Mach Learn
DOI 10.1007/s10994-010-5227-2

TECHNICAL NOTE

An experimental test of Occam's razor in classification

Jan Zahálka · Filip Železný

Received: 31 May 2010 / Revised: 5 November 2010 / Accepted: 8 November 2010

Abstract A widely persisting interpretation of Occam's razor is that given two classifiers with the same training error, the simpler classifier is more likely to generalize better. Within a long-lasting debate in the machine learning community over Occam's razor, Domingos (Data Min. Knowl. Discov. 3:409–425, 1999) rejects this interpretation and proposes that model complexity is only a confounding factor usually correlated with the number of models from which the learner selects. It is thus hypothesized that the risk of overfitting (poor generalization) follows only from the number of model tests rather than the complexity of the selected model. We test this hypothesis on 30 UCI data sets using polynomial classification models. The results confirm Domingos' hypothesis on the 0.05 significance level and thus refutes the above interpretation of Occam's razor. Our experiments however also illustrate that decoupling the two factors (model complexity and number of model tests) is problematic.

Keywords Model complexity · Generalization · Empirical evaluation

1 Introduction

In philosophy of science, the thesis "Entities should not be multiplied beyond necessity" articulated by the 14th century logician William of Occam and known as the *Occam's razor* states that the simplest theory should be chosen from among all theories which equally well explain observed data (Nolan 1997). While Occam's razor is obviously relevant to machine learning (Gamberger and Lavrač 1997), the exact way it translates to this context has been subject to a long-lasting debate, pertaining mainly to the usefulness, provability, and

Editor: Johannes Fürnkranz.

J. Zahálka · F. Železný (✉)
Faculty of Electrical Engineering, Czech Technical University in Prague, Praha, Czech Republic
e-mail: zelezny@fel.cvut.cz

J. Zahálka
e-mail: zahalja1@fel.cvut.cz

Published online: 18 November 2010

empirical validity of various interpretations of the razor. A thorough review of theoretical and empirical aspects of Occam's razor in knowledge discovery was provided in by Domingos (1999). Subsequent discussions appeared in various venues, including The NIPS 2001 workshop "Foundations of Occam's razor and parsimony in learning".[1]

Domingos (1999) distinguishes two interpretations of Occam's razor. One of them, which we refer to as the *strong interpretation*, states that "given two models with the same training-set error, the simpler one should be preferred because it is likely to have lower generalization error." Empirical evidence against the strong interpretation appeared in the early work of Murphy and Pazzani (1994) which found some learned simple decision trees to generalize worse than slightly more complex trees. Further concurring evidence was then provided by Webb (1996). On the other hand, Esmeir and Markovitch (2007) determined generalization performance to correlate with model simplicity, and later Needham and Dowe (2001) argued that the findings of Murphy and Pazzani (1994) were due to the adoption of an inappropriate model complexity measure.

In agreement with a parallel study of Jensen and Cohen (2000), Domingos (1999) offers a clarification of the contradictory arguments by accounting for another aspect, the number of models tested during learning. In fact, an early study by Quinlan and Cameron-Jones (1995) already reported on the adverse effect of overly extensive model search on generalization. Domingos (1999) rejects the strong interpretation of Occam's razor by pointing out that overfitting is indeed due to multiple model testing rather than complexity per se. A confusion between the two factors arises from the fact that a learning algorithm usually conducts a greater amount of testing to fit a more complex model. This explanation represents an experimentally testable hypothesis. To our surprise, through literature search we did not find an attempt to validate it.

As convincing as the hypothesis appears, its correctness is not obvious. Strictly speaking, Domingos' explanation does not logically refute the strong interpretation of Occam's razor. Indeed, the fact that model complexity is confounded by another factor leading to overfitting does not mean that complexity itself would not entail poor generalization as well. Better generalization of simpler models can be argued also on basis completely unrelated to the amount of testing in learning. For example, Piatetsky-Shapiro (1996) believes the strong interpretation to be mostly true on real-life data due to human involvement in the creation of the data. To resolve the standing question, it is thus important to conduct experiments on a large collection of real-life data sets. This is what we contribute through the present study.

In particular, our experiments test the following assertions.

1. Models selected from a larger set of tested candidate models overfit more than those selected from a smaller set (assuming constant model complexity).
2. More complex models overfit more than simpler models (assuming a constant number of candidate models tested).

According to Domingos' hypothesis, the first assertion should be true and the second should be false.

The rest of the paper is organized as follows. In the next section we describe the data, algorithms and protocol of our experiments. Section 3 reveals the results. In Sect. 4 we conclude the paper.

[1] http://rii.ricoh.com/~stork/OccamWorkshop.html.

2 Experimental material

2.1 Data

We downloaded 30 classification learning benchmark datasets from the UCI Machine Learning repository.[2] Table 1 shows the name, the number of attributes, types of attributes involved, and the number of instances for each of these data sets.

Due to the requirements of the chosen classification strategy (binary polynomial separation in R^n, described below), a few pre-processing steps were conducted for all the data sets.

Table 1 Datasets used in the experiment. Attribute types are C (categorical), I (integer), R (real). (*) The original "Poker Hand" dataset contains more than a million entries, and was trimmed to 1000 entries for computational feasibility

#	Name	# Attributes	Att. Types	# Instances
1	*Abalone*	8	C, I, R	4177
2	*Acute Inflammations*	6	C, I	120
3	*Adult*	14	C, I	48842
4	*Balance Scale*	4	C	625
5	*Balloons*	4	C	16
6	*Blood Transfusion*	5	R	748
7	*Car Evaluation*	6	C	1728
8	*Congressional Voting*	16	C	435
9	*Contraceptive Method*	9	C, I	1473
10	*Echocardiogram*	12	C, I, R	132
11	*Glass Identification*	10	R	214
12	*Haberman's Survival*	3	I	306
13	*Hepatitis*	19	C, I, R	155
14	*Horse Colic*	27	C, I, R	368
15	*Iris*	4	R	150
16	*Lenses*	4	C	24
17	*Letter Recognition*	16	I	20000
18	*MAGIC Gamma*	11	R	19020
19	*Mammographic Mass*	6	I	961
20	*MONK's Problems*	7	C	432
21	*Mushroom*	22	C	8124
22	*Pima Indians Diabetes*	8	I, R	768
23	*Poker Hand*	11	C, I	1000(∗)
24	*Post-Operative Patient*	8	C, I	90
25	*Statlog (Credit)*	15	C, I, R	690
26	*Statlog (Heart)*	13	C, R	270
27	*Statlog (Shuttle)*	9	I	58000
28	*Vehicle Silhouettes*	18	I	946
29	*Wine*	13	I, R	178
30	*Zoo*	17	C, I	101

[2] http://archive.ics.uci.edu/ml/datasets.html, accessed 16.4.2010.

First, we removed all instances containing at least one missing value. Data sets involving more than two classes were binarized by merging semantically similar classes. Categorical attributes were expanded into binary attribute tuples according to the 1-in-n scheme (Hastie et al. 2001). Finally, we normalized all attributes by a linear projection from the original attribute's range to the [0; 1] interval.

2.2 Learning algorithms

For class separation, we adhered to the elementary method of polynomial boundaries in the Euclidean vector space R^n of data attributes. An advantage of this approach is that the degree d of the separating polynomial is a hardly-disputable measure of its complexity, unlike in other model types where complexity indicators are trickier (Grünwald 2001).

In particular, the classification boundary of degree d for n attributes has the form

$$\sum_{i=0}^{d} \sum_{j=1}^{n} c_{i,j} A_j^i \tag{1}$$

where $c_{i,j}$ are learned real coefficients ($\sum_{j=1}^{n} c_{0,j}$ representing the offset constant) and A_j are real variables corresponding to attributes. For computational feasibility we do not include cross-products (such as $A_1^2 A_2$) in the polynomial; even without the cross-products, growing d entails growing complexity of the boundary both in terms of irregularity and in terms of the number of parameters.

We considered two ways to learn the parameters $c_{i,j}$.

Selection from random models We first considered a simple method where m models are generated randomly (coefficients $c_{i,j}$ are assigned values sampled from the uniform distribution on the integer interval $[-100; 100]$), and the model with the least error on training data is selected from these m models. This method is ideal from the viewpoint of our experiment since it does not introduce dependence between the two inspected factors, complexity d and number of tested models m.

Gradient descent learning While the method above is experimentally transparent, it does not really remind of learning. We therefore also considered an alternative method where m models are first generated randomly just as above, but their coefficients are subsequently tuned through an algorithm performing a gradient descent on an error function parameterized by the coefficients $c_{i,j}$. Following the basis-expansion method (Hastie et al. 2001), the algorithm learns a linear boundary in the expanded space spanned by dimensions A_j^i ($1 \leq i \leq d$, $1 \leq j \leq n$), which in turn corresponds to a polynomial boundary in the original space spanned by dimensions A_j ($1 \leq j \leq n$). Out of the m models tuned this way, the model possessing the smallest training error is selected.

This approach requires extra prudence since it may introduce statistical dependence between the number of model tests on one hand, and complexity on the other hand. Clearly, each iteration of the learning algorithm represents a model test. For growing degree d of the polynomial, it will usually take more iterations of the learning algorithm to converge. It would not be correct to express the number of model tests as $\sum_{k=1}^{a} i_k$ where i_k is $1+$ the number of iterations on the k'th model. Indeed, e.g. the two cases ($\alpha \in N$)

- $a = \alpha$, $i_k = 1$ $(1 \leq k \leq a)$
- $a = 1$, $i_1 = \alpha$

would yield the same $\sum_{k=1}^{a} i_k = \alpha$ but the former case corresponds to α mutually independent model tests whereas the latter case counts α strongly dependent model tests. We therefore rather decided to modify the stopping criterion of the gradient-descent algorithm so that learning ends after a fixed number of iterations that is independent of d and is rather equal to the number of training data instances. The learning rate, that is, the magnitude of the adjustment of coefficients at each iteration, was kept constant (0.1) for all iterations and all experiments.

2.3 Measurements and reproducibility

We performed a 5-fold cross-validation for each combination of the following variables

- data set: one of the 30 described in Sect. 2.1
- learning algorithm: one of the two described in Sect. 2.2
- $m \in \{10, 20, 30, 40, 50\}$ (number of model candidates)
- $d \in \{1, 2, 3, 4, 5\}$ (polynomial boundary degree, i.e. complexity)

thus amounting to 7500 sessions of learning and testing. For each combination of the above 4 factors we averaged the 5 cross-validation results, obtaining 1500 measurement tuples consisting of the values of the above variables, the averaged training error $\hat{E}$, testing error E, and the *overfitting quantifier* defined as $E - \hat{E}$.

The program codes and consolidated data sets needed to reproduce the experiments along with the full tabulation of results and the R-system scripts for their analysis are available at http://ida.felk.cvut.cz/occam.zip.

3 Results

The resulting measurements were analyzed separately for the two different learning methods. Our approach was to compare the overfitting variable $(E - \hat{E})$ of observations corresponding to extremal values of the two inspected factors.[3] Thus we compare all observations where $m = 10$ against those where $m = 50$ to test the null hypothesis that overfitting has the same median for both values of the factor against the alternative hypothesis that overfitting is stronger for $m = 50$. Similarly, we compare all $d = 1$ observations against all $d = 5$ observations to test the second null hypothesis that overfitting has the same median for both values of the factor against the second alternative hypothesis that overfitting is stronger for $d = 5$.

To obtain the p-values of the null hypotheses, we used the one-tailed paired Wilcoxon test. For the first hypothesis we always paired two observations having the respective extremal values of m and identical values of the other two factors (dataset and complexity d). We proceeded similarly for the second hypothesis. The resulting p-values are shown in Table 2.

For both learning approaches, the results reject the first null hypothesis on the 0.05 significance level, i.e. overfitting is stronger when models are selected from 50 candidate models rather than 10. On the other hand, for both learning approaches, the second null hypotheses

[3] Therefore, measurements in which simultaneously $10 < m < 50$ and $1 < d < 5$ are not used in the analysis. We leave them in the downloadable supplemental material so that extended statistical tests (such as regression based tests) can be conducted in possible follow-up work.

Table 2 P-values corresponding to the null hypotheses that the overfitting variable $(E - \hat{E})$ has the same median for both extremal values of m (first row), and that it has the same median for both extremal values of d (second row). Analyzed separately for each of the two learning algorithms (columns). P-values smaller than 0.05 are shown in bold face

p-values (Wilcoxon 1-tailed, paired)	Select/Rand	Grad/Desc
No. of Candidate Models m (10 vs. 50)	**0.0346**	**0.0004256**
Selected Model Complexity d (1 vs. 5)	0.5505	0.8406

Table 3 Means and standard deviations (in parentheses) of the overfitting variable $(E - \hat{E})$ over all observations corresponding to the extremal values of m (number of model tests) and d (model complexity). Shown separately for the two learning algorithms. For readability, actual values have been multiplied by 100 and then rounded to two decimal places

Means and std. deviations	Select/Rand		Grad/Desc	
No. of Candidate Models m	10	50	10	50
Overfitting $(E - \hat{E}) \cdot 100$	28.95 (8.46)	44.24 (11.48)	38.03 (9.84)	49.29 (10.78)
Model Complexity d	1	5	1	5
Overfitting $(E - \hat{E}) \cdot 100$	33.20 (10.27)	31.26 (8.55)	45.29 (9.69)	46.81 (10.01)

must be maintained that overfitting has the same median whether the model complexity (i.e., degree of the polynomial class boundary) is 1 or 5.

For completeness, Table 3 shows the means and standard deviations of the overfitting variable $(E - \hat{E})$ corresponding to all the inspected cases.

4 Conclusions

The experiments conducted with 30 UCI datasets lead us to the conclusion that overfitting indeed follows from the number of models tested in the course of learning, and not from the complexity of the selected model. This observation confirms the hypothesis worded by Domingos (1999), which has so far been left untested.

Our paper of course inherits all limitations of empirical studies. Strictly speaking, the conclusions are constrained to the particular choice of the experimental datasets, learning algorithms and the ranges of parameters. Nevertheless, the question whether model complexity entails overfitting in real-life datasets is naturally empirical and could hardly be resolved in a deductive manner.

The present results are probably not immediately useful for the developers of deterministic learning algorithms where complex models typically require more model testing, thus intermingling the two factors which we isolated in this study. They are however highly significant to less orthodox, randomized learning approaches such as those based on stochastic local search (Rückert and Kramer 2003; Železný et al. 2006; Paes et al. 2007). Indeed, one interpretation of our results is that starting model refinement in a randomly chosen point of the model space perhaps corresponding to a complex model does not systematically incur more overfitting than starting the refinement in the simplest model.

We would also like this study to serve as supporting material helping to clearly understand the overfitting phenomenon in the research and instruction of machine learning.

Acknowledgements This work was supported by the Czech Science Foundation through project P103/10/1875 "Learning from Theories".

References

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*, 409–425.

Esmeir, S., & Markovitch, S. (2007). Occam's razor just got sharper. In M. Veloso (Ed.), *IJCAI'07: proceedings of the 20th international joint conference on artificial intelligence* (pp. 768–773). San Mateo: Morgan Kaufmann.

Gamberger, D., & Lavrač, N. (1997). Conditions for Occam's razor applicability and noise elimination. In M. van Someren & G. Widmer (Eds.), *ECML'97: proceedings of the 9th European conference on machine learning* (pp. 108–123). Berlin: Springer.

Grünwald, P. (2001). Occam, Bayes, MDL and the real world (presentation slides). In *NIPS 2001 workshop on Occam's razor*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Berlin: Springer.

Jensen, D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*, 308–338.

Murphy, P. M., & Pazzani, M. J. (1994). Exploring the decision forest: an empirical investigation of Occam's razor in decision tree induction. *Journal of Artificial Intelligence Research*, *1*(1), 257–275.

Needham, S. L., & Dowe, D. L. (2001). Message length as an effective Ockham's razor in decision tree induction. In T. Richardson & T. Jaakkola (Eds.), *Proceedings of the 8th international workshop on artificial intelligence and statistics* (pp. 253–260).

Nolan, D. (1997). Quantitative parsimony. *The British Journal for the Philosophy of Science*, *48*(2), 329–343.

Paes, A., Železný, F., Zaverucha, G., Page, D., & Srinivasan, A. (2007). ILP through propositionalization and stochastic k-term DNF learning. In S. Muggleton, R. Otero, & A. Tamaddoni-Nezhad (Eds.), *ILP'06: proceedings of the 16th international conference on inductive logic programming* (pp. 379–393).

Piatetsky-Shapiro, G. (1996). Editorial comments. *KDD Nuggets*, *96*, 28.

Quinlan, J. R., & Cameron-Jones, R. (1995). Oversearching and layered search in empirical learning. In L. P. Kaelbling & A. Saffiotti (Eds.), *IJCAI'95: proceedings of the 14th international joint conference on artificial intelligence* (pp. 1019–1024). San Mateo: Morgan Kaufmann.

Rückert, U., & Kramer, S. (2003). Stochastic local search in k-term DNF learning. In T. Fawcett & N. Mishra (Eds.), *ICML 2003: proceedings of the 20th international conference on machine learning* (pp. 648–655). New York: AAAI Press.

Železný, F., Srinivasan, A., & Page, D. (2006). Randomised restarted search in ILP. *Machine Learning*, *64*(1–3), 183–208.

Webb, G. I. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, *4*, 397–417.