A New MDL Measure for Robust Rule Induction (Extended Abstract)

Bernhard Pfahringer *

Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
E-mail: bernhard@ai.univie.ac.at

Abstract. We present a generalization of a particular Minimum Description Length (MDL) measure that so far has been used for pruning decision trees only. The generalized measure is applicable to (propositional) rule sets directly. Furthermore the new measure also does not suffer from problems reported for various MDL measures in the ML literature. The new measure is information-theoretically plausible and yet still simple and therefore efficiently computable. It is incorporated in a propositional FOIL-like learner called KNOPF.

1 Introduction

The *Minimum Description Length (MDL) Principle* [Rissanen 78], sometimes also called the *Minimum Message Length (MML) Principle*, has been successfully applied in Machine Learning, both for inducing decision trees [Quinlan 93, Forsyth 93], for constructing new attributes [Pfahringer 94a], and in ILP [Muggleton et al. 92]. But recently also some problems with MDL were discovered [Quinlan 94]. Section 2 will describe these problems. A new MDL measure applicable to propositional rule learning aimed at overcoming these problems will be introduced in section 3. Section 4 very briefly discusses a propositional FOIL-like learning algorithm using this new MDL measure as both a stopping criterion for rule induction and as a criterion to choose between different rule sets, especially to choose between sets of pruned rules and sets of unpruned rules. Section 5 lists open problems and further research directions.[2]

2 MDL in Rule Learning and its Problems

Empirical induction is always faced with the problem of *overfitting* the data, especially in the presence of noise or irrelevant attributes. The MDL principle is a possible solution as it measures both the simplicity and the accuracy of a particular rule set in a common currency, namely in terms of the number of

* Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry of Science and Research. I would like to thank Gerhard Widmer and Johannes Fürnkranz for help with this paper.

[2] For a long version of this paper including empirical results see [Pfahringer 94b].

bits needed for encoding. The precise MDL formula used by [Quinlan 93] for simplifying rule sets is:

$$Cost = TheoryCost + log_2\left(\binom{C}{FP}\right) + log_2\left(\binom{NC}{FN}\right)$$

In this formula *TheoryCost* is an estimate for the number of bits needed to encode the theory. C is the total number of training examples covered by the theory, FP is the number of false-positive examples, NC is the total number of training examples not covered by the theory, and FN is the number of false-negative examples.

So what are the problems with that formula? [Quinlan 94] states two and we would like to add an additional one: (1) the formula for computing exception cost is symmetric; (2) if the class to be learned is significantly in the majority (minority), induced theories tend to under-generalize (over-generalize), especially in the presence of noise and with small numbers of learning examples; (3) when learning from lots of examples in the presence of noise, there is still a tendency to fit the noise.

All these problems with the above formula stem from the fact that this formula is just an approximation of the generic MDL principle as defined above. It does not estimate encoding cost of *all* examples with respect to a given theory but instead computes a kind of penalty for wrong classifications only! So the remedy would be to look for a formula that is more faithful to the MDL principle.

3 An alternative MDL formula

[Forsyth 93] introduces a well-performing formula for encoding decision trees:

$$\begin{aligned} cost(tree) &= \Sigma cost(leaf_i) \\ cost(leaf_i) &= d_i + e_i * n_i \end{aligned}$$

where d_i is the depth of the leaf in the tree, n_i is the number of examples covered by the leaf, and e_i is average entropy of the outcome at that leaf defined by:

$$e_i = -(p * log(p) + (1 - p) * log(1 - p))$$

where p is the proportion of positive examples covered by $leaf_i$.[3] Note that $e_i * n_i$ is the number of bits needed by an optimal or 'Huffman' coding of the classifications at $leaf_i$ in terms of the relative frequencies of positive and negative examples at $leaf_i$.

We have modified this formula for coding sets of propositional rules. The essential differences are a cost estimate for examples not covered by the rule set and an information-theoretically plausible encoding cost for the rules themselves. Note that the ordering of rules is significant in this encoding, meaning that an example is covered by the first of all the rules matching it. We define the cost of a rule set as follows:

[3] $0 * log(0)$ is defined to equal 0.

$$cost(ruleset) = n_{nc} * e_{nc} + \Sigma cost(rule_i)$$
$$cost(rule_i) = rc_i + e_i * n_i$$

where n_{nc} is the total number of examples not covered by the rule set and e_{nc} is the according entropy of this set. The complexity of a single rule is accounted for by rc_i. This is an estimate of the coding cost for the body of the rule. Assuming a total number N_{pt} of tests that could possibly be used by a rule and adopting Quinlan's idea for encoding exceptions we can define the cost for encoding the body of a rule as follows. The cost for choosing $Length_i$ tests out of N_{pt} possible can be estimated as:

$$rc_i = log_2 \left(\binom{N_{pt}}{Length_i} \right)$$

The new estimate certainly solves problem 3 as coding cost for *all* examples is estimated (remember that the entropy of a rule e_i is multiplied by the total number of examples covered by that rule n_i as part of the cost of a rule). Problem 2 is only partially solved as errors are penalized in a totally different way. We do not get consistently over- or under-generalizing behavior, but with too small training sets the empty theory can result from induction. But this is a consequence of using MDL itself: enough positive data has to support a rule, otherwise the intrinsic cost of the rule will outweigh the classification advantage gained by this rule. Regarding the so-called problem 1, the new formula is even more symmetric in the sense that in principle positive and negative rules could be freely mixed in an induced theory. For practical reasons one would have to add one more bit per rule for encoding the decision part (positive or negative) of each rule, if one wanted to take advantage of that property.

To summarize, the new formula measures cost for encoding all the training examples in terms of the theory (the single rules), classification errors are accounted for at a per-rule basis using local entropies, and complexity of rules is estimated in an information-theoretically plausible way. Furthermore this formula still is symmetric with respect to *negative* theories.

4 Algorithmic Usage of the new Formula

For empirical testing of the new formula we have implemented a kind of propositional FOIL [Quinlan & Cameron-Jones 93] called KNOPF. Right now KNOPF is restricted to purely symbolic 2-class learning problems. It is completely free of user-settable parameters. The MDL principle is used in two ways: firstly as a stopping criterion when inducing a single rule set and secondly for choosing the final rule set out of a number of induced rule sets.

The first pruning strategy is *correctness preserving*: the pruned rule will not cover more negative examples than the unpruned rule. The second strategy just maximizes the difference $p - n$ of positive and negative examples covered by the rule. MDL also implicitly judges the presence of noise: When a set of correctness preservingly pruned rules is chosen as the final result, we can assume noise-free data and vice versa.

5 Conclusions, Related Work, and Further Research

We have defined a new MDL measure for rule sets and incorporated it into the inductive learner KNOPF. This new measure is information-theoretically plausible in the way it encodes the theory and the examples and it also gives good experimental results. But there are still a lot of open questions and opportunities for improvement, e.g. finding better coding schemas, improving search, generalize the formula, so that it will be applicable to numbers and variables, and take into account new attributes produced by constructive induction.

In summary, the new MDL measure proposed in this paper is a generalization of the formula given in [Forsyth 93] applicable to sets of rules, it overcomes the deficiences of the formula used in C4.5, and it is simpler (and may also be more reliable for small training sets) than the coding scheme used by [Muggleton et al. 92].

References

[Forsyth 93] Forsyth R.S.: Overfitting Revisited: An Information-Theoretic Approach to Simplifying Discrimination Trees, in JETAI 6(3), 1994.

[Muggleton et al. 92] Muggleton S., Srinivasan A., Bain M.: Compression, Significance, and Accuracy, in Sleeman D. and Edwards P.(eds.), Machine Learning: Proceedings of the Ninth International Workshop (ML92), Morgan Kaufmann, San Mateo, CA, pp.338-347, 1992.

[Pfahringer 94a] Pfahringer B.: CiPF 2.0: A Robust Constructive Induction System, Proceedings of the Workshop on Constructive Induction and Change of Representation, 11th International Conference on Machine Learning (ML-94/COLT-94), New Brunswick, New Jersey., 1994.

[Pfahringer 94b] Pfahringer B.: A New MDL Measure for Robust Rule Induction, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, TR-94-29, 1994. (also available electronically as ftp://ftp.ai.univie.ac.at/papers/oefai-tr-94-29.ps.Z)

[Quinlan & Cameron-Jones 93] Quinlan J.R., Cameron-Jones R.M.: FOIL: A Midterm Report, in Brazdil P.B.(ed.), Machine Learning: ECML-93, Springer, Berlin, pp.3-20, 1993.

[Quinlan 93] Quinlan J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[Quinlan 94] Quinlan J.R.: The Minimum Description Length Principle and Categorical Theories, in Cohen W.W. and Hirsh H.(eds.), Machine Learning: Proceedings of the Eleventh International Conference (ML94), Morgan Kaufmann, San Mateo, CA, 1994.

[Rissanen 78] Rissanen J.: Modeling by Shortest Data Description, in Automatica, 14:465-471, 1978.