No-Free-Lunch and the Minimum Description Length

Nicholas Piël
Universiteit van Amsterdam

2007-11-15

1 Introduction

The No-Free-Lunch theorem (NFL) states that no learning algorithm exists for the complete domain of problems that will outperform any other algorithm. Or in other words, every learning algorithm will perform equally well when averaged on the complete problem domain [3]

The minimum description length (MDL) is a formalization of Occam's Razor in which the best hypothesis for a given set of data is the one that leads to the largest compression of the data. This will help us against overfitting as this compressed result is a tradeoff between the complexity of the hypothesis and the data given this hypothesis. [1]

In the following sections i will discuss how both views relate and differ and how they influence each other along 4 different arguments.

2 MDL and NFL operate on a different subset of the domain

The original argument to be discussed was: "No-Free-Lunch is only about UNSEEN data. MDL is about the domain as a whole." from this argument it feels as if the domain on which MDL operates in some way encompasses the NFL-domain. In reality it is more the other way around.

In his paper Wolpert mentions that when we want to do something interesting with our learning algorithm we have to test our hypothesis h on unseen data. For this argument he even quotes a different article [2]. However, the theorems for NFL strictly state that NFL applies only if we iterate over all possible target functions f and hypothesis h for all data points. From the NFL theorem we can deduct that we can only have a good generalization performance if the learned hypothesis on our observed data aligns with the target function on the data to be observed.

MDL in its two part code approach, minimizes the data given the hypothesis (this can be loosely interpreted as error count) and next to that makes sure that it will not overfit by also keeping the length of the hypothesis small.

From this we could say that *NFL operates on the complete domain of all possible hypothesises* and that *MDL tries to generalize its hypothesis build from the observed data* by limiting its complexity accordingly. Thus, the domain on which MDL operates is in essence a subset of the domain on which NFL operates and not the other way around.

However, they still operate on the complete field of hypothesises. You could compare it with the financial trading market. Both 'algorithms' operate on the same market, one group of apologists try to warn all investors with the common sense that if one investor wins something there has to be another one that looses something. Thus the NFL people live along the slogan: "Past Performance is No Guarantee of Future Results". The other group states that they do not really care about other investors as long as their investment is doing well, and yes there are still groups that make heaps of money. Their slogan is: "A trend is your friend", people who apply MDL for their learning algorithm belong to this latter group.

3 MDL picks the best hypothesis

While the statement in the previous section makes it sound like that MDL will fall in the same pit-falls as other learning methods, in the same way as the Maximum Likelihood Estimation puts too much faith in its own data and is therefore unable to handle unseen events. This is not really the case.

MDL is aware that there is other data out there, and that the hypothesis that just limits the error rate on the observed data may not perform well on unobserved data. Roughly said we could say that MDL first generates the Bayes optimal hypothesis in order to limit the error rate on the observed data, then applies Occam's razor as it believes that by doing so it will decrease the error on unobserved data[1]. And yes, this anti-overfitting measurement will increase the error on the observed data.

So if we are only interested in the error rate on the observed data and we know that new data will not differ in any way. We are better of applying the Bayesian method then we would with applying MDL.

4 Because of NFL all algorithmic research is futile

This argument can be countered rather intuitively. We are not living in a Kolmogorov random world, there are patterns we can use. And every action we as a human[2] take in someway exploit these recurring patterns. Think of the muscle contractions needed for movement or speech. We can specialize to our specific problem domain whether you are a researcher of machine learning or a professional soccer player. If we where God and our goal was to win the world championship 2008 of soccer we would not put 'van Someren' as one of the attackers. On the other hand if we wanted to publish many papers regarding machine learning we would not ask 'van Nistelrooy' to write them up. Now, if we are not God but a mere researcher of machine learning trying to sell products we would still need to choose the algorithm that exploits domain specific knowledge.

5 MDL works because the problems we observe in the real world are more likely to be simple

MDL is all based around Occam's razor. And we all know that when playing with sharp things it can get bloody very easily. There isn't any real proof behind Occam's razor and if we look at nature as our prime example we quickly start wondering why we humans have an appendix or two kidneys when one is enough. Why didn't nature apply the razor here? Well it is exactly because of the complexity of Nature that we need to apply the razor if we do not want to get lost in complexity. Remember that the main reason behind MDL is to counter over-fitting. In other words we have observed the data, we could establish an hypothesis that has exactly zero error on this dataset. But MDL is smarter than this, and knows that nature is bound to have some surprises up her sleeve and counters natures complex problems with simple solutions. These simple solutions are forced to generalize over the observed data and are therefore more robust against unseen data that is different or more complex.

References

[1] Peter Grünwald. A tutorial introduction to the minimum description principle. 2005.

[2] S.M. Weiss and C.A. Kulikowski. *Computer Systems that Learn*. Morgan Kauffman, 1991.

[3] D.H. WOLPERT. The supervised learning no-free-lunch theorems. *Online World Conference of Soft-computing*, 6, 2001.

[1] The real machinery behind the MDL algorithm differs from this example, but i think it perfectly shows the difference between the MDL and Bayesian approach

[2] This NFL arguments applies to us humans as well if you are of the opinion that a human is in essence nothing more then complex but completely deterministic algorithm