Boosting Theory Towards Practice: Recent Developments in Decision Tree Induction and the Weak Learning Framework

Michael Kearns
AT&T Research
mkearns@research.att.com

Difficulties in Comparing Machine Learning Heuristics

One of the original goals of computational learning theory was that of formulating models that permit meaningful comparisons between the different machine learning heuristics that are used in practice [Kearns *et al.*, 1987]. Despite the other successes of computational learning theory, this goal has proven elusive. Empirically successful machine learning algorithms such as **C4.5** and the backpropagation algorithm for neural networks have not met the criteria of the well-known Probably Approximately Correct (PAC) model [Valiant, 1984] and its variants, and thus such models are of little use in drawing distinctions among the heuristics used in applications. Conversely, the algorithms suggested by computational learning theory are usually too limited in various ways to find wide application.

The Theoretical Status of Decision Tree Learning

As an illustration, let us review what has been discovered about decision tree learning algorithms in the computational learning theory literature. Consider the simple framework in which a learning algorithm receives random examples, uniformly drawn from the hypercube $\{0,1\}^n$, that are assigned binary labels according to some decision tree T that has at most s nodes. A natural goal would be to find an algorithm that can infer a good approximation to T in time and sample complexity that is bounded by a polynomial in n and s. [1]

The existence of such an algorithm remains an apparently challenging open problem, so even with the various favorable and unrealistic assumptions (uniform input distribution, no noise or missing attributes in the data, the existence of a small "target" tree, and so on), computational learning theory has so far not provided vast advances in algorithm design for decision tree induction from random examples. On the other hand, in the framework under consideration, the heuristics for decision tree learning that are in wide experimental use do not fare much better. It is rather easy to show that **CART** and **C4.5** will fail to meet the stated criteria, and for the usual reasons: if the target decision tree computes the parity of just two out of the n variables, top-down heuristics like **CART** and **C4.5** may simply build a complete binary tree of depth n before achieving non-trivial error. Of course, this particular construction does not rule out the possibility that slight *modifications* of the standard heuristics might succeed — but a recent result [Blum *et al.*, 1994] demonstrated that small decision trees can not be learned by *any* algorithm that works solely by "estimating conditional probabilities" [Kearns, 1993]. The precise definition of this notion is slightly technical, but suffice to say that **CART** and **C4.5** — which operate primarily by estimating the probabilities of reaching certain nodes in a decision tree, or the conditional distribution of the label given that a node is reached — are canonical examples of the notion. Thus, although computational learning theory has yet to suggest powerful algorithms for decision tree learning from random examples, we can assert that if such algorithms exist, they will look nothing like the standard heuristics. Perhaps the more likely outcome is that the problem is simply intractable. This would mean that the assumption that a small decision tree is labeling the data is not especially helpful when examining decision tree learning algorithms, and we must seek alternative assumptions if we wish to account for the empirical success of **CART** and **C4.5**.

Provably efficient algorithms become available if we are willing to assume that the learning algorithm is provided with black-box access to the unknown target decision tree (that is, *membership queries*, which let the learner actively choose the instances to be labeled). A number of rather simple and elegant learning algorithms have recently been proposed in this setting [Bshouty, 1993; Kushilevitz and Mansour, 1991] that will infer the unknown tree in polynomial time,

[1] Here we are in the PAC model, where there is no noise in the sample data, with the additional restriction that the input distribution is uniform.

in strong contrast to the case where only random examples are available. However, because of the requirement for a source of information rarely available in real applications, these algorithms seem unlikely to replace the top-down heuristics, and their analysis sheds no light on why such heuristics succeed.

Viewing Top-Down Decision Tree Heuristics as Boosting Algorithms

The preceding summary indicates that some of the models of computational learning theory are unable to provide nontrivial insights into the behavior of **CART** and **C4.5**. One might be tempted to attribute this state of affairs to an inevitable chasm between theory and practice — that is, to claim that the standard heuristics succeed in practice due to some favorable structure possessed by real problems that simply cannot be captured by theory as we currently know it. Fortunately, some recent developments seem to demonstrate that such a defeatist position is not necessary.

The *weak learning* or *boosting* model is a descendant of the PAC model in which, rather than directly assuming that the target function can be represented in a particular fashion, we instead assume that there is always a "simple" function that is at least weakly correlated with the target function. We refer the reader to the literature for the precise technical definition, but for our informal purposes here, it suffices to assume that on any input distribution, there is an attribute whose value is correlated with the label.

In this setting, nontrivial performance bounds have recently been proven for both **CART** and **C4.5** [Kearns and Mansour, 1996]. More precisely, if we assume that there is always an attribute whose value correctly predicts the binary label with probability $1/2 + \gamma$ (thus, the attribute provides an advantage γ over random guessing), then for **CART** it suffices to grow a tree of size

$$\left(\frac{1}{\epsilon}\right)^{c/(\gamma^2\epsilon^2\log(1/\epsilon))} \tag{1}$$

in order to achieve error less than ϵ (where $c > 0$ is a constant), and for **C4.5**, a tree of size

$$\left(\frac{1}{\epsilon}\right)^{c\log(1/\epsilon)/\gamma^2} \tag{2}$$

suffices (see [Kearns and Mansour, 1996] for detailed statements and proofs). These bounds imply, among other things, that if we assume that the advantage γ is a fixed constant, then both algorithms will drive the error below any fixed ϵ in a constant number of splits. Until the result of [Schapire, 1990], the existence of *any* algorithm — much less a standard heuristic — possessing this "boosting" behavior was not known. The results given by Equations (1) and (2) provide nontrivial peformance guarantees for **CART** and **C4.5** in an independently motivated theoretical model.

A Framework for Comparisons

The theoretical results for **CART** and **C4.5** in the weak learning model do more than simply reassure us that these empirically successful algorithms can in fact be *proven* successful in a reasonable model. As one might have hoped, these results also provide a technical language in which one can attempt to make detailed comparisons between algorithms. Developing this language further has been the focus of our recent experimental efforts [Dietterich *et al.*, 1996], which we now summarize.

First of all, notice that the bounds of Equations (1) and (2) predict that the performance of **C4.5** should be superior to that of **CART**. In the analysis of [Kearns and Mansour, 1996], there are good technical reasons for this difference that are beyond our current scope, but that have to do with the differing concavity of the information gain splitting criterion used by **C4.5** and the Gini splitting criterion used by **CART**. Furthermore, again based on concavity arguments, they also suggest a new splitting criterion that enjoys an even better bound of

$$\left(\frac{1}{\epsilon}\right)^{c/\gamma^2} \tag{3}$$

on the tree size required to achieve error ϵ. In [Dietterich *et al.*, 1996] we demonstrate experimentally that this new splitting criterion results in small but statistically significant improvements in accuracy and tree size over **C4.5**, so the weak learning analysis seems to have pointed us to some modest improvements to the standard algorithms.

Another intriguing issue raised by the theoretical results emerges if one compares any of Equations (1), (2) and (3) to the bounds enjoyed by the recently introduced **Adaboost** algorithm due to [Freund and Schapire, 1995], which requires only

$$\frac{1}{2\gamma^2}\ln\frac{1}{\epsilon} \tag{4}$$

"rounds" (where each round is roughly analogous to a single split made by a top-down decision tree algorithm) to achieve error ϵ. The naive interpretation of this bound, which is only the logarithm of the best bound achieved by a top-down decision tree algorithm given by Equation (3), would lead us to predict that **Adaboost** should vastly outperform, for instance, **C4.5**. In practice, the two algorithms are in fact rather comparable [Freund and Schapire, 1996; Dietterich *et al.*, 1996]. In the latter citation, we provide extensive experimental evidence that this discrepancy between the disparate theoretical bounds and the parity of the algorithms on real problems can be explained by our interpretation of the advantage parameter γ. Briefly, while theoretical boosting results often assume for convenience that there is a simple function with a predictive advantage of γ over random guessing on *any* input distribution, in reality this advantage varies from distribution to distribution (possibly

degrading to the trivial value of zero on "hard" distributions). Since **Adaboost** and **C4.5** explore very different spaces of input distributions as they grow their hypotheses, and since the theoretical bounds are valid only for the smallest advantage γ that holds on the distributions actually explored by the algorithm in question, γ has different meaning for the two algorithms. In [Dietterich *et al.*, 1996], we plot the advantages for each algorithm and demonstrate that while the theoretical bounds for a fixed advantage γ may be *worse* for **C4.5** than for **Adaboost**, the value of γ achieved on real problems is *better*. This empirical fact largely reconciles the theoretical statements with the observed behavior.

Thus, although the weak learning model provides what seems to be the right parameter to study (namely, the advantage γ), experimental examination of this parameter was required for real understanding of what the theory was saying and not saying. This kind of interaction — where the theory suggests improvements to the popular algorithms, and experimentation with these algorithms modifies our interpretation of the theory — seems like a good first step towards the goal mentioned at the outset. There is of course still much work to be done to further close the gap between theory and practice; but at least in the case of decision tree learning, the weak learning framework seems to have provided some footholds that were missing in previous models.

In the bibliography, we provide some additional references on the topics discussed here.

References

Aslam, J. A. and Decatur, S. E. 1993. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. In *Proceedings of the 35th IEEE Symposium on the Foundations of Computer Science.* IEEE Computer Society Press, Los Alamitos, CA. 282–291.

Blum, A.; Furst, M.; Jackson, J.; Kearns, M.; Mansour, Y.; and Rudich, S. 1994. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the 26th ACM Symposium on the Theory of Computing.* ACM Press, New York, NY.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Wadsworth International Group.

Bshouty, N. and Mansour, Y. 1995. Simple learning algorithms for decision trees and multivariate polynomials. In *Proceedings of the 36th IEEE Symposium on the Foundations of Computer Science.* IEEE Computer Society Press, Los Alamitos, CA. 304–311.

Bshouty, N. H. 1993. Exact learning via the monotone theory. In *Proceedings of the 34th IEEE Symposium on the Foundations of Computer Science.* IEEE Computer Society Press, Los Alamitos, CA. 302–311.

Dietterich, Tom; Kearns, Michael; and Mansour, Yishay 1996. Applying the weak learning framework to understand and improve C4.5. In *Machine Learning: Proceedings of the Thirteenth International Conference.* Morgan Kaufmann.

Drucker, H.; Schapire, R.; and Simard, P. 1992. Improving performance in neural networks using a boosting algorithm. In Hanson, S.J.; Cowan, J.D.; and Giles, C.L., editors 1992, *Advances in Neural Information Processing Systems.* Morgan Kaufmann, San Mateo, CA. 42–49.

Freund, Yoav and Schapire, Robert E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *Second European Conference on Computational Learning Theory.* Springer-Verlag. 23–37.

Freund, Y. and Schapire, R. 1996. Some experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference.* Morgan Kaufmann.

Freund, Yoav 1995. Boosting a weak learning algorithm by majority. *Information and Computation* 121(2):256–285.

Jackson, J. 1994. An efficient membership query algorithm for learning DNF with respect to the uniform distribution. In *Proceedings of the 35th IEEE Symposium on the Foundations of Computer Science.* IEEE Computer Society Press, Los Alamitos, CA.

Kearns, M. and Mansour, Y. 1996. On the boosting ability of top-down decision tree learning algorithms. In *Proceedings of the 28th ACM Symposium on the Theory of Computing.* ACM Press, New York, NY.

Kearns, Michael J. and Vazirani, Umesh V. 1994. *An Introduction to Computational Learning Theory.* The MIT Press.

Kearns, M.; Li, M.; Pitt, L.; and Valiant, L. 1987. Recent results on boolean concept learning. In Langley, Pat, editor 1987, *Proceedings of the Fourth International Workshop on Machine Learning.* Morgan Kaufmann, San Mateo, CA. 337–352.

Kearns, M. 1993. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25th ACM Symposium on the Theory of Computing.* ACM Press, New York, NY. 392–401.

Kushilevitz, E. and Mansour, Y. 1991. Learning decision trees using the Fourier spectrum. In *Proc. of the 23rd Symposium on Theory of Computing.* ACM Press, New York, NY. 455–464.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning* 5(2):197–227.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134–1142.