On Autistic Interpretations of Occam's Razor

Ismael García* José Hernández†

Extended Abstract

After the initial relevance of deduction in AI, nowadays, inductive learning, and all its varieties (abductive reasoning, reasoning by analogy, connectionism, ILP, grammatical inference, HMM, EBR, etc.), are beginning to play a more central and still agglutinative role in the proper subset of AI devoted to make intelligent machines. The issue has been much clearer in discovery science, where induction has been the prominent inference process.

In this trend, new unified frameworks for understanding reasoning have been appearing, with the aim of integrating all the different inference mechanims [25]. In particular, a radical approach has been undertaken by Wolff, with the claim that *"all kinds of computing and formal reasoning may usefully be understood as information compression by pattern matching, unification and search"* [46].

In this paper, we will discuss critically a very influential and now classical issue in this line, which is based on the view of unsupervised learning as compression [38], its famous operative Minimum Description Length (MDL) Principle [32] and, mainly, its formal justification. The MDL [32] principle is closely related with the Minimum Message Length (MML) principle [44] and Maximum Likelihood Estimators, but the idea is *formally* older [38]. However, they all are "fresh interpretations" [15] under Algorithmic Information Theory of a much older idea attributed to William of Ockham 1290?-1349? which is usually known as the Occam's Razor.

The principle was rejected by Popper because he said that there is no objective criterion for simplicity. But Algorithmic Complexity $C(x)$ or Kolmogorov Complexity $K(x)$ [24] is an objective criterion for simplicity. This is precisely what R.J.Solomonoff proposed as *a perfect theory of induction [43].* Algorithmic Complexity inspired J. Rissanen in 1978 to use it as a general modelling method, giving the popular MDL principle [32]: *The best model to explain a set of data is the one which minimises the sum of: the length, in bits, of the description of the theory; and, the length, in bits, of data when encoded with the help of the theory. Then, we enclose the exceptions, if any.*

*Universitat Politècnica de València, Institut Tecnològic d'Informàtica. Camí de Vera 14, Aptat. 22.012 E-46071, València, Spain. E-mail:ivarea@iti.upv.es

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

Recently, this idea of two-part code has been corrected to a one-part code [34], but the same problem can appear intrinsically, some part can be very compressed (the main rule) and other parts are quoted as exceptions.

Since the principle is not computable in general, it is usually approximated or used in restricted descriptive mechanisms, like attribute languages. It has been shown in practice [31] [14] [17] [16] [27] [48] [29] [15] that the MDL principle avoids over-generalisation (underfitting) and, usually, under-generalisation (overfitting).

Our discussion is motivated from the apparent contradiction between the so-called "no-free-lunch" theorems about induction [47] [36] stating that one learner cannot be better than another when performance is averaged uniformly over all possible problems. These results only allow that a learner could be better than another for a particular distribution of problems. Vitanyi & Li [43] show that the MDL principle is almost optimal for the universal distribution $2^{-K(x)}$. Of course, the universal distribution (i.e. Occam's Razor formalised) is just a choice when you have not any information at all about the real origin of the information, but is this the case in real applications of machine learning, scientific discovering or even cognition? Are we always so autistic about the source of the information that we pretend to discover?

Using the same information-theoretic approach, we study the case for finite and short data and we arrive to a slightly different result: MDL is a good principle but not the best one for finite data and/or perfect hypotheses. The argument is based upon recently introduced variants and definitions around the idea of Intensional Complexity, which intrinsically penalise or 'simply' do not allow exceptions, seen these as *extensional* descriptions. The idea is just to distribute more uniformly the compression ratio between the model and the data, avoiding that, for the sake of maximum compression, the model results in a very compressed part plus some cases that are not compressed at all, (i.e. quoted extensionally). This extensional part is not validated, making the whole theory weak. An ontology is difficult to construct from here if the exceptions are unrelated (not explained) with the other facts. That is to say, the point lays between the anomaly and the expected noise.

Using intensional complexity, a parametrised Shortest Intensional Description (SID) is defined. This changes the statement that *"optimal compression (Minimum Description Length (MDL)) gives you the best hypothesis provided the data are random with respect to the hypothesis, the data are not completely perfect and the data grow to infinity"* [43] into the following one *"the SID criterion gives you a more robust hypothesis when the data are perfect, ensuring and not supposing that the data are random to the hypothesis."* Moreover, it does not require that "the data grow to infinity," so it can be used to undertake finite real problems. More importantly, our definitions are free from the "MDL's principle paradox", since the shortest hypothesis is *never* random to the data. To solve this problem, we use time-space considerations or a resource-bounded randomness to avoid paradoxes. In addition, this yiedls our criterion

computable.

Encompassing the ideas of compression there were presented different models of learning: identification in the limit [19], PAC model [41], Query-Learning [2] and others, all based on the ideas of 'identification', defined as the moment (the limit) where no 'mind change' is possible. In the framework of incremental learning, it is shown that our intensional criterion is less conservative that the MDL principle, and consequently it minimises the number of whole 'mind changes' (although these changes are usually more radical). Loosely, we should say that the MDL principle complies with Kuhn's philosophy of changing paradigms; when the number of exceptions is too great, the paradigm must be changed. In contrast, the SID usually anticipates this necessity since any exception forces the revision of the model.

This engages with the classical dilemma between informative and probable hypotheses. It is clear that an explanation must have some degree of plausibility to avoid fantastic hypotheses, but in many applications, like scientific discovery or abduction, we must regard an explanation as an investment, even a "risky bet" that could be soon falsified. This is merely Popper's criterion of falsifiability [30]: one does not always want the most likely explanation, because sometimes it is the less informative too.

The issue is clear when the data are random (and this usually happens with short data because it makes no worthy any compression). The MDL principle just gives the data themselves, which does not correspond to the idea of 'model'. More importantly, no learning has taken place. By forcing intensionality, different informative hypotheses can be induced. This gives clues to the enigma of "hyper-learning" or "poverty of stimulus" in those cases where the MDL principle cannot give the "intuitive" hypothesis in most of the bias.

Finally, we define a new formal notion for the "intensional value" of a hypothesis, namely as a quotient between the computational effort that has been made from the data to the hypothesis divided by the computational effort that is made form the hypothesis to the data. In this way, the connection with learning and Levin's "Universal Search Problems" is made explicitly. The complexity of discovering is equal to the complexity of increasing the "intensional value", which is proven to be NP-hard.

From here, and very far from the classical notion of 'identification', we propose a different notion of learning (or discovering): *the more a system learns the more intensionally valuable the description is with respect to the data.* Consequently the blurry notions of underfitting and overfitting may be better understood.

In conclusion, the MDL principle works well in those environments where the bias does not allow extensional descriptions or where the data are huge and from statistical or imperfect sources. But, when faced to a concrete learning problem, we have to tune length, computational time, robustness (or intensionality) and 'informativeness' of descriptions according to the expectation we have about the source of knowledge. In our view, Occam's Razor should be understood in this

non-autistic way.

Keywords: Knowledge Discovery, Model Formation, Hypotheses Selection, Occam's Razor, Minimum Description Length (MDL) Principle, Machine Learning, Intensional Complexity, Overfitting.

References

[1] Abe, N. "Towards Realistic Theories of Learning". *New Generation Computing, 15*, 1997.

[2] Angluin, D. "Queries and concept learning". *Machine Learning 2, No. 4*, pp. 319-342, 1988.

[3] Balasubramanian, V. "Statistical Inference, Occam's Razor, and Statical Mechanics on the Space of Probability Distributions". *Neural Computation, 9*, pp. 349-368, 1997.

[4] Blum, M. "A machine-independent theory of the complexity of recursive functions". *J. ACM 14, 4*, 322-6, 1967.

[5] Blum, L.; Blum, M. "Towards a mathematical theory of inductive inference". *Inform. and Control 28*, pp. 125-155, 1975.

[6] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M.K. "Occam's razor". *Inf.Proc.Lett. 24*, pp. 377-380, 1987.

[7] Blumer, A.; Ehrenfeucht, A.; Haussler, D.; Warmuth, M. "Learnability and the Vapnik-Chervonenkis Dimension". *Journal of ACM, 36*, pp. 929-965, 1989.

[8] Board, R.; Pitt, L. "On the necessity of Occam algorithms". *In Proc., 22nd ACM Symp. Theory of Comp.*, 1990.

[9] van den Bosch "Simplicity and Prediction" Master Thesis, department of Science, Logic & Epistemology of the Faculty of Philosophy at the University of Groningen, 1994.

[10] Carnap, R. "Meaning and necessity: A study in semantics and modal logic". *Chicago, University of Chicago Press*, 1947.

[11] Case J.; Smith, C. "Comparison of identification criteria for machine inductive inference". *Theoret. Comput. Sci. 25*, pp. 193-220, 1983.

[12] Chaitin, G.J. "Algorithmic Information Theory". *Fourth printing, Cambridge University Press*, 1992.

[13] Chen, K. "Tradeoffs in inductive inference of nearly minimal sized programs". *Inform. and Control 52*, pp. 68-86, 1982.

[14] Cheeseman, P. "On finding the most probable model". *In Shrager, J. and Langley, P. "Computational models of scientific discovery and theory formation"*, chapter 3, Morgan Kaufmann, 1990.

[15] Conklin, D.; Witten, I. H. "Complexity-Based Induction". *Machine Learning, 16*, pp. 203-225, 1994.

[16] Derthick, M. "The Minimum Description Length Principle Applied to Feature Learning and Analogical Mapping". *MCC Technical Report Number ACT-CYC-234-90*, 1990.

[17] Freivalds, R. "Inductive inference of minimal size programs". *In M. Fulk and J. Case (eds) "Proceedings of the third Annual Workshop on Computational Learning Theory"*, pp. 1-20, Morgran Kaufman, San Mateo, CA, 1990.

[18] Freivalds, R.; Kinber, E.; Smith, C.H. "On the Intrinsic Complexity of Learning". *Inf. and Control 123*, pp. 64-71, 1995.

[19] Gold, E. M. "Language Identification in the Limit". *Inform and Control, 10*, pp. 447-474, 1967.

[20] Gull, S.F. "Bayesian inductive inference and maximum entropy". *In Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol. 1: Foundations, de. by G.J. Erickson and C.R. Smith, pp. 53-74. Dordrecht: Kluwer 1988.

[21] Hinton, G.; Zemel, R. "Autoencoders, minimum description length, and Helmholtz free energy". *In Cowan, J., Tesauro, G., and Alspector, J. (eds.) Advances in Neural Information Processing Systems 6*, Morgan Kaufmann Publishers, San Francisco, CA., 1994.

[22] Kuhn, T.S. "The Structure of Scientific Revolution" University of Chicago, 1970.

[23] Levin, L.A. "Universal search problems" *Problems Inform. Transmission, 9*, pp. 265-266, 1973.

[24] Li, M.; Vitanyi, P. "An Introduction to Kolmogorov Complexity and its Applications" *2nd Ed. Springer-Verlag*, 1997.

[25] Michalski, R.S. "Inferential Theory of Learning as a Conceptual Basis for Multistrategy Learning" *Machine Learning, 11*, 111-151, 1993.

[26] Muggleton, S. "Inductive Logic Programming" *New Generation Computing, 8, 4*, pp. 295-318, 1991.

[27] Muggleton, S.; Srinivasan, A.; Bain, M. "Compression, significance and accuracy". *In D. Sleeman and P. Edwards (eds.) Machine Learning: Proceedings of the Ninth International Conference (ML92)*, pp. 523-527, Wiley 1992.

[28] Muggleton, S. "Inverse Entailment and Progol" *New Generation Computing, 13*, pp. 245-286, 1995.

[29] Pfahringer, Bernhard. "Controlling Constructive Induction in CiPF: An MDL Approach". *In F. Bergadano and L. de Raedt (eds) Machine Learning, Proceedings of the European Conference on Machine Learning (ECML-94)*, pp. 242-256, Lecture Notes in AI 784, Springer-Verlag 1994.

[30] Popper, K.R. "Conjectures and Refutations: The Growth of Scientific Knowledge". Basic Books, New York, 1962.

[31] Quinlan, J.; Rivest. R. "Inferring decision trees using the minimum description length principle". *Information and Computation*, vol. 80, pp. 227-248., 1989.

[32] Rissanen, J. "Modelling by the shortest data description" *Automatica-J.IFAC, 14*, pp. 465-471, 1978.

[33] Rissanen, J. "Stochastic Complexity and modelling" *Annals Statist. vol. 14*, pp. 1080-1100, 1986.

[34] Rissanen, Jorma J. "Fisher Information and Stochastic Complexity" IEEE Transactions on Information Theory, Vol.42, No.1, January 1996.

[35] Rivest, R.L.; Sloan, R. "A Formal Model of Hierarchical Concept Learning". *Inf. and Comp. 114*, pp. 88-114, 1994.

[36] Schaffer, C. i "A conservation law for generalization performance". *In Proceedings of the Eleventh International Conference on Machine Learning*, pp. 259-265, Morgan Kaufmann, 1994.

[37] Shapiro, E. "Inductive Inference of Theories form Facts". *RR 192, D. Comp. Science, Yale Univ., 1981, in Lassez, J.; Plotking, G. (eds.) "Computational Logic"*, The MIT Press 1981.

[38] Solomonoff, R.J. "A formal theory of inductive inference". *Inf. Control.* Vol. 7, 1-22, Mar., pp. 224-254, June 1964.

[39] Solomonoff, R.J. "Complexity-based induction systems: comparisons and convergence theorems". *IEEE Trans. Inform. Theory, IT-24*, pp. 422-432, 1978.

[40] Stolcke, A.; Omohundro, S. "Inducing Probabilistic Grammars by Bayesian Model Merging". *In R.C. Carrasco and J. Oncina (Eds.) Grammatical Inference and Applications, Lecture Notes in Artificial Intelligence, 862*, pp. 106-118, Springer-Verlag 1994.

[41] Valiant, L. "A theory of the learnable". *Communication of the ACM 27(11)*, pp. 1134-1142, 1984.

[42] Vitanyi, P.; Li, M. "Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity".

[43] Vitnyi, P.; Li, M. "On Prediction by Data Compression" Proc. 9th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, Vol. 1224, Springer-Verlag, Heidelberg, 14-30, 1997.

[44] Wallace, C.S; Boulton, D.M. "An information measure for classification". *Computing Journal 11*, pp. 185-195, 1968.

[45] Watanabe, S. "Pattern Recognition as Information Compression". *In Watanabe (ed.) Frontiers of Pattern Recognition New York: Academic Press*, 1972.

[46] Wolff, J.G. "Computing as Compression: An Overview of the SP Theory and System". *New Gen. Computing 13*, pp. 187-214, 1995.

[47] Wolpert, D. "On the connection between in-sample testing and generalization error". *Complex Systems, 6*, pp. 47-94, 1992.

[48] Zemel, R. "A minimum description length framework for unsupervised learning". *Ph. D. Thesis, Dept. of Computer Science, University of Toronto*, Toronto, Canada, 1993.