, , 1–19 ()

The Role of Occam's Razor in Knowledge Discovery

PEDRO DOMINGOS pedrod@cs.washington.edu
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195

Abstract. Many KDD systems incorporate an implicit or explicit preference for simpler models, but this use of "Occam's razor" has been strongly criticized by several authors (e.g., Schaffer, 1993; Webb, 1996). This controversy arises partly because Occam's razor has been interpreted in two quite different ways. The first interpretation (simplicity is a goal in itself) is essentially correct, but is at heart a preference for more comprehensible models. The second interpretation (simplicity leads to greater accuracy) is much more problematic. A critical review of the theoretical arguments for and against it shows that it is unfounded as a universal principle, and demonstrably false. A review of empirical evidence shows that it also fails as a practical heuristic. This article argues that its continued use in KDD risks causing significant opportunities to be missed, and should therefore be restricted to the comparatively few applications where it is appropriate. The article proposes and reviews the use of domain constraints as an alternative for avoiding overfitting, and examines possible methods for handling the accuracy–comprehensibility trade-off.

Keywords: Model selection, overfitting, multiple comparisons, comprehensible models, domain knowledge

1. Occam's Two Razors

Occam's razor is often considered one of the fundamental tenets of modern science. In its original form, it states that "Nunquam ponenda est pluralitas sin necesitate," which, approximately translated, means "Entities should not be multiplied beyond necessity" (Tornay, 1938). It was formulated by William of Occam in the late Middle Ages as a criticism of scholastic philosophy, whose theories grew ever more elaborate without any corresponding improvement in predictive power. 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 is trickier than might appear at first. One difficulty is that many different definitions of simplicity are possible. In practice, simplicity is typically equated (loosely) with the syntactic size of the model: for example, the number of nodes in a decision tree, the number of conditions in a rule set, the number of weights in a neural network, or, in general, the number of parameters in a model. Definitions of this type have obvious limitations, and we will consider some more sophisticated ones below. Unfortunately, no fully satisfactory computable definition of simplicity exists, and perhaps none is possible. Thus, for the most part this article will be concerned with the heuristic view of simplicity above, since it is this view that is typically implicit in the practical use of Occam's razor. 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 of the razor 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 article 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, some consequences are discussed, and alternatives to the use of the second razor are proposed.

2. Theoretical Arguments for the Second Razor

2.1. 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.

For the present purposes, the results can be summarized thus. Suppose that the generalization error of a hypothesis is greater than ϵ. Then the probability that the hypothesis is correct on m independent examples is smaller than $(1-\epsilon)^m$. If there are $|H|$ hypotheses in the hypothesis class considered by a learner, the probability that at least one is correct on all m training examples is smaller than $|H|(1-\epsilon)^m$, since the probability of a disjunction is smaller than the sum of the probabilities of the disjuncts. Thus, if a model with zero training-set error is found within a sufficiently small set of models, it is likely to 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.

2.2. 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 (BMA) (Bernardo & Smith, 1984; Chickering & Heckerman, 1997). In BMA, no single "best" model is selected; rather, every model in the chosen space is retained, and its posterior probability given the data is computed. Predictions are made by voting among all models, with each model weighted by its posterior probability. If the "true" model space and prior distribution are known, this is the optimal prediction procedure in the sense of minimizing generalization error (or, in general, a given loss function). However, for most model spaces of interest to KDD the number of models is far too large for the model average to be efficiently computable, and there is also no closed form solution. Several approximation methods have been proposed to address this problem. A number of of these are based on the assumption that, for sufficiently large samples, the distribution of model parameters given a model structure is approximately multivariate Gaussian. If we retain only those terms in this approximation that increase with the sample size, and approximate the maximum *a posteriori* parameter settings by the maximum likelihood ones, we obtain an expression for the (logarithm of) the probability of a model structure that is the sum of two terms: a likelihood term (reflecting performance on the training data), and a term penalizing the complexity of the model (the number of parameters). Criteria of this type include the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Akaike Information Criterion (AIC) (Akaike, 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., a "large enough" 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.

2.3. The Information-Theoretic Argument

The minimum description length (MDL) principle (Rissanen, 1978) is one of the forms 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 given the model. 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 from Bayes' theorem and coding theory. Let h be a hypothesis, and let X be the training set. Then the most probable hypothesis given X is the one for which

$$P(h|X) = \frac{P(h)P(X|h)}{P(X)}$$

is maximum. Taking logarithms of both sides and multiplying by -1, this is equivalent to minimizing

$$-log_2 P(h|X) = -log_2 P(h) - log_2 P(X|h) + log_2 P(X)$$

where the $log_2 P(X)$ term can be ignored when choosing the most probable hypothesis, since it is the same for all hypotheses. According to the source coding theorem (Shannon's first theorem) (Cover & Thomas, 1991), when we need to transmit symbols s from an alphabet S across a channel, the most efficient binary encoding of the symbols has length $-log_2 P(s)$ for each s (i.e., more probable symbols have shorter codes). Thus the above result can be interpreted as minimizing the length in bits of the optimal coding for the hypothesis, plus the length in bits of the optimal coding for the data given the hypothesis. 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 a result of circular reasoning: models are more probable *a priori* because they have shorter codes, but they have shorter codes because they are more probable *a priori*. Care should be taken to avoid confusion between assigning the shortest codes to the *a priori* most probable hypotheses and considering that the syntactically simplest models in the representation being used (e.g., the decision trees with fewest nodes) are the *a priori* 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. For example, if the model with highest prior is a decision tree with a million nodes, it can be assigned a 1-bit code, leaving longer codes for *a priori* less likely trees; since many of these will have fewer nodes than the "best" tree, no preference for small trees is implied.

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.

3. Theoretical Arguments Against the Second Razor

3.1. "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 no argument about which is preferable in general can be made. This derives from the simple fact that, in the absence of further information, knowledge of the target values for training examples implies nothing about the target values for unseen examples, and so all models consistent with the training data (i.e., all possible combinations of assignments of target values to unseen examples) are equally likely to be correct. Although the "no free lunch" 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 et al., 1995). This is a matter for empirical study, which the next two sections address.

3.2. The Vapnik-Chervonenkis Dimension

Vapnik's (1995) theory of structural risk minimization shows that the generalization ability of a class of models is not a function of its number of parameters, but of its VC dimension. The VC dimension of a hypothesis class H over an instance space X is the size of the largest subset of X for which H can generate all possible binary labelings. Although the VC dimension and the number of parameters 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).

3.3. 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, drawing on the study of multiple comparison problems in statistics (Miller, 1981), Jensen and Cohen (1999) 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. One way to do this is through the use of Bonferroni adjustments (Jensen & Schmill, 1997), although this may cause underfitting when the hypotheses being tested are not independent. A more general but computationally intensive solution is randomization testing (Edgington, 1980; Jensen, 1992). An approach that explicitly takes dependences into account is proposed in Domingos (1998a) and Domingos (1999).

3.4. Bias-Variance

Bias and variance are two useful concepts in characterizing the generalization behavior of learning algorithms (Geman et al., 1992). Bias is the systematic component of generalization error, that is incurred even if an infinite sample is available; variance is the additional error incurred given a finite sample, produced by overresponsiveness to random fluctuations. 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 then systematic underfitting or overfitting will result. Thus these methods can only be optimal in very specific cases. In particular, commonly-used complexity-penalty methods assume that variance increases relatively slowly with model complexity (compared with bias). Whenever this is not the case (e.g., in regression with wide-tailed input distributions), complexity-penalty methods are prone to catastrophic mistakes, choosing hypotheses that are much worse than the best available one.

4. 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.

4.1. 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 (1999) 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 related papers, Jensen and Schmill (1997) and Oates and Jensen (1998) have shown empirically that correcting for multiple testing when pruning leads to better results than MDL and related methods.

4.2. 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.

4.3. Other Low-Variance Algorithms

More generally, several pieces of recent work (e.g., Domingos & Pazzani, 1997; Friedman, 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.

4.4. 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.

5. Empirical Evidence Against the Second Razor

Empirical evidence against the second razor comes from two main sources: experiments that have been carried out specifically to test one aspect or another of the relationship between simplicity and accuracy, and practical systems that learn complex models and outperform systems that learn simple ones. We consider each in turn.

5.1. Experiments Testing 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 consistent 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) varied the width of the beam search conducted by the CN2 rule learner, and found that excessive search often leads to models that are simultaneously simpler and less accurate. Murthy and Salzberg (1995) made similar observations when varying the depth of lookahead in decision tree induction. Webb (1996) 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.

5.2. Systems that Learn Complex Models

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 have outperformed simpler state-of-the-art models in handwritten digit recognition (Schölkopf et al., 1995), text classification (Joachims, 1998) and other tasks (Schölkopf et al., 1998). Cestnik and Bratko (1988), Gams (1989) and Datta and Kibler (1995) show how redundancy can improve noise resistance and therefore accuracy. Schuurmans (1997) has proposed a geometric evaluation measure that markedly outperforms complexity-penalty ones in polynomial regression tasks. Webb's (1997) decision-tree grafting procedure, developed as a result of the experiments mentioned above, improves C4.5's accuracy in most commonly-used datasets by producing larger trees. Another example is Domingos' (1996b) RISE system, which consistently outperforms CN2 and C4.5/C4.5RULES on common datasets by inducing substantially more complex rule sets. RISE typically produces a few short, general rules covering many training examples and a large number of longer, more specific ones covering few. If the latter are removed, RISE's accuracy and complexity both fall to the level of CN2 and C4.5 (Domingos, 1996a). Thus the reason that CN2 and C4.5 do not select more complex rule sets is not that accurate ones do not exist, but that they are unable to find them using their simplicity-biased search.

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.

Perhaps the single most significant piece of evidence against the second razor is the broad success of multiple-model approaches. Methods like bagging (Breiman, 1996), boosting (Freund & Schapire, 1996), stacking (Wolpert, 1992) and error-correcting output codes (Kong & Dietterich, 1995) learn several models by varying the training set or other factors, and combine them to make predictions. Compared to learning a single model, this almost always results in reductions in error rate, often large ones (e.g., Drucker et al., 1994; Quinlan, 1996; Maclin & Opitz, 1997). The model ensemble is effectively equivalent to a single model, but a much more complex one. This is verified in Domingos (1997b), where, for each domain, the ensemble produced by bagging is explicitly converted to a single model using the original base learner (C4.5RULES). In the overwhelming majority of cases, the result is both more complex and more accurate than the model obtained using the base learner directly. Rao and Potts (1997) also show how bagging builds increasingly accurate and complex frontiers from simpler, less accurate ones obtained by CART. Thus the success of model ensembles shows that large error reductions can systematically result from significantly increased complexity.

All of this evidence supports 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.

6. The First Razor Revisited

There is a good reason for preferring simpler models: they are easier for people 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 may be larger than a similarly accurate decision tree, but more easily understood because all lines in the table use the same attributes (Langley, 1996; Kohavi & Sommerfield, 1998). Induced models are also more comprehensible if they are consistent with previous knowledge, even if this makes them more complex (Pazzani, 1991; Pazzani et al., 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 and empirical research (e.g., Kononenko, 1990; Pazzani, 1991; Kohavi & Sommerfield, 1998).

7. Discussion

All the evidence reviewed in this article 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.

7.1. Constrained Knowledge Discovery

If we abandon the preference for simpler hypotheses, in what other ways can we avoid overfitting when learning very flexible models? One answer lies in constraining discovery by the use of domain knowledge, or of educated guesses about the domain. This does not create a knowledge acquisition bottleneck, because constraints reflecting only weak knowledge or quite general assumptions are often sufficient. Domain constraints can appear in many forms: Boolean expressions over the presence or absence of certain types of items in discovered rules (Srikant et al., 1997); restrictions on the sign of inequalities in antecedents of rules (Pazzani et al., 1997); qualitative models that induced rules should be consistent with (Clark & Matwin, 1993); knowledge of which variables are causes and which are effects (Cooper, 1997; Domingos, 1998b); knowledge of determinations (i.e., which variables determine which others) (Russell, 1986); and many others. Donoho and Rendell (1996) list a large number of such types of available "fragmentary knowledge," and successfully use a subset of them in a bankruptcy prediction task (for example, features in incompatible units should not be added when forming derived features). Abu-Mostafa (1989) similarly describes several different types of "hints" that can be given to a learner, and methods for incorporating them into the learning process. Bishop (1995, Section 8.7) reviews several different ways in which knowledge of symmetries and invariances in the domain can be incorporated into neural network learning. The RL rule induction system (Clearwater & Provost, 1990) allows the user to specify a wide range of information on the attributes, and preferences and constraints on the rules to be induced. This has been instrumental in the success of its application to carcinogenicity prediction and other problems (Lee et al., 1998).

The use of constraints on the form of discovered rules is widespread in the field of inductive logic programming, under the name of "declarative bias" (Nédellec et al., 1996). For example, GRENDEL(Cohen, 1994) allows the user to specify a grammar for the language in which the rule antecedents will be expressed. This facility is also available in RIPPER, a very efficient propositional rule learner (Cohen, 1995). Similar ideas have been used in the LAGRAMGE equation discovery system, and successfully applied to modeling ecosystems (Todorovski & Džeroski, 1997).

Domain constraints are often used in association rule discovery, typically in the form of extensions to SQL that allow the user to specify the form of the rules to be discovered (Han et al., 1996; Meo et al., 1996; Shen et al., 1996; Kamber et al., 1997). KDD systems can also be constrained to look for rules that contradict prior knowledge, in the hope that these rules will represent interesting new deviations (Liu et al., 1997; Padmanabhan & Tuzhilin, 1998). Constraints can be accumulated in an interactive manner, by letting the user reject rules produced by the system and generalizing from his/her justifications (Provost & Jensen, 1998), or by letting the user give advice to the system as he/she observes its learning in progress (Maclin

& Shavlik, 1996). Domain constraints can be combined with a simplicity bias, if the latter is believed to be appropriate (Djoko et al., 1995).

In addition to accuracy, domain constraints can also improve comprehensibility and speed. Accuracy and speed are improved by reducing the search needed to find an accurate model. Comprehensibility is improved by making the results of induction consistent with previous knowledge.

One consequence of the theory of structural risk minimization (Vapnik, 1995) is that any restriction on the model space that limits its VC dimension is an *a priori* valid way to combat overfitting. In order to obtain the type of continuous trade-off between error and complexity that is found in typical implementations of the second razor, a sequence of progressively less restricted model spaces (or *structure*) is required. This can potentially be provided by many different factors (called "luckiness functions" by Shawe-Taylor et al. (1996). For classification learners where class predictions are obtained by thresholding a continuous output, one such factor is the *margin*, or minimum distance by which the continuous output is on the correct side of the threshold. Requiring progressively larger margins correspondingly reduces the learner's VC dimension, and thus its ability to overfit. Support vector machines are an example of a *large margin classifier* (Smola et al., 1998), where the margin is the distance of the separating hyperplane to the closest instances of either class (the support vectors). Multiple-model methods like boosting can also be seen as large margin classifiers, where the continuous output is the sum of the models' weighted votes (Schapire et al., 1997) (but see also Grove & Schuurmans (1998)).

The general Bayesian method for combatting overfitting is the attachment of prior probabilities to models. It is unfortunate that these are often used only to penalize complexity. Many other types of information can be incorporated into the prior distribution; for example, models that differ more from the "*a priori* best guess" model can be assigned lower priors (Heckerman et al., 1995).

Structural risk minimization and prior distributions only take into account the model space searched by the learner, not the way it is searched. Process-oriented evaluation (Domingos, 1998a; Domingos, 1999) instead computes how the difference between the current best model's training-set error and expected generalization error increases as search progresses. Intuitively, the more models have been attempted, the less likely the observed error of the "best" model found so far is to be close to its true error, and so the higher the expected true error for a given observed error. Using this method, training error can be continuously traded off against the amount of search conducted, even in the absence of VC-dimension results or informative priors.

7.2. Separating Accuracy and Comprehensibility

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, seem a promising avenue. Many methods have been developed to first learn a neural network, on the assumption that this is the appropriate representation, and then extract a comprehensible model from it, in the form of a set of rules or a decision tree (Craven, 1996; Andrews & Diederich, 1996). Recently, approaches for extracting single models from model ensembles have also been developed (Domingos, 1997a; Breiman & Shang, 1997). The more comprehensible models obtained by these methods are typically less accurate than the "full" models, as might be expected, but still more accurate than models of the same type learned directly from the data. In many cases (e.g., Craven, 1996; Domingos, 1997a), it is also possible to trade off accuracy and comprehensibility by varying parameters of the translation system.

Comprehensible output from a complex model can also be obtained by extracting only those parts that are relevant to the user's immediate goals. Association rule systems that allow the user to sift through the (often very many) rules produced are an example of this approach (Imielinski et al., 1996; Shen et al., 1996). This second stage of "mining the model" can also be accomplished visually (Kohavi & Kunz, 1997; Brunk et al., 1997).

For applications where the goal is not immediate insight, the extraction of comprehensible output can often be delayed until performance time. It then takes the form of explanations for the decisions made (Hasling et al., 1984). Generating an explanation for a specific decision is likely to be much easier than producing a comprehensible global model.

8. 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 article 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.

Andrews, R., & Diederich, J. (Eds.). (1996). *Proceedings of the NIPS-96 Workshop on Rule Extraction from Trained Artificial Neural Networks*. Snowmass, CO: NIPS Foundation.

Bernardo, J. M., & 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., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, *24*, 377–380.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L., & Shang, N. (1997). Born again trees. Technical report, Statistics Department, University of California at Berkeley, Berkeley, CA.

Brunk, C., Kelly, J., & Kohavi, R. (1997). MineSet: An integrated system for data mining. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 135–138). Newport Beach, CA: AAAI Press.

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

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

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

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

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

Cohen, W. W. (1994). Grammatically biased learning: Learning logic programs using an explicit antecedent description language. *Artificial Intelligence*, *68*, 303–366.

Cohen, W. W. (1995). Fast effective rule induction. *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 115–123). Tahoe City, CA: Morgan Kaufmann.

Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, *1*, 203–224.

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

Craven, M. W. (1996). *Extracting Comprehensible Models from Trained Neural Networks*. PhD thesis, Department of Computer Sciences, University of Wisconsin – Madison, Madison, WI.

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

Djoko, S., Cook, D. J., & Holder, L. B. (1995). Analyzing the benefits of domain knowledge in substructure discovery. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 75–80). Montréal, Canada: AAAI Press.

Domingos, P. (1996a). Two-way induction. *International Journal on Artificial Intelligence Tools*, *5*, 113–125.

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

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

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

Domingos, P. (1998a). A process-oriented heuristic for model selection. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 127–135). Madison, WI: Morgan Kaufmann.

Domingos, P. (1998b). When (and how) to combine predictive and causal learning. *Proceedings of the NIPS-98 Workshop on Integrating Supervised and Unsupervised Learning*, Breckenridge, CO: NIPS Foundation.

Domingos, P. (1999). Process-oriented estimation of generalization error. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: Morgan Kaufmann.

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

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

Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., & Vapnik, V. (1994). Boosting and other machine learning algorithms. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 53–61). New Brunswick, NJ: Morgan Kaufmann.

Edgington, E. S. (1980). *Randomization Tests*. New York, NY: Marcel Dekker.

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

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

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148–156). Bari, Italy: Morgan Kaufmann.

Friedman, J. H. (1997). On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*, 55–77.

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

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Grove, A. J., & Schuurmans, D. (1998). Boosting in the limit: Maximizing the margin of learned ensembles. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 692–699). Madison, WI: AAAI Press.

Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Stefanovic, N., Xia, B., & Zaiane, O. (1996). DBMiner: a system for mining knowledge in large relational databases. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 250–255). Portland, OR: AAAI Press.

Hasling, D. W., Clancey, W. J., & Rennels, G. (1984). Strategic explanations for a diagnostic consultation system. In M. J. Coombs (Ed.), *Developments in Expert Systems* (pp. 117–133). London, UK: Academic Press.

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

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*, 197–243.

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

Imielinski, T., Virmani, A., & Abdulghani, A. (1996). DataMine: application programming interface and query language for database mining. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 256–262). Portland, OR: AAAI Press.

Jensen, D. (1992). *Induction with Randomization Testing: Decision-Oriented Analysis of Large Data Sets*. PhD thesis, Washington University, Saint Louis, MO.

Jensen, D., & Cohen, P. R. (1999). Multiple comparisons in induction algorithms. *Machine Learning*. To appear.

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

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the Tenth European Conference on Machine Learning* (pp. 137–142). Chemnitz, Germany: Springer-Verlag.

Kamber, M., Han, J., & Chiang, J. Y. (1997). Metarule-guided mining of multi-dimensional association rules using data cubes. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 207–210). Newport Beach, CA: AAAI Press.

Kohavi, R., & Kunz, C. (1997). Option decision trees with majority votes. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 161–169). Nashville, TN: Morgan Kaufmann.

Kohavi, R., & Sommerfield, D. (1998). Targeting business users with decision table classifiers. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 249–253). New York, NY: AAAI Press.

Kong, E. B., & Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 313–321). Tahoe City, CA: Morgan Kaufmann.

Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga (Ed.), *Current Trends in Knowledge Acquisition*. Amsterdam, The Netherlands: IOS Press.

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

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

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

Liu, B., Hsu, W., & Chen, S. (1997). Using general impressions to analyze discovered classification rules. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 31–36). Newport Beach, CA: AAAI Press.

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

Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, RI: AAAI Press.

Maclin, R., & Shavlik, J. (1996). Creating advice-taking reinforcement learners. *Machine Learning*, *22*, 251–281.

Meo, R., Psaila, G., & Ceri, S. (1996). A new SQL-like operator for mining association rules. *Proceedings of the Twenty-Second International Conference on Very Large Databases* (pp. 122–133). Bombay, India: Morgan Kaufmann.

Miller, Jr., R. G. (1981). *Simultaneous Statistical Inference* (2nd ed.). New York, NY: Springer-Verlag.

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, Computer Science Department, Rutgers University, New Brunswick, NJ.

Murphy, P., & 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., & Salzberg, S. (1995). Lookahead and pathology in decision tree induction. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1025–1031). Montréal, Canada: Morgan Kaufmann.

Nédellec, C., Rouveirol, C., Adé, H., Bergadano, F., & Tausend, B. (1996). Declarative bias in ILP. In L. de Raedt (Ed.), *Advances in Inductive Logic Programming* (pp. 82–103). Amsterdam, the Netherlands: IOS Press.

Oates, T., & Jensen, D. (1998). Large datasets lead to overly complex models: An explanation and a solution. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 294–298). New York, NY: AAAI Press.

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

Padmanabhan, B., & Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 94–100). New York, NY: AAAI Press.

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

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 416–432.

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*.

Provost, F., & Jensen, D. (1998). *KDD-98 Tutorial on Evaluating Knowledge Discovery and Data Mining*. New York, NY: AAAI Press.

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

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

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

Rao, J. S., & Potts, W. J. E. (1997). Visualizing bagged decision trees. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 243–246). Newport Beach, CA: AAAI Press.

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

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

Russell, S. J. (1986). Preliminary steps towards the automation of induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 477–484). Philadelphia, PA: AAAI Press.

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

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

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. *Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, TN: Morgan Kaufmann.

Schölkopf, B., Burges, C., & Smola, A. (Eds.). (1998). *Advances in Kernel Methods: Support Vector Machines*. Cambridge, MA: MIT Press.

Schölkopf, B., Burges, C., & Vapnik, V. (1995). Extracting support data for a given task. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 252–257). Montréal, Canada: AAAI Press.

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

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

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

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1996). Structural risk minimization over data-dependent hierarchies. Technical Report NC-TR-96-053, Department of Computer Science, Royal Holloway, University of London, Egham, UK.

Shen, W.-M., Ong, K., Mitbander, B., & Zaniolo, C. (1996). Metaqueries for data mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 375–398). Menlo Park, CA: AAAI Press.

Smola, A., Bartlett, P., Schölkopf, B., & Schuurmans, D. (Eds.). (1998). *Proceedings of the NIPS-98 Workshop on Large Margin Classifiers*. Breckenridge, CO: NIPS Foundation.

Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 67–73). Newport Beach, CA: AAAI Press.

Todorovski, L., & Džeroski, S. (1997). Declarative bias in equation discovery. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 376–384). Nashville, TN: Morgan Kaufmann.

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., & 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. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 846–851). Nagoya, Japan: Morgan Kaufmann.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.

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