Modern MDL meets Data Mining Insights, Theory and Practice

Jilles Vreeken
CISPA Helmholtz Center for Information Security
Saarland Informatics Campus, Saarbrücken, Germany
jv@cispa.saarland

Kenji Yamanishi
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
yamanishi@mist.i.u-tokyo.ac.jp

ABSTRACT

When considering a data set it is often unknown how complex it is, and hence it is difficult to assess how rich a model for the data should be. Often these choices are swept under the carpet, ignored, left to the domain expert, but in practice this is highly unsatisfactory; domain experts do not know how to set k, what prior to choose, or how many degrees of freedom is optimal any more than we do.

The Minimum Description Length (MDL) principle can answer the model selection problem from an intuitively appealing and clear viewpoint of information theory and data compression. In a nutshell, it asserts that the best model is the one that best compresses both the data *and* that model. It does not only imply the best strategy for model selection, but also gives a unifying viewpoint of designing optimal data mining algorithms for a wide range of issues, and has been very successfully applied to a wide range of data mining tasks, ranging from pattern mining, clustering, classification, text mining, graph mining, anomaly detection, up to causal inference.

In this tutorial we give an introduction to the basics of model selection, show important properties of MDL-based modelling, successful examples as well as pitfalls for how to apply MDL to solve data mining problems, but also introduce advanced topics on important new concepts in modern MDL (e.g, normalized maximum likelihood (NML), sequential NML, decomposed NML, and MDL change statistics) and emerging applications in dynamic settings.

KEYWORDS

minimum description length, data mining, machine learning, information theory

ACM Reference Format:

Jilles Vreeken and Kenji Yamanishi. 2019. Modern MDL meets Data Mining Insights, Theory and Practice. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332284

1 TUTORIAL OUTLINE

Selecting a model for a given set of data is at the heart of what data analysts do, whether they are statisticians, machine learners or data miners. However, the philosopher Hume already pointed out that the "*Problem of Induction*" is unsolvable; there are infinitely many functions that touch any finite set of points. So, it is not surprising that there are many different principled approaches to guide the search for a good model. Well-known examples are Bayesian Statistics and Statistical Learning Theory.

In the last decade information theoretic methods for selecting the best model slowly but surely became popular in the data mining community, and have led to state-of-the-art solutions in areas as diverse as pattern based modelling, change detection, and causal inference. In this tutorial we will review the state-of-the-art in information-theoretic model selection based on the Minimum Description Length principle and its implication and applications in data mining.

The tutorial consists of four parts: (I) Introduction to MDL; (II) MDL in Action; (III) Modern MDL: Stochastic Complexity and Normalized Maximum Likelihood; (IV) Dynamic Model Selection and Change-Detection by MDL. In parts I and III we will introduce basic concepts and give insight in theory, whereas in parts II and IV we will show how these insights can used to solve open problems data mining and machine learning. Below we give the outline of our tutorial, including references to publications we will cover, the slides and full reference lists will be made available online.[1]

Part I. Introduction to MDL

- Model Selection and Occam's Razor [5]
- Two-Part MDL [20, 23]
- MDL, AIC, BIC, and Kolmogorov Complexity [7, 8, 27]
- Strengths and Weaknesses of 2-part MDL [1]
- Refined MDL [7, 23]

Part II. MDL in Action-Static Data

- Pattern mining and Pattern-based Modelling [2, 14, 28]
- Denoising, Clustering, Anomaly Detection [9, 24, 25]
- Regression and Causal Inference [4, 15]
- Independence Testing and Graphical Modelling [16, 18]
- Rank Estimation for NMF [17]
- Deep Learning [3]

Part III. Stochastic Complexity

- Normalized Maximum Likelihood (NML) [13, 21, 22]
- Theoretical basis for Consistency [7, 26]
- Estimation Optimality and Rate of Convergence [7, 26]
- Latent Variable Models [10, 29, 34]
- Luckiness and High-Dimensional Sparse Models [19]

Part IV. MDL in Action-Dynamic Settings

- Change Statistics [30, 33]

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332284

[1] http://eda.mmci.uni-saarland.de/mdldm/

- Dynamic Model Selection [6, 12, 32]
- Structural Entropy for Change Sign Detection [11]
- Failure Detection, and Emergent Market Detection [31]

2 TUTORS' BIOGRAPHIES

Jilles Vreeken is faculty at the CISPA Helmholtz Center on Information Security, where he leads the Exploratory Data Analysis group. He is particularly interested in developing well-founded theory and efficient methods for extracting informative models from large data. He defended his PhD thesis titled *Making Pattern Mining Useful* in 2009, and has authored 3 book chapters and over 75 conference and journal papers – 9 of which at KDD. He received three best paper awards, the ACM SIGKDD 2010 Doctoral Dissertation Runner-Up Award, and the IEEE ICDM 2018 Tao Li Early Career Award. He is member of the steering committee of ECML PKDD, while prior he was panel chair for SIAM SDM 2019, tutorial chair for SDM 2017, program co-chair for ECML PKDD 2016, and workshop co-chair of IEEE ICDM 2012. He co-organised nine workshops and co-lectured five tutorials.

Kenji Yamanishi is Full Professor in Computer Science at the University of Tokyo. His research interests include information-theoretic machine learning and data mining. He received his ME degree from the University of Tokyo in 1987, and his Dr. Eng degree from the same University in 1992. In 1987 he joined the NEC Corporation where he rose to department head of the Data Mining Technology Center and fellow in the Internet Systems Research Laboratories. He joined the Graduate School of Information Science and Technology of the University of Tokyo in 2009. He published over 85 publications on MDL-based learning theory, of which 14 at KDD, and is a regular (senior) member of the KDD program committee, and is one of authors of the book "*Advances in minimum description length: Theory and applications*" edited by Grünwald, Myung, and Pitt). A number of his MDL-based data mining tools on text mining and anomaly detection have been deployed for business use.

REFERENCES

[1] Pieter Adriaans and Paul Vitányi. 2009. Approximation of the Two-Part MDL Code. *IEEE Transactions on Information Theory* 55, 1 (2009), 444–457.
[2] Apratim Bhattacharyya and Jilles Vreeken. 2017. Efficiently Summarising Event Sequences with Rich Interleaving Patterns. In *Proceedings of the SIAM International Conference on Data Mining (SDM'17)*. SIAM.
[3] Léonard Blier and Yann Ollivier. 2018. The Description Length of Deep Learning models. In *Advances in Neural Information Processing Systems 31*. 2216–2226.
[4] Kailash Budhathoki and Jilles Vreeken. 2018. Origo: Causal Inference by Compression. *Knowledge and Information Systems* 56, 2 (2018), 285–307.
[5] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience New York.
[6] T. Erven, P. Grunwald, and S. Rooij. 2012. Catching up by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma. *Jr.Royal Stat.Soc.Ser.B* 74, Issue 3 (2012), 361–417.
[7] Peter Grunwald. 2007. *The Minimum Description Length Principle*. MIT Press.
[8] Peter Grünwald and Paul M. B. Vitányi. 2004. *Shannon information and Kolmogorov complexity*. Technical Report cs.IT/0410002. arXiv.
[9] S. Hirai and K. Yamanishi. 2012. Detecting changes of clustering structures using normalized maximum likelihood coding. In *Proceedings of 2012 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 343–351.
[10] S. Hirai and K. Yamanishi. 2013. Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering. *IEEE Transactions on Information Theory* 59, 11 (2013), 7718–7727.
[11] S. Hirai and K. Yamanishi. 2018. Detecting latent structure uncertainty with structural entropy. In *Proceedings of 2018 IEEE International Conference on BigData*. IEEE, 26–35.
[12] R. Kaneko, K. Miyaguchi, and K. Yamanishi. 2017. Detecting changes in streaminig data with information-theoretic windowing. In *Proceedings of 2017 IEEE International Conference on BigData*. IEEE, 646–655.
[13] P. Kontkanen and P. Myllymäki. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inform. Process. Lett.* 103, 6 (2007), 227–233.
[14] Danai Koutra, U Kang, Jilles Vreeken, and Christos Faloutsos. 2014. VoG: Summarizing Graphs using Rich Vocabularies. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*. SIAM, 91–99.
[15] Alexander Marx and Jilles Vreeken. [n.d.]. Telling Cause from Effect by MDL-based Local and Global Regression.
[16] Alexander Marx and Jilles Vreeken. 2019. Testing Conditional Independence on Discrete Data using Stochastic Complexity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.
[17] Pauli Miettinen and Jilles Vreeken. 2014. MDL4BMF: Minimum Description Length for Boolean Matrix Factorization. *ACM Transactions on Knowledge Discovery from Data* 8, 4 (2014), A18:1–31.
[18] K. Miyaguchi, S Matsushima, and K. Yamanishi. 2017. Sparse graphical modeling via stochastic complexity. In *Proceedings of 2017 International Conference on Data Mining*. SIAM, 723–731.
[19] K. Miyaguchi and K. Yamanishi. 2018. High-dimensional penalty selection via minimum description length principle. *Machine Learning* 107, 8-10 (2018), 1283–1302.
[20] Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica* 14, 1 (1978), 465–471.
[21] Jorma Rissanen. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics* 11, 2 (1983), 416–431.
[22] Jorma Rissanen. 1996. Fisher information and stochastic complexity. *IEEE Transactions on Information Technology* 42, 1 (1996), 40–47.
[23] Jorma Rissanen. 2012. *Optimal Estimation of Parameters*. Cambridge University Press. https://doi.org/10.1017/CBO9780511791635
[24] Arno Siebes and René Kersten. 2012. Smoothing Categorical Data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Bristol, UK*. Springer, 42–57.
[25] Koen Smets and Jilles Vreeken. 2011. The Odd One Out: Identifying and Characterising Anomalies. In *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*. Society for Industrial and Applied Mathematics (SIAM), 804–815.
[26] A. Suzuki and K. Yamanishi. 2018. Exact calculation of normalized maximum likelihood code length using Fourier analysis. In *Proceedings of 2018 IEEE International Symosium on Information Theory*. IEEE, 1211–1215.
[27] N.K. Vereshchagin and P.M.B. Vitanyi. 2004. Kolmogorov's Structure functions and model selection. *IEEE Transactions on Information Technology* 50, 12 (2004), 3265– 3290.
[28] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. 2011. Krimp: Mining Itemsets that Compress. *Data Mining and Knowledge Discovery* 23, 1 (2011), 169–214.
[29] Tyani Wu, Shinya Sugawara, and Kenji Yamanishi. 2017. Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1165–1174.
[30] K. Yamanishi and S. Fukushima. 2018. Model change detection with the MDL principle. *IEEE Transactions on Information Theory* 64, 9 (2018), 6115–6126.
[31] K. Yamanishi and Y. Maruyama. 2005. Dynamic syslog mining for network failure monitoring. In *Proceedings of 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 499–508.
[32] K. Yamanishi and Y. Maruyama. 2007. Dynamic model selection with its applications to novelty detection. *IEEE Transactions on Information Theory* 53, 6 (2007), 2180–2189.
[33] K. Yamanishi and K. Miyaguchi. 2016. Detecting gradual changes from data stream using MDL-change statistics. In *Proceedings of 2016 IEEE International Conference on BigData*. IEEE, 156–163.
[34] K. Yamanishi, W. Tianyi, S. Sugawara, and M. Okada. 2019. The decomposed normalized maximum likelihood codelength criterion for hierarchical latent variable models. *Data Mining and Knowledge Discovery* (2019).