Occam's Two Razors: The Sharp and the Blunt

Pedro Domingos
Artificial Intelligence Group
Instituto Superior Técnico
Lisbon 1096, Portugal
pedrod@gia.ist.utl.pt

Abstract

Occam's razor has been the subject of much controversy. This paper argues that this is partly because it has been interpreted in two quite different ways, the first of which (simplicity is a goal in itself) is essentially correct, while the second (simplicity leads to greater accuracy) is not. The paper reviews the large variety of theoretical arguments and empirical evidence for and against the "second razor," and concludes that the balance is strongly against it. In particular, it builds on the case of (Schaffer, 1993) and (Webb, 1996) by considering additional theoretical arguments and recent empirical evidence that the second razor fails in most domains. A version of the first razor more appropriate to KDD is proposed, and we argue that continuing to apply the second razor risks causing significant opportunities to be missed.[1]

Occam's Two Razors

William of Occam's famous razor states that "Nunquam ponenda est pluralitas sin necesitate," which, approximately translated, means "Entities should not be multiplied beyond necessity" (Tornay 1938). It was born in the late Middle Ages as a criticism of scholastic philosophy, whose theories grew ever more elaborate without any corresponding improvement in predictive power. In the intervening centuries it has come to be seen as one of the fundamental tenets of modern science, and today it is often invoked by learning theorists and KDD practitioners as a justification for preferring simpler models over more complex ones. However, formulating Occam's razor in KDD terms it trickier than might appear at first. Leaving aside for the moment the question of how to measure simplicity, let *generalization error* of a model be its error rate on unseen examples, and *training-set error* be its error on the examples it was learned from. Then the formulation that is perhaps closest to Occam's original intent is:

First razor: Given two models with the same generalization error, the simpler one should be preferred because simplicity is desirable in itself.

On the other hand, within KDD Occam's razor is often used in a quite different sense, that can be stated as:

Second razor: Given two models with the same training-set error, the simpler one should be preferred because it is likely to have lower generalization error.

We believe that it is important to distinguish clearly between these two versions of Occam's razor. The first one is largely uncontroversial, while the second one, taken literally, is false. Several theoretical arguments and pieces of empirical evidence have been advanced to support it, but each of these is reviewed below and found wanting. The paper also reviews previous theoretical arguments against the "second razor," and, more importantly, mounting empirical evidence that it fails in practice. Finally, the first razor is revisited and refined, and some consequences are discussed.

Theoretical Arguments for the Second Razor

The PAC-Learning Argument

Although a large fraction of the computational learning theory literature is concerned with the relationship between accuracy and simplicity (at least superficially), the basic argument is neatly encapsulated in Blumer *et al.*'s (1987) paper "Occam's razor." While the mathematical results in this paper are valid, they have only a very indirect relationship to the second razor, and do not "prove" it. In short, this paper shows that, if a model with low training-set error is found within a sufficiently small set of models, it is likely to also have low generalization error. This model, however, could be arbitrarily complex. The only connection of this result to Occam's razor is provided by the information-theoretic notion that, if a set of models is small, its members can be distinguished by short codes. But this in no way endorses, say, decision trees with fewer nodes over trees with many. By this result, a decision tree with one million nodes extracted from a set of ten such trees is preferable to one with ten nodes extracted from a set of a million, given the same training-set error.

Put another way, the results in (Blumer *et al.* 1987) only say that if we select a sufficiently small set of models prior to looking at the data, and by good fortune one of those models closely agrees with the data, we can be confident that it will also do well on future data. The theoretical results give no guidance as to how to select that set of models.

The Bayesian Argument

Claims of a general theoretical foundation for preferring simple models can also be found in the statistical and pattern recognition literature. While the details vary, they typically take the form of an approximation to the optimal prediction procedure of Bayesian model averaging (Bernardo & Smith 1994; Chickering & Heckerman 1997) that results in evaluating candidate models according to a sum of two terms: an error or likelihood term, and a term penalizing the complexity of the model. Criteria of this type include AIC (Akaike 1978), BIC (Schwarz 1978), and many others. Similar criteria with an information-theoretic interpretation, like MML (Wallace & Boulton 1968) and MDL (Rissanen 1978) are discussed below.

Consider BIC, the first criterion to be explicitly proposed as an approximation to Bayesian model averaging. Leaving aside the fact that BIC involves a sequence of approximations and assumptions that may or may not be valid in practice (e.g., an infinite sample), its use of a complexity penalty does not imply that simpler models are more probable, because BIC computes probabilities for model *structures*, as opposed to models. This distinction is important. $ax+b$ and ax^2+bx+c are model structures; each can be instantiated by many different models, for example $5x+2$ and $3x^2+x+10$. BIC approximates the marginal likelihood of a model structure, which is the average of the likelihoods of all the models that instantiate it (weighted by their prior probabilities given the model structure). BIC penalizes the model structure's dimension because higher-order spaces effectively contain many more models than lower-order ones, and thus contain many more low-likelihood models along with the "best" one(s). (In precise terms, higher-order model structures have a higher VC dimension (Haussler 1988); or, considering finite-precision numbers, they literally contain more models.) For example, the model space defined by ax^2+bx+c contains many more models than the one defined by $ax+b$. Thus, if the correct model is $3x^2+x+10$, the quadratic structure is correspondingly the correct one, but it may still appear less likely than the linear structure, because the high likelihood of the $3x^2+x+10$ model will be averaged with a large number of vanishingly small likelihoods corresponding to the many poor quadratic models that are possible. However, this has no bearing on the likelihood of the $3x^2+x+10$ model; it will still be more likely than any linear model, irrespective of its quadratic degree. Thus, choosing a model structure according to BIC and then instantiating the parameters can lead to a suboptimal model.

Similar remarks apply to the more recent work of MacKay (1992). The "Occam factors" that appear in his evidence framework penalize *model structures* with many parameters, as opposed to models, and can also lead to suboptimal choices.

The Information-Theoretic Argument

The minimum description length (MDL) principle (Rissanen 1978) is perhaps the form in which the second razor is most often applied (e.g., (Quinlan & Rivest 1989)). According to this principle, the "best" model is the one which minimizes the total number of bits needed to encode the model and the data. The MDL principle is appealing because it reduces two apparently incommensurable attributes of a model (error rate and complexity) to the same form: bits of information. However, there is no guarantee that it will select the most accurate model. Rissanen simply proposes it as a fundamental principle. The closely-related minimum message length (MML) principle (Wallace & Boulton 1968) is derived by taking the logarithm of Bayes' theorem and noting that, according to information theory (Cover & Thomas 1991), logarithms of probabilities can be seen as (minus) the lengths of the most efficient codes for the corresponding events. This has led some researchers to believe that a trade-off between error and complexity is "a direct consequence of Bayes' theorem, requiring no additional assumptions" (Cheeseman 1990). However, this belief is founded on a confusion between assigning the shortest codes to the most probable hypotheses and *a priori* considering that the syntactically simplest models in the representation being used (e.g., the decision trees with fewest nodes) are the most probable ones. If they have higher priors, more complex models can be assigned shorter codes, but this obviously does not imply any preference for simpler models in the original representation (e.g., if the model with highest prior is a decision tree with a million nodes, it can be assigned a 1-bit code, without this implying any preference for small trees.)

Information theory, whose goal is the efficient use of a transmission channel, has no direct bearing on KDD, whose goal is to infer predictive and comprehensible models from data. Having assigned a prior probability to each model in the space under consideration, we can always recode all the models such that the most probable ones are represented by the shortest bit strings. However, this does not make them more predictive, and is unlikely to make them more comprehensible.

Theoretical Arguments Against the Second Razor

"Zero-Sum" Arguments

A number of well-known theoretical results have been established which imply that the second razor cannot be true. These results include Schaffer's (1994) conservation law of generalization performance and Wolpert's

(1996) "no free lunch" theorems, and are in turn implicit in Mitchell's (1980) demonstration that bias-free learning is impossible. In essence, they imply that, for every domain where a simpler model is more accurate than a more complex one, there exists a domain where the reverse is true, and thus that no argument about which is preferable in general can be made. Although these results negate the second razor in a mathematical sense, they still leave open the possibility that it will apply in most (or all) real-world domains (Rao, Gordon, & Spears 1995). This is a matter for empirical study, which the next two sections address.

The Vapnik-Chervonenkis Dimension

In his sophisticated theory of learning, Vapnik (1995) shows that the generalization ability of a class of models is not a function of its number of parameters, but of its VC dimension. Although the two are sometimes related, in general they are not. Model structures with a very large number of parameters can generalize quite reliably, if constrained in other ways. The model structure $class = sign(\sin ax)$, with a single parameter, has an infinite VC dimension, because it can discriminate an arbitrarily large, arbitrarily labeled set of points on the x axis (Vapnik 1995, p. 78).

Overfitting Is Due to Multiple Testing

According to conventional wisdom, overfitting is caused by overly complex models, and Occam's razor combats it by introducing a preference for simpler ones. However, Cohen and Jensen (1997) have shown that overfitting in fact arises not because of complexity *per se*, but because attempting a large number of models leads to a high probability of finding a model that fits the training data well purely by chance. Attempting 10 complex models incurs a smaller risk of overfitting than attempting 100 simple ones. Overfitting is thus best combatted not by the second razor, but by taking this multiple testing phenomenon into account when scoring candidate models (Jensen & Schmill 1997; Domingos 1998).

Bias-Variance

Schuurmans *et al.* (1997) have shown that complexity-penalty methods assume a particular bias-variance profile, and that if the true profile does not correspond to the postulated one systematic underfitting or overfitting will result. Thus these methods can only be optimal in very specific cases.

Empirical Evidence for the Second Razor

Arguably, most KDD researchers who routinely apply the second razor do not believe that it is universally true, but simply that it generally applies in practice. For example, Piatetsky-Shapiro (1996) argues that "Occam's razor is not 'ALWAYS' true – but is mostly true in most real-world situations." This section and the next attempt to determine if this is indeed the case.

Pruning

A simple empirical argument for the second razor might be stated as "Pruning works." Indeed, pruning often leads to models that are both simpler and more accurate than the corresponding unpruned ones (Mingers 1989). However, it can also lead to lower accuracy (Schaffer 1993). It is easy to think of simple problems where pruning can only hurt accuracy (e.g., applying a decision tree algorithm like C4.5 to learning a noise-free, diagonal frontier). More importantly, as mentioned above, Cohen and Jensen (1997) have shown persuasively that pruning should not be seen as a correction of overly complex models, but as an effective reduction of the number of models attempted. In a related paper, Jensen and Schmill (1997) have shown empirically that correcting for multiple testing when pruning leads to better results than MDL and related methods.

The 1R Algorithm

In an oft-cited paper, Holte (1993) observes that a decision tree containing a single node can sometimes come reasonably close to C4.5 in accuracy. However, in Holte's experiments his 1R ("1-rule") algorithm was on average 5.7% less accurate than C4.5, which is hardly negligible. The closest results to C4.5 were obtained by the 1R$*$ measure, which finds the accuracy of the best possible 1-rule by looking at the test set. These results appear to have led to some confusion. As Holte points out, 1R$*$ is a *measure*, not an algorithm; it makes no sense to consider its accuracy "competitive" with C4.5's. A similar measure for decision trees would always achieve the Bayes rate (lowest error possible). At most, these experiments suggest that the advantage of going to more complex models is small; they do not imply that simpler models are better (Elomaa 1994). However, as we shall see below, more recent results call even this conclusion into question.

Other Low-Variance Algorithms

More generally, several pieces of recent work (e.g., (Friedman 1996; Domingos & Pazzani 1997)) have suggested that simple learners like the naive Bayesian classifier or the perceptron will often do better than more complex ones because, while having a higher systematic error component (the bias), they are less prone to random fluctuations (the variance). Again, these results do not imply a preference for simpler models, but for restricting search. Suitably constrained, decision-tree or rule induction algorithms can be as stable as simpler ones, and more accurate. Theory revision systems (e.g., (Ourston & Mooney 1994)) are an example of this: they can produce accurate theories that are quite complex with comparatively little search, by making incremental changes to an initial theory that is already complex.

Physics, Etc.

The second razor is often justified by pointing to its success in the "hard" sciences. (Although these arguments are fuzzier, they should still be addressed, because they form a large part of the razor's appeal.) A popular example comes from astronomy, where it favors Copernicus' model of the solar system over Ptolemy's. Ironically, in terms of predictive error the two models are indistinguishable, since they predict the same trajectories. Copernicus's model is preferable on the intrinsic merits of simplicity (first razor). An alternative, slightly humorous example is provided by flat earth vs. spherical earth. The second razor clearly favors the flat earth theory, being a linear model, while the spherical one is quadratic and no better at explaining everyday observations in the Middle Ages.

Another favorite example is relativity vs. Newton's laws. The following passage is from (Cover & Thomas 1991):

> In the end, we choose the simplest explanation that is consistent with the observed data. For example, it is easier to accept the general theory of relativity than it is to accept a correction factor of c/r^3 to the gravitational law to explain the precession of the perihelion of Mercury, since the general theory explains more with fewer assumptions than does a "patched" Newtonian theory.

In fact, the general theory of relativity makes more assumptions than Newton's gravitational law, and is far more complex, so this cannot be the reason for preferring it. The preference comes from the fact that the c/r^3 factor *is* a patch, applied to (over)fit the theory to a particular observation. As Pearl (1978) insightfully notes:

> It would, therefore, be more appropriate to connect credibility with the nature of the selection procedure rather than with properties of the final product. When the former is not explicitly known ... simplicity merely serves as a rough indicator for the type of processing that took place prior to discovery.

Yet another example is Maxwell's four concise and elegant equations of electromagnetism. In fact, these equations are concise and elegant only in the notation of differential operators that was introduced many years after his death. In their original form, they were long and unwieldy, leading Faraday to complain of their incomprehensibility, which precluded him from empirically testing them.

The list goes on. In any case, the fact that comparatively simple equations have proved successful in modeling many physical phenomena is no indication that the same will be true in the large variety of areas KDD is being applied to—medicine, finance, earth sensing, molecular biology, marketing, process control, fault detection, and many others.

Empirical Evidence Against the Second Razor

Several authors have carried out experiments that directly or indirectly investigate the relationship between simplicity and accuracy, and obtained results that contradict the second razor. Fisher and Schlimmer (1988) observed that concept simplification only sometimes improved accuracy in the ID3 and COBWEB systems, and that this was dependent on the training set size and the degree of dependence of the concept on the attributes. Murphy and Pazzani (1994) induced all consistent decision trees for a number of small, noise-free domains, and found that in many cases the smallest trees were not the most accurate ones. Schaffer (1993) conducted a series of experiments showing that pruning can increase error, and that this effect can increase with the noise level. Quinlan and Cameron-Jones (1995) and Murthy and Salzberg (1995) found that excessive search often leads to models that are simultaneously simpler and less accurate. Webb (1996; 1997) showed that, remarkably, the accuracy of decision trees on common datasets can be consistently increased by grafting additional nodes onto the tree, even after the data has been perfectly fit. Chickering and Heckerman (1997) compared several different methods for approximating the likelihood of simple Bayesian model structures, and found that the BIC/MDL approach was almost always the least accurate one. Lawrence *et al.* (1997) conducted experiments with backpropagation in synthetic domains, and found that training neural networks larger than the correct one led to lower errors than training networks of the correct size.

Another source of evidence against the second razor is the growing number of practical machine learning systems that achieve reductions in error by learning more complex models. Vapnik's (1995) support vector machines learn polynomials of high degree (and resulting feature spaces of dimension up to 10^{16}), and outperformed simpler state-of-the-art models in the USPS handwritten digit recognition database. Cestnik and Bratko (1988), Gams (1989) and Datta and Kibler (1995) show how redundancy can improve noise resistance and therefore accuracy. Domingos' (1996) RISE system consistently outperforms CN2 and C4.5/C4.5RULES on common datasets by inducing substantially more complex rule sets. Webb's abovementioned decision-tree grafting is another example. Schuurmans (1997) has proposed a geometric evaluation measure that markedly outperforms complexity-penalty ones in polynomial regression tasks.

Arguably, practical experience with MDL-based systems themselves provides evidence against the second razor. For example, after spending considerable effort to find a good coding for trees and examples, Quinlan and Rivest (1989) found that better results were obtained by introducing an *ad hoc* coefficient to reduce the penalty paid by complex decision trees.

Finally, the success of multiple-model approaches in almost all commonly-used datasets (e.g., (Quinlan 1996)) shows that large error reductions can systematically result from sharply increased complexity. In particular, Rao and Potts (1997) show how bagging builds accurate frontiers from CART trees that approximate them poorly, and Domingos (1997b) shows how a single model learned by emulating the behavior of a bagged ensemble is both more complex and more accurate than a model induced directly from the data by the same learner (C4.5RULES).

All of this evidence points to the conclusion that not only is the second razor not true in general; it is also typically false in the types of domains KDD has been applied to.

The First Razor Revisited

The true reason for preferring simpler models is that they are easier for us to understand, remember and use (as well as cheaper for computers to store and manipulate). Thus the first razor is justified. However, simplicity and comprehensibility are not always equivalent. For example, a decision table (Langley 1996) may be larger than a similarly accurate decision tree, but more easily understood because all lines in the table use the same attributes. Induced models are also more comprehensible if they are consistent with previous knowledge, even if this makes them more complex (Pazzani, Mani, & Shankle 1997). A better form of the first razor would perhaps state that given two models with the same generalization error, the more comprehensible one should be preferred. What exactly makes a model comprehensible is largely domain-dependent, but also a matter for cognitive research.

Discussion

All the evidence reviewed in this paper shows that, contrary to the second razor's claim, greater simplicity does not necessarily (or even typically) lead to greater accuracy. Rather, care must be taken to ensure that the algorithm does not perform more search than the data allows, but this search can (and often should) be performed among complex models, not simple ones.

The second razor can be trivially made true by, after the fact, assigning the simplest representations to the most accurate models found. However, this is of no help in finding those models in the first place. Using "simple model" as just another way of saying "probable model" or "model from a small space," as is often done in the literature, constitutes a multiplication of entities beyond necessity, and thus runs afoul of the first razor, which is as applicable to KDD research as to other branches of science. More importantly, it can lead to the misconception that simpler models in the initial, commonly-used representation (e.g., a decision tree or a list of rules) are for some reason more likely to be true.

The second razor will be appropriate when we really believe that the phenomenon under study has a simple model in the representation language used. But this seems unlikely for the domains and representations KDD typically deals with, and the empirical evidence bears this out. More often, the second razor seems to function as a poor man's substitute for domain knowledge—a way of avoiding the complexities of adjusting the system to the domain before applying it to the data. When this happens, overfitting may indeed be avoided by use of the second razor, but at the cost of detectable patterns being missed, and unnecessarily low accuracy being obtained. The larger the database, the likelier this is. Databases with millions or tens of millions of records potentially contain enough data to discriminate among a very large number of models. Applying the second razor when mining them risks rediscovering the broad regularities that are already familiar to the domain experts, and missing the second-order variations that are often where the payoff of data mining lies.

Systems that allow incorporation of domain constraints (e.g, (Clearwater & Provost 1990; Clark & Matwin 1993; Lee, Buchanan, & Aronis 1998)) are an alternative to blind reliance on simplicity. Incorporating such constraints can simultaneously improve accuracy (by reducing the search needed to find an accurate model) and comprehensibility (by making the results of induction consistent with previous knowledge). Weak constraints are often sufficient ((Abu-Mostafa 1989; Donoho & Rendell 1996; Pazzani, Mani, & Shankle 1997); see also (Bishop 1995), Section 8.7). If we accept the fact that the most accurate models will not always be simple or easily understandable, we should allow an explicit trade-off between the two. Systems that first induce the most accurate model they can, and then extract from it a more comprehensible model of variable complexity (e.g., (Craven 1996; Domingos 1997a)) seem a promising avenue.

Conclusion

Occam's razor can be interpreted in two ways: as favoring the simpler of two models with the same generalization error because simplicity is a goal in itself, or as favoring the simpler of two models with the same training-set error because this leads to lower generalization error. This paper found the second version to be provably and empirically false, and argued that in the first version simplicity is only a proxy for comprehensibility. A resulting prescription for KDD research and applications is to prefer simpler models only when we honestly believe the target phenomenon to be simple. Given that this is seldom the case in practice, we should instead seek to constrain induction using domain knowledge, and decouple discovering the most accurate (and probably quite complex) model from extracting comprehensible approximations to it.

References

Abu-Mostafa, Y. S. 1989. Learning from hints in neural networks. *Journal of Complexity* 6:192–198.

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* 30A:9–14.

Bernardo, J. M., and Smith, A. F. M. 1994. *Bayesian Theory.* New York, NY: Wiley.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition.* Oxford, UK: Oxford University Press.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1987. Occam's razor. *Information Processing Letters* 24:377–380.

Cestnik, B., and Bratko, I. 1988. Learning redundant rules in noisy domains. In *Proceedings of the Eighth European Conference on Artificial Intelligence*, 348–356. Munich, Germany: Pitman.

Cheeseman, P. 1990. On finding the most probable model. In Shrager, J., and Langley, P., eds., *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, CA: Morgan Kaufmann. 73–95.

Chickering, D. M., and Heckerman, D. 1997. Efficent approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning* 29:181–212.

Clark, P., and Matwin, S. 1993. Using qualitative models to guide inductive learning. In *Proceedings of the Tenth International Conference on Machine Learning*, 49–56. Amherst, MA: Morgan Kaufmann.

Clearwater, S., and Provost, F. 1990. RL4: A tool for knowledge-based induction. In *Proceedings of the Second IEEE International Conference on Tools for Artificial Intelligence*, 24–30. San Jose, CA: IEEE Computer Society Press.

Cohen, P. R., and Jensen, D. 1997. Overfitting explained. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, 115–122. Fort Lauderdale, FL: Society for Artificial Intelligence and Statistics.

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory.* New York, NY: Wiley.

Craven, M. W. 1996. *Extracting Comprehensible Models from Trained Neural Networks.* Ph.D. Dissertation, Department of Computer Sciences, University of Wisconsin – Madison, Madison, WI.

Datta, P., and Kibler, D. 1995. Learning prototypical concept descriptions. In *Proceedings of the Twelfth International Conference on Machine Learning*, 158–166. Tahoe City, CA: Morgan Kaufmann.

Domingos, P., and Pazzani, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29:103–130.

Domingos, P. 1996. Unifying instance-based and rule-based induction. *Machine Learning* 24:141–168.

Domingos, P. 1997a. Knowledge acquisition from examples via multiple models. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 98–106. Nashville, TN: Morgan Kaufmann.

Domingos, P. 1997b. Why does bagging work? A Bayesian account and its implications. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 155–158. Newport Beach, CA: AAAI Press.

Domingos, P. 1998. A process-oriented heuristic for model selection. In *Proceedings of the Fifteenth International Conference on Machine Learning.* Madison, WI: Morgan Kaufmann.

Donoho, S., and Rendell, L. 1996. Constructive induction using fragmentary knowledge. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 113–121. Bari, Italy: Morgan Kaufmann.

Elomaa, T. 1994. In defense of C4.5: Notes on learning one-level decision trees. In *Proceedings of the Eleventh International Conference on Machine Learning*, 62–69. New Brunswick, NJ: Morgan Kaufmann.

Fisher, D. H., and Schlimmer, J. C. 1988. Concept simplification and prediction accuracy. In *Proceedings of the Fifth International Conference on Machine Learning*, 22–28. Ann Arbor, MI: Morgan Kaufmann.

Friedman, J. H. 1996. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. Technical report, Department of Statistics and Stanford Linear Accelerator Center, Stanford University, Stanford, CA. ftp://playfair.stanford.edu/pub/friedman/kdd.ps.Z.

Gams, M. 1989. New measurements highlight the importance of redundant knowledge. In *Proceedings of the Fourth European Working Session on Learning*, 71–79. Montpellier, France: Pitman.

Haussler, D. 1988. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence* 36:177–221.

Holte, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11:63–91.

Jensen, D., and Schmill, M. 1997. Adjusting for multiple comparisons in decision tree pruning. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 195–198. Newport Beach, CA: AAAI Press.

Langley, P. 1996. Induction of condensed determinations. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 327–330. Portland, OR: AAAI Press.

Lawrence, S.; Giles, C. L.; and Tsoi, A. C. 1997. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 540–545. Providence, RI: AAAI Press.

Lee, Y.; Buchanan, B. G.; and Aronis, J. M. 1998. Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning* 30:217–240.

MacKay, D. 1992. Bayesian interpolation. *Neural Computation* 4:415–447.

Mingers, J. 1989. An empirical comparison of pruning methods for decision tree induction. *Machine Learning* 4:227–243.

Mitchell, T. M. 1980. The need for biases in learning generalizations. Technical report, Rutgers University, Computer Science Department, New Brunswick, NJ.

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

Murthy, S., and Salzberg, S. 1995. Lookahead and pathology in decision tree induction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1025–1031. Montréal, Canada: Morgan Kaufmann.

Ourston, D., and Mooney, R. J. 1994. Theory refinement combining analytical and empirical methods. *Artificial Intelligence* 66:273–309.

Pazzani, M.; Mani, S.; and Shankle, W. R. 1997. Beyond concise and colorful: Learning intelligible rules. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 235–238. Newport Beach, CA: AAAI Press.

Pearl, J. 1978. On the connection between the complexity and credibility of inferred models. *International Journal of General Systems* 4:255–264.

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

Quinlan, J. R., and Cameron-Jones, R. M. 1995. Oversearching and layered search in empirical learning. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1019–1024. Montréal, Canada: Morgan Kaufmann.

Quinlan, J. R., and Rivest, R. L. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80:227–248.

Quinlan, J. R. 1996. Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 725–730. Portland, OR: AAAI Press.

Rao, J. S., and Potts, W. J. E. 1997. Visualizing bagged decision trees. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. Newport Beach, CA: AAAI Press.

Rao, R. B.; Gordon, D.; and Spears, W. 1995. For every action, is there really an equal and opposite reaction? Analysis of the conservation law for generalization performance. In *Proceedings of the Twelfth International Conference on Machine Learning*, 471–479. Tahoe City, CA: Morgan Kaufmann.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Schaffer, C. 1993. Overfitting avoidance as bias. *Machine Learning* 10:153–178.

Schaffer, C. 1994. A conservation law for generalization performance. In *Proceedings of the Eleventh International Conference on Machine Learning*, 259–265. New Brunswick, NJ: Morgan Kaufmann.

Schuurmans, D.; Ungar, L. H.; and Foster, D. P. 1997. Characterizing the generalization performance of model selection strategies. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 340–348. Nashville, TN: Morgan Kaufmann.

Schuurmans, D. 1997. A new metric-based approach to model selection. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 552–558. Providence, RI: AAAI Press.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6:461–464.

Tornay, S. C. 1938. *Ockham: Studies and Selections.* La Salle, IL: Open Court.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory.* New York, NY: Springer-Verlag.

Wallace, C. S., and Boulton, D. M. 1968. An information measure for classification. *Computer Journal* 11:185–194.

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

Webb, G. I. 1997. Decision tree grafting. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 846–851. Nagoya, Japan: Morgan Kaufmann.

Wolpert, D. 1996. The lack of a priori distinctions between learning algorithms. *Neural Computation* 8:1341–1390.