ELSEVIER

Artificial Intelligence 82 (1996) 157–179

Artificial Intelligence

Noise modelling and evaluating learning from examples

Ray J. Hickey*

Faculty of Informatics, University of Ulster at Coleraine, Ulster, Co. Londonderry BT52 1SA, N. Ireland, UK

Received July 1993; revised July 1994

Abstract

The means of evaluating, using artificial data, algorithms, such as ID3, which learn concepts from examples is enhanced and referred to as the method of artificial universes. The central notions are that of a class model and its associated representations in which a class attribute is treated as a dependent variable with description attributes functioning as the independent variables. The nature of noise in the model is discussed and modelled using information-theoretic ideas especially that of majorisation. The notion of an irrelevant attribute is also considered. The ideas are illustrated through the construction of a small universe which is then altered to increase noise. Learning curves for ID3 used on data generated from these universes are estimated from trials. These show that increasing noise has a detrimental effect on learning.

1. Introduction

Supervised learning of concepts from classified examples remains a problem of major interest. Amongst the many approaches that have been taken to this task are: induction of decision trees (Quinlan [24]), or high-level rules (Clark and Niblett [8]); instance-based learning (see Aha [2] and Cost and Salzberg [9]); artificial neural networks (Rumelhart [26]); genetic classifiers (see Booker, Goldberg and Holland [4]); Bayesian classifiers (see Cheeseman [6] and Langley, Iba and Thompson [17]).

Typically the examples presented to the algorithm are representative in some sense of a set of possible examples and often constitute a very small subset. Each example consists of a description in an appropriate representation language (often just a simple attribute–value formalism) together with the assigned class or concept offered by the teacher. The learning task is therefore one of induction of a general concept description from the particular cases provided.

* Telephone: +44-(0)1265-44141. E-mail: rj.hickey@ulst.ac.uk.

SSDI 0004-3702(94)00094-8

In real-world settings the task is complicated by the presence of noise of various forms such as errors in recording attribute values or in classification by the teacher. This not only makes the work of the learning algorithm more difficult but also complicates the evaluation of its performance.

Performance of an algorithm or comparison of several algorithms is usually assessed by one of the following means:

(1) *Empirical analysis using real data.* From a database of examples (such as one of the standard sets kept in the "Machine Learning Repository" (Murphy and Aha [21])) subsets are drawn at random and used for learning. Further subsets are then drawn and used for evaluation of the classification performance of the learned description.

(2) *Empirical analysis using artificial data.* In order to simulate the effects of noise, data is generated according to a given prescription and then aspects of the example description or the class are altered using a mechanism involving known probabilities. Irrelevant attributes may be introduced.

(3) *Average-case analysis.* Examples are generated according to a known probabilistic prescription. The expected value behaviour of the learning algorithm is then derived.

(4) *PAC analysis.* The probably approximately correct (PAC) theory of Valiant [31] provides a theoretical basis for assessing performance. A concept is said to be PAC-learnable by an algorithm if with probability $1-\delta$ the learned concept description has a probability $1-\varepsilon$ of classifying almost correctly on subsequent trials. Although the analysis is probabilistic, no underlying distribution for examples is assumed and it is therefore "worst case" over all possible distributions.

In the first three methods a learning curve showing performance (typically classification accuracy) against number of training examples can be derived. For methods (1) and (2) this curve is estimated from data over many trials. Kibler and Langley [16] argue that method (2) provides greater opportunity for systematic investigation than method (1) particularly with regard to the controlled administration of noise; in method (1) the naturally occurring noise cannot be quantified satisfactorily. Amongst the most well-known artificial data sets containing an element of noise is the LED domain of Breiman et al. [5] where components of an LED display for digits are inverted with a small probability. Aha [1] has introduced a variation of the empirical approach by taking a database and altering it in a random fashion to produce a variant for experimentation (called a case) which retains the essential characteristics of the original.

In method (3) the theoretical learning curve can, mathematics permitting, be derived from the underlying setup (into which noise may be introduced); see Pazzani and Sarrett [22] or Langley et al. [17]. Unfortunately this approach seems feasible only for simple algorithms. Analysis of more sophisticated algorithms such as ID3 (Quinlan [24, 25]) or backpropagation (Rumelhart [26]) would appear to be too difficult.

Unlike methods (1)–(3), the PAC approach offers very little insight into the performance of an algorithm in typical circumstances and is overly pessimistic

about its capabilities. Pazzani and Sarrett [22] show learning curves derived from the average case and PAC approaches. For noise modelling in the PAC approach see Valiant [32], Angluin and Laird [3] or Sakakibara [27].

1.1. Modelling noise and extending the use of artificial data: the method of artificial universes

Although, as indicated above, the introduction of noise in a systematic way can be undertaken in methods (2), (3) and (4), in practice the modelling has often been ad hoc and lacking in any underlying theory. The resulting artificial domains are then not sufficiently realistic with regard to complexity of noise. Also it may not be clear how much noise overall has been introduced.

The purpose of this work is to enhance the capabilities of method (2) by introducing a unified theory of noise which makes it possible to measure and modify with ease the amount of noise in an artificial system. Amongst the benefits of this approach are:

(1) a simple means of producing data sets which possess the required amount of noise and which provide challenging tasks for a wide range of algorithms;
(2) greater clarity concerning the relative importance of different sources of noise;
(3) greater insight into the nature of noise and information and how these may be explicated.

The approach involves specifying a complete probabilistic model for the attributes used in the example description and the class. This will be referred to as an artificial universe. The class model of the universe will declare relationships between descriptions and class distributions—as distinct from individual classes. The class model can be represented in many different forms from an exhaustive table to a comparatively small set of very general rules.

An artificial universe can be used to generate examples for use by any of the types of learning algorithm mentioned above and, after learning has taken place, can provide a true indication of the performance of a learned concept description. Since the latter is a random variable its expected value behaviour can be approximated from a series of trials involving generated examples.

Noise will be modelled using the class distributions in the class model. The amount of noise will be manipulated principally using the majorisation relation which is concerned with the relative degree of inequality amongst elements in a real vector, in this case a vector of probabilities constituting a distribution.

Although what is proposed here is merely an extension or elaboration of method (2) above it will be referred to as the *method of artificial universes*. Specifying a complete probability model for the generation of data is hardly a new idea. Many of the artificial domains in the literature such as the LED domain [5] are such models. General-purpose generators offering a limited noise modelling capability have also been developed; see, for example Lounis and Bisson [18]. In essence the contribution here is to provide a simple yet general means of

prescribing a particular amount of noise and in a way which is largely independent of its physical source.

The ideas in this paper were outlined by Hickey [14] where the use of the method was illustrated through the construction of a small universe and subsequent experimentation using ID3 on generated examples. The paper concentrates on the underlying theory although a similar set of experiments to those in [14] is performed but with a much larger number of trials.

1.2. Plan of the paper

The definition of an artificial universe and its class model are given in Section 2 and a running example is introduced. The representation of the class model is also discussed. The modelling of different types of noise is addressed in Section 3. Two types of irrelevant attribute (pure noise and redundant) are defined.

The role of majorisation in explicating noise and its relationship to other information-theoretic ideas are reviewed in Section 4. This is applied to produce information statistics for a universe.

In Section 5 the means of assessing the performance of a deterministic classifier, acquired through learning, is discussed. Experimental results obtained from using ID3 on data generated from universes with varying degrees of noise are reported.

2. Artificial universes

An object or situation to be classified is described using attributes (referred to as the description or condition attributes) which will be labelled $a, b, c, \ldots$. These attributes may be discrete (usually finite) or continuous. The description attribute set together with the values that attributes may take is called the *description schema*. A vector of values $(a_1, b_1, \ldots)$ where there is one value from each attribute in the schema is called a *description vector*. More generally a partially instantiated description vector is called a *condition* or *complex*. The classes that may be assigned to an object or situation will be labelled *class*1, *class*2, ... and will be regarded as values of an attribute called *class* (the class attribute).

An artificial universe together with the notion of a class model and its representation are then defined as follows:

Definition 2.1. An (*artificial*) *universe* consists of a description schema, a class attribute and the joint distribution of the class and description attributes.

Definition 2.2. The function which maps a description vector to the distribution of the class attribute conditional on that vector is called the *class model* of the universe.

Definition 2.3. Any statement specifying the class model is said to be a *representation* of the class model.

The class model is analogous to a statistical model, such as is used in regression analysis: *class* is treated as the dependent variable with the description schema providing the independent variables. Likewise the class distributions play the role that error distributions play in a statistical model. It is important to note that it is the *observed* schema vectors and class and their probabilistic relationship that are being modelled. The artificial universe makes no statement about whether, for example, the class associated with a particular description is the correct one. This point will be further elaborated in the next section when the physical sources of noise are discussed.

A representation is usually a set of rules of the form:

if ⟨complex⟩ then ⟨*class* distribution⟩

where the complexes form a partition of the set of description vectors. The class model and joint distribution of the description attributes are sufficient to completely define the universe.

To illustrate these ideas a small universe, called universe 1, will be built. This is a variation on the universe defined in [14]: it has the same description schema and joint distribution of the description attributes however the number of classes has been increased from three to eight. The schema involving description attributes a, b, c, d, and e is:

a: a_1, a_2
b: b_1, b_2, b_3, b_4
c: c_1, c_2, c_3
d: d_1, d_2, d_3
e: e_1, e_2
class: $1, 2, 3, 4, 5, 6, 7, 8$

The rule set specifying the class model is shown in Table 1. Note that d and e do not appear. The joint distribution of the description attributes is provided in Table 2. Here (a, b), c and e are defined to be mutually independent; d is dependent on a and c. The probability of any description occurring is then obtained as:

$$P(a = v_1, b = v_2, c = v_3, d = v_4, e = v_5)$$
$$= P(a = v_1, b = v_2) \cdot P(c = v_3) \cdot P(d = v_4 \mid a = v_1, c = v_3) \cdot P(e = v_5) . \quad (1)$$

The right-hand side of (1) is an example of a *generating expression* for a description vector.

The rule set in Table 1 offers a fairly compressed representation involving general complexes. Clearly if all attributes in the description schema are finite discrete it is possible to tabulate the mapping between individual description

Table 1
A representation of the class model for universe 1

Rule number	Description	Class distribution
1	$a = a_1$ and $b = b_1$	(0.5, 0.5, 0, 0, 0, 0, 0, 0)
2	$a = a_2$ and $b = b_1$ and $c = c_1$	(0.55, 0, 0, 0.45, 0, 0, 0, 0)
3	$a = a_2$ and $b = b_1$ and $c = c_2$	(0, 0, 0.6, 0, 0, 0, 0, 0.4)
4	$a = a_2$ and $b = b_1$ and $c = c_3$	(0, 0, 0, 0, 0, 0, 0.4, 0.6)
5	$b = b_2$ and $c = c_1$	(0, 0, 0, 0.3, 0, 0, 0.7, 0)
6	$b = b_2$ and $c = c_2$	(0, 0, 0, 0, 0, 0.55, 0.45, 0)
7	$b = b_2$ and $c = c_3$	(0, 0, 0.4, 0, 0.6, 0, 0, 0)
8	$a = a_1$ and $b = b_3$	(0.3, 0.7, 0, 0, 0, 0, 0, 0)
9	$a = a_2$ and $b = b_3$ and $c = c_1$	(0, 0, 0, 0.6, 0, 0.4, 0, 0)
10	$a = a_2$ and $b = b_3$ and $c = c_2$	(0, 0, 0.6, 0, 0, 0.4, 0, 0)
11	$a = a_2$ and $b = b_3$ and $c = c_3$	(0, 0, 0, 0, 0, 1, 0, 0)
12	$b = b_4$ and $c = c_1$	(0, 0, 0, 0.35, 0, 0.65, 0, 0)
13	$a = a_1$ and $b = b_4$ and $c = c_2$	(0, 0, 0, 0, 0.5, 0, 0, 0.5)
14	$a = a_2$ and $b = b_4$ and $c = c_2$	(0.4, 0, 0, 0.6, 0, 0, 0, 0)
15	$a = a_1$ and $b = b_4$ and $c = c_3$	(0, 0.6, 0, 0, 0, 0.4, 0, 0)
16	$a = a_2$ and $b = b_4$ and $c = c_3$	(0, 0, 0, 0, 0, 0.5, 0.5, 0)

vectors and their class distributions; this will be referred to as the *enumerated* representation.

One of the motivations for this work is to simulate in an artificial setting some of the complications of the real world. Making the description attributes dependent on one another in some way is one aspect of this. In other approaches to artificial data modelling mutual independence is often assumed.

Table 2
Probabilities for the joint distribution of the description attributes in universe 1; (a, b), c and e are mutually independent; d is dependent on (a, c)

		b			
		b_1	b_2	b_3	b_4
a	a_1	0.05	0.02	0.4	0
	a_2	0.15	0.03	0.25	0.1
	c			e	
	c_1	c_2	c_3	e_1	e_2
	0.6	0.05	0.35	0.3	0.7

		d		
		d_1	d_2	d_3
Conditional values of a and c	a_1 and (c_1 or c_3)	0.4	0.2	0.4
	a_1 and c_2	0.2	0.5	0.3
	a_2 and c_1	0.1	0.2	0.7
	a_2 and c_2	0.25	0.7	0.05
	a_2 and c_3	0.3	0.3	0.4

Once a universe is specified as above, examples can be generated from it as follows:

(1) Generate a description vector from the joint distribution of the description attributes using the generating expression.
(2) Look up the generated vector in the class model to find the class distribution.
(3) Generate a class from the class distribution.
(4) The description vector and the class thus obtained constitute the example.

It may appear that the use of a rule set to define the class model is somehow biased towards learning algorithms which induce high-level rules. This is not the case however. It is just a convenient way to produce a specification for a universe. Moreover the choice of one representation rather than another, for example the use of general rules versus the enumerated representation, is of no consequence for the generation of examples.

The marginal, i.e. unconditional, distribution of class (referred to as the *default* distribution) can be obtained from the universe specification. For universe 1 this is:

$$(0.1215, 0.1510, 0.0400, 0.2115, 0.0445, 0.2675, 0.1270, 0.0370)\,. \tag{2}$$

The majority class, i.e. most probable, in the default distribution is called the *default* class. Here it is class 6 with a 26.75% chance of occurring.

2.1. Sub-universes

There are universes within universes. These will be called *sub-universes*. If several attributes in the example description are removed this results in a sub-universe with a reduced description schema. This will be referred to as the *marginal* sub-universe w.r.t. the remaining description attributes. There are also *conditional* sub-universes relative to a given complex having one or more attributes uninstantiated. These latter attributes can be used to formulate a sub-inverse conditional on the complex. Such a complex may involve partial instantiation of one or more attributes, e.g. $b = b_1$ or b_2. In such a case the partially instantiated attributes are included in the schema for the sub-universe with their values being restricted to those specified in the instantiation.

Class models and generating expressions can be constructed for sub-universes from the specification of the universe itself.

3. Noise and the class model

In learning from examples, noise is anything which obscures the relationship between description and class. There are three major physical sources of noise:

(1) insufficiency of the description schema,
(2) corruption of attribute values in the example description,
(3) erroneous classification of training examples.

If there are non-degenerate distributions appearing in the class model this can be attributed to the presence of one or more of these noise sources. The user of the artificial universe method can adopt one of two attitudes to the origin of noise. Firstly she can ignore it, i.e. say nothing about it, and just investigate how the noise affects learning, perhaps varying the degree of noise in the manner to be described below. Secondly she can declare that the noise originates from some of the sources above. It would be difficult, though, to specify how much came from each source without building a more elaborate model. For the most part, however, the origin of noise does not affect the analysis of learning and the first approach is often all that is required. To see this consider the sources separately and how they might be catered for within the class model.

The class distributions in the class model describe the uncertainty in the association between observed description vector and the class it is assigned and can therefore be taken to model the first of these, the insufficiency of the description schema (what the statisticians call residual error).

These distributions can also be used to model—or rather eliminate the separate need for—attribute noise. To explicitly allow for corruption of attribute values a distinction can be made between true attributes (which are not observed) and actual attributes which are observed. Usually, as, for example, in the LED domain an actual attribute takes values from the same set as its corresponding true attribute according to a given probability distribution (called the *corruption distribution*) which, in the most general case, is conditional upon the true description vector.

The definition of a universe can be augmented to provide the joint distribution of both *true* and *actual* attributes together with the class attribute. From this can be obtained a true and an actual class model (class distributions conditional on the true and actual attributes respectively).

If it is assumed that the mechanism which corrupts example descriptions is at work both when examples are being obtained for learning and also when the learned concept description is being tested (on fresh unclassified examples) then the true class model has no role to play: the actual class model can be used to generate examples and to assess the learned concept.

It may be argued that the purpose of introducing separate true and actual attributes is that it allows the affect of attribute noise to be determined. Schaffer [28], for example, increases the probability of corruption in the LED domain and investigates the consequences for learning. In general, though, it is difficult to state the overall effect on noise resulting from a particular set of corrupting distributions. It does not follow, paradoxical though it may seem, that corrupting the attribute distributions necessarily increases noise. In fact it is possible to define a universe in which the true class model has distributions which are non-degenerate (so that class is uncertain) whereas the actual class model has each of its distributions degenerate at a single class!

Errors in example classification do, however, require more elaborate modelling. Here the class distributions reflect teacher miss-classification (and perhaps other sources too), so that some of the classes present in the examples are actually the wrong class. After learning, however, the acquired concept description is used

to make classifications. Thus the classification error mechanism is not present. Evaluation of learning would use the true class distributions—not those used to generate the examples. Two universes are therefore required: an actual universe (used for example generation) and a true universe. It appears from the literature that there is little awareness of this aspect of classification noise since the data for testing is usually generated according to the same prescription as that for learning.

3.1. Irrelevant attributes

In addition to noise there are often "irrelevant" attributes present in the schema, i.e. those that contribute little or nothing to classification. There are two cases to distinguish here, namely those of redundancy and pure noise.

Definition 3.1. An attribute having the property that there is a representation of the class model in which it does not appear is said to be *redundant* relative to those attributes which do appear.

Definition 3.2. An attribute, a, which satisfies

$$P(class = c \mid S, A) = P(class = c \mid S)$$

for all c, S and A, where S is a subset of the vector description space of all the description attributes excluding a and A is a subset of the values taken by a, is called a *pure noise* attribute. A pure noise attribute is said to be *uninformative* (about class); all other attributes are said to be *informative*. A universe in which at least one attribute is informative is said to be *informative*, otherwise it is said to be *uninformative*.

The redundancy of an attribute may be relative to a particular subset of attributes, for example, when it is functionally dependent on those in the subset. Pure noise implies redundancy but not vice versa: the omission of an attribute from a representation of the universe does not guarantee that it is pure noise as will be seen below.

In universe 1, d is redundant (it is not instantiated in any of the rules in the given representation of the class model) but it is not pure noise. The class distribution for the condition $b = b_3$ and $c = c_1$ is:

$$(0.113, 0.263, 0, 0.375, 0, 0.250, 0, 0),$$

indicating that class 4 is the most likely. If, however, $d = d_1$ is also instantiated the class distribution changes to

$$(0.212, 0.494, 0, 0.176, 0, 0.118, 0, 0),$$

which makes class 2 the most likely.

Some authors, for example Pazzani and Sarrett [22], use the term "irrelevant" to mean redundancy as defined here. Since redundant attributes may contain useful information about class the use of the term "irrelevant" is unfortunate. Only a pure noise attribute is truly irrelevant in the sense that there are no circumstances in which knowledge of its value can influence classification.

The definition for pure noise appears similar to that of conditional independence of class and the pure noise attribute given all the other description attributes—abbreviated in this discussion to "conditional independence" (see Pearl [23] for a general discussion of conditional independence). In the latter definition, however, the conditioning description vector must be fully instantiated whereas that is not the case for pure noise. Thus pure noise implies conditional independence but not conversely. The specification of a class model in which an attribute does not appear, as was made for universe 1, is equivalent to the assertion of conditional independence for that attribute (so that redundancy is essentially conditional independence).

The implementation of a pure noise attribute can be achieved with the further requirement (in addition to omission from the concept description) that it be independent of the vector of all the other description attributes.[1] In universe 1, e is pure noise. Additional pure noise attributes can be added very easily to a universe already defined simply by specifying a marginal distribution for each one and adding a corresponding term to the generating expression for a description vector.

By taking S to be the empty set in Definition 3.2 it follows that class and a pure noise attribute are independent.

Clearly a universe is uninformative if and only if all the class distributions in the class model are identical. In this case they are all equal to the default distribution for class. In any representation of the class model, the complexes that appear in the rules define uninformative conditional sub-inverses. The notion of informativeness of an attribute is quite separate from that of noise in the universe class distributions. If an attribute is informative then there will be occasions when knowledge of its value will be of some use to classification regardless of how much noise is in the universe although, of course, the extent of the latter limits just how informative an attribute can be.

It is also possible for a universe with no noise in the class distributions to have several pure noise attributes. The assessment of how informative individual attributes are will be pursued further in the next section when noise in the universe is further explicated using information-theoretic concepts.

4. Assessing the degree of noise in the class distributions

The question remains as to how to manipulate the class distribution to achieve a required amount of noise. In statistical modelling of residual error, it is common practice to employ normal distributions with the variance parameter indicating the

[1] Let a' denote the vector of all description attributes excluding a. The assertion: "independence of a and a' together with the conditional independence of *class* and a given a' implies that a is pure noise" follows easily from elementary probability theory. This sufficient condition for pure noise is also readily seen to be equivalent to the independence of $(class,a')$ and a.

extent of noise. For class distributions where there are at most a small number of discrete (often nominal) classes this is not appropriate. What is needed is a means of explicating the degree of noise using only the probabilities in the class distribution and not the classes themselves. For example, in universe 1 the class distribution for the first rule for the class model (Table 1) is (0.5, 0.5, 0, 0, 0, 0, 0, 0). Is there more noise in this distribution than that of the next rule which is (0.55, 0, 0, 0.45, 0, 0, 0, 0)? Such a means of comparison is provided by the majorisation relation applied to probability vectors. Majorisation underpins much of the theory of measurement of information and uncertainty.

4.1. Majorisation and noise

Measures of uncertainty such as Shannon's entropy function

$$entropy(P) = -\sum_{i=1}^{n} p_i \ln p_i ,$$

where $P = (p_1, \ldots, p_n)$ and $\sum_{i=1}^{n} p_i = 1$ is the probability distribution on the n possible classes, provide an assessment of noise based on probabilities only.

Entropy, though, is only one of many possible measures of uncertainty. The well-known additivity of information property of entropy which renders it unique up to positive multiples (Shannon and Weaver [30]) and which has meaning in communication theory is not relevant here. Another popular measure is provided by the Gini index of diversity:

$$gini(P) = 1 - \sum_{i=1}^{n} p_i^2 .$$

Information and uncertainty are usually regarded as dual, i.e. the greater the information the lesser the uncertainty, and will be treated as such here. Noise will be identified with uncertainty.

Uncertainty measures are usually defined to be strictly Schur-concave (see Hickey [11, 12]) that is they respect the pre-ordering afforded by the majorisation relation between probability distributions.

Majorisation as developed in the mathematical theory of inequalities provides a pre-ordering amongst vectors of real numbers having the same total of their elements (but not necessarily having the same number of elements) interpreted as: "the elements in this vector are less equal than those in that vector". It is a pre-ordering rather than a partial ordering because it lacks antisymmetry: vectors which differ only in permutation of elements majorise each other but are not necessarily identical.

Applied to discrete probability distributions or, indeed, relative frequency distributions, strict majorisation explicates the notion "less noisy than", i.e. the distribution whose probabilities are less equal is the less noisy or, equivalently, the more informative. Here the background to majorisation will be outlined for probability distributions only. For the general theory of majorisation see [19].

Majorisation can be defined using the notion of an *equalising transfer* which involves re-distributing some (larger) probability from one event to another having a smaller probability (which may be zero) in such a way as to render the two probabilities involved more equal. Formally, if in $P = (p_1, \ldots, p_n)$, there are i and j such that $p_i > p_j$ then some of the excess in p_i over p_j is transferred to p_j such that a new distribution, P', is created with all except the ith and jth probabilities unchanged and the latter being replaced by p'_i and p'_j where

$$p'_i = cp_i + (1-c)p_j\,, \qquad p'_j = (1-c)p_i + cp_j \tag{3}$$

for some c, $0 \leqslant c \leqslant 1$. This leads to a definition of majorisation:

Definition 4.1. If discrete distributions P and Q are such that Q can be transformed into P by a finite number of equalising transfers then Q is said to *majorise* P (or P *de-majorises* Q) and this is written $P \preccurlyeq Q$. If $P \preccurlyeq Q$ and Q is not a permutation of P then Q is said to *strictly majorise* P (or P *strictly de-majorises* Q) and this is written $P \prec Q$.

It follows that majorisation is transitive. The case $c = 0$ corresponds to swapping the ith and jth probabilities. Thus if Q is a permutation of P then $P \preccurlyeq Q$ and $Q \preccurlyeq P$. The expression in (3) can be reformulated as

$$P = QS\,, \tag{4}$$

where S is the doubly stochastic matrix with

$$s_{kk} = 1\,, \quad k \neq i, j\,,$$
$$s_{ii} = c\,, \qquad s_{ji} = 1-c\,, \qquad s_{ij} = 1-c\,, \qquad s_{jj} = c\,,$$

and all other elements zero.

It is shown in [19] that $P \preccurlyeq Q$ if and only if (4) holds for some (more complex) doubly stochastic matrix S.

If $P \prec Q$ then P will be regarded as *more uncertain* or *noisier* or *less informative* than Q. With the majorisation interpretation of noise, low noise is synonymous with high concentration of probability on a small number of events whereas high noise corresponds to a spread of probability across a large number of events.

The first two class distributions in universe 1 are related by majorisation:

$$(0.5, 0.5, 0, 0, 0, 0, 0, 0, 0) \preccurlyeq (0.55, 0, 0, 0.45, 0, 0, 0, 0)$$

since a single transfer of 0.05 from 0.55 to 0.45 produces a permutation of the left-hand side.

A more useful reformulation of the definition of majorisation can be given in terms of partial sums of the sorted probabilities in each vector (see [19]). Let $(p_{[1]}, \ldots, p_{[n]})$ denote the decreasing rearrangement of P, i.e. $p_{[1]} \geqslant \cdots \geqslant p_{[n]}$, then $P \preccurlyeq Q$ holds if and only if

$$q_{[1]} + \cdots + q_{[k]} \geqslant p_{[1]} + \cdots + p_{[k]} \tag{5}$$

for all k, $1 \leqslant k \leqslant n$. This is the most suitable form for computation.

The following useful properties of majorisation ($P \preccurlyeq Q$) are immediate consequences of (5):

(1) $\max(P) \leq \max(Q)$ where max is the maximum function.
(2) The size of the support of P, i.e. the number of non-zero probabilities, must be at least as great as that of Q.
(3) The smallest non-zero probability in P must be at least as large as that in Q.

It can also be seen from (5) that majorisation is defined as "modulo zero probabilities": distributions $(0.3, 0.7)$ and $(0.3, 0.7, 0, 0, 0)$ are equally noisy. (Since, in this application, noise will usually be modelled over a fixed number of classes this point is of little consequence.) Amongst n classes the uniform distribution $(1/n, \ldots, 1, n)$ is the most noisy and is majorised by every other distribution on n classes. At the other end of the scale any degenerate distribution is least noisy in the sense that it majorises every other distribution.

If classes in a universe are combined, the resulting class distribution in the class model majorises the original since combining probabilities amounts to the reverse of an equality transfer, i.e. an *inequality* transfer.

A concrete example of the use of majorisation is provided by the commonly used device of introducing noise through inversion of classes with a known probability. This is the mechanism in the classification noise process of Angluin and Laird [3]. In a two-class problem suppose $Q = (q_1, q_2)$ is a class distribution. Suppose further that although a class is generated according to Q it is then inverted with probability α. This results in a new class distribution $P = (p_1, p_2)$ where

$$p_1 = (1-\alpha)q_1 + \alpha q_2\,, \qquad p_2 = \alpha q_1 + (1-\alpha)q_2\,, \tag{6}$$

which is just (3). Thus $P \preccurlyeq Q$. More generally, in an n-class problem, the inversion probability, α, can be split equally amongst the remaining $n-1$ classes. This generalizes (6) to (4) where S has diagonal elements $(1-\alpha)$ and off-diagonal elements $\alpha/(n-1)$ and so again $P \preccurlyeq Q$.

4.2. *Increasing noise in universe* 1

To illustrate the use of majorisation, universe 1 will be made noisier (thereby creating universe 2) by de-majorising each class distribution. This will be achieved by leaving the majority class probability in each distribution unchanged and spreading the remaining probability more evenly over the other classes instead of it being concentrated on a single class as is the case in universe 1. This is the strategy that was employed in [14]. The joint distribution of the description attributes will not be altered.

The new class distributions are shown in Table 3 alongside those for universe 1. Notice that only the distribution for rule 11 is unchanged—there is no residual probability to re-distribute here. In the other cases the complement of the majority class probability has been spread fairly evenly across most of the other classes.

Table 3
Class distributions for universes 1 and 2

Rule number	Universe 1	Universe 2
1	(0.5, 0.5, 0, 0, 0, 0, 0, 0)	(0.5, 0.1, 0.1, 0, 0, 0.1, 0.1, 0.1)
2	(0.55, 0, 0, 0.45, 0, 0, 0, 0)	(0.55, 0.05, 0.1, 0.1, 0.05, 0, 0.1, 0.05)
3	(0, 0, 0.6, 0, 0, 0, 0, 0.4)	(0.1, 0.1, 0.6, 0, 0.05, 0.05, 0.05, 0.05)
4	(0, 0, 0, 0, 0, 0, 0.4, 0.6)	(0.1, 0.1, 0.05, 0.05, 0, 0.05, 0.05, 0.6)
5	(0, 0, 0, 0.3, 0, 0, 0.7, 0)	(0.05, 0.05, 0.05, 0.05, 0, 0.05, 0.7, 0.05)
6	(0, 0, 0, 0, 0, 0.55, 0.45, 0)	(0, 0.05, 0.05, 0.05, 0.05, 0.55, 0.15, 0.1)
7	(0, 0, 0.4, 0, 0.6, 0, 0, 0)	(0.1, 0.05, 0.05, 0.1, 0.6, 0, 0.05, 0.05)
8	(0.3, 0.7, 0, 0, 0, 0, 0, 0)	(0.1, 0.7, 0.05, 0.05, 0, 0.05, 0.05, 0)
9	(0, 0, 0, 0.6, 0, 0.4, 0, 0)	(0.1, 0.05, 0.05, 0.6, 0, 0.1, 0.05, 0.05)
10	(0, 0, 0.6, 0, 0, 0.4, 0, 0)	(0.05, 0.05, 0.6, 0.05, 0.1, 0.05, 0.05, 0.05)
11	(0, 0, 0, 0, 0, 1, 0, 0)	(0, 0, 0, 0, 0, 1, 0, 0)
12	(0, 0, 0, 0.35, 0, 0.65, 0, 0)	(0.05, 0.05, 0, 0.05, 0, 0.65, 0.1, 0.1)
13	(0, 0, 0, 0, 0.5, 0, 0, 0.5)	(0.1, 0.1, 0.05, 0.05, 0.5, 0.1, 0.05, 0.05)
14	(0.4, 0, 0, 0.6, 0, 0, 0, 0)	(0.1, 0.05, 0.05, 0.6, 0.05, 0.05, 0.05, 0.05)
15	(0, 0.6, 0, 0, 0, 0.4, 0, 0)	(0.05, 0.6, 0.05, 0.05, 0.05, 0.05, 0.05, 0.1)
16	(0, 0, 0, 0, 0, 0.5, 0.5, 0)	(0.1, 0.1, 0.1, 0.1, 0, 0.05, 0.5, 0.05)

Universe 2 has a different default class distribution. It is

$$(0.1364, 0.1699, 0.0594, 0.1378, 0.0531, 0.2124, 0.1534, 0.0778)\,, \tag{7}$$

whereas that for universe 1, given in (2), is:

$$(0.1215, 0.1510, 0.0400, 0.2115, 0.0445, 0.2675, 0.1270, 0.0370)\,.$$

The default class is still class 6 but its probability has dropped by over 5%. It is not generally the case, though, that increasing noise will reduce the probability of the default class. In the next section, universe 2 will be altered slightly to produce an increase in default class probability over that of universe 1.

4.3. Measures of information and uncertainty

As a step towards defining these measures a function which is monotone w.r.t. majorisation is needed:

Definition 4.2. A continuous real-valued function, ϕ, on the space of finite discrete probability distributions is *Schur-convex* if $\phi(P) \leqslant \phi(Q)$ whenever $P \preccurlyeq Q$. If $\phi(P) < \phi(Q)$ whenever $P \prec Q$ then ϕ is *strictly Schur-convex*. If ϕ is (strictly) Schur-convex then $-\phi$ is (*strictly*) *Schur-concave*.

Schur-convex functions are necessarily symmetric. Ordinary convexity together with symmetry implies Schur-convexity (see [19]). Although not essential in an arbitrary Schur-convex function, convexity has an important meaning for information measures as it guarantees (and is in fact equivalent to) the property, familiar from entropy, that expected conditional information is always at least as

great as that of an unconditional distribution. Expressed in terms of random variables X and Y this property is:

$$\phi(X \mid Y) \geq \phi(X) . \tag{8}$$

If ϕ in (8) is strictly convex then equality occurs if and only if X and Y are independent. Accordingly, the following definition is made:

Definition 4.3. A real-valued continuous function, ϕ, on the space of discrete probability distributions is a *measure of information* (*uncertainty*) if it is symmetric and strictly convex (concave). If the function lacks strictness, the measure is said to be *weak*.

Entropy and the Gini index are both symmetric and strictly concave and are therefore measures of uncertainty. The maximum function, $\max(P)$, is a weak information measure: its convexity is not strict. The commonly-used error or miss-classification function, $error(P) = 1 - \max(P)$ is a weak measure of uncertainty.

The development given above can be extended to continuous distributions where the idea of majorisation carries over in a natural way (see [13]). The observation that strict convexity was desirable in a measure of information was also made by Breiman et al. [5] in the context of selection measures for the CART learning algorithm.

All information measures render the same ordering of informativeness between two distributions which are related by majorisation (in fact majorisation can be defined in terms of this property [19]). On the other hand if two distributions are not related by majorisation then it is always possible to find two information measures which will order them differently. Thus majorisation between distributions is a stronger condition than "has greater information" as assessed by a real-valued measure. The latter inequality, without majorisation holding, could be just an artefact of the measure used and not indicative of any material difference in information. It is for this reason that it is generally better to manipulate the amount of noise using majorisation itself rather than, say, by increasing entropy.

4.4. Information in a universe

Measures of information are useful, however, for providing overall summaries of the information content of a universe, that is, information about class.

Definition 4.4. The *rule information* in a universe with respect to an information measure is its expectation over the class distributions in the class model of the universe. The *default information* is that of the default distribution with respect to the measure. The *information gain* is:

| rule information − default information | .

A dual definition holds for uncertainty measures. By virtue of the strict convexity in Definition 4.3 the information gain is zero if and only if the universe is uninformative.

The rule information assesses the contribution of the attributes to identifying class. The benefit of the attributes is, however, relative to the default information as provided by the information gain. Care must be taken when comparing universes in terms of their noise content using rule information. Suppose in a universe all class distributions are permutations of one distribution, P, but not all identical. The uninformative universe with all class distributions equal to P has the same rule information but zero information gain.

When the information measure is the max function, the default and rule informations have interpretations as classification rates:

Definition 4.5. When the information measure is max, the default and rule informations are called the *default* and *universe classification rates* (*DCR* and *UCR*) respectively.

Increasing the noise in a class distribution in any representation of the universe can never increase the rule information (and will decrease it unless the information measure is weak). In particular the UCR can only decrease. The information gain, though, may increase or decrease.

Table 4 shows information statistics, using entropy and max, for universes 1 and 2. It is clear that the description attributes facilitate the identification of class as there is substantial information over the default. Also shown is the information about class provided by each attribute on its own. For both universes, b is the most informative attribute as judged by both entropy and max. No single attribute, however, can classify satisfactorily on its own. Because universe 2 was

Table 4
Information statistics relating to entropy and max for universes 1 and 2; ranks of each attribute are shown in brackets

Information	Universe 1		Universe 2	
	Entropy	Max	Entropy	Max
Default	1.8742	0.2675	1.9847	0.2124
Rules	0.5976	0.6540	1.1849	0.6540
Gain	1.2766	0.3865	0.7998	0.4416
Attributes:				
a	1.6728 (3)	0.3622 (2)	1.8853 (2)	0.2922 (2)
b	1.3936 (1)	0.4250 (1)	1.6761 (1)	0.3649 (1)
c	1.6420 (2)	0.3395 (3)	1.8921 (3)	0.2316 (3)
d	1.8526 (4)	0.2675 (4)	1.9728 (4)	0.2151 (4)
e	1.8742 (5)	0.2675 (5)	1.9847 (5)	0.2124 (5)

constructed from universe 1 by leaving the majority probability in each class distribution unaltered, its UCR of 65.4% is the same as that for universe 1.

Even a small universe such as this can provide a substantial task for learning algorithms. Holte [15] has observed that many of the real data sets in the "Machine Learning Repository" [21] appear to possess a single attribute which is very informative with the others contributing little. With the method of artificial universes it is comparatively simple to build a small manageable universe and yet in which several attributes are needed for effective classification.

5. Evaluating learning

The result of concept learning from examples will usually be a mechanism for classification, i.e. a means of deciding, given *any* description vector, what the corresponding class is. Some learning algorithms may produce a non-deterministic classification, i.e. offer a choice of possible classes with an indication of the degree of uncertainty attached to each. For example a trained neural network may give the strengths of its output units or a decision tree may have relative frequencies for observed classes in its leaves. Some rule induction algorithms such as CN2 (Clark and Boswell [7]) induce overlapping rules. The user of such a system can decide on a rule for selection of a particular class thus creating a deterministic classifier. It will be assumed here that classifiers produce a definite class on each occasion.

Definition 5.1. A (*deterministic*) *classifier* is a mapping from the set of description vectors to the set of classes. A statement of this mapping is called a *representation* of the classifier.

The universe itself provides a best classifier, i.e. one which has the optimal probability of correct classification.

Definition 5.2. A *best classifier* associated with a universe maps each description vector to a *majority class* in its associated class distribution in the class model, i.e. one which has maximum probability in this distribution. A *default classifier* for a universe assigns each description vector a majority class in the default distribution.

Because of ties in the class distribution, the best and the default classifiers may not be unique.

Concept learning can be characterised as estimation of the best classifier from examples. With some forms of learning from examples, though, such as neural networks and instance-based learning, the classifier obtained is implicit rather than explicit. To obtain an extensional form of the classifier all possible description vectors must be supplied to it and the resulting classifications noted. For small universes with discrete attributes this does not present a problem. For

very large cases or if continuous attributes are involved this would present difficulties. One possibility here is to estimate the classifier using a second stage of learning. A large sample of description vectors are supplied to the classifier for classification. The resulting (noiseless) example set is then passed to a learning algorithm which generates rules. The rule set obtained is an approximation to the true classifier.

Evaluating an extensionally available classifier is straightforward. The classifier provides a set of deterministic rules of the form

if ⟨complex⟩ then ⟨class⟩

from which, using the universe specification, the classification rate can be calculated.

Definition 5.3. The probability that an induced classification rule correctly classifies a randomly selected example from the universe which satisfies its condition is called the *actual classification probability* of the rule. The expected value of the actual classification probability in a classifier is its *actual classification rate* (*ACR*).

The ACR of a classifier can never exceed the UCR (Definition 4.5) of the universe. It can, though, if the classifier is sufficiently bad, be less than the DCR, which is the rate for the default classifier. The UCR is the classification rate for the best classifier.

An ACR associated with a classifier learned from data randomly generated from a universe is a random variable. The expected ACR over example sets of a particular size is a useful indicator of the effectiveness of the learning algorithm. The relationship between this expected ACR and size of example set is the *learning curve*. The expected ACR for an example set size, while difficult to compute for algorithms such as ID3, can be estimated from a number of trials of learning.

5.1. Experiments with ID3 on universes 1 and 2

To examine the effect of increased noise on learning with ID3, a number of trials were performed using universes 1 and 2. Each trial consisted of generating an example set of a particular size, inducing a tree with ID3, producing the deterministic classifier obtained by adopting the majority class in each leaf of the tree (a mild form of pruning) and computing, from the probabilities in the universe, the ACR of the classifier.

To estimate expected ACRs and the learning curve, a number of trials were carried out for a range of example set sizes from 5 to 5000. Because of the high variability in ACR obtained from small sizes, a large number of replications were carried out for these (4000 for size 5) with the number of replications decreasing as the size increased (down to 25 for a size of 5000).

The results of these experiments are shown in Table 5 and the learning curves

Table 5
Estimates of expected ACR, expressed as a percentage, for rules sets induced by ID3 from example sets generated by universes 1 and 2; estimated standard errors are also given as percentages

Size	No. of trials	Universe 1	Universe 2
		Est. expected ACR (Est. standard error) DCR = 26.75%, UCR = 65.4%	Est. expected ACR (Est. standard error) DCR = 21.24%, UCR = 65.4%
5	4000	29.2 (0.12)	23.5 (0.11)
8	4000	34.8 (0.13)	27.8 (0.13)
10	2000	37.5 (0.18)	30.0 (0.18)
12	2000	39.8 (0.17)	31.7 (0.18)
15	2000	42.1 (0.16)	33.6 (0.18)
20	1000	45.1 (0.20)	35.7 (0.24)
30	1000	48.2 (0.16)	39.1 (0.21)
50	500	51.0 (0.16)	42.5 (0.24)
75	500	52.4 (0.12)	44.4 (0.18)
100	200	53.6 (0.16)	46.6 (0.25)
200	100	56.0 (0.19)	51.2 (0.29)
300	100	57.3 (0.20)	53.8 (0.23)
500	100	59.1 (0.13)	57.7 (0.14)
750	100	60.7 (0.09)	60.4 (0.11)
1000	100	61.6 (0.10)	61.6 (0.08)
1250	100	62.2 (0.09)	62.5 (0.07)
1500	50	62.4 (0.11)	63.1 (0.07)
2000	50	63.1 (0.09)	63.9 (0.06)
3000	50	63.9 (0.08)	64.6 (0.04)
5000	25	64.5 (0.06)	64.9 (0.04)

up to size 100 are displayed in Fig. 1. The expected ACR for universe 1 is consistently above that for universe 2 until about size 750 showing that the increased noise in universe 2 presents considerable difficulties for ID3. For larger example set sizes, universe 2 appears to be slightly better with the differences being statistically significant for sizes of 1500 and above. The explanation for this

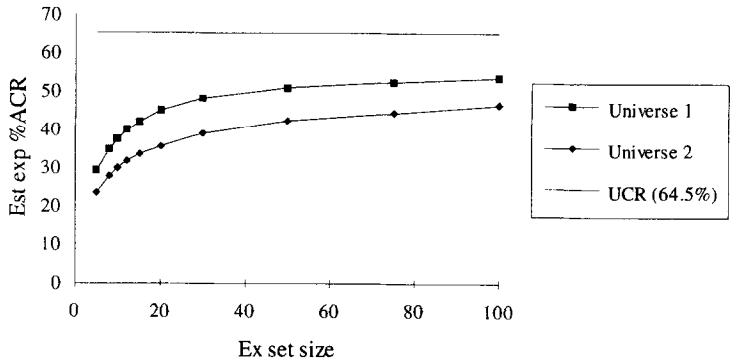

Fig. 1. ID3 learning curves for example sets of up to size 100 from universes 1 and 2.

Table 6
Class distributions for universe 3

Rule number	Class distribution
1	(0.5, 0.1, 0.05, 0, 0, 0.2, 0.05, 0.1)
2	(0.55, 0.05, 0.1, 0.05, 0.05, 0.15, 0, 0.05)
3	(0.05, 0.05, 0.6, 0, 0.05, 0.15, 0.05, 0.05)
4	(0.05, 0.05, 0.05, 0.05, 0, 0.15, 0.05, 0.6)
5	(0.05, 0.05, 0.05, 0.05, 0, 0.05, 0.7, 0.05)
6	(0, 0.05, 0.05, 0.05, 0.05, 0.55, 0.15, 0.1)
7	(0.05, 0.05, 0.05, 0, 0.6, 0.15, 0.05, 0.05)
8	(0.05, 0.7, 0.05, 0.05, 0, 0.1, 0.05, 0)
9	(0.05, 0.05, 0.05, 0.6, 0, 0.15, 0.05, 0.05)
10	(0.05, 0.05, 0.6, 0.05, 0.05, 0.1, 0.05, 0.05)
11	(0, 0, 0, 0, 0, 1, 0, 0)
12	(0.05, 0.05, 0, 0.05, 0, 0.65, 0.1, 0.1)
13	(0.1, 0.05, 0.05, 0.05, 0.5, 0.15, 0.05, 0.05)
14	(0.05, 0.05, 0.05, 0.6, 0.05, 0.1, 0.05, 0.05)
15	(0.05, 0.6, 0.05, 0.05, 0.05, 0.1, 0.05, 0.05)
16	(0.05, 0.1, 0.05, 0.05, 0, 0.2, 0.5, 0.05)

may be that the single minority class in universe 1 is emerging on occasions as the majority in the data—something that is unlikely to happen in universe 2 with its small residual probabilities.

It might be argued that the better expected ACRs obtained for universe 1 for smaller sample sizes are due to a superior default classification rate (DCR): from Table 4 the DCR for universe 1 is 26.75% whereas for universe 2 it is 21.24% so that universe 1 has a considerable head start on universe 2. To investigate this, universe 2 was modified slightly to produce a DCR close to that of universe 1. This was achieved by transferring a small amount of probability to the majority class (class 6) in the class distribution of most rules.

The new distributions which, together with the joint distributions of the description attributes, define universe 3 are shown in Table 6. The information statistics for universe 3 are shown in Table 7. The default distribution for universe

Table 7
Information statistics for universe 3; ranks of each attribute are shown in brackets

Information	Entropy	Max
Default	1.9453	0.2705
Rules	1.1692	0.6540
Gain	0.7761	0.3835
Attributes:		
a	1.8391 (2)	0.3293 (2)
b	1.6217 (1)	0.3825 (1)
c	1.8412 (3)	0.2736 (3)
d	1.9326 (4)	0.2705 (4)
e	1.9453 (5)	0.2705 (5)

Table 8
Estimates of expected ACR, expressed as a percentage, for rules sets induced by ID3 from example sets generated by universes 1 and 3; estimated standard errors are also given as percentages

Size	No. of trials	Universe 1	Universe 3
		Est. expected ACR (Est. standard error) DCR = 26.75%, UCR = 65.4%	Est. expected ACR (Est. standard error) DCR = 27.05%, UCR = 65.4%
5	4000	29.2 (0.12)	24.8 (0.12)
8	4000	34.8 (0.13)	29.2 (0.13)
10	2000	37.5 (0.18)	31.3 (0.18)
12	2000	39.8 (0.17)	33.0 (0.18)
15	2000	42.1 (0.16)	35.0 (0.17)
20	1000	45.1 (0.20)	36.7 (0.23)
30	1000	48.2 (0.16)	40.3 (0.19)
50	500	51.0 (0.16)	43.5 (0.22)
75	500	52.4 (0.12)	45.7 (0.19)
100	200	53.6 (0.16)	47.4 (0.22)
200	100	56.0 (0.19)	51.8 (0.24)
300	100	57.3 (0.20)	55.3 (0.16)
500	100	59.1 (0.13)	58.0 (0.14)
750	100	60.7 (0.09)	60.5 (0.10)
1000	100	61.6 (0.10)	62.0 (0.08)
1250	100	62.2 (0.09)	62.7 (0.07)
1500	50	62.4 (0.11)	63.3 (0.08)
2000	50	63.1 (0.09)	64.0 (0.07)
3000	50	63.9 (0.08)	64.6 (0.03)
5000	25	64.5 (0.06)	65.0 (0.02)

3 is

$$(0.1129, 0.1666, 0.0551, 0.1245, 0.0525, 0.2705, 0.1419, 0.0760)$$

and thus the DCR is 27.05%, slightly above that for universe 1 which is 26.75%.

Trials similar to those described above for universes 1 and 2 were carried out to estimate the ID3 learning curve for universe 3. The results are shown in Table 8 (alongside those for universe 1 for comparison). It can be seen that the pattern for universe 3 follows that for universe 2: ACRs are well below those for universe 1 for sample sizes up to 750 and then universe 3 creeps ahead by a small but statistically significant amount. Thus the results for universe 3 are comparable with those from universe 2 to which it is similar in all regards except for DCR. The lower ACRs for universes 2 and 3, therefore, appear to be caused by the de-majorising of the class distributions.

6. Conclusion

The method of artificial universes has the class model as its central notion. By focusing on the class distributions of the model as the means of describing noise

together with the inclusion of redundant and irrelevant attributes, the process of setting up learning tasks of varying degrees of difficulty is greatly facilitated. Smallish universes with 5 to 10 attributes can provide quite challenging tasks yet remain comprehensible to the experimenter.

The majorisation relation affords a simple practical way of manipulating noise levels and through its link with information and uncertainty measures gives a unified account of several seemingly different indicators.

It would appear from experimental results that, at least for ID3, increasing noise by de-majorising class distributions leads to poorer learning curves.

References

[1] D.W. Aha, Generalising from case studies: a case study, in: D. Sleeman and P. Edwards, eds., *Proceedings Ninth International Conference on Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1992) 1–10.

[2] D.W. Aha, D. Kibler and M. Albert, Instance-based learning algorithms, *Mach. Learning* **6** (1991) 37–66.

[3] D. Angluin and P. Laird, Learning from noisy examples, *Mach. Learning* **2** (1988) 343–370.

[4] L.B. Booker, D.A. Goldberg and J.H. Holland, Classifier systems and genetic algorithms, *Artif. Intell.* **40** (1989) 235–282.

[5] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984).

[6] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor and D. Freeman, AUTOCLASS: a Bayesian classification system, in: *Proceedings Fifth International Conference on Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1988) 54–64.

[7] P. Clark and R. Boswell, Rule induction with CN2: some recent improvements, in: Y. Kodratoff, ed., *EWSL-91* (Springer-Verlag, Berlin, 1991) 151–163.

[8] P. Clark and T. Niblett, The CN2 induction algorithm, *Mach. Learning* **3** (1989) 261–283.

[9] S. Cost and S. Salzberg, A weighted nearest neighbour algorithm for learning with symbolic features, *Mach. Learning* **10** (1993) 57–78.

[10] D. Haussler, Quantifying inductive bias: AI learning algorithms and Valiant's learning framework, *Artif. Intell.* **36** (1986) 177–221.

[11] R.J. Hickey, A note on the measurement of randomness, *J. Appl. Prob.* **19** (1982) 229–232.

[12] R.J. Hickey, Majorisation, randomness and some discrete distributions, *J. Appl. Prob.* **20** (1983) 897–902.

[13] R.J. Hickey, Continuous majorisation and randomness, *J. Appl. Prob.* **21** (1984) 924–929.

[14] R.J. Hickey, Artificial universes: towards a systematic approach to evaluating algorithms which learn from examples, in: D. Sleeman and P. Edwards, eds., *Proceedings Ninth International Conference on Machine Learning* (Morgan Kaufman, San Mateo, CA, 1992) 196–205.

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

[16] D. Kibler and P. Langley, Machine learning as an experimental science, in: D. Sleeman, ed., *Proceedings Third European Working Session on Learning* (Pitman, Glasgow, Scotland, 1988) 81–92.

[17] P. Langley, W. Iba and K. Thompson, An analysis of Bayesian classifiers, in: *Proceedings AAAI-92*, San Jose, CA (MIT Press, Cambridge, MA, 1992) 223–228.

[18] H Lounis and G.M. Bisson, Evaluation of learning systems: an artificial databased approach, in: Y. Kodratoff, ed., *EWSL-91* (Springer-Verlag, Berlin, 1991) 463–481.

[19] A.W. Marshall and I. Olkin, *Inequalities: The Theory of Majorisation and its Applications* (Academic Press, New York, 1979).

[20] J. Mingers, An empirical comparison of pruning methods for decision tree induction, *Mach. Learning* **4** (1989) 227–243.
[21] P.M. Murphy and D.W. Aha, UCI repository of machine learning databases, Maintained at the Department of Information and Computer Science, University of California, Irvine, CA (1992).
[22] M.J. Pazzini and W. Sarrett, A framework for average case analysis of conjunctive learning algorithms, *Mach. Learning* **9** (1992) 349–372.
[23] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).
[24] J.R. Quinlan, Induction of decision trees, *Mach. Learning* **1** (1986) 81–106.
[25] J.R. Quinlan, Simplifying decision trees, *Int. J. Man-Mach. Stud.* **27** (1987) 221–234.
[26] D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning internal representations by error propagation, in: D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing* **1** (MIT Press, Cambridge, MA, 1986) 318–362.
[27] Y. Sakakibara, Noise-tolerant occam algorithms and their applications to learning decision trees, *Mach. Learning* **11** (1993) 37–62.
[28] C. Schaffer, Sparse data and the effect of overfitting avoidance in decision tree induction, in: *Proceedings AAAI-92*, San Jose, CA (MIT Press, Cambridge, MA, 1992) 147–152.
[29] C. Schaffer, Overfitting avoidance as bias, *Mach. Learning* **10** (1993) 153–178.
[30] C.E. Shannon and W. Weaver, *Mathematics of Communication Theory* (University of Illinois Press, Urbana, IL, 1964).
[31] L.G. Valiant, A theory of the learnable, *Commun. ACM* **27** (1984) 1134–1142.
[32] L.G. Valiant, Learning conjunctions of disjunctions, in: *Proceedings IJCAI-85*, Los Angeles, CA (Morgan Kaufmann, San Mateo, CA, 1985) 560–566.