkeh[1][0000−0002−3581−8262]

Campus Institute for Dynamics of Biological Networks, Georg-August University Goettingen

Abstract. The theory of partial information decomposition (PID) is an extension of classical information theory that allows to describe how information about some target variable is distributed over a range of source variables. In this way PID distinguishes fundamental types of uncertainty reduction. In the simplest case of two information sources the total reduction in our uncertainty about the target may be partly redundant to both information sources, partly unique to one of them, and partly synergistic, i.e. some aspect of our uncertainty about the target may only be reduced if we have access to both information sources at the same time. The problem of disentangling these components necessarily arises in any inductive inference with more than one "premise" (information source, observation, etc.). We argue, therefore, that PID theory is of great relevance to the foundations of inductive inference in particular as it pertains to the allocation of epistemic value over multiple information sources. Our formulation of the theory draws from principles of mereology and formal logic thereby placing PID on the foundation of two of the most elementary concepts of human thought: the part-whole relationship and the relation of logical implication.

Keywords: Information Theory · Mereology · Logic · Inductive Inference.

The quest for a decomposition of information in terms of unique, redundant, and synergistic components goes back to the 1950s, but it was only in the seminal 2010 paper by Williams and Beer [6] that the mathematical structure of the problem was described in its entirety and the first full solution was proposed. Since then there has been a flurry of alternative approaches aiming to overcome some of the shortcomings of the original solution (for instance [1, 4, 2]). However, no consensus could be reached thus far. Doing so would be of wide-ranging practical significance since problems of the PID type are ubiquitous in virtually all fields of quantitative research. In neuroscience, for instance, PID pertains to the question of how stimuli are encoded in a network of multiple neurons. Do the neurons encode the stimulus synergistically? Or are differerent aspects of the stimulus encoded uniquely by different neurons? How redundant is the encoding? In machine learning one may ask how information about the output is allocated over different potential features. This analysis can then be used for efficient feature selection [7].

It should be noted that the problem soon becomes very complex as a larger number of information sources are considered because ever more complicated types of information arise (think for example of information shared by some variables yet at the same time synergistic with respect to some other variables). In order to deal with this complexity, our own work focuses, on the one hand, on the mathematical structure and conceptual foundations of PID [3], and on the other, on concrete numerical measures of the different PID components [5]. In the former line of research we show how partial information decomposition can be fully explained in terms of parthood relationships between information contributions of the source variables about the target variable. In doing so, we arrive at a unique solution to the hierarchical organization of information contributions that must be respected if the information decomposition is to embody the intuitive notion of one entity being part of another one. In the second line of research we establish a connection between PID theory and formal logic. In particular, we describe how the different PID components can be measured by the local information provided by a special class of statements about the observed values of the information sources: the class of statements with monotonic truth-tables. The defining feature of such statements is that changing the truth value of an atomic statement from false to true, cannot make the statement false if it was previously true.

Here, we would like to suggest that PID could play an important role in the analysis of inductive inferences in the sense of empirical / statistical inference from data to hypothesis. Especially from a Bayesian perspective, one may consider the target variable as representing some parameter Θ of interest and the information sources as multiple observations or experimental outcomes $X_1, \dots, X_n$ (the "premises" of the inductive inference). Our goal is to use these observations to reduce our uncertainty about the parameter. In the most simple case of two observations, PID aims to decompose the joint mutual information into the four basic components of redundancy, uniqueness, and synergy:

$$I(X_1, X_2 : \Theta) = Red(X_1, X_2 : \Theta) + U(X_1 : \Theta) + U(X_2 : \Theta) + Syn(X_1, X_2 : \Theta) \quad (1)$$

Viewed in this light, PID distinguishes between different types of uncertainty reduction and, importantly, allows us to describe these types in a quantitative way. Is our reduction in uncertainty about Θ uniquely due to a particular observation? To what degree does it result from synergistically combining information obtained from multiple observations? Perhaps, our uncertainty reduction is even purely synergistic so that we cannot learn anything about the parameter by looking at an individual observation in isolation. Or is some part of our uncertainty reduction redundant to multiple observations? By answering all possible questions of this nature, PID provides a detailed picture of the epistemic structure of inductive inferences with multiple premises. It tells us which observations, or collections of observations, matter to what degree when it comes to reducing our uncertainty about the parameter.

References

1. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., Ay, N.: Quantifying Unique Information. Entropy **16**(4), 2161–2183 (apr 2014). https://doi.org/10.3390/e16042161, http://www.mdpi.com/1099-4300/16/4/2161
2. Finn, C., Lizier, J.T.: Pointwise partial information decomposition using the specificity and ambiguity lattices. Entropy **20**(4), 297 (2018)
3. Gutknecht, A.J., Wibral, M., Makkeh, A.: Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. Proceedings of the Royal Society A **477**(2251), 20210110 (2021)
4. Ince, R.A.A.: Measuring multivariate redundant information with pointwise common change in surprisal. Entropy **19**(7), 318 (2017)
5. Makkeh, A., Gutknecht, A.J., Wibral, M.: Introducing a differentiable measure of pointwise shared information. Physical Review E **103**(3), 032149 (2021)
6. Williams, P.L., Beer, R.D.: Nonnegative decomposition of multivariate information. arXiv preprint arXiv:1004.2515 (2010)
7. Wollstadt, P., Schmitt, S., Wibral, M.: A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. arXiv preprint arXiv:2105.04187 (2021)