Introduction to Minimum Message Length, applications and related issues

David L. Dowe

School of Computer Science and Software Engineering,
Clayton School of I.T., Monash University, Clayton, Vic 3800, Australia;
e-mail: dld@brucedotcsse dot monash . edu dot au www.csse.monash.edu.au/~dld

1 Introduction - Brief History

Minimum Message Length (MML) machine learning statistical (or inductive) inference, "data mining" trade-off between simplicity of hypothesis (H) and goodness of fit to the data (D) (Wallace & Boulton, 1968 [45, p185 sec. 2]) [3][4, p64 col. 1][1][6, sec. 1 col. 1][5][7, sec. 1 col. 1] (Wallace & Boulton, 1975 [46, sec. 3]) [2][52][51][48] (Wallace 2005 book, "*Statistical and Inductive Inference by Minimum Message Length*" [44]) (Comley & Dowe, 2003 [8]) (Comley & Dowe, M.I.T. Press, April 2005 [9, secs. 11.1 and 11.4.1]) (Dowe, Gardner & Oppy, Brit. J. Phil. Sci. 2007 [16]) (Dowe, 2008a, "Foreword re C. S. Wallace", Christopher Stewart WALLACE (1933-2004) memorial special issue, *Computer Journal*, Oxford Univ Press [11, sec. 0.2.4, p535 col. 1 and elsewhere]).

MML is Bayesian, advocates *two-part* messages (H, then D given H), substantially before (Rissanen 1978 [34]) Minimum Description Length (MDL).

Statistical invariance (x, y), polar: "same"
Most classical statistical methods statistically invariant

MML statistically invariant [46], but most other Bayesian methods in use not statistically invariant

Statistical consistency Converge to the right answer as the amount of data increases

Neyman-Scott problem (1948 [33])

1. the heights $\mu_1, ..., \mu_N$ of each of the N people,
2. the accuracy (σ) of the measuring instrument.

We have JN measurements from which we need to estimate $N + 1$ parameters.
$JN/(N + 1) \leq J$, so the amount of data per parameter is bounded above (by J).

It turns out that $\hat{\sigma}^2_{\mathrm{MaximumLikelihood}} \rightarrow \frac{J-1}{J}\sigma^2$, and so for fixed J as $N \rightarrow \infty$

Akaike's AIC, Schwarz's BIC (1978), Rissanen's MDL (1978, [34]) all statistically inconsistent for the Neyman-Scott problem
MML statistically consistent (Dowe & Wallace, 1997 [22]) (Wallace, 2005 [44])

General form of Neyman-Scott problem: amount of data per parameter bounded above

E.g., aptitude tests and IQs; testing petrols on many engines and octane ratings; etc.

Statistical inconsistency in rival methods but no known case yet of MML being statistically inconsistent

Conjecture(s) (Dowe, Baxter, Oliver & Wallace 1998 [13]) (Wallace & Dowe, 1999a [48]) (Comley & Dowe, MIT Press, 2005 [9]) (Dowe, Gardner & Oppy, Brit J Phil Sci 2007 [16]) (Dowe 2008a, "Foreword re C. S. Wallace" [11]) :

Only MML and closely related Bayesian methods will be both statistically invariant and statistically consistent in general for problems where the amount of data per parameter is bounded above;

If the above conjecture is wrong and there are any non-Bayesian methods, then they will converge to the true answer more slowly than MML does.

Slight variant of Conjecture(s) for model misspecification

Elusive model paradox (Dowe 2008a [11], Dowe 2008b [12])

Consider two processes: one generates a sequence of numbers (or bits), the other tries to guess the sequence.

First - or generating - sequence is like a soccer player taking a penalty kick or a tennis player serving a ball. It tries to get different to what the guesser will guess.

Second - or guessing - sequence is like soccer goalie or tennis receiver, and tries to guess generated sequence.

If both use methods that are statistically consistent, then first can eventually anticipate guessing sequence and change it while second can eventually accurately home in on first sequence.
Paradox?

Only one known way out of elusive model paradox.

Probabilistic prediction, uniqueness of log-loss

(Good 1952 [24]) introduces log-loss for the binomial distribution
Score: $-\log(p)$ or $-\log(1-p)$

(Dowe & Krusel 1993 [20, p4, Table 3]) uses log-loss for (8-state) multinomial distribution

Introduced by Dowe et al. (1996) for Normal/Gaussian distribution, for margins on Australian Football League (AFL) games [20, p4, Table 3][21, 14, 15, 19][13, sec. 3][32, Figs. 3-5][36, sec. 4][31, Table 2][8, sec. 9][37, sec. 5.1][9, sec. 11.4.2][38, sec. 3.1][29, Tables 2-3][30][39, secs. 4.2 - 4.3] (and possibly also [40, sec. 4.3]), [10](Dowe 2008a [11, sec. 0.2.5, footnotes 170-176 and accompanying text])(Dowe 2008b [12, pp437-438])

Uniqueness (Dowe, 2008a [11] and 2008b [12])
Log-loss shown to be *the* unique scoring system for probabilistic predictions which is invariant to framing of questions

Generalised Bayesian net and other applications

Following (Dowe & Wallace 1998 [23]), (Comley & Dowe June 2003 [8]) give first application of MML to Bayesian networks using both discrete (multi-valued) and continuous-valued attributes.
Repeated and refined in (Comley & Dowe MIT Press April 2005 [9], camera-ready version submitted in Oct 2003).

Many other applications of MML - including, e.g., clustering and mixture modelling (Wallace & Dowe 1994 [47]), (Wallace & Dowe 2000 [50]) and spatial correlation (Wallace 1998 [43], Visser & Dowe 2007 [41]) - and, in turn, to (e.g.) climate modelling (Visser, Dowe & Uotila 2009 [42]).

Relationship between MML and Kolmogorov complexity (Wallace & Dowe 1999a, "Minimum Message Length and Kolmogorov complexity", *Computer J* [48]) highlights the universality of MML in modelling problems.
Statistical consistency keeps all in order.
But poor approximations don't always work - several criticisms of MML and/or Ockham's razor (e.g., Kearns, Mansour, Ng & Ron 1997 [28]) are premised on inefficient or unreliable coding schemes

Invariant "priors"

Sir H. Jeffreys (1946) [27] notes that the square root of the expected Fisher information has the same mathematical form as a Bayesian prior *and* that it is statistically invariant

Although Jeffreys himself never actually advocated its use, Rissanen (1996 [35]) uses it as what he calls a "prior" in some of his later MDL work.

Chris Wallace and others have argued against its use on philosophical grounds - e.g., (Wallace & Dowe 1999b [49]). Basically, it comes from the data, not *prior* to it.

That said, for the fun of it, I have used MML and Bayesian invariance principles to create a multitude of invariant "objective" priors (whose use in practice I do not necessarily advocate).

MML and "intelligence"

In addressing an audience from a Centre for Research in *Intelligent* Systems, ...

Inductive learning = two-part compression (Dowe & Hajek, 1997a, 1997b, [17]) (Dowe & Hajek, 1998, [18])

See also related slightly later work by J. Hernandez-Orallo [26, 25].

References

1. D. M. Boulton. Numerical classification based on an information measure. Master's thesis, M.Sc. thesis, Basser Computing Dept., University of Sydney, Sydney, Australia, 1970.
2. D. M. Boulton. *The Information Measure Criterion for Intrinsic Classification.* PhD thesis, Dept. Computer Science, Monash University, Clayton, Australia, August 1975.
3. D. M. Boulton and C. S. Wallace. The information content of a multistate distribution. *J. Theor. Biol.*, 23:269–278, 1969.
4. D. M. Boulton and C. S. Wallace. A program for numerical classification. *Computer Journal*, 13(1):63–69, February 1970.
5. D. M. Boulton and C. S. Wallace. A comparison between information measure classification. In *Proc. of the Australian & New Zealand Association for the Advancement of Science (ANZAAS) Congress*, August 1973. abstract.
6. D. M. Boulton and C. S. Wallace. An information measure for hierarchic classification. *Computer Journal*, 16(3):254–261, 1973.
7. D. M. Boulton and C. S. Wallace. An information measure for single link classification. *Computer Journal*, 18(3):236–238, 1975.
8. Joshua W. Comley and David L. Dowe. General Bayesian networks and asymmetric languages. In *Proc. Hawaii International Conference on Statistics and Related Fields*, 5-8 June 2003.
9. Joshua W. Comley and David L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, M. A. Pitt, and I. J. Myung, editors, *Advances in Minimum Description Length: Theory and Applications (MDL Handbook)*, pages 265–294. M.I.T. Press, April 2005. Chapter 11, ISBN 0-262-07262-9. Final camera-ready copy submitted in October 2003. [Originally submitted with title: "Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages".].
10. D. L. Dowe. Discussion following "Hedging predictions in machine learning, A. Gammerman and V. Vovk". *Computer Journal*, 2(50):167–168, 2007.

11. D. L. Dowe. Foreword re C. S. Wallace. *Computer Journal*, 51(5):523 – 560, September 2008. Christopher Stewart WALLACE (1933-2004) memorial special issue.
12. 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.
13. D. L. Dowe, R. A. Baxter, J. J. Oliver, and C. S. Wallace. Point estimation using the Kullback-Leibler loss function and MML. In X. Wu, Ramamohanarao Kotagiri, and K. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 87–95, Berlin, April 15–17 1998. Springer.
14. D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. *3rd Conf. on Maths and Computers in Sport*, pages 233–241, 1996. See also Technical Report TR 96/297, Dept. Computer Science, Monash University, Australia 3168, Dec 1996.
15. D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. Technical report TR 96/297, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1996.
16. D. L. Dowe, S. Gardner, and G. R. Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, 58(4):709 – 754, December 2007.
17. D. L. Dowe and A. R. Hajek. A computational extension to the Turing test. Technical Report 97/322, Dept. Computer Science, Monash University, Australia 3168, October 1997.
18. D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing test. In *Proceedings of the International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98)*, pages 101–106, Gippsland, Australia, February 1998.
19. D. L. Dowe, A.J. Hurst, K.L. Lentin, G. Farr, and J.J. Oliver. Probabilistic and Gaussian football prediction competitions - Monash. *Artificial Intelligence in Australia Research Report*, June 1996.
20. D. L. Dowe and N. Krusel. A decision tree model of bushfire activity. Technical report TR 93/190, Dept. of Computer Science, Monash University, Clayton, Vic. 3800, Australia, September 1993.
21. D. L. Dowe, K.L. Lentin, J.J. Oliver, and A.J. Hurst. An information-theoretic and a Gaussian footy-tipping competition. *FCIT Faculty Newsletter, Monash University, Australia*, pages 2–6, June 1996.
22. D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics - 28th Symposium on the interface*, volume 28, pages 614–618, 1997.
23. D. L. Dowe and C. S. Wallace. Kolmogorov complexity, minimum message length and inverse learning. In W Robb, editor, *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 144, Queensland, Australia, July 1998.
24. I. J. Good. Rational decisions. *J. Roy. Statist. Soc. B*, B 14:107–114, 1952.
25. José Hernández-Orallo. Beyond the Turing test. *Journal of Logic, Language and Information*, 9(4):447–466, 2000.
26. José Hernandez-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems, ICSC Press*, pages 146–163, 1998.

27. H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. of the Royal Soc. of London A*, 186:453–454, 1946.
28. M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.
29. L. Kornienko, D. W. Albrecht, and D. L. Dowe. A preliminary MML linear classifier using principal components for multiple classes. In *Proc. 18th Australian Joint Conference on Artificial Intelligence (AI'2005)*, volume 3809 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 922–926, Sydney, Australia, Dec 2005. Springer.
30. L. Kornienko, D. W. Albrecht, and D. L. Dowe. A preliminary MML linear classifier using principal components for multiple classes. Technical report CS 2005/179, School of Computer Sci. & Softw. Eng., Monash Univ., Melb., Australia, 2005.
31. Lara Kornienko, David L. Dowe, and David W. Albrecht. Message length formulation of support vector machines for binary classification - A preliminary scheme. In *Lecture Notes in Artificial Intelligence (LNAI), Proc. 15th Australian Joint Conf. on Artificial Intelligence*, volume 2557, pages 119–130. Springer-Verlag, 2002.
32. S. L. Needham and D. L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In *Proc. 8th Int. Workshop on Artif. Intelligence and Statistics (AI+STATS 2001)*, pages 253–260, Jan. 2001.
33. J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrika*, 16:1–32, 1948.
34. J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
35. J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, 42(1):40–47, January 1996.
36. P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins. In R. McKay and J. Slaney, editors, *Proc. 15th Australian Joint Conference on Artificial Intelligence - Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Germany, ISSN: 0302-9743, Vol. 2557*, number 2557 in Lecture Notes in Artificial Intelligence (LNAI), pages 131–142. Springer Verlag, 2002.
37. P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Lecture Notes in Artificial Intelligence (LNAI) 2903 (Springer), Proc. 16th Australian Joint Conf. on Artificial Intelligence*, pages 269–281, Perth, Australia, Dec. 2003.
38. P. J. Tan and D. L. Dowe. MML inference of oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 3339 (Springer), Proc. 17th Australian Joint Conf. on Artificial Intelligence*, volume 3339, pages 1082–1088, Cairns, Australia, Dec. 2004.
39. P. J. Tan and D. L. Dowe. Decision forests with oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 5th Mexican International Conf. Artificial Intelligence*, pages 593–603, Apizaco, Mexico, Nov. 2006.
40. P. J. Tan, D. L. Dowe, and T. I. Dix. Building classification models from microarray data with tree-based classification algorithms. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 20th Australian Joint Conf. on Artificial Intelligence*, Dec. 2007.
41. Gerhard Visser and D. L. Dowe. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In *Proc. 6th IEEE International Conf. on Computer and Information Science (ICIS) 2007*, pages 17–22, July 2007.

42. Gerhard Visser, D. L. Dowe, and J. Petteri Uotila. Enhancing MML clustering using context data with climate applications. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*. Springer, December 2009. To appear, in press.
43. C. S. Wallace. Intrinsic classification of spatially correlated data. *Computer Journal*, 41(8):602–611, 1998.
44. C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer Verlag, May 2005. ISBN 0-387-23795X.
45. C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
46. C. S. Wallace and D. M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.
47. 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.
48. C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.
49. C. S. Wallace and D. L. Dowe. Refinements of MDL and MML coding. *Computer Journal*, 42(4):330–337, 1999.
50. 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.
51. C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B*, 49(3):240–252, 1987. See also Discussion on pp252-265.
52. C. S. Wallace and M. P. Georgeff. A general objective for inductive inference. Technical Report #83/32, Department of Computer Science, Monash University, Clayton, Australia, March 1983. Reissued in June 1984 as TR No. 44.