MDL in the Limit

Pieter Adriaans and Wico Mulder

[1] Theory of Computer Science Group, IVI
University of Amsterdam, Science Park 107,
1098 XG Amsterdam, The Netherlands
[2] Logica, Prof. W.H. Keesomlaan, 1183 DJ Amstelveen, The Netherlands
P.W.Adriaans@uva.nl, wico.mulder@logica.com

Abstract. We show that within the Gold paradigm for language learning an informer for a superfinite set can cause an optimal MDL learner to make an infinite amount of mind changes. In this setting an optimal learner can make an infinite amount of wrong choices without approximating the right solution. This result helps us to understand the relation between MDL and identification in the limit in learning: MDL is an optimal model selection paradigm, identification in the limit defines recursion theoretical conditions for convergence of a learner.

1 Introduction

In a landmark paper Gold [1] introduced the idea of identification in the limit as a paradigm to study language learning. We start with a student and a teacher. At the beginning of the learning process they select a class of languages $\mathcal{L}$. The teacher consequently selects an element $L_i \in \mathcal{L}$ and starts to produce example sentences from L_i. After each example the student is allowed to update his guess for the language the teacher has selected. We expect the teacher to be an informer for the language i.e. each sentence from L_i will be produced in the limit. The class of languages $\mathcal{L}$ is considered to be *identifiable in the limit from positive information* if the student can for each language in $\mathcal{L}$ reach a stable guess in a finite amount of time on the basis of only positive examples. A well-known border case for this form of learning are the so-called superfinite sets. We give an example:

Definition 1. *Let* $L_\infty = \{a\}^*$ *and let* $L_k = \{a^j \mid 0 \leq j \leq k\}$. *We define the superfinite class of languages* $\mathcal{L} = \{L_\infty\} \cup \{L_k | k \in \mathbb{N}\}$

One can prove ([4], pg. 203) that for the set of finite languages $\{L_k | k \in \mathbb{N}\}$ the student can deploy a lazy learning strategy, i.e. take the longest sentence a^k the teacher has produced so far to be an indication of the intended language L_k. As soon as we add L_∞ to the set this does not work anymore. If L_∞ is the intended language this strategy will lead to an infinite chain of mind changes of the student. On the other hand there is no point in this process where the student can, with certainty, guess that L_∞ is the target language. We saw that superfinite sets are not identifiable in the limit from positive information.

J.M. Sempere and P. García (Eds.): ICGI 2010, LNAI 6339, pp. 258–261, 2010.

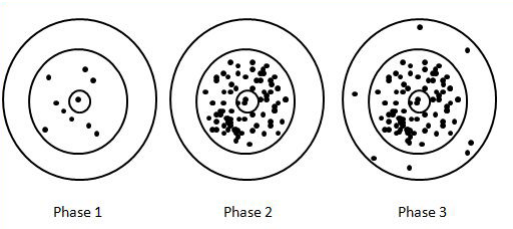

Fig. 1. The main idea behind the proof. Suppose the circles define some complexity borders in a sample space. Examples further removed from the center are more complex. In phase 1 the distribution is sparse and we prove that the best MDL model is infinite. In phase 2 the distribution is dense and the best MDL model is finite. In phase 3 the best model is infinite again. This can go on indefinitely.

A totally different paradigm for learning is the so-called Minimum Description Length (MDL) principle [2]: the best theory M to explain the data x is the one that minimizes: the sum of 1) the length of the description in bits of M (the model code) and 2) the length of the description in bits of x given M (the data-to-model code). We want to investigate how MDL performs in terms of identification in the limit. As a preliminary exercise we investigate the performance of MDL on super-finite sets. Since MDL is technically an optimal model selection strategy we do not expect MDL to settle for either L_∞ or a finite language L_k. In some cases it might make an infinite number of mind changes.

2 Outline of the Proof of the Main Theorem

The prefix-free Kolmogorov complexity of a binary string can (following [3]) be defined as: $K(x) = \min_{i,p}\{|\bar{i}| + |p| : T_i(p) = x\}$ where $i \in \{1, 2, ...\}$ and $p \in \{0, 1\}^*$. Here $|\bar{i}|$ is the length of a self-delimiting code of an index (see [3], pg. 79) and T is a universal Turing machine that runs program p after interpreting $|\bar{i}|$. The length of $|\bar{i}|$ is limited for practical purposes by $n + 2\log n + 1$, where $n = |i|$. Let the universal distribution be $m(x) = \sum_x 2^{-K(x)}$. Let $\mathcal{M}$ be the set of prefix-free programs. Using Bayes' law, the optimal computational model under this distribution would be: $M_{map}(x) = argmax_{M\in\mathcal{M}} \frac{m(M)m(x|M)}{m(x)}$ which can be rewritten as: $= argmin_{M\in\mathcal{M}} - \log m(M) - \log m(x|M)$ Here $-\log m(M)$ can be interpreted as the length of the optimal *data-code* in Shannon's sense and $-\log m(x|M)$ as the length of the optimal *data-to-model code*. Using Levin's coding theorem ([3], pg. 273) this can be rewritten as:

$$M_{map}(x) = argmin_{M\in\mathcal{M}} K(M) + K(x|M) \tag{1}$$

This gives optimal *two-part code compression* of x. We now give the central theorem:

Theorem 1. *An informer for a super-finite set can cause an optimal MDL learner to make an infinite amount of mind changes.*

Outline of proof.

- Suppose the language chosen by the informer is L_∞. We need to estimate the optimal model code for M and the optimal data to model for x given M. We can interpret the string a^k as the unary representation of the number k. Model and data set can now be interpreted as (sets of) natural numbers.
- Optimal Model Code: any finite model L_k can be coded as a natural number k, i.e. $K(L_k) = \log k + O(1)$. The code for the infinite model is of small constant length and is given by $L_\infty = \{a\}^*$, i.e. $K(L_\infty) = O(1)$.
- Optimal Data to Model Code: the data produced by the informer at time t can be effectively coded as a set of natural numbers. There are two optimal coding techniques:
  1. *A self-delimiting list of numbers for sparse sets.* Let D be the data set at time t coded in terms of natural numbers. Given the fact that we need not more than $2 \log \log n$ additional bits to make the number n self delimiting the optimal code length for D is limited by $K(D) \leq \Sigma_{i \in D} \log i + \log \log i$. Note that in this code the largest sentence produced by the observer is included in self-delimiting in the data-to-model code. The sparse coding scheme is therefore associated with the selection of L_∞ as optimal model, since this adds the least additional bits. The self-delimiting representation does not take in to account the mutual information between the elements of D. For sparse sets where the numbers have no mutual information this is optimal.
  2. *Subset coding for dense sets.* Here the data is encoded as elements of a subset using Newton's binomial formula: $K(D) = \log d + \log m + \log \binom{m}{d}$. Here $m = |M|$ and $d = |D|$. Since $\log k! \approx \int_1^k logxdx$ we can approximate $\log \binom{m}{d} \approx \int_d^m logxdx - \int_1^d logxdx$. Note that in this case we can interpret M as the model with $\log m$ as model code. The dense coding is associated with a finite language as optimal model. The subset coding is optimal if the numbers are so dense that there is a lot of mutual information.
- We now have two estimates for the MDL code:

$$K(L_\infty) + K(D|L_\infty) = O(1) + \Sigma_{i \in D} \log i + \log \log i$$

$$K(L_m) + K(D|L_m) = \log m + \log d + \int_d^m logxdx - \int_1^d logxdx + O(1)$$

- The potential oscillating behavior of an MDL learner is proved by the following observation: For large enough m there is always a set of natural numbers D such that:

$$K(L_\infty) + K(D|L_\infty) \approx K(L_m) + K(D|L_m)$$

The proof is as follows: suppose that the best MDL code for a certain D is $K(L_m) + K(D|L_m)$. This implies that the elements of D have a lot of mutual

information. Start adding large new elements a^k where $k >> m$ such that the numbers have low mutual information. Since there is an infinite amount of sentences this is always possible. Very soon (if you do it right after the first added element) the MDL code $K(L_\infty) + K(D|L_\infty)$ starts to be more efficient. Suppose on the other hand that $K(L_\infty) + K(D|L_\infty)$ is the best model. Now start to add new elements smaller then a^m to D, where a^m is the biggest sentence you have seen so far. After some point the mutual information will be so big that $K(L_m) + K(D|L_m)$ is a better model. The 'tipping point' is defined by a comparison between the two MDL estimates:

$$O(1) + \Sigma_{i \in D} \log i + \log \log i = \log m + \log d + \int_d^m logxdx - \int_1^d logxdx + O(1)$$

– This shows that at each point in time the informer has the power to steer an MDL learning process in the direction of L_∞ or in the direction of some L_k. Note that the optimal model in practice often will be a mixture of the two approaches: i.e. dense model for an initial segment, and a sparse model for the rest of the data, but this does not affect the main point of the proof: the optimal model for the largest sentences seen so far determines the final model estimate.

3 Discussion

MDL, as a criterium for optimal model selection, is in a way orthogonal to identification in the limit. Note also that in order to generate the infinite amount of mind shifts for the student the teacher also has to make an infinite amount of changes of strategy. As soon as the teacher uses a probability distribution over the whole of L_∞ the MDL learner will with high probability stabilize on this guess early in the process and never change its mind. The fact that this class of languages can be interpreted as consisting of sets of unary representations of natural makes it easy to calculate the MDL scores. It might be difficult to generalize these results to more complex languages.

References

[1] Gold, E.M.: Language Identification in the Limit. Information and Control 10(5), 447–474 (1967)
[2] Grünwald, P.D.: The Minimum Description Length Principle, 570 pages. MIT Press, Cambridge (2007)
[3] Li, M., Vitányi, P.M.B.: An Introduction to Kolmogorov Complexity and Its Applications, 3rd edn. Springer, New York (2008)
[4] Zeugmann, T., Lange, S.: A Guided Tour Across the Boundaries of Learning Recursive Languages. In: Lange, S., Jantke, K.P. (eds.) GOSLER 1994. LNCS (LNAI), vol. 961, pp. 190–258. Springer, Heidelberg (1995)