IPI Letters

ISSN 2976 - 730X
IPI Letters 2024,Vol 2 (2):104-106
https://doi.org/10.59973/ipil.128

Received: 2024-09-23
Accepted: 2024-09-24
Published: 2024-09-25

Communication

Information, Entropy, and the Zeta Function

Barry Robson[1,*]

[1] The Dirac Foundation, Oxfordshire, United Kingdom

*Corresponding author: robsonb@aol.com

The relation between the entropy (in thermodynamics) of Boltzmann and of Shannon information (in communication), has recently been discussed by Vopson (e.g. Refs [1,2]). Amongst several interesting similarities and differences pointed out by Vopson, one is that writing information to a computational device increases entropy. This assertion appears counterintuitive because entropy is often described as negative information. Nonetheless, it appears much more natural when expressed in terms of a definition of information proposed by the present author BR. There are, of course, other kinds of information, e.g., Kullback-Leibler, Fano mutual, and Chaitin information [3]. These lie closer to the author's approach discussed below, but there are fundamental differences. Some have been touched upon by others (e.g., Refs [4-6]), but these are relatively recent and lack aspects important to the present discussion. The original publications [7-9] were the consequences of a student project given to the present author, namely, to assess what information about three-dimensional structures of proteins were conveyed in their amino acid sequences when data was extremely limited. As data became more plentiful, a Theory of Expected information emerged that could also be applied more generally [10]. As a predictive method, it resembles but preempted the Bayes Net [11].
The relevance to the relationship between information and entropy is clearer in later applications that included the analysis of biomedical data, given as large spreadsheet [12-15]. One wished to address the information required to write and (primarily) extract information as knowledge from a spreadsheet, e.g. each record (row) is a patient. The important thing is that the individual data elements are attributes of form 'attribute type':='attribute value', e.g. 'systolic Blood pressure (mmHg)':=140. There are 8 main points; point 6 is important to the write-read-issue.

(1) The number of attribute types each called say A, B, C,... and hence number of columns in a spreadsheet, is the so-called "explicit dimensionality", say N.

(2) Each unique attribute, an individual data element such as 'systolic Blood pressure (mmHg)':=140, may be encoded by a distinct prime number 2, 3, 5, 7,..., the product of which, i.e. a composite number, encodes all the information in each row (record) analysed in turn. Note that attributes in a record can occur more than once with the same number code. This leads to many insights and useful algorithms, not least to relate the method of counting

combinations of states and to the Riemann zeta function [12,13].

(3) The prime factorization theorem ensures that we can always recover all the information in a record (row). The number theory function called τ indicates that we can divide each record composite number by 1, 2 ,3, 4, 5,. . . and when the result is an integer it is a prime encoding e.g. the prime number for (A) or the composite number for each combination (A, B), (A, C,) (B, D, F), (G, L, P, S, V). (Only 2 divisions up to the square root is of the record composite number is needed and the rest can be deduced).

(4) Summing over the combinatorial equations for the factorization into the above combinations shows that number of number of combinations ("joint events") to extract per row for association analysis, e.g. (A), (A, B), (A, C,) (B, D, F), (G, L, P, S, V) etc., is 2^N combinations if one had huge computer resources. 2^N includes the empty state ( ), so one typically writes the combinatorial sum as $2^N - 1$. This 2^N has nothing to do with whether the values in the spreadsheet are binary or not.

(5) Having extracted those combinations as joint events per record (row), one wishes to count how many times each of e.g. (A), (A, B), (A, C,) (B, D, F), (G, L, P, S, V) etc. occur in the whole spreadsheet, to do statistics and notably compute e.g. log odds or Fano mutual information for each combination. That depends on the number of distinguishable attribute values per column, called the cardinality, which need not be binary. The cardinality does not have to be the same for each column, but here we say it is C, so C^N is the "effective dimensionality". The method will allow unknowns, so C can be exactly the maximum number of distinguishable values over all attributes in principle, even if not all examples might be seen and entered on the spreadsheet.

(6) N and C are then sufficient to describe the information-entropy issue, with an important caveat (size of counts) below. Increasing N and C obviously increases the information in your spreadsheet, but it also increases the entropy that matters to analysis because it gives more ways to pick out any combinatorial state (A), (A, B) etc., and test, e.g., its Fano information to see if it is significant. This entropy is "The Dragon on the Gold" of high dimensional data mining [12]. The logo for this American Chemical Society paper [12] was as follows.

(7) The actual information measure developed in natural units is the integral: $\int P\,(p|D)\,log\,(p)\,dp$, (natural log) given data D, an expected value reminiscent of Shannon information, but now p is the probability that that we believe that n out of n_{total} observations accurately reflects the underlying or "true" probability "out there", and one will be less sure if n and n_{total} are small numbers like 2/3 than if large e.g. 2000/3000. So, for each n/n_{total} we hold a distribution of posterior degrees of belief $P\,(p|D)$; the entropy of that is larger for

smaller data. The integral is easiest to solve for Bayes' Equation to give expected predictive odds $\int P(p|D)\, log\,(p/(1-p))\, dp$, but it always ends in expressions in partially summed zeta functions ξ, i.e. ξ(s = 1, n) = 1 + 1/2 + 1/3 +...+ 1/n here n is a number of occurrences. Choosing s > 1 gives higher moments of information as other surprise measures. (e.g. s = ∞ gives 0 or 1. We may speak of the Riemann zeta function because the value of s can also be complex (including an imaginary part), including split-complex, which leads to a large variety of applications utilizing Dirac notation and algebra in quantum mechanics, probabilistic knowledge nets, and extending Bayes Nets to bidirectional general graphs including cyclic paths without iteration [16]. For any event X occurring n[X] times, e.g. (A) or joint event X e.g. (A,B) etc. the integration gives I(X) = ξ (s = 1, n[X]) - ξ (s = 1, n_{total} – n[X]).
Note that we should write E(X:not-X | D) where E indicates the expected value of prior odds X:not-X, conditional on data D. However, we may hold the new information measure to be axiomatic of information that depends on amount of data. It is assumed here there is no prior belief for simplicity (though it is incredibly simple to include it by adding to n a virtual frequency υ).

(8) The expected Fano information in natural units can be written as I(X;Y) where each of X and Y can be one event or a joint event: I(X; Y) = ξ(s = 1, n[X,Y]) - ξ (s = 1, e[X,Y]). Note that n[ ] is the observed frequency and e is the expected frequency, e.g. in the chi-squared test sense.

References

[1] Melvin M. Vopson, (2024), "The second law of infodynamics and its implications for the simulated universe hypothesis", AIP Advances, 13, 10.

[2] Melvin M. Vopson, (2024), "Introducing the Second Law of Infodynamics and its Implications, IPI Lecture, 21st September. https://youtu.be/Wi8OYvqQWUU

[3] W. H. Zurek (Editor), (2023), " Complexity, Entropy, and the Physics of Information", Vols I and II, Santa Fe Institute Press.

[4] M. Aoyagi and S.Watanabe,(2009) "The zeta function for learning theory and resolution of singularities", More Progresses in Analysis. May 2009, 279-288.

[5] 5. P. Tempesta, (2011), "Group entropies, correlation laws and zeta functions", ARXIV 1105.1935.

[6] N. Combe, Y.I. Manin, M. Marcolli, (2022), "Geometry of information: Classical and quantum aspects", Theoretical Computer Science, 908, 2-27.

[7] Pain, R.H., and Robson, B. (1970) "Analysis of the Code Relating Sequence to Conformation in Globular Proteins", Nature (1970), 227, 62-63.

[8] Robson, B, and Pain, R. H. (1971) ."Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Possible Implications for the Mechanism of Formation of Helical Regions", J. Mol. Biol. 58, 237-256.

[9] 9. Robson, B. and Pain, R. H (1972) Directional Information Transfer in Protein Helices", Nature New Biology (1972) 238, 107-108.

[10] Robson, B. (1974) "Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Theory and Application of Expected Information", B. Robson, Biochem. J. (1974) 141, 853-867.

[11] J. Pearl, (1985), "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA. pp. 329–334.

[12] Robson, B. (2004) "The Dragon on the Gold: Myths and Realities for Data Mining in Biotechnology using Digital and Molecular Libraries", J. Proteome Res. (Am. Chem. Soc.) 3 (6), 1113 - 9.

[13] 13. Robson, B (2005) . "Clinical and Pharmacogenomic Data Mining: 3. Zeta Theory As a General Tactic for Clinical Bioinformatics" J. Proteome Res. (Am. Che. Soc.) 4(2); 445-455.

[14] B. Robson and O.K. Baek, (2024), "Use of a theory of expected information for sparse data and adverse events in clinical trials and other biomedical studies", Information Sciences, Volume 680, October, 121027.

[15] B. Robson and R. Cooper, (2024) "Glass Box and Black Box Machine Learning Approaches to Exploit Compositional Descriptors of Molecules in Drug Discovery and Aid the Medicinal Chemist", ChemMedChem,(Wiley), e202400169.

[16] B. Robson (2019), "Bidirectional General Graphs for inference. Principles and implications for medicine, Computers in Biology and Medicine, 10,382-399.