Machine-Learning Applications of Algorithmic Randomness

Volodya Vovk, Alex Gammerman, Craig Saunders
Computer Learning Research Centre and Department of Computer Science
Royal Holloway, University of London, Egham, Surrey TW20 0EX, England
{vovk,alex,craig}@dcs.rhbnc.ac.uk

Abstract

Most machine learning algorithms share the following drawback: they only output bare predictions but not the confidence in those predictions. In the 1960s algorithmic information theory supplied universal measures of confidence but these are, unfortunately, non-computable. In this paper we combine the ideas of algorithmic information theory with the theory of Support Vector machines to obtain practicable approximations to universal measures of confidence. We show that in some standard problems of pattern recognition our approximations work well.

1 INTRODUCTION

Two important differences of most modern methods of machine learning (such as statistical learning theory, see Vapnik [21], 1998, or PAC theory) from classical statistical methods are that:

- machine learning methods produce bare predictions, without estimating confidence in those predictions (unlike, eg, prediction of future observations in traditional statistics (Guttman [5], 1970));
- many machine learning methods are designed to work (and their performance is analysed) under the general iid assumption (unlike the classical parametric statistics) and they are able to deal with extremely high-dimensional hypothesis spaces; cf Vapnik [21] (1998).

In this paper we will further develop the approach of Gammerman et al [4] (1998) and Saunders et al [17] (1999), where the goal is to obtain confidences for predictions under the general iid assumption in high-dimensional situations. Figure 1 demonstrates the desirability of confidences. The main contribution of this paper is embedding the approaches of Gammerman et al [4] (1998) and Saunders et al [17] (1999) into a general scheme based on the notion of algorithmic randomness.

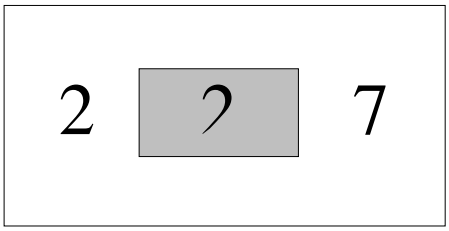


Figure 1: If the training set only contains clear 2s and 7s, we would like to attach much lower confidence to the middle image than to the right and left ones

As will become clear later, the problem of assigning confidences to predictions is closely connected to the problem of defining random sequences. The latter problem was solved by Kolmogorov [8] (1965), who based his definition on the existence of the Universal Turing Machine (though it became clear that Kolmogorov's definition does solve the problem of defining random sequences only after Martin-Löf's paper [15], 1966); Kolmogorov's definition moved the notion of randomness from the grey area surrounding probability theory and statistics to mathematical computer science.

Kolmogorov believed his notion of randomness to be a suitable basis for *applications* of probability. Unfortunately, the fate of this idea was different from Kolmogorov's 1933 axioms (Kolmogorov [7], 1933), which

are universally accepted as the basis for the *theory* of probability. The algorithmic notion of randomness has mainly remained of a purely mathematical interest and has not become the leading paradigm in statistics or machine learning. Some of the reasons why this happened are:

- algorithmic measures of randomness are non-computable;
- little work has been done on computable approximations to Kolmogorov's randomness (one of the exceptions is Longpré [14], 1992);
- the algorithmic theory of randomness has been mainly concerned with the case of binary sequences, which is far too restrictive for any practical applications.

Remark 1 It is interesting that the situation with the notion of Kolmogorov complexity is somewhat different, despite the fact that the notions of Kolmogorov complexity and randomness are extremely closely connected[1]: despite being non-computable, Kolmogorov complexity inspired the MDL and MML principles [16, 26] (and their generalization, Complexity Approximation Principle [24]), which have many practical applications. (It should be noted, however, that algorithmic randomness has been used in the discussions of the MDL principle; see Li and Vitanyi [13], 1997, and [12], 1995.)

The main advantage of Kolmogorov's notion of randomness in comparison with the earlier definitions (eg, von Mises's) is that it is applicable to finite sequences and that it provides degrees of randomness; this is its crucial feature which makes practical applications possible. Later Kolmogorov's definition was developed by, among others, Martin-Löf [15] (1966), Levin [11] (1973) and Gács [2] (1980).

The main goal of this paper is to study computable approximations to algorithmic randomness and to apply those approximations to some benchmark datasets. The main technical tool will be Vapnik's [21] (1998) theory of Support Vector machines, but in principle it is possible to find useful approximations based on other techniques, such as ridge regression (see, eg, Saunders et al [18], 1998). The approach of the algorithmic theory of randomness, as presented in this paper, provides a unified view of the results in Gammerman et al [4] (1998) and Saunders et al [17] (1999).

For excellent reviews of algorithmic information theory, see V'yugin [25] (1994) and Li and Vitanyi [13] (1997).

[1]For example, one of the definitions of randomness deficiency of a finite binary sequence is the length of this sequence minus its Kolmogorov complexity.

2 ALGORITHMIC THEORY OF RANDOMNESS

Typically we will be interested in randomness of a sequence $z = (z_1, \dots, z_n)$ of elements $z_i \in Z$ of some *sample space* Z. (In the traditional algorithmic information theory $z_i \in \{0, 1\}$; in our typical applications z is a sequence

$$(x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, y_{1+1}), \dots, (x_{l+k}, y_{l+k})$$

containing the training classified examples and the test examples with their provisional classifications; l is the number of training examples and k is the number of test examples.) We will assume that Z is equipped with some computability structure which allows us to speak of, say, computable functions on Z (but we do not assume that Z is, say, discrete).

We will consider two kinds of randomness deficiency, which we call Martin-Löf deficiency (this is a universal version of the standard statistical notion of p-values) and Levin deficiency (which is close to being a universal version of Bayes factors). Let $\mathcal{P} = \mathcal{P}_1, \mathcal{P}_2, \dots$ be a sequence of statistical models such that, for every $n = 1, 2, \dots$, $\mathcal{P}_n$ is a set of probability distributions in Z^n. In this paper we will only be interested in specific $\mathcal{P}$ (namely, the iid models and exchangeability models) which are computable, and our definitions of randomness deficiency will only make intuitive sense for computable $\mathcal{P}$.

We say that a function $t : Z^* \to \overline{\mathbb{N}}$ (where $\overline{\mathbb{N}}$ is the set $\mathbb{N} = \{0, 1, \dots\}$ of nonnegative integers extended by adding the infinity ∞) is a *Martin-Löf test for $\mathcal{P}$-randomness* if

1. for all $n \in \mathbb{N}$ and $m \in \mathbb{N}$ and all $P \in \mathcal{P}_n$,
$$P\{z \in Z^n : t(z) \geq m\} \leq 2^{-m};$$
2. t is semicomputable from below, in the sense that there exists a computable sequence of computable functions $t_i : Z^* \to \mathbb{N}$, $i = 1, 2, \dots$, such that $t(z) = \sup_i t_i(z)$ for all $z \in Z^*$.

Intuitively, a test for randomness is a device for finding unusual features in the data $z \in Z^n$. Item 1 says that

the amount of unusual features is measured in bits (and every extra bit halves the amount of sequences exhibiting the unusual features); item 2 says that the device should be implementable on a computer (we are interested in universal tests for randomness, which find **all** unusual features in our data; therefore, our tests for randomness are allowed to work forever, all the time finding new regularities in the data; technically, they are only required to be semicomputable from below rather than computable).

A useful modification of the previous definition is where item 1 is strengthened as follows. We say that a function $t : Z^* \to \overline{\mathbb{Z}}$ (where $\overline{\mathbb{Z}}$ is the set of all integers $\mathbb{Z}$ extended by adding the infinity ∞) is a *Levin test for $\mathcal{P}$-randomness* if

1. for all $n \in \mathbb{N}$ and all $P \in \mathcal{P}_n$,
$$\int_{Z^n} 2^{t(z)} P(dz) \le 1;$$
2. t is semicomputable from below.

(To see that item 1 of this definition implies item 1 of the previous definition, apply Markov's inequality.)

We will say that a Martin-Löf test for $\mathcal{P}$-randomness T is *universal* if it is largest to within an additive constant, in the sense that for any other Martin-Löf test for $\mathcal{P}$-randomness t there exists a constant C such that, for all $z \in Z^*$, $T(z) \ge t(z) - C$. Analogously, a largest to within an additive constant Levin test for $\mathcal{P}$-randomness will also be called *universal*.

Lemma 1 (Kolmogorov, Martin-Löf, Levin) *If $\mathcal{P}$ is computable, there exist a universal Martin-Löf test for $\mathcal{P}$-randomness and a universal Levin test for $\mathcal{P}$-randomness.*

For every computable $\mathcal{P}$ we will fix some universal Martin-Löf test for $\mathcal{P}$-randomness $d^{\mathrm{ML}}_{\mathcal{P}}$ and some universal Levin test for $\mathcal{P}$-randomness $d^{\mathrm{L}}_{\mathcal{P}}$; the value $d^{\mathrm{ML}}_{\mathcal{P}}(z)$ will be called the *Martin-Löf $\mathcal{P}$-randomness deficiency* of z and the value $d^{\mathrm{L}}_{\mathcal{P}}(z)$ will be called the *Levin $\mathcal{P}$-randomness deficiency* of z.

The standard notions of the algorithmic theory of randomness use the "logarithmic scale" (in which randomness deficiency is defined to within an additive constant). In applications, it is often more convenient to use the "direct scale". Now we will reformulate some of the previous definitions in the "direct scale".

We say that a function $t : Z^* \to [0,1]$ is a *p-value function wr to $\mathcal{P}$* if

1. for all $n \in \mathbb{N}$, all $r \in [0,1]$ and all $P \in \mathcal{P}_n$,
$$P\{z \in Z^n : t(z) \le r\} \le r;$$
2. t is semicomputable from above, in the sense that there exists a computable sequence of computable functions $t_i : Z^* \to [0,1]$, $i = 1, 2, \ldots$, such that $t(z) = \inf_i t_i(z)$ for all $z \in Z^*$.

This definition is practically equivalent to the standard statistical notion: in practice, item 2 is completely irrelevant, since the p-value functions of any interest in applications of statistics are always computable.

Notice that:

- if t is a Martin-Löf test for $\mathcal{P}$-randomness, 2^{-t} is a p-value function wr to $\mathcal{P}$;
- if t is a p-value function wr to $\mathcal{P}$, $\lfloor -\log t \rfloor$ (in this paper, log always stands for the base 2 logarithm) is a Martin-Löf test for $\mathcal{P}$-randomness.

We will call the function $2^{-d^{\mathrm{ML}}_{\mathcal{P}}}$ the *Martin-Löf $\mathcal{P}$-randomness level*; this function is a smallest, to within a constant factor, p-value function wr to $\mathcal{P}$.

We say that a function $t : Z^* \to [0, \infty]$ is a *$\mathcal{P}$-lottery* if

1. for all $n \in \mathbb{N}$ and all $P \in \mathcal{P}_n$,
$$\int_{Z^n} t(z) P(dz) \le 1;$$
2. t is semicomputable from below.

The connection with the notion of Levin test of $\mathcal{P}$-randomness is obvious. We will call the function $2^{-d^{\mathrm{L}}_{\mathcal{P}}}$ the *Levin $\mathcal{P}$-randomness level*; the inverse $2^{d^{\mathrm{L}}_{\mathcal{P}}}$ of this function is a largest, to within a constant factor, $\mathcal{P}$-lottery.

Remark 2 The difference between Martin-Löf and Levin randomness levels is analogous to the difference between Bayes factors and p-values. The latter is discussed in, eg, Schervish [19] (Section 4.6.2) and Vovk [23]. It is always possible (though not advisable in practice) to use Levin randomness level as Martin-Löf randomness level.

Since randomness deficiencies are defined to within an additive constant and randomness levels are defined to within a constant factor, the following notation will

turn out to be very useful: $=^+$, $\le^+$ and $\ge^+$ stand for the equality and inequalities to within an additive constant; $=^\times$, $\le^\times$ and $\ge^\times$ stand for the equality and inequalities to within a constant factor.

It turns out that Levin and Martin-Löf deficiencies are closely connected (cf the relation between plain and prefix complexity (3.4) in Li and Vitanyi [13], 1997):

Theorem 1 *For any computable $\mathcal{P}$ and z ranging over Z^*,*

$$d^{\mathrm{ML}}_{\mathcal{P}}(z) =^+ d^{\mathrm{L}}_{\mathcal{P}}(z \mid d^{\mathrm{ML}}_{\mathcal{P}}(z)).$$

(This theorem involves "conditional" variants of randomness deficiency; such variants are defined in a natural way.) This immediately implies

Corollary 1 *Levin and Martin-Löf randomness deficiency coincide to within* log.

Remark 3 Our definition of randomness is not the only possible: we can define "uniform randomness deficiency" and define, eg, the iid deficiency as the minimum of the deficiencies over all iid measures. The difference is not big when the sample space Z is compact; this follows from the following elaboration of Levin's [11] (1973) result: for any compact class $\mathcal{P}$ of probability measures,

$$d^{\mathrm{L}}_{\mathcal{P}}(x) =^+ \inf_{P\in\mathcal{P}} d^{\mathrm{L}}_{P}(x) \tag{1}$$

and

$$d^{\mathrm{ML}}_{\mathcal{P}}(x) =^+ \inf_{P\in\mathcal{P}} d^{\mathrm{ML}}_{P}(x). \tag{2}$$

However, in applications $\mathcal{P}$ is often not compact (although it is always "constructively closed"). It is an important problem to study whether (1) and (2) hold true for the iid distributions; intuitively, we would expect that a sequence is iid random if and only if it is random wr to some iid distribution.

3 THE IDEAL PREDICTION IN THE IID CASE

In this paper we will be mostly interested in $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2, \ldots)$ for which $\mathcal{P}_n$ is the set of all distributions P^n, P ranging over all probability distributions in Z. This $\mathcal{P}$ will be called the *iid model* and $\mathcal{P}$-randomness deficiency will be called *iid deficiency* and denoted d_{iid} (with upper index ML or L). From this point on we will be mainly interested in the iid deficiency and we will also use the less awkward expression "randomness deficiency" in place of "iid deficiency" (retaining the lower index "iid" in d_{iid}). This terminology agrees with that accepted in nonparametric statistics (see, eg, Fraser [1], 1957).

Remark 4 An important alternative to the iid model is the stationarity model. In this paper, however, we will restrict our attention to the iid assumption, which is more widely used in machine learning.

Suppose for a minute that the randomness deficiency (either Martin-Löf or Levin) is computable. Then our prediction problem will become trivial (if we accept the iid assumption and ignore computation time). Assuming we have training set $(x_1, y_1), \ldots, (x_l, y_l)$ and test set $x_{l+1}, \ldots, x_{l+k}$ and our goal is to predict the classifications $y_{l+1}, \ldots, y_{l+k}$ for $x_{l+1}, \ldots, x_{l+k}$, we can act as follows:

1. Consider all possible values $Y_1, \ldots, Y_k$ for labels $y_{l+1}, \ldots, y_{l+k}$ and compute (in practice, approximate from above) the randomness level of every possible completion
$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_1), \ldots, (x_{l+k}, Y_k).$$
2. Predict the set $Y_1, \ldots, Y_k$ corresponding to the completion with the largest randomness level.
3. Output as the *confidence* in this prediction one minus the second largest randomness level.
4. Output as the *credibility* the randomness level of the output prediction $Y_1, \ldots, Y_k$ (ie, the largest randomness level for all possible predictions).

To understand the intuition behind confidence, let us tentatively choose a conventional "significance level" such as 1%. If the confidence in our prediction exceeds 99% and the prediction is wrong, the actual data sequence belongs to an *a priori* chosen set of probability less than 1% (namely, the set of all data sequences with randomness level less than 1%).

Intuitively, low credibility means that either the training set is non-random or the test examples are not representative of the training set (say, in the training set we have images of digits and in the test set we have those of letters).

Remark 5 A common belief in algorithmic information theory is that terms of order $O(\log n)$ are not important in inequalities between complexities or randomness deficiencies for sequences of length n. It is interesting that, under our approach to prediction under the iid assumption, randomness deficiency of the

order of magnitude $\log n$ becomes the best we can realistically hope for in the problem of pattern recognition with one unclassified example. Indeed, suppose we have a training set

$$(x_1, y_1), \ldots, (x_l, y_l), \quad y_i \in \{-1, 1\}$$

and a new unclassified example x_{l+1}; we want to predict the label $y_{l+1} \in \{-1, 1\}$ of x_{l+1}. **Claim:** *If the true sample*

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1})$$

is random, the maximum randomness deficiency of the wrong sample

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, -y_{l+1})$$

is $\log(l+1)$ *(as usual, up to an additive constant).* To see why this is true, notice that if the true sample is random with respect to some iid distribution P^n, the wrong sample will have randomness deficiency at most $\log(l+1)$ wr to P^n corrupted by changing the label of every example with probability $1/(l+1)$. It is easy to see that this upper bound, $\log(l+1)$, is precise: if there is a very simple algorithm for finding y_i from x_i, the last example will be the "strangest" one in the wrong sample, and so the randomness deficiency of the wrong sample will jump to $\log(l+1)$.

4 PERMUTATION DEFICIENCY

In this section we consider a fundamental lower bound on randomness deficiency, which we call permutation deficiency. In practice we will only be able to find a lower bound on the randomness deficiency of some sequence z by first finding a lower bound L on the permutation deficiency of z and then using L as a lower bound on the randomness deficiency of z. We will see that the difference between permutation deficiency and randomness deficiency can be quite big, and it remains an open problem whether one can find easily computable and natural lower bounds for randomness deficiency that are not simultaneously lower bounds for permutation deficiency.

First we will define the exchangeability model (which is very popular in the foundations of Bayesian statistics; see, eg, Schervish [19], 1995). We say that a measure P on a product set Z^n is *exchangeable* if the distribution of the vector $z_1 \ldots z_n$ under P equals the distribution of the vector $z_{\pi(1)} \ldots z_{\pi(z)}$ for any permutation π on the set $\{1, \ldots, n\}$ (here $z_1, \ldots, z_n$ are the coordinate random variables). The *exchangeability model* is defined to be the sequence $\mathcal{P}_1, \mathcal{P}_2, \ldots$ of the following statistical models: every $\mathcal{P}_n$ is the set of all exchangeable distributions on Z^n. The $\mathcal{P}$-randomness deficiency, where $\mathcal{P}$ is the exchangeability model, will be called the *permutation deficiency*. The following theorem will give a more explicit representation of permutation deficiency, but before we can state it we will need several definitions. A *bag* is a set to each element of which is assigned a nonnegative integer called its *arity* (intuitively, how many times this element occurs in the bag). The *size* $|b|$ of a bag b is the sum of the arities of its elements. The *configuration* of a sequence $z = z_1 \ldots z_n$ is the bag which consists of all distinct elements in z, the arity of each element being the number of times it occurs in the sequence. For any sequence z, $\mathrm{conf}(z)$ stands for the configuration of z and $\Xi(z)$ stands for the set of all sequences of the same length and with the same configuration as z. If K is prefix complexity and C is plain Kolmogorov complexity (see Li and Vitanyi [13], 1997), the Martin-Löf and Levin deficiency of randomness of an element z of a set A can be defined as

$$d_A^{\mathrm{ML}}(z) = \log|A| - C(z \,|\, A), \; d_A^{\mathrm{L}}(z) = \log|A| - K(z \,|\, A),$$

respectively.

Theorem 2 *If z ranges over Z^*, $d^{\mathrm{ML}}_{\mathrm{exch}}(z) =^+ d^{\mathrm{ML}}_{\Xi(z)}(z)$ and $d^{\mathrm{L}}_{\mathrm{exch}}(z) =^+ d^{\mathrm{L}}_{\Xi(z)}(z)$.*

This theorem shows that Kolmogorov's [9] (1968) "Bernoulli sequences" are exactly the sequences with a small permutation deficiency in the binary case.

To establish a relation between randomness deficiency and permutation deficiency we will use the notion of randomness deficiency of a bag. Let $\mathcal{P}$ be the iid model. We will say that a test for $\mathcal{P}$-randomness (either Martin-Löf or Levin) is *bag-invariant* if it takes the same value for any two sequences with the same configuration. There exists a universal bag-invariant test which will be called *randomness deficiency* (for bags) and denoted $d_{\mathrm{iid}}(\mathrm{conf}(z))$ with a suitable upper index.

The following theorem generalizes Theorem 1 in Vovk [22] (1986):

Theorem 3 *The randomness deficiency of a sequence equals the sum of its permutation deficiency and the randomness deficiency of its configuration:*

$$d^{\mathrm{L}}_{\mathrm{iid}}(z) =^+ d^{\mathrm{L}}_{\mathrm{iid}}(\mathrm{conf}(z)) + d^{\mathrm{L}}_{\mathrm{exch}}\left(z \,|\, d^{\mathrm{L}}_{\mathrm{iid}}(\mathrm{conf}(z))\right).$$

Theorem 2 in Vovk [22] (1986) gives a simple characterization of Levin randomness deficiency for bags in

terms of prefix complexity in the binary case. That result shows that in the binary case the randomness deficiency of bags of size n is at most $\log n$. Unfortunately, in the case of infinite sample space the randomness deficiency of bags of size n can be as large as n:

Theorem 4 *If the sample space is constructively infinite (meaning that it contains an infinite computable sequence of distinct elements),*

$$\sup_{|b|=n} d^{\mathrm{ML}}_{\mathrm{iid}}(b) \stackrel{+}{\geq} \sup_{|b|=n} d^{\mathrm{L}}_{\mathrm{iid}}(b) \stackrel{+}{\geq} n \log e - \frac{1}{2} \log n,$$

where n ranges over the positive integers and b ranges over the bags.

Remark 6 Kolmogorov ([9], 1968) believed that, in the binary case, Bernoulli sequences should be defined as sequences with small permutation deficiency rather than sequences with small randomness deficiency. Note [22] (Vovk, 1986) was written in an attempt to understand the difference between Kolmogorov's definition and the definition accepted in this paper. The main result of [22] is that, in the binary case, these two definitions are close but different.

Remark 7 In the case of infinite sequences, de Finetti's theorem (see, eg, Schervish [19], 1995) says that the two models of iid and exchangeability are equivalent; therefore, for the infinite sequences, $d_{\mathrm{iid}} \stackrel{+}{=} d_{\mathrm{exch}}$. There are finite variants of de Finetti's theorem (see, eg, Schervish [19], 1995, Theorem 1.70 due to Diaconis and Freedman); similarly to the results of Vovk [22] (1986) (but contrary to Theorems 3 and 4 above) they say that the iid and exchangeability models are close.

5 PRACTICABLE APPROXIMATIONS

Let us concentrate on the problem of pattern recognition, in which the set Y of possible labels is $\{-1, 1\}$, and the case where $k = 1$ (there is only one test example). Following Vapnik [21] (1998), we consider the quadratic optimization problem

$$\Phi(w, \xi) = \frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l+1} \xi_i\right) \to \min \qquad (3)$$

$$\left(w \in H,\ \xi = (\xi_1, \ldots, \xi_{l+1}) \in \mathbb{R}^{l+1}\right),$$

where C is an *a priori* fixed positive constant, subject to the constraints

$$y_i\left((w \cdot F(x_i)) + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, l+1, \quad (4)$$

$$\xi_i \geq 0, \quad i = 1, \ldots, l+1, \qquad (5)$$

where F is some (typically non-linear) transformation applied to the data and taking values in a Hilbert space H.

Using Lagrange multipliers α_i corresponding to constraints (4) we can approximate from below both d^{ML} and d^{L}. The latter was done in Gammerman et al [4] (1998) and the former was done in the recent paper Saunders et al [17] (1999). The approach of Gammerman et al [4] (1998) suffered from the "distortion phenomenon" (see Subsection 8.2 of that paper); the solution suggested (implicitly) in [4] was to use the function

$$p(z_1, \ldots, z_{l+1}) = \frac{f(\alpha_1) + \cdots + f(\alpha_{l+1})}{f(\alpha_{l+1})(l+1)}, \qquad (6)$$

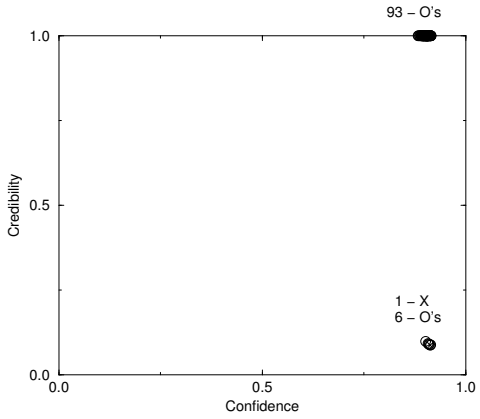
where f is some monotonic non-decreasing function with $f(0) = 0$, as an upper bound for the Levin permutation level. The specific function $f(\alpha)$ suggested in Gammerman et al [4] (1998) was $f(\alpha) = \operatorname{sign} \alpha$ (that is, $f(0) = 0$ and $f(\alpha) = 1$ when $\alpha > 0$). The results reported in [4] correspond to using the SV method for prediction and using function (6) for estimating confidence and credibility. Those results are reproduced here as Figure 2. In that figure (and in the figures below) it is easy to identify two clusters; one of the clusters contains those examples which are support vectors in both "pictures" (in the terminology of Gammerman et al [4]; in other words, which are support vectors and remain support vectors when the classification of the last example is changed), and the other cluster contains those examples which are support vectors in only one "picture".

Those experiments and all experiments described in this paper are done for a simple pattern recognition problem of identifying handwritten digits using a database of US postal data of 9300 digits, where each digit is a 16×16 vector (cf LeCun et al [10], 1990). The experiments are conducted for a subset of these data (a training set of 400 examples and 100 test sets of 1 example each), and include a construction of two-class classifier to separate digit "2" from digit "7". The constant C in (3) was set to ∞ (we felt that this way we would obtain good approximations to randomness deficiency); therefore, we actually solved the quadratic optimization problem

$$\frac{1}{2}(w \cdot w) \to \min \quad \left(w \in H,\ \xi = (\xi_1, \ldots, \xi_{l+1}) \in \mathbb{R}^{l+1}\right),$$

subject to the constraints

$$y_i\left((w \cdot F(x_i)) + b\right) \geq 1, \quad i = 1, \ldots, l+1. \qquad (7)$$



Cluster	1	2
Average confidence	0.902	0.910
Average credibility	1	0.091

Figure 2: Experimental results for SV predictions and for confidences and credibilities corresponding to $f(\alpha) = \operatorname{sign}\alpha$; characteristics of the two clusters (which can be identified by their average credibility); the correctly predicted examples are marked with O and the errors with X

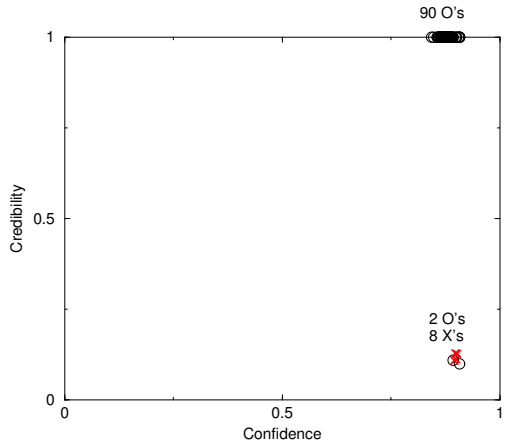
Using results of preliminary experiments we chose as F in (7) a function implicitly given by the polynomial (of degree 3) kernel $K(x, x') = (x \cdot x')^3/256$.

The "distortion phenomenon" leads to much poorer predictive performance (ie, the number of mistakes made) for the "pure" algorithm of Gammerman et al [4] (1998) (corresponding to $f(\alpha) = \operatorname{sign}\alpha$); see Figure 3. The predictive performance, confidences and credibilities corresponding to the functions $f(\alpha) = \alpha$ and $f(\alpha) = \alpha^2$ are shown in Figures 4 and 5.

Remark 8 In the spirit of Remark 5, it is instructive to compare our approximations to randomness deficiency (typical order of magnitude $\log(l+1)$, assuming there is only one example to be classified) with the usual results of statistical learning theory (see, eg, Vapnik [21], 1998, Chapters 4 and 10) and PAC theory (see, eg, Haussler et al [6], 1994). Say, Vapnik's ([21], Theorem 10.5; Theorems 10.6 and 10.7 are of a similar form) denominator $l+1$ in

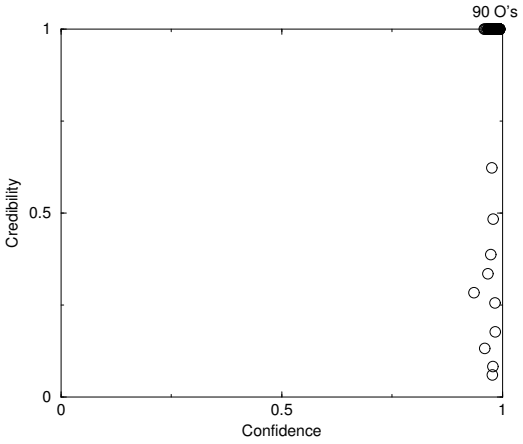
$$\text{prob(error)} \le \frac{\mathbf{E}\mathcal{K}_{l+1}}{l+1}$$

corresponds to our $\log(l+1)$ (when the "direct scale" is used). Of course, the advantage of our approach



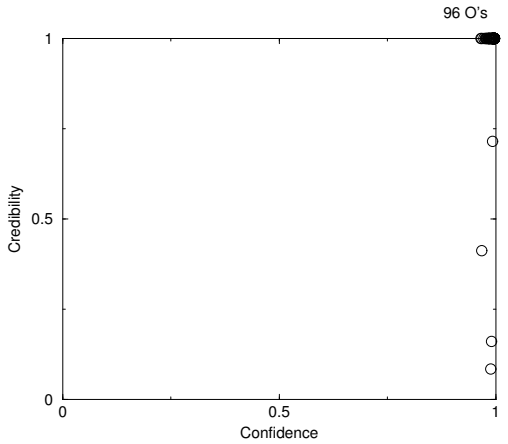
Cluster	1	2
Average confidence	0.899	0.900
Average credibility	1	0.120

Figure 3: Experimental results for $f(\alpha) = \operatorname{sign}\alpha$



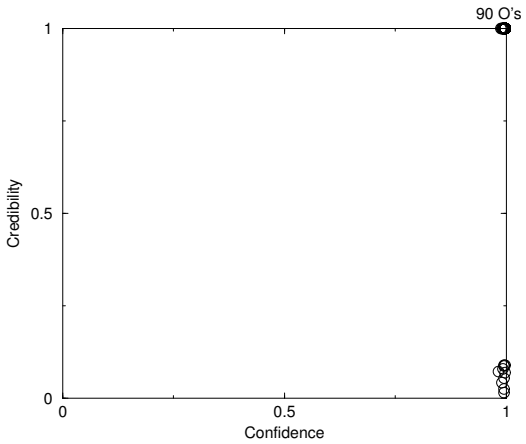
Cluster	1	2
Average confidence	0.993	0.980
Average credibility	1	0.282

Figure 4: Experimental results for $f(\alpha) = \alpha$



Cluster	1	2
Average confidence	0.997	0.990
Average credibility	1	0.343

Figure 5: Experimental results for $f(\alpha) = \alpha^2$



Cluster	1	2
Average confidence	0.998	0.996
Average credibility	1	0.063

Figure 6: Experimental results for Martin-Löf randomness level

is that our measures of credibility and confidence are applicable to individual examples and not just describe the results of hypothetical repetitions of the prediction task.

Our preliminary studies show that in the problem of regression lotteries are very convenient, but in the problems of pattern recognition considered in this paper it is not clear which function f to choose in (6). Our solution is to use the following p-value function p as an upper bound for the Martin-Löf permutation level (and, *a forteriori*, randomness level): arrange all Lagrange multipliers α_i in decreasing order and define $p(z_1 \dots z_n)$ to be the rank of α_{l+1} divided by $l+1$. The results for this approach (which is also described in Saunders et al [17], 1999) are given in Figure 6.

6 SOME IMPOSSIBILITY RESULTS

6.1 ON-LINE PREDICTION

Another interesting application of the algorithmic theory of randomness is the following explanation why the bulk of work in machine learning under the general iid assumption (such as statistical learning theory and PAC theory) has been done in the batch setting. (The on-line setting has also been very popular, eg, in the theory of prediction with expert advice, but it uses different assumptions.) The algorithmic theory of randomness implies that on-line prediction under the iid assumption is impossible, at least at random moments in time. We will say that $n \in \mathbb{N}$ is "locally random" if n is a random element of the set

$$A(n) = \left\{2^{\lfloor \log n \rfloor}, \dots, 2^{\lfloor \log n \rfloor + 1} - 1\right\};$$

formally, the *local randomness deficiency* of n is $d^{\mathrm{ML}}_{A(n)}(n)$, or $d^{\mathrm{L}}_{A(n)}(n)$. Analogously to Martin-Löf and Levin randomness deficiencies defined above, we can define randomness deficiencies for infinite data sequences in Z^∞. The next theorem formalizes the impossibility of on-line prediction under the iid assumption at random time (notice that it is possible to predict at time n with $K(n) \ll n$: this situation differs little from the batch setting, where n is given in advance).

Theorem 5 *Let $(x_1, y_1), \dots, (x_n, y_n)$, $y_i \in \{-1, 1\}$, range over all finite data sequences (with the first $l = n-1$ elements interpreted as the training set and (x_n, y_n) interpreted as the test example) and ζ range over all continuations of the sequence*

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, -y_n) \tag{8}$$

(the test example is classified wrongly). Then

$$\inf_\zeta d^{\mathrm{L}}_{\mathrm{iid}}(\zeta) \leq^+ d^{\mathrm{L}}_{\mathrm{iid}}\left((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\right) + d^{\mathrm{L}}_{A(n)}(n). \tag{9}$$

(It is clear that a similar inequality holds for Martin-Löf deficiency as well.) Inequality (9) shows that we cannot exclude the wrong classification if the data sequence is random and the time n is also random: the wrong sequence (8) also has a random continuation.

6.2 DENSITY ESTIMATION

According to Vapnik [20] (1995) (see also Vapnik [21], 1998) there are three main problems of statistical learning theory: pattern recognition; regression estimation; density estimation. As we have seen earlier, the problem of pattern recognition can be efficiently solved for typical real-world data sets, in the sense that we can obtain measures of confidence and credibility which are valid under the iid assumption (without need of any other assumptions) and which work well in practice. On the other hand, in typical high-dimensional cases the problem of density estimation can only be solved under assumptions essentially stronger than the iid assumption. To see why, assume that the unlabelled examples x_i are taken from some discrete space X, that $y_i \in \{-1, 1\}$, and we are asked to estimate the probability that $y_{l+1} = 1$. If all unlabelled examples in the training and test sets are distinct (which is typical when the number of attributes is big as compared to the number of examples), no non-trivial estimate (such as an interval containing neither 0 nor 1) of this probability is possible. Indeed, if the full sample $(x_1, y_1), \ldots, (x_{l+1}, y_{l+1})$ is random wr to an iid distribution P, it will also be random wr to a distribution P^* randomly generated by the following stochastic process: for all $c \in X$, $P^*(x = c) = P(x = c)$ and $P^*(y \mid x = c)$ is concentrated on $y = -1$ or $y = 1$, the latter with probability $P(y = 1 \mid x = c)$ and the former with probability $P(y = -1 \mid x = c)$. Of course, density estimation becomes possible when additional assumptions are made. In low-dimensional situations, informative confidence intervals for density estimation are obtained in, eg, Gammerman and Thatcher [3] (1992).

6.3 REGRESSION

There are several possible understanding of the term "regression". One understanding is "regression estimation": we assume that the examples (x_i, y_i) are generated by some iid distribution, and our goal is to estimate the conditional expectation of y_{l+1} given x_{l+1} (we are assuming that there is only one test example). This problem coincides with that of density estimation when $y_i \in \{-1, 1\}$, and so, according to the previous subsection, is infeasible when our only assumption is iid.

If "regression" is understood as estimating y_{l+1} when $y_i \in \mathbb{R}$ are not restricted to a finite set like $\{-1, 1\}$, the problem of regression can be efficiently solved under the iid assumption in high-dimensional cases (work in progress).

There is one more popular statement of the regression (and pattern recognition) problem, where only y_i are generated stochastically given x_i; x_i themselves are not generated stochastically and are just given constants. It is easy to see that in this case even pattern recognition (and *a forteriori* regression) is impossible without making additional assumptions.

7 CONCLUSION

This paper answers the question why one should want to use the algorithmic theory of randomness: in practice, we still use "non-algorithmic" notions such as p-values or lotteries. As we have shown, using the algorithmic theory of randomness we can ask (and answer) questions about relationships between

- universal p-values and universal lotteries (Martin-Löf vs Levin randomness level);
- exchangeability and randomness.

The second item raises the open question (already mentioned): is it possible to make use of the randomness deficiency of the configuration (which can, according to Theorem 4, be quite big)?

Besides the "positive" results discussed in the previous paragraph, the algorithmic theory of randomness also allows us to prove impossibility of prediction in certain situations; as shown in the previous section, such important problems as density estimation in high-dimensional spaces, regression estimation in high-dimensional spaces, and on-line prediction (where it is required that valid measures of confidence are output at every step) cannot be solved if our only assumption is iid.

Acknowledgements

The referees' insightful comments helped us to improve the presentation. We thank EPSRC for providing financial support through grants GR/L35812 ("Support Vector and Bayesian learning algorithms"), GR/M14937 ("Predictive complexity: recursion-theoretic variants") and GR/M16856 ("Comparison of Support Vector Machine and Minimum Message Length methods for induction and prediction").

References

[1] D A S Fraser. *Nonparametric methods in statistics.* Wiley, New York, 1957.

[2] Peter Gács. Exact expressions for some randomness tests. *Z Math Logik Grundl Math*, 26:385–394, 1980.

[3] Alex Gammerman and A R Thatcher. Bayesian diagnostic probabilities without assuming independence of symptoms. *Yearbook of Medical Informatics*, pages 323–330, 1992.

[4] Alex Gammerman, Vladimir Vapnik, and Volodya Vovk. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–156, San Francisco, CA, 1998. Morgan Kaufmann.

[5] Irwin Guttman. *Statistical tolerance regions: classical and Bayesian.* Griffin, London, 1970.

[6] David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994.

[7] Andrei N Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933.

[8] Andrei N Kolmogorov. Three approaches to the quantitative definition of information. *Problems Inform Transmission*, 1:1–7, 1965.

[9] Andrei N Kolmogorov. Logical basis for information theory and probability theory. *IEEE Trans Inform Theory*, IT-14:662–664, 1968.

[10] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L J Jackel. Handwritten digit recognition with back-propagation network. *Advances in Neural Information Processing Systems*, pages 396–404, 1990.

[11] Leonid A Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14:1413, 1973.

[12] Ming Li and Paul Vitányi. Computational machine learning in theory and praxis. In J van Leeuwen, editor, *Computer Science Today, Recent Trends and Developments*, volume 1000 of *Lecture Notes in Computer Science*, pages 518–535. Springer, 1995.

[13] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer, New York, 2nd edition, 1997.

[14] L Longpré. Resource-bounded Kolmogorov complexity and statistical tests. In O Watanabe, editor, *Kolmogorov Complexity and Computational Complexity*, pages 66–84. Springer, 1992.

[15] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[16] Jorma Rissanen. Stochastic complexity (with discussion). *J R Statist Soc B*, 49:223–239 and 252–265, 1987.

[17] Craig Saunders, Alex Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of IJCAI'99.* To appear.

[18] Craig Saunders, Alex Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In J Shavlik, editor, *Machine Learning, Proceedings of the Fifteenth International Conference*, pages 515–521, San Francisco, CA, 1998. Morgan Kaufmann.

[19] Mark J Schervish. *Theory of statistics.* Springer, New York, 1995.

[20] Vladimir N Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

[21] Vladimir N Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

[22] Volodya Vovk. On the concept of the Bernoulli property. *Russ Math Surv*, 41:247–248, 1986.

[23] Volodya Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *J R Statist Soc* B, 55:317–351, 1993.

[24] Volodya Vovk and Alex Gammerman. Complexity Approximation Principle. *The Computer Journal*, 1999. To appear.

[25] Vladimir V V'yugin. Algorithmic entropy (complexity) of finite objects and its applications to defining randomness and amount of information. *Selecta Mathematica Sovietica*, 13:357–389, 1994.

[26] Chris S Wallace and P R Freeman. Estimation and inference by compact coding (with discussion). *J R Statist Soc* B, 49:240–265, 1987.