Compression and intelligence: social environments and communication

David L. Dowe[1] José Hernández-Orallo[2] Paramjit K. Das[1]

[1] Computer Science and Software Engineering, Clayton School of Information Technology, Monash University, Vic. 3800, Australia.
{david.dowe@infotech.monash.edu.au, pdas3@student.monash.edu.au}
[2] DSIC, Universitat Politècnica de València, València, Spain.
jorallo@dsic.upv.es

Abstract. Compression has been advocated as one of the principles which pervades inductive inference and prediction - and, from there, it has also been recurrent in definitions and tests of intelligence. However, this connection is less explicit in new approaches to intelligence. In this paper, we advocate that the notion of compression can appear again in definitions and tests of intelligence through the concepts of 'mind-reading' and 'communication' in the context of multi-agent systems and social environments. Our main position is that two-part Minimum Message Length (MML) compression is not only more natural and effective for agents with limited resources, but it is also much more appropriate for agents in (co-operative) social environments than one-part compression schemes - particularly those using a posterior-weighted mixture of all available models following Solomonoff's theory of prediction. We think that the realisation of these differences is important to avoid a naive view of 'intelligence as compression' in favour of a better understanding of how, why and where (one-part or two-part, lossless or lossy) compression is needed.

Keywords: two-part compression, Minimum Message Length (MML), Solomonoff theory of prediction, tests of intelligence, communication.

1 Compression, inference, prediction and intelligence

Several authors [1, 5, 6, 11, 7, 9] have suggested the relevance of compression to intelligence, especially the inductive inferential (or inductive learning) part of intelligence. M. Hutter even proposed a compression contest (the Hutter prize) which was "motivated by the fact that being able to compress well is closely related to acting intelligently" (http://prize.hutter1.net) [2, footnote 180]. However, many compression algorithms are able to compress data in a much better way than humans (either lossless or lossy compression). Humans are better at compressing information which is relevant to their goals (or rewards). So, many agree that compression must have a role, but it is not clear which kind of compression must be considered.

One position advocated is that two-part Minimum Message Length (MML) compression [26, 28, 25, 4], which states the theory in the first part, gives the inductive inference part of intelligence [5, 6]. Other authors have considered the one-part Solomonoff predictive compression [20] to be the appropriate way of using the data for modelling, perhaps due to its emphasis on prediction rather than explanation and its presumed consequent superiority in predicting the future.

The relationship between MML and Kolmogorov complexity, the similarities between Wallace's MML inference/explanation work and Solomonoff's predictive work – and the subtle difference between inference/explanation and prediction – have been discussed in [28][25, chap. 2]. In short, Solomonoff will take a posterior-weighted mixture of all available models, and so his predictive approach will typically involve something which is not one of the available models - whereas the Wallace MML approach will use the single best available model. Technically, a mixture of models may not compress at all, since encoding all (or a great number of) the possible models may require more bits than the data itself.

In addition, there seems to be confusion amongst many authors about the distinction between one-part and (MML) two-part compression. In one-part compression, we simply wish to encode the data. In two-part (MML) compression, we wish to encode the model in the first part of the message and then we encode the data given the model in the second part of the message [28][25, chap. 2]. An alternative way of describing the two-part coding is that a (possibly Universal) Turing machine (TM) could read the first part of the message, whereupon it would write nothing but rather go into an "educated" state or become an Educated Turing Machine (ETM) [25, chap. 2][28]. Upon reading the second part of the message (which encodes the data), the (now Educated) TM would perform a decoding and then write out the data.

However, in terms of a single agent operating in some environment, it will clearly predict better (even if only slightly) when using the Solomonoff predictive distribution. Nonetheless, if the agent is time-limited – as it typically will be in a realistic environment – then there will be disadvantages to using the entire Solomonoff posterior predictive distribution. Indeed, this will typically involve infinite summations and – further – the uncomputability of the Halting problem. It is worth mentioning, though, that some approximations can work in practice (such as Monte Carlo AIXI [24]) by reducing the number of models in the mixture.

Partly in response to Searle's "Chinese room" argument [19], we also raise the issue of compression as a non-behavioural (introspective) indicator of intelligence - i.e., given two agents who have scored equally well on a test and one of which compresses better than the other, which should we prefer [5, sec. 5.1][6, sec. 5][4, sec. 7.3]?[3] We compare this to other purely behavioural ways of assessing and detecting intelligence.

[3] We certainly note [16, sec. 5.2] that human society gives Nobel prizes and various other accolades to those who give a good single theory (or MML explanation) for observed data. Examples include (e.g.) special/general relativity, Helicobacter pylori as the cause of stomach ulcers, etc.

The rest of the paper analyses the relation between the several views and applications of the notion of compression and intelligence, focussing on social environments and communication.

2 Social environments and communication

Social environments and multi-agent systems generally include competition and co-operation. For competition, it is necessary to have mind-reading abilities in order to anticipate what other agents might do (predator-preys, games such as the *prisoners' dilemma*, etc.). While we could perhaps use a mixture of models for these social environments as well, the other agents are resource-bounded, and they will generally act according to a reduced number of models – or a single one. Consequently, using a large mixture of models to explain and predict the behaviour of other agents seems inefficient and unrealistic.

Nonetheless, it is in co-operation where the different approaches to inductive inference and prediction perhaps become more apparent. First, co-operation implies communication. In order to communicate a concept, we need an efficiently compressed expression of the concept. We do not expect to transmit a mixture of models but a single model. Second, in order to transmit (i.e., understand) the concept, we need descriptions which are clearly separated from the data. Here, a two-part compression seems to have advantages over a one-part compression, since with the former it is easier to extract the concept or model we want to communicate. Third, in co-operation, agents need to share models and procedures. In other words, agents should share the same ontology. This is only possible if the ontology can be isolated from the data – and if it is the same for all.

Let us elaborate upon the points from the above paragraph with some examples. The creation of language is about developing a set of (hierarchical) concepts for the purposes of concise description of the observed world and correspondingly concise communication. Elaborating upon the ideas outlined in [25, chap. 9] (and [2, footnote 128][4, sec. 7.2]), this can be thought of as a problem of (hierarchical) intrinsic classification or (hierarchical) mixture modelling (or clustering), where we might identify classes such as (e.g.) animal, vegetable, mineral, animal-dog, animal-cat, vegetable-carrot, vegetable-potato, vegetable-fruit, mineral-metal, mineral-salt, animal-dog-labrador, animal-dog-collie, animal-dog-labrador-black, animal-dog-labrador-golden, etc. Following these principles of MML mixture modelling [26, 27, 29, 25] enables us to arrive at a *single* theory, which is the first part of an MML message and which describes the concepts or classes. The data of all the various individual animals, vegetables and minerals (or things) on the planet (such as their heights and weights, etc.) is encoded in the second part of the message. Users of the language are free to communicate the concepts from this single best MML theory.

Knowledge (and human knowledge especially) in a social environment is all about this, about sharing models. And this shared knowledge makes co-operation possible. For humans (elevated in knowledge), science is a type of knowledge where we typically use one theory to explain the evidence, and not hundreds.

Despite the rationale that one model (or a small set of models) is better for resource-bounded agents which need to communicate their concepts, there are some other issues around compression and intelligence that are more difficult to dissect.

2.1 Lossless and lossy compression

In other areas of computer science (image, audio and video processing in particular), we clearly distinguish between lossless and lossy compression [17, 15]. In inductive inference, this distinction is less clear. Prediction and inference can also be defined and performed in noisy environments, where some details have to be lost to avoid overfitting (see, e.g., [25, sec. 4.9]). This is, of course, one of the rationales behind two-part codes, where the theory part could be seen as the lossy compression and the other part could be seen as the detail which (optionally) is used to cover the rest of the data. In fact, some compression schemes may have more than two parts, with each part adding more detail to the previous part, in a hierarchical way (although the MML message could be re-structured so that this is again in two parts). Perception is a clear example of this as well, especially because the world deals with continuous (non-discrete) sources of data.

One issue which is difficult to isolate is the 'distortion criterion' [17] for lossy compression. In image, audio and video compression, the distortion and quality criteria are set by human perception - i.e., what kind of loss is acceptable depending on the application. If this external reference is lost, it is much more difficult to distinguish the information that can be lost from the information that should be preserved. Perception and intelligence must be able to determine the details which are relevant to an agent's actions and those which are completely irrelevant - i.e., agents must perform selectively lossy compression. Memory and everyday linguistic concepts must also be able to drop details and keep the essential. The mechanisms and principles which should guide all this are yet to be discovered. In many codings which are used in reinforcement learning (e.g. [21, 22]), compression is used to code future rewards efficiently, so any detail which is irrelevant to predict future rewards can be dropped. In fact, this link between compression and reinforcement can be made explicit [8]. Again, compression is required, but the precise formulation and application is crucial.

2.2 The elusive model paradox and (human) unpredictability

The interaction between predator and prey, between sellers and buyers, or the behaviour which takes place in board or mind games (such as the *prisoners' dilemma*) has been analysed in ethology, economics, game theory, artificial intelligence and other disciplines. We can discuss all this in terms of prediction and compression.

For example, Scriven discusses the notion of (human) predictability [18] in one of the simplest possible social environments: an iterated game of two humans with one trying to do what the other does and the other trying to avoid this happening. Scriven finds an apparent logical paradox that both should be able to predict the other, while Lewis and Shelby Richardson [13] note Scriven's assumption that the calculations done by each agent in modelling the other are required to terminate. Indeed, whether one looks at doing two-part MML inferential modelling or Solomonoff predictive modelling, one ultimately runs into the Halting problem (or Entscheidungsproblem) [2, footnote 211][3, p455][4, sec. 7.5] - and (the paradox is circumvented by the fact that) the relevant calculations will not terminate. The ability to recognise "*other minds*" and engage in "mind-reading" is clearly advantageous in general in social environments. It is presumably of little surprise that two competing agents of equal computational power and equal inference (modelling) or predictive technique have no advantage over one another.

3 Detecting and assessing intelligence

The understanding of compression as a necessary trait of intelligence has led to some approaches for detecting and assessing intelligence where compression plays a fundamental role. Some of these approaches are non-behavioural, i.e., introspective, and require an analysis of the models the agent is using. In fact, the analysis of the level of compression in the models was used as a response to Searle's "Chinese Room" argument [19]. In [5, sec. 2.1][6, sec. 2][4, sec. 7.3], compression was advocated as a non-behavioural way of assessing and detecting intelligence. This required measuring the bits of the model the agents are using, if we are comparing them. This idea is even more explicit in the *Hutter prize* (`http://prize.hutter1.net`) [2, footnote 180]. In general, however, it is not possible to precisely measure the length of a model by introspection, since the inner knowledge representation may not be accessible. Even for artificial agents, this might be impractical as agents become more and more complex.

One possible way to overcome this limitation is through the use of language. Through language we can ask and communicate models and see whether the explanation for a phenomenon (or an action) given by an agent is shorter than the explanation given by another agent. In fact, interviews, exams and other kinds of tests commonly tell between rote learning and full comprehension by requesting an explanation for the answers, which can then be compared to the right model. This is also recurrent in the Turing Test [23, 14] and its implementations, where the artificial agents frequently fail when they are asked to give explanations. This is well-known in psychology as well, where there are many introspective techniques based on asking the right questions.

The other possible way is to stick to purely behavioural tests, which are completely independent from the nature of the agents. Psychometric tests are generally behavioural, since subjects only need to guess answers right or wrong. Many evaluation settings in artificial intelligence are also behavioural, such as game contests, robot competitions, reinforcement learning evaluation, etc. Although behavioural tests seem to be disconnected from the notion of compression, the links arise again in many and diverse ways. Firstly, since prediction and compression are linked, performance is better for those systems which are able to compress the evidence (in a goal-oriented way). Secondly, the difficulty of the exercises or tasks which are used to detect intelligence can be approximated using notions which are closely related to compression, such as many variants of Kolmogorov complexity. Finally, the distribution of tasks can be obtained using some kind of universal distribution. All this has been explored by [11, 7, 9, 12, 10], where the original static (sequence-prediction) tests have evolved into more interactive and adaptive tests.

Finally, it is insightful (as an extreme case) to see whether (and how) intelligence can be detected through a (slow) uni-directional form of communication - where, rather than having interactive conversation, instead we send a message conveying some information which we hope is understood. When no previous knowledge is shared, this seems impossible due to the lack of common references. However, compression is again advocated as a possibility to make this feasible, even in the case of uni-directional messages[4].

4 Conclusions

In this paper we have discussed the role that compression might have in intelligence, with an emphasis on communication and language, and the exchange and evaluation of models.

We have argued that the ability to do *two*-part (MML) compression is (in general) an advantage in *social* environments. It is an advantage firstly for the same reasons that it is an advantage in an isolated environment of one agent, including the fact that the MML-inferred theory is a good predictor. But, secondly, it will also typically be an advantage in the (co-operative) social environment, where we can teach (or tell or show) our theories to others. One interesting area of research would be to follow the ideas in Monte Carlo AIXI [24] and construct MML agents, and see whether the latter behave better (with the same resources) in social environments.

Hence, while agreeing that both the optimal Solomonoff predictor and the Wallace MML inference are both relevant to at least the inductive inference (or inductive learning) part of intelligence, we take the position here of suggesting that – at least in the context of social agents in a multi-agent environment – MML is perhaps more pertinent to what we (as social humans in our multi-human environment) might commonly refer to as ‘intelligence’.

Acknowledgments

We thank the anonymous reviewers for their helpful comments, and we thank Kurt Kleiner for some challenging and ultimately very helpful questions in the

[4] Indeed, we can use similar principles to construct a message to send in search of an alien intelligence [2, sec. 0.2.5, p542].

broad area of this work. We also acknowledge the funding from the Spanish MEC and MICINN for projects TIN2009-06078-E/TIN, Consolider-Ingenio CSD2007-00022 and TIN2010-21062-C02, and Generalitat Valenciana for Prometeo/2008/051.

References

1. G. J. Chaitin. Godel's theorem and information. *International Journal of Theoretical Physics*, 21(12):941–954, 1982.
2. D. L. Dowe. Foreword re C. S. Wallace. *Computer Journal*, 51(5):523 – 560, Sept 2008. Christopher Stewart WALLACE (1933-2004) memorial special issue.
3. D. L. Dowe. Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference - from (medical) "evidence". *Social Epistemology*, 22(4):433 – 460, October - December 2008.
4. D. L. Dowe. MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In P. S. Bandyopadhyay and M. R. Forster, editor, *Handbook of the Philosophy of Science - Volume 7: Philosophy of Statistics*, pages 901 – 982. Elsevier, 2011.
5. D. L. Dowe and A. R. Hajek. A computational extension to the Turing Test. *Technical Report #97/322, Dept Computer Science, Monash University, Melbourne, Australia, 9pp*, 1997.
6. D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing Test. In *Intl. Conf. on Computational Intelligence & multimedia applications (ICCIMA'98), Gippsland, Australia*, pages 101–106, February 1998.
7. J. Hernández-Orallo. Beyond the Turing Test. *J. Logic, Language & Information*, 9(4):447–466, 2000.
8. J. Hernández-Orallo. Constructive reinforcement learning. *International Journal of Intelligent Systems*, 15(3):241–264, 2000.
9. J. Hernández-Orallo. On the computational measurement of intelligence factors. In A. Meystel, editor, *Performance metrics for intelligent systems workshop*, pages 1–8. National Institute of Standards and Technology, Gaithersburg, MD, U.S.A., 2000.
10. J. Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508 – 1539, 2010.
11. J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proc. Intl Symposium of Engineering of Intelligent Systems (EIS'98)*, pages 146–163. ICSC Press, 1998.
12. S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.
13. D. K. Lewis and J. Shelby-Richardson. Scriven on human unpredictability. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 17(5):69 – 74, October 1966.
14. G. Oppy and D. L. Dowe. The Turing Test. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy.* Stanford University, 2011. http://plato.stanford.edu/entries/turing-test/.
15. D. Salomon, G. Motta, and D. C. O. N. Bryant. *Handbook of data compression.* Springer-Verlag New York Inc, 2009.

16. P. Sanghi and D. L. Dowe. A computer program capable of passing I.Q. tests. In *4th International Conference on Cognitive Science (and 7th Australasian Society for Cognitive Science Conference)*, volume 2, pages 570–575, Univ. of NSW, Sydney, Australia, Jul 2003.
17. K. Sayood. *Introduction to data compression.* Morgan Kaufmann, 2006.
18. M. Scriven. An essential unpredictability in human behavior. In B. B. Wolman and E. Nagel, editors, *Scientific Psychology: Principles and Approaches*, pages 411–425. Basic Books (Perseus Books), 1965.
19. J. R. Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.
20. R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964.
21. R. S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in neural information processing systems*, pages 1038–1044, 1996.
22. R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction.* The MIT Press, 1998.
23. A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.
24. J. Veness, K.S. Ng, M. Hutter, and D. Silver. A Monte Carlo AIXI Approximation. *Journal of Artificial Intelligence Research, JAIR*, 40:95–142, 2011.
25. C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length.* Springer-Verlag, 2005.
26. C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
27. C. S. Wallace and D. L. Dowe. Intrinsic classification by MML - the Snob program. In *Proc. 7th Australian Joint Conf. on Artificial Intelligence*, pages 37–44. World Scientific, November 1994.
28. C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999. Special issue on Kolmogorov complexity.
29. C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, January 2000.