On-line Confidence Machines are well-calibrated

Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
vovk@cs.rhul.ac.uk

Technical Report

CLRC-TR-02-01

23rd April 2002
revised 29th April 2002

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

Transductive Confidence Machine (TCM) and its computationally efficient modification, Inductive Confidence Machine (ICM), are ways of complementing machine-learning algorithms with practically useful measures of confidence. We show that when TCM and ICM are used in the on-line mode, their confidence measures are well-calibrated, in the sense that predictive regions at the confidence level $1 - \delta$ will be wrong with relative frequency at most δ (approaching δ in the case of randomised TCM and ICM) in the long run. This is not just an asymptotic phenomenon: we also show that the error probability of randomised TCM and ICM is δ at every trial and that errors happen independently at different trials.

1 Introduction

The bulk of work in computational learning theory is done under the *i.i.d. assumption*: the random examples fed to the learning algorithm are independent and identically distributed. The modern theory was started by Vapnik and Chervonenkis (see [16] for a recent review) and, much later but independently, by Valiant [15]; nowadays, this theory is often referred to as the theory of PAC learning.

A serious drawback of PAC learning theory is the weakness of error bounds it produces in practice, even for relatively clean data sets. The notion of Transductive Confidence Machine (TCM) was introduced in [12, 17] in an attempt to overcome this drawback. It is argued in [9] that TCM indeed produces more useful results: for example, in the case of the standard USPS data set (see, e.g., [16], p. 496) typical PAC bounds on error probability exceed 1 whereas TCM is able to predict the vast majority (approximately 95%) of test examples at the confidence level 99%. (See also Figures 1 and 2.)

The properties of PAC learning algorithms and TCM are usually studied in the batch setting. This paper studies on-line properties of TCM, showing that TCM's confidence measures are well-calibrated under the i.i.d. assumption (and the weaker exchangeability assumption). To my knowledge, this is the first result establishing that some on-line algorithm is well-calibrated without using any assumptions beyond i.i.d.

Figures 1 and 2 provide empirical evidence that TCM is well-calibrated. For every new hand-written digit TCM predicts a set of possible labels (0 to 9) for this digit (called the *predictive region*); for a vast majority of examples the predictive region has one label. It can be seen that the number of errors made grows linearly, and the slope is approximately 5% for the confidence level 95% and 1% for the confidence level 99%.

A popular alternative to PAC learning is Bayesian learning. The latter requires a complete specification of the probability distribution generating the data and so is beyond the scope of this paper; see [7] for a comparison between the TCM and Bayesian approaches.

2 Region predictors

Our basic protocol is as follows. Nature outputs pairs

$$(x_1, y_1), (x_2, y_2), \dots \tag{1}$$

called *examples*. Each example (x_i, y_i) consists of an *object* x_i and its *label* y_i; e.g., the objects can be hand-written digits and y_i their classifications (numbers from 0 to 9). The objects are elements of a measurable space $\mathbf{X}$ called the *object space* and the labels are elements of a measurable space $\mathbf{Y}$ called the *label space*. We will use the notation $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ for the *example space*; therefore, the infinite data sequence (1) will be an element of the measurable space $\mathbf{Z}^\infty$.

We will be assuming that the data sequence (1) is output according to some probability distribution P in $\mathbf{Z}^\infty$. The usual further assumption made in PAC theory is that P is an *i.i.d. distribution*, i.e., $P = Q^\infty$, where Q is a distribution in $\mathbf{Z}$ (in other words, that individual examples are generated by Q independently of each other). For us, it will be sufficient to make the weaker assumption that P is *exchangeable*: for every positive integer n, every permutation $\pi : \{1,\ldots,n\} \to \{1,\ldots,n\}$, and every measurable set $E \subseteq \mathbf{Z}^n$,

$$\begin{aligned} &P\left\{(z_1,z_2,\ldots) \in \mathbf{Z}^\infty \,|\, (z_1,\ldots,z_n) \in E\right\} \\ &= P\left\{(z_1,z_2,\ldots) \in \mathbf{Z}^\infty \,|\, (z_{\pi(1)},\ldots,z_{\pi(n)}) \in E\right\}. \end{aligned}$$

The difference is almost non-existent, however (in the current context of infinite data sequences): according to De Finetti's representation theorem (see, e.g., [13], Theorem 1.49), every exchangeable distribution is a mixture of i.i.d. distributions.

We are interested in algorithms for predicting, at every trial n, the label y_n given the object x_n and all the previous examples, from (x_1,y_1) to (x_{n-1},y_{n-1}). Since we are interested in prediction with confidence, our algorithms are given an extra input $(1-\delta) \in (0,1)$, which we call the *confidence level*. Formally, we define a *region predictor* to be a function

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0,1) \to 2^{\mathbf{Y}} \tag{2}$$

($2^{\mathbf{Y}}$ is the set of all subsets of $\mathbf{Y}$) which, for every confidence levels $1-\delta_1 \le 1-\delta_2$, every positive integer n, and every *incomplete data sequence*

$$x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n \tag{3}$$

(we often ignore unnecessary parentheses, such as those around (x_i, y_i)) satisfies

$$\begin{aligned} &\Gamma\left(x_1,y_1,\ldots,x_{n-1},y_{n-1},x_n,1-\delta_1\right) \\ &\subseteq \Gamma\left(x_1,y_1,\ldots,x_{n-1},y_{n-1},x_n,1-\delta_2\right). \end{aligned} \tag{4}$$

Intuitively, given the incomplete data set (3) and a confidence level $1-\delta$, the region predictor Γ predicts that

$$y_n \in \Gamma\left(x_1,y_1,\ldots,x_{n-1},y_{n-1},x_n,1-\delta\right),$$

and the larger $1-\delta$ the more emphatic the prediction; condition (4) is a natural requirement of consistency. Formally, for any infinite data sequence

$$\omega = (x_1,y_1,x_2,y_2,\ldots), \tag{5}$$

confidence level $1-\delta$, and positive integer n, we define the number of errors that Γ makes at the confidence level $1-\delta$ on the sequence ω during the first n trials to be

$$\begin{gathered} \mathrm{Err}_n(\Gamma,1-\delta,\omega) := \\ \#\left\{i=1,\ldots,n \,|\, y_i \notin \Gamma\left(x_1,y_1,\ldots,x_{i-1},y_{i-1},x_i,1-\delta\right)\right\}, \end{gathered}$$

where $\#B$ stands for the size of the set B. Sometimes we will also need the individual prediction results

$$\mathrm{err}_n(\Gamma, 1-\delta, \omega) := \begin{cases} 1 & \text{if } y_n \notin \Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and the whole infinite sequence of prediction results

$$\mathrm{err}(\Gamma, 1-\delta, \omega) := (\mathrm{err}_1(\Gamma, 1-\delta, \omega), \mathrm{err}_2(\Gamma, 1-\delta, \omega), \ldots).$$

In §4 we will also consider *randomised region predictors*, which depend, additionally, on an element of an auxiliary probability space.

3 Transductive Confidence Machine

Transductive Confidence Machine (TCM) is a way to define a region predictor from a "bare predictions" algorithm. Formally, it is a way of transition from what we call an "individual strangeness measure" to a region predictor; first we will give formal definitions and then give a simple example of an individual strangeness measure.

A family of measurable functions $\{A_n \mid n \in \mathbb{N}\}$, where $A_n : \mathbf{Z}^n \to \mathbb{R}^{\mathbb{N}}$ for all n, $\mathbb{N}$ is the set of all positive integers and $\mathbb{R}$ is the set of all real numbers (equipped with the Borel σ-algebra), is called an *individual strangeness measure* if, for any $n \in \mathbb{N}$, any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, any $(z_1, \ldots, z_n) \in \mathbf{Z}^n$, and any $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{\mathbb{N}}$,

$$(\alpha_1, \ldots, \alpha_n) = A_n(z_1, \ldots, z_n) \Longrightarrow (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \ldots, z_{\pi(n)}). \tag{7}$$

In other words,

$$A_n : (z_1, \ldots, z_n) \mapsto (\alpha_1, \ldots, \alpha_n) \tag{8}$$

is called an individual strangeness measure if every α_i is determined by the (unordered) bag $\wr z_1, \ldots, z_n \int$ and z_i. (The difference between the bag $\wr z_1, \ldots, z_n \int$ and the set $\{z_1, \ldots, z_n\}$ is that the former can contain several copies of the same element.)

The *TCM associated with the individual strangeness measure* A_n is the following region predictor:

$$\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta) \tag{9}$$

is defined to be the set of all labels $y \in \mathbf{Y}$ such that

$$\frac{\#\{i = 1, \ldots, n \mid \alpha_i \geq \alpha_n\}}{n} > \delta, \tag{10}$$

where

$$(\alpha_1, \ldots, \alpha_n) := A_n((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y)). \tag{11}$$

In general, a *TCM* is the TCM associated with some individual strangeness measure.

The definition of TCM can be illustrated by the following simple example of an individual strangeness measure, the one used in producing Figures 1–7: mapping (8) can be defined, in the spirit of the 1-Nearest Neighbour Algorithm, as (assuming the objects are vectors in a Euclidean space)

$$\alpha_i := \frac{\min_{j \neq i: y_j = y_i} d(x_i, x_j)}{\min_{j \neq i: y_j \neq y_i} d(x_i, x_j)}, \tag{12}$$

where d is the Euclidean distance (i.e., an object is considered strange if it is in the middle of objects labelled in a different way and is far from the objects labelled in the same way).

Let us say that a set $E \subseteq \{0,1\}^\infty$ is *monotonic* if, for any two infinite binary sequences $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$,

$$\left.\begin{array}{r} (a_1, a_2, \ldots) \in E \\ a_i \leq b_i, \forall i \end{array}\right\} \Longrightarrow (b_1, b_2, \ldots) \in E. \tag{13}$$

The following result shows that, as far as upper bounds on $P\{\mathrm{err} \in E\}$ for monotonic E are concerned, we can assume that the error probability of Γ is δ at every trial and errors happen independently at different trials.

Theorem 1 *For any confidence level* $1 - \delta$*, any exchangeable probability distribution* P *in* $\mathbf{Z}^\infty$*, and any monotonic* $E \subseteq \{0,1\}^\infty$*, any TCM* Γ *satisfies*

$$P\{\omega \,|\, \mathrm{err}(\Gamma, 1-\delta, \omega) \in E\} \leq B_\delta^\infty(E), \tag{14}$$

where B_δ *is the Bernoulli distribution in* $\{0,1\}$ *with the parameter* δ*:* $B_\delta\{1\} = \delta$ *and* $B_\delta\{0\} = 1 - \delta$*.*

Corollary 1 *Every TCM* Γ *is* conservatively well-calibrated *in the sense that, for any exchangeable probability distribution* P *in* $\mathbf{Z}^\infty$ *and any confidence level* $1 - \delta$*,*

$$\limsup_{n \to \infty} \frac{\mathrm{Err}_n(\Gamma, 1-\delta, \omega)}{n} \leq \delta \tag{15}$$

for P*-almost all* $\omega \in \mathbf{Z}^\infty$*.*

This corollary immediately follows from the usual strong law of large numbers and Theorem 1 since the complement of (15) is monotonic. Using, instead, the law of the iterated logarithm, we can strengthen (15) to

$$\limsup_{n \to \infty} \frac{\mathrm{Err}_n(\Gamma, 1-\delta, \omega) - n\delta}{\sqrt{2\delta(1-\delta) n \ln \ln n}} \leq 1.$$

We will also state two finite-sample implications of Theorem 1: Hoeffding's inequality (see, e.g., [4], Theorem 8.1) implies that, for any positive integer N and any constant $\epsilon > 0$,

$$P\{\omega \,|\, \mathrm{Err}_N(\Gamma, 1-\delta, \omega) \geq N(\delta + \epsilon)\} \leq e^{-2N\epsilon^2};$$

the central limit theorem implies that, for any constant c,

$$\limsup_{N \to \infty} P\left\{\omega \,|\, \mathrm{Err}_N(\Gamma, 1-\delta, \omega) \geq N\delta + c\sqrt{N}\right\} \leq \frac{1}{\sqrt{2\pi}} \int_{\frac{c}{\sqrt{\delta(1-\delta)}}}^{\infty} e^{-u^2/2} du.$$

4 Randomised Transductive Confidence Machine

In this section we introduce a modification of TCM which will allow us to simplify, strengthen, and prove easily Theorem 1. The *randomised Transductive Confidence Machine (rTCM) associated with the individual strangeness measure* A_n is the following randomised region predictor Γ: for any label $y \in \mathbf{Y}$,

1. if $\#\{i = 1, \ldots, n \mid \alpha_i > \alpha_n\}/n > \delta$ (as before, the αs are defined by (11)), the label y is included in (9);
2. if $\#\{i = 1, \ldots, n \mid \alpha_i \geq \alpha_n\}/n \leq \delta$, y is not included in (9);
3. otherwise, y is included in (9) with probability

$$\frac{\#\{i = 1, \ldots, n \mid \alpha_i \geq \alpha_n\} - n\delta}{\#\{i = 1, \ldots, n \mid \alpha_i = \alpha_n\}}. \tag{16}$$

In the typical case where all or almost all $\alpha_1, \ldots, \alpha_n$ are different, there is very little difference between TCM and rTCM (provided n is not too small).

To make the definition of rTCM more formal, we introduce the auxiliary probability space $([0,1]^\infty, U^\infty)$, where $[0,1]^\infty$ is equipped with the standard σ-algebra and U is the uniform probability distribution in $[0,1]$; intuitively, $(\tau_1, \tau_2, \ldots) \in [0,1]^\infty$ are random numbers for use at trials $1, 2, \ldots$, respectively, produced by a random number generator. The rTCM Γ is a function of the type $\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0,1) \times [0,1] \to 2^{\mathbf{Y}}$ (cf. (2)) where the dependence on the extra argument $\tau \in [0,1]$ (random number) arises because of item 3 of the definition of rTCM; for concreteness, we interpret it as: y is included in

$$\Gamma\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1 - \delta, \tau_n\right)$$

if

$$\tau_n < \frac{\#\{i \mid \alpha_i \geq \alpha_n\} - n\delta}{\#\{i \mid \alpha_i = \alpha_n\}}. \tag{17}$$

Notice that functions such as err depend on the extra argument $\tau \in [0,1]^\infty$ in the case of rTCM.

Theorem 2 *For any rTCM Γ, any confidence level $1-\delta$, and any exchangeable probability distribution P in $\mathbf{Z}^\infty$, the image of $P \times U^\infty$ under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto \operatorname{err}(\Gamma, 1 - \delta, \omega, \tau)$$

is the probability distribution B_δ^∞ of independent Bernoulli trials with parameter δ.

This theorem may appear too strong to be true: it is generally believed that to make categorical assertions about error probabilities some Bayesian-type assumptions are needed and that the general i.i.d. assumption is not sufficient. For example, in the theory of PAC learning an error probability ϵ is only asserted with some probability $1-\delta$. It should be remembered, however, that Theorem 2 does not assert that the probability of error, $\mathrm{err}_n = 1$, is δ conditionally on knowing the whole past (3); it is only asserted that it is δ unconditionally and conditionally on knowing $\mathrm{err}_1, \ldots, \mathrm{err}_{n-1}$. (Actually, it is quite obvious that the probability of error is often not equal to δ if the whole past is known: if the predictive region is empty, the conditional probability of error is 1; to balance this, the conditional probability that a non-empty predictive region is wrong will tend to be less than δ.)

Theorem 2 immediately implies Theorem 1: if an rTCM Γ and a TCM $\Gamma^\dagger$ are constructed from the same individual strangeness measure, the latter's errors $\mathrm{err}_n^\dagger$ never exceed the former's errors err_n, $\mathrm{err}_n^\dagger \le \mathrm{err}_n$. Theorem 2 also implies

Corollary 2 *Every rTCM Γ is* precisely well-calibrated *in the sense that*

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma, 1-\delta, \omega, \tau)}{n} = \delta$$

for P-almost all $\omega \in \mathbf{Z}^\infty$ and U^∞-almost all $\tau \in [0,1]^\infty$.

5 Inductive Confidence Machine

For large data sets, TCMs can be computationally inefficient. Inductive Confidence Machine (ICM) is a modification of TCM which sacrifices (in typical cases) some predictive accuracy for computational efficiency (for details of ICM in the batch setting, see [11]).

In the case of ICMs, the role of an individual strangeness measure will be played by a pair consisting of an "inductive algorithm" and "discrepancy measure". Let $\hat{\mathbf{Y}}$ be a *prediction space* (an arbitrary measurable space). An *inductive algorithm* D is a measurable function that maps every bag $\wr z_1, \ldots, z_n \int$ (of any size) of elements of $\mathbf{Z}$ to a function $D_{\wr z_1,\ldots,z_n \int} : \mathbf{X} \to \hat{\mathbf{Y}}$. The usual interpretation of $D_{\wr z_1,\ldots,z_n \int}$ is that it is a decision rule, found from the training set $\wr z_1, \ldots, z_n \int$, which computes the predicted label $\hat{y} := D_{\wr z_1,\ldots,z_n \int}(x)$ for any new object x. Usually, but not always, $\hat{\mathbf{Y}} = \mathbf{Y}$. A *discrepancy measure* is a measurable function $\Delta : \mathbf{Y} \times \hat{\mathbf{Y}} \to \mathbb{R}$; it will be used to measure the discrepancy between the predicted label $\hat{y}$ and the true label y.

Given an inductive algorithm D and a discrepancy measure Δ, we can define an individual strangeness measure $\{A_n\}_{n=1}^\infty$ as follows: for any $((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbf{Z}^*$, the values

$$(\alpha_1, \ldots, \alpha_n) = A_n((x_1, y_1), \ldots, (x_n, y_n)) \tag{18}$$

can be defined by the formula

$$\alpha_i := \Delta\left(y_i, D_{\wr (x_1,y_1),\ldots,(x_n,y_n) \int}(x_i)\right) \tag{19}$$

or the formula

$$\alpha_i := \Delta\left(y_i, D_{\wr (x_1,y_1),\ldots,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\ldots,(x_n,y_n) \int}(x_i)\right). \tag{20}$$

This shows that with every inductive algorithm and discrepancy measure we can associate a TCM. Formula (20) is more natural than (19) but leads to less computationally efficient TCMs.

At a crude level, one can divide inductive algorithms into two classes: "proper inductive algorithms" and "transductive algorithms" (see [16]; sometimes transductive algorithms are called "instance-based"). For proper inductive algorithms, $D_{\wr z_1,\ldots,z_n \int}$ can be computed, in some sense: e.g., $D_{\wr z_1,\ldots,z_n \int}$ may be described by a polynomial, and computing $D_{\wr z_1,\ldots,z_n \int}$ may mean computing the coefficients of the polynomial; as soon as $D_{\wr z_1,\ldots,z_n \int}$ is computed, computing $D_{\wr z_1,\ldots,z_n \int}(x)$ for a new object x takes very little time. For transductive algorithms (such as the Nearest Neighbours Algorithms), little can be done before seeing the new object x; even allowing considerable time for pre-processing $\wr z_1, \ldots, z_n \int$, computing $D_{\wr z_1,\ldots,z_n \int}(x)$ will be a difficult task.

Notice that, even when formula (19) rather than (20) is used and D is a proper inductive algorithm, the TCM based on the individual strangeness measure (19) will still be inefficient: for every new example (x_n, y_n), computing $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta)$ will require constructing a new decision rule.

To define an ICM from an inductive algorithm D and a discrepancy measure Δ, first fix a finite or infinite sequence of positive integer parameters $m_1, m_2, \ldots$ (called *training trials*); it is required that $m_1 < m_2 < \cdots$. If the sequence $m_1, m_2, \ldots$ is finite, $(m_1, m_2, \ldots) = (m_1, \ldots, m_l)$, we set $m_i := \infty$ for $i > l$. The ICM based on D, Δ, and the sequence $m_1, m_2, \ldots$ of training trials is defined to be the region predictor Γ such that $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta)$ is computed as follows:

- ◇ if $n \le m_1$, $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta)$ is found using a fixed TCM;
- ◇ otherwise, find the k such that $m_k < n \le m_{k+1}$ and set

$$\begin{gathered} \Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta) := \\ \left\{ y \in \mathbf{Y} \mid \frac{\#\{j = m_k + 1, \ldots, n \mid \alpha_j \ge \alpha_n\}}{n - m_k} > \delta \right\}, \end{gathered} \tag{21}$$

where the αs are defined by

$$\begin{gathered} \alpha_j := \Delta\left(y_j, D_{\wr (x_1,y_1),\ldots,(x_{m_k},y_{m_k}) \int}(x_j)\right), \quad j = m_k + 1, \ldots, n-1, \\ \alpha_n := \Delta\left(y, D_{\wr (x_1,y_1),\ldots,(x_{m_k},y_{m_k}) \int}(x_n)\right). \end{gathered} \tag{22}$$

We can see that ICM requires recomputing the decision rule being used not at every trial but only at the training trials $m_1, m_2, \ldots$; the rate of growth of m_i determines the chosen balance between predictive accuracy and computational efficiency. The most important case is perhaps where there is only one training trial m_1. Randomised ICM (rICM) can be defined analogously to rTCM.

Theorem 3 *Theorems 1 and 2 continue to hold in the case of ICMs and rICMs, respectively.*

Let a and b be positive numbers such that either $a \geq 1$ and $b \geq 1$ or $a > 1$. If an individual strangeness measure A_n is computable in time $\Theta(n^a \log^b n)$, the TCM associated with A_n spends time $\Theta(n^{a+1} \log^b n)$ on the computations needed for the first n trials. On the other hand, if an inductive algorithm D is computable in time $\Theta(n^a \log^b n)$, a discrepancy measure Δ is computable in constant time, and the sequence m_i is infinite and grows exponentially, the ICM based on D, Δ, and m_i spends the same, to within a constant factor, time $\Theta(n^a \log^b n)$. (We have been assuming that the TCM or ICM is given A_n or D as an oracle and the label space $\mathbf{Y}$ is finite and fixed.) In the case where the sequence m_i is finite, the ICM's computation time becomes $\Theta(n \log n)$ (e.g., use red-black trees for storing α_is; [1], Chapters 14 and 15).

The performance of the Nearest Neighbour ICM on the USPS data set with training trial 4649 (the middle of the data set) is shown in Figures 6 and 7. It can be seen from these figures (and is obvious anyway) that the ICM's performance (measured by the number of uncertain predictions) deteriorates sharply after training trials m_i. Perhaps in practice there should be short spells of "learning" after each training trial, when the ICM is provided with fresh "training examples" and its predictions are not used or evaluated.

6 Variable confidence level

For any region predictor Γ, infinite data sequence (5), and trial n, we define the *nth p-value* as

$$p_n(\Gamma, \omega) := \sup \{\delta \mid y_n \in \Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta)\}$$

(in the case of randomised region predictors, p_n will also depend on the result τ_n of the auxiliary randomisation).

We will use *Confidence Machine (CM)* as the name for region predictors that are either TCM or ICM; *rCM* will mean either rTCM or rICM. It is clear that the definitions of CM and rCM can be extended to the case where $1-\delta = 1-\delta_n$ depends on the trial n; $\mathrm{err}(\Gamma, 1-\delta)$ can be naturally extended to this case as well. The proofs of Theorems 2 and 3 will show that the following slightly stronger assertion holds.

Theorem 4 *For any rCM Γ, any sequence $1-\delta_n$ of confidence levels, and any exchangeable probability distribution P in $\mathbf{Z}^\infty$, the image of $P \times U^\infty$ under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto \mathrm{err}(\Gamma, (1-\delta_n), \omega, \tau)$$

is the probability distribution $B_{\delta_1} \times B_{\delta_2} \times \cdots$ of independent Bernoulli trials with parameters δ_n.

This immediately implies the following seemingly stronger statement.

Corollary 3 *For any rCM* Γ *and any exchangeable probability distribution* P *in* $\mathbf{Z}^\infty$*, the image of* $P \times U^\infty$ *under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto (p_1(\Gamma, \omega, \tau_1), p_2(\Gamma, \omega, \tau_2), \ldots)$$

is the uniform probability distribution in $[0,1]^\infty$*.*

A *variable confidence level* is any measurable function $f : [0,1]^* \to (0,1)$; intuitively, f chooses a confidence level $1-\delta$ to use from the information $(p_1, \ldots, p_n)$ about the p-values (in particular, the errors made) at the previous trials. The prediction made by Γ with the variable confidence level f at trial n is

$$\begin{aligned} &\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, f) \\ &:= \Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, f(p_1, \ldots, p_{n-1})). \end{aligned}$$

The simplest special case where $\delta_n := f(p_1, \ldots, p_{n-1})$ depends only on n was already used in Theorem 4; we will call such f *deterministic confidence levels.*

Corollary 3 immediately implies the following generalization of Theorem 4:

Corollary 4 *For any rCM* Γ*, any variable confidence level* f*, any exchangeable probability distribution* P *in* $\mathbf{Z}^\infty$*, and any trial* $n = 1, 2, \ldots$*,*

$$\begin{aligned} \mathbb{P}(\mathrm{err}_n(\Gamma, f(p_1(\Gamma, \omega, \tau_1), \ldots, p_{n-1}(\Gamma, \omega, \tau_{n-1})), \omega, \tau_n) = 1 \\ \mid p_1(\Gamma, \omega, \tau_1), \ldots, p_{n-1}(\Gamma, \omega, \tau_{n-1})) \\ = 1 - f(p_1(\Gamma, \omega, \tau_1), \ldots, p_{n-1}(\Gamma, \omega, \tau_{n-1})), \omega, \tau_n) \end{aligned}$$

$\mathbb{P}$*-almost surely, where* $\mathbb{P}$ *is the probability distribution* $P \times U^\infty$ *on* $(\omega, \tau) \in (\mathbf{Z}^\infty \times [0,1]^\infty)$*.*

7 Relations to other approaches

PAC learning

In this section we will briefly discuss the "batch" case where we are given two sets of examples: the training set

$$(x_1, y_1, \ldots, x_l, y_l) \tag{23}$$

of size l and the test set

$$(x_{l+1}, y_{l+1}, \ldots, x_{l+n}, y_{l+n}) \tag{24}$$

of size n. We are willing to assume that the examples z_i are generated by an i.i.d. distribution $P = Q^\infty$ in Z^∞ but do not want to make any further assumptions about P; and we would like to evaluate the quality of an inductive algorithm D.

The standard approach in PAC learning theory is, given the training set (23), to upper bound the error probability

$$L_l := Q\left\{(x,y) \in (\mathbf{X}\times\mathbf{Y}) \mid D_{\{x_1,y_1,\ldots,x_l,y_l\}}(x) \neq y\right\}.$$

We cannot guarantee that such an upper bound, say ϵ, will hold always, but quite often it is possible to show that an upper bound of ϵ will hold with probability $1-\delta$ over the random choice of the training set from Q^l.

Paper [6] studies the expected value $\mathbb{E}\, L_l$, where the expectation is over the random choice of the training set, but it is generally believed that knowing $\mathbb{E}\, L_l$ would only be marginally useful; as Devroye, Györfi, and Lugosi ([4], p. 2) put it, "this number would indicate the quality of an average data sequence [(23)], not *your* data sequence". Suppose that n is very large. One pragmatic interpretation of this objection (we will call it the "PAC objection") against using $\mathbb{E}\, L_l$ is as follows:

- If $L_l < \epsilon$, we can be sure that the percentage of errors D makes on the test set is close to or less than ϵ.
- On the other hand, if we only know that $\mathbb{E}\, L_l < \epsilon$, the percentage of errors D makes on the test set can be considerably greater than ϵ. It is true that if we additionally randomise over the training sets, the percentage of errors will be close to or less than ϵ, but we are interested in our particular training set (23), not in what might have happened but did not happen.

The PAC objection is also applicable to the batch version of our approach. Consider, for concreteness, using an ICM Γ to output predictive region corresponding to the objects x_{l+i}, $i = 1,\ldots,n$, in the test set (as in [11]). The training set (23) is split into two parts: the *proper training set*

$$(x_1, y_1, \ldots, x_m, y_m) \tag{25}$$

of size m and the *calibration set*

$$(x_{m+1}, y_{m+1}, \ldots, x_l, y_l) \tag{26}$$

of size $l-m$. For every test object x_i, $i = l+1,\ldots,l+n$, compute the predictive region

$$\begin{gathered}\Gamma(x_1,y_1,\ldots,x_l,y_l,x_i,1-\delta) := \\ \left\{y \in \mathbf{Y} \mid \frac{\#\{j = m+1,\ldots,l,i \mid \alpha_j \geq \alpha_i\}}{l-m+1} > \delta\right\},\end{gathered} \tag{27}$$

where the αs are defined by

$$\begin{gathered}\alpha_j := \Delta\left(y_j, D_{\{(x_1,y_1),\ldots,(x_m,y_m)\}}(x_j)\right), \quad j = m+1,\ldots,l, \\ \alpha_i := \Delta\left(y, D_{\{(x_1,y_1),\ldots,(x_m,y_m)\}}(x_i)\right).\end{gathered} \tag{28}$$

(Cf. (21) and (22).) It is true that

$$P\left\{(x_1,y_1,x_2,y_2,\ldots) \mid y_i \notin \Gamma(x_1,y_1,\ldots,x_l,y_l,x_i,1-\delta)\right\} \leq \delta \tag{29}$$

for every $i = l+1, \ldots, l+n$, but the events in (29) are not independent, and so

$$\frac{\#\{i = l+1, \ldots, l+n \mid y_i \notin \Gamma(x_1, y_1, \ldots, x_l, y_l, x_i, 1-\delta)\}}{n} \tag{30}$$

can be significantly above δ even when n is very large.

To meet the PAC objection (as interpreted above), we can slightly modify the application of the ICM constructed from the training set to the test set: after processing each test example (x_i, y_i), $i = l+1, \ldots, l+n$, the corresponding α_i should be added to the pool of αs used in generating the predictive regions for the following test examples. Formally, redefine

$$\begin{gathered}\Gamma(x_1, y_1, \ldots, x_l, y_l, x_i, 1-\delta) := \\ \left\{ y \in \mathbf{Y} \mid \frac{\#\{j = m+1, \ldots, i \mid \alpha_j \geq \alpha_i\}}{i-m} > \delta \right\},\end{gathered} \tag{31}$$

where the αs are defined by

$$\begin{aligned}\alpha_j &:= \Delta\left(y_j, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_j)\right), \quad j = m+1, \ldots, i-1, \\ \alpha_i &:= \Delta\left(y, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_i)\right).\end{aligned} \tag{32}$$

Theorem 3 says that, under this modification, (30) will not exceed δ, up to statistical fluctuations.

This "semi-batch" mode of learning is in one respect even safer than the standard PAC mode: in the PAC mode, there is always a non-zero probability δ (if we are unlucky with the training set) that the generalization error exceeds the chosen ϵ, and so the long-term percentage of wrong predictions will exceed the chosen threshold; whereas in the semi-batch mode the long-term percentage of wrong predictions is guaranteed to be at most the chosen threshold δ, no matter how unlucky we are with the training set (the worst that can happen if we are unlucky with the training set is that the percentage of uncertain predictions will increase).

Notice that in the case $n \ll k$ the semi-batch mode differs so slightly from the batch mode of applying ICM that the errors on the test set produced by the batch mode (27), (28) can be expected to be "nearly independent".

ICM applied in both batch and semi-batch modes is computationally quite efficient. In the case of bare predictions, the application of the inductive algorithm D to the test set requires time

$$O\left(T_{\text{train}} + nT_{\text{appl}}\right),$$

where T_{train} is the time required for computing the decision rule $D_{\wr x_1, y_1, \ldots, x_l, y_l \int}$ and T_{appl} is the time needed to apply this decision rule to a new object. The batch ICM (see (27) and (28)) requires time

$$O\left(T^{\dagger}_{\text{train}} + (l-m+n)T^{\dagger}_{\text{appl}} + (l-m)\log(l-m) + n\log(l-m)\right),$$

where $T^{\dagger}_{\text{train}}$ is the time required for computing the decision rule $D_{\{x_1,y_1,\ldots,x_m,y_m\}}$ and $T^{\dagger}_{\text{appl}}$ is the time needed to apply this decision rule to a new object (we assume that computing Δ is fast); we allow time $(l-m)\log(l-m)$ for sorting the αs obtained from the calibration set ([1], Part II) and time $\log(l-m)$ for finding the rank of a test α in the set of the calibration αs ([1], Chapters 14 and 15). In the case of semi-batch ICM ((31), (32)), the required time increases only slightly (in typical cases) to

$$O\left(T^{\dagger}_{\text{train}} + (l-m+n)T^{\dagger}_{\text{appl}} + (l-m)\log(l-m) + n\log(l-m+n)\right).$$

There are other versions of the PAC objection, but they border on controversial issues in the philosophy of statistics and will not be discussed in this paper.

Predicting on randomly chosen points

A precursor of TCM was considered by Haussler, Littlestone, and Warmuth [6][1]. The most important difference is that the version of TCM considered in [6] is only valid under a subclass of the exchangeable distributions, namely the exchangeable distributions consistent with a class of $\{0,1\}$-valued functions on $\mathbf{X}$. The formal definition is as follows. Let F be a class of $\{0,1\}$-valued functions on $\mathbf{X}$. The *F-TCM* associated with an individual strangeness measure A_n is the following region predictor: (9) is defined to be the set of all labels $y \in \mathbf{Y}$ such that

- ⋄ the sample $(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, y)$ is *consistent with* F, in the sense that there exists $f \in F$ such that $f(x_i) = y_i$, $i = 1, \ldots, n-1$, and $f(x_n) = y$,
- ⋄ and (10) holds, where the αs are defined by (11).

The randomised version of this definition can be given and analogues of Theorems 1 and 2 proved. Theorem 2.3 in [6] can be restated in the following way:

Theorem 5 (Haussler, Littlestone, and Warmuth) *Let F be a class of $\{0,1\}$-valued functions on $\mathbf{X}$ of VC dimension V; consider the deterministic confidence level $1-\delta_n := 1 - V/n$. There exists an F-TCM Γ such that for any sequence $(x_1, y_1, x_2, y_2, \ldots)$ of examples,*

$$|\Gamma(x_1, y_1, \ldots, x_n, 1 - V/n)| \le 1, \ \ \forall n. \tag{33}$$

Of course, the statement of the theorem is of interest only for sequences of examples that are consistent with F: there is $f \in F$ such that $y_i = f(x_i)$, $i = 1, 2, \ldots$.

[1] Of course, there are many other precursors, some of which are even closer to our version than [6]: see, e.g., [5].

The restatement given by Theorem 5 sheds new light on the discussion at the end of §2 in [6] (such as Corollary 2.2): the independence of the errors at different trials allows us to use the powerful machine of probability theory. We will only give two examples.

Let F be a class of binary functions on $\mathbf{X}$ of finite VC-dimension V and the exchangeable distribution P generating the data be consistent with F. The on-line F-TCM Γ whose existence is asserted in Theorem 5 never makes uncertain predictions at the confidence level $1 - V/n$, so the only measure of its performance is the number of errors

$$\mathrm{Err}_n := \mathrm{Err}_n(\Gamma, 1 - V/n)$$

it makes. The law of the iterated logarithm implies that, P-almost surely,

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n - V \ln n}{\sqrt{2V \ln n \ln \ln \ln n}} \le 1.$$

Poisson's theorem (see, e.g., [14], Theorem III.3.4; the presence of "identically distributed" there is a typo) shows that the distribution of the cumulative number of errors $\mathrm{Err}_N - \mathrm{Err}_{\lceil N/2 \rceil}$ in the second half of the first N trials is asymptotically dominated, as $N \to \infty$, by the Poisson distribution with parameter $\ln 2$.

8 Conclusion

The main advantages of this paper's approach are:

- As compared to the standard theory of PAC learning, our error bounds are practically meaningful (see §1 and Figures 1–7).
- As compared to the theory of Bayesian learning, we do not assume anything beyond the exchangeability of the underlying probability distribution.
- The usual justification ("validity") of TCM with confidence level $1 - \delta$ is the fact that the error probability at any trial does not exceed δ (see, e.g., [8], Theorem 1). This paper adds the crucial observation that, in the case of rTCM, the events "error at trial n", $n = 1, 2, \ldots$, not only have probability δ, but also are independent of each other. Very little can be said about a sequence of events of probability δ, but combined with independence this gives us a plethora of known properties for a random walk.

This paper is essentially an application of Dawid's [2, 3] prequential dictum: to evaluate the quality of even a batch procedure, put it to work in the on-line (or *prequential* = predictive sequential) mode. Both TCM and ICM pass this test.

These are the open problems about on-line TCM/ICM that appear most interesting to me:

1. Is it possible to prove a result analogous to Theorem 2 or Theorem 3 (or, more realistically, Corollary 2) for a wider class of probability distributions than the exchangeable distributions, such as the stationary distributions?
2. Is it possible to show that rTCMs are the most powerful, in some sense, region predictors for which Theorem 2 holds (cf. [10])?
3. Give a natural common generalisation of TCM and ICM and prove a theorem containing both Theorem 2 and 3.
4. Suppose we suspect that the true exchangeable distribution generating examples $z_n \in \mathbf{Z}$ is P (but still want an error bound of δ whatever the true distribution). What is the TCM optimal under P (e.g., in the sense that $\mathbb{E}_P \operatorname{Unc}_N$ attains its minimum, where N is a known in advance number of trials and Unc_N is the number of trials between 1 and N where the TCM outputs a predictive region of size at least two)?

Acknowledgments

The following people have helped me greatly in work on this paper: Alex Gammerman, Ilia Nouretdinov, Leo Gordon, and Kostas Proedrou.

This work was partially supported by EPSRC (grant GR/R46670/01 "Complexity Approximation Principle and Predictive Complexity: Analysis and Applications"), BBSRC (grant 111/BIO14428 "Pattern Recognition Techniques for Gene and Promoter Identification and Classification in Plant Genomic Sequences"), and EU (grant IST-1999-10226 "EurEdit: The Development and Evaluation of New Methods for Editing and Imputation").

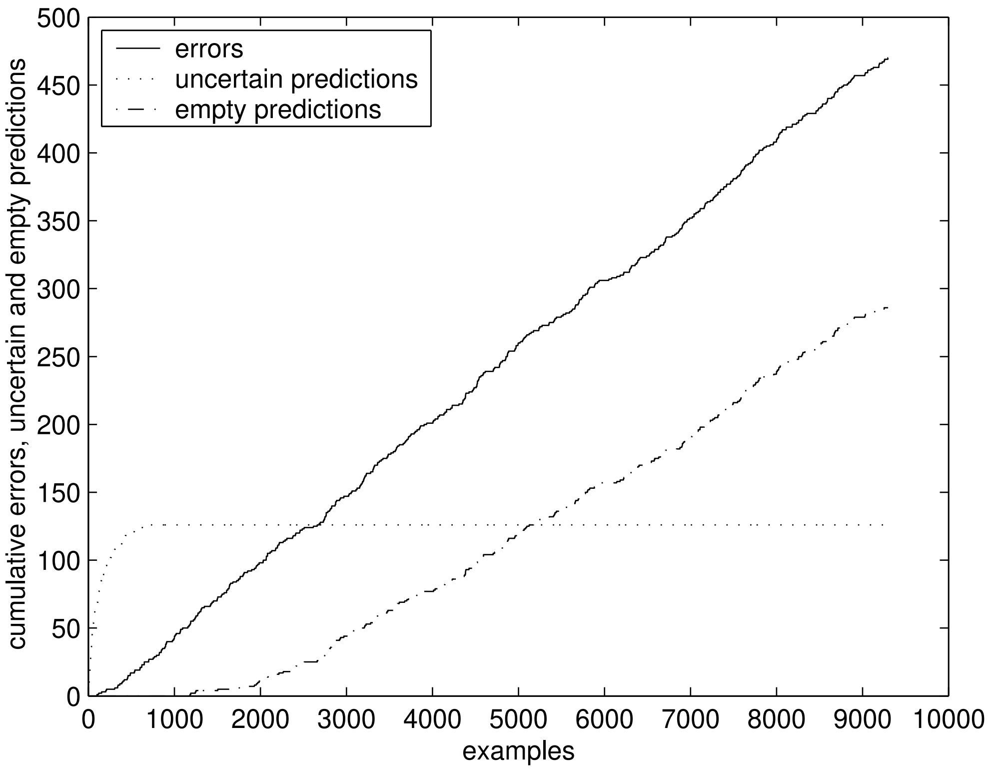

Figure 1 On-line performance of the Nearest Neighbour TCM on the USPS data set (9298 hand-written digits, randomly permuted) for the confidence level 95%. The solid line shows the cumulative number of errors, dotted the cumulative number of *uncertain predictions* (i.e., predictive regions with more than one label), and dashdot the cumulative number of empty predictive regions (inevitably leading to an error). This figure and Figures 2–7 are not significantly affected by statistical variation (due to the random choice of the permutation of the data set).

By default, the random permutation is done using MatLab's function `randperm` without setting the initial state of the generator (this is equivalent to setting the initial state to 0). The exceptions are: Figure 3 (the initial state is 1), Figure 4 (the initial state is 2), and Figure 5 (the initial state is 3).

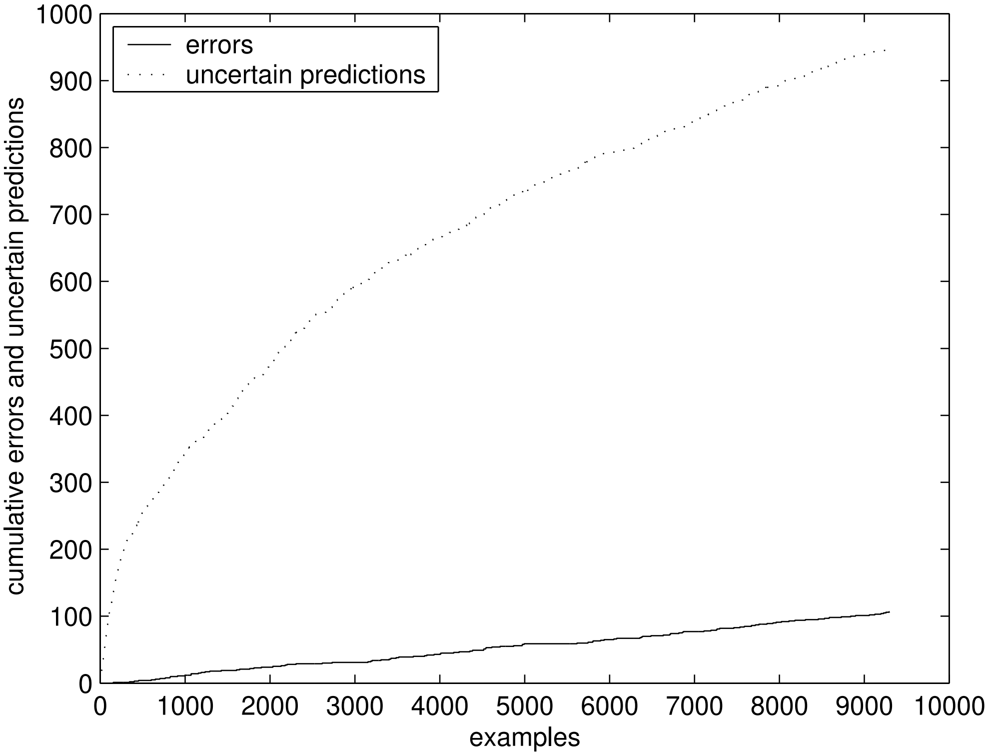

Figure 2 On-line performance of the Nearest Neighbour TCM on the USPS data set for the confidence level 99%. The dashdot line coincides with the horizontal axis (there are no empty predictions) and so is invisible.

These are some summary statistics:

- ⋄ the number of uncertain predictions at the confidence level 99% in the full data set is 946, which gives the success rate

$$\frac{9298 - 946}{9298} = 89.83\%;$$

- ⋄ the number of uncertain predictions at the confidence level 99% in the second half of the data set is $946 - 711 = 235$, which gives the success rate

$$\frac{4649 - 235}{4649} = 94.95\%.$$

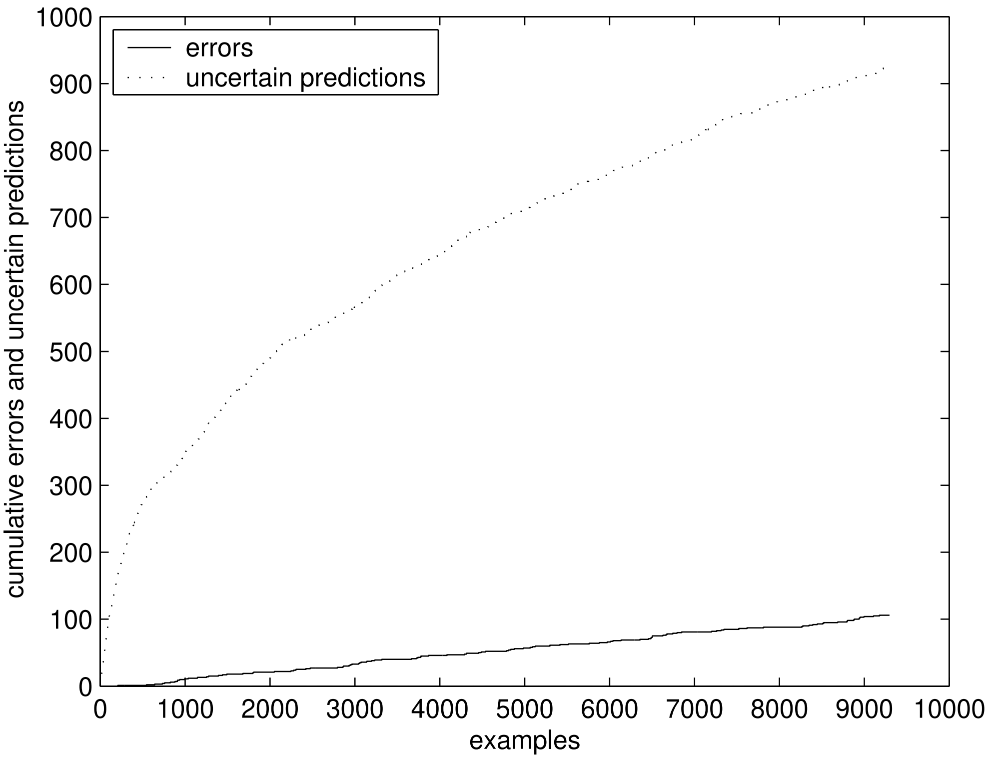

Figure 3 On-line performance of the Nearest Neighbour TCM on a different random permutation of the USPS data set for the confidence level 99%. There are no empty predictions.

Summary statistics:

- the number of uncertain predictions at the confidence level 99% in the full data set is 925, which gives the success rate

$$\frac{9298 - 925}{9298} = 90.05\%;$$

- the number of uncertain predictions at the confidence level 99% in the second half of the data set is $925 - 693 = 232$, which gives the success rate

$$\frac{4649 - 232}{4649} = 95.01\%.$$

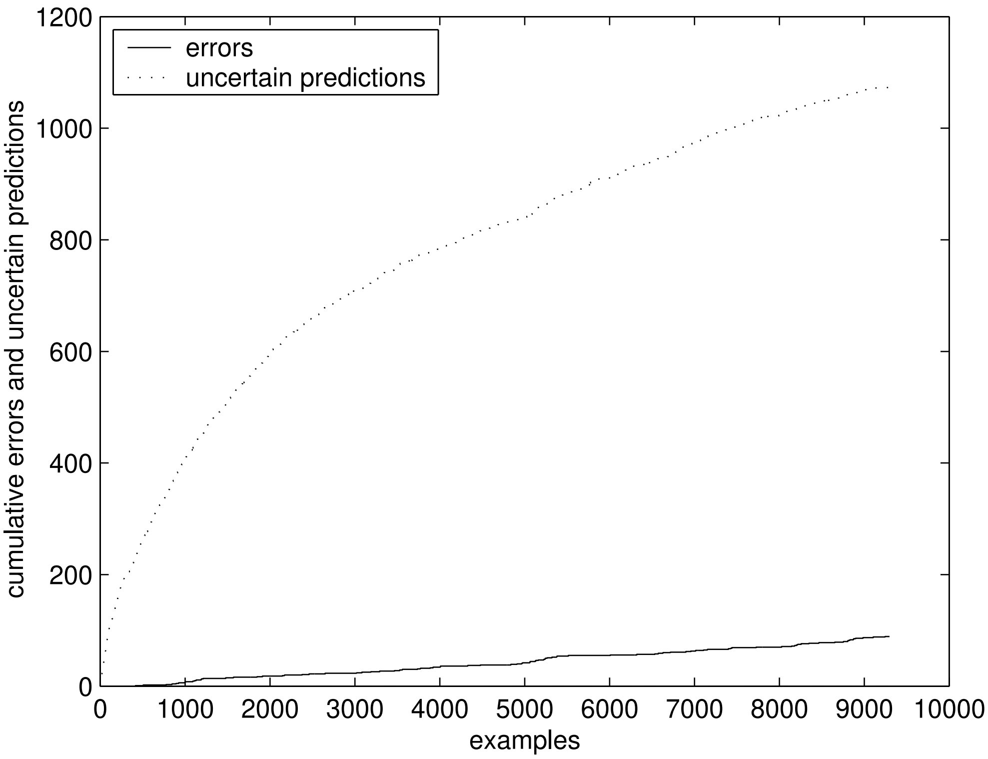

Figure 4 On-line performance of the Nearest Neighbour TCM on another random permutation of the USPS data set for the confidence level 99%. There are no empty predictions.

Summary statistics:

- the number of uncertain predictions at the confidence level 99% in the full data set is 1073, which gives the success rate

$$\frac{9298 - 1073}{9298} = 88.46\%;$$

- the number of uncertain predictions at the confidence level 99% in the second half of the data set is $1073 - 824 = 249$, which gives the success rate

$$\frac{4649 - 249}{4649} = 94.64\%.$$

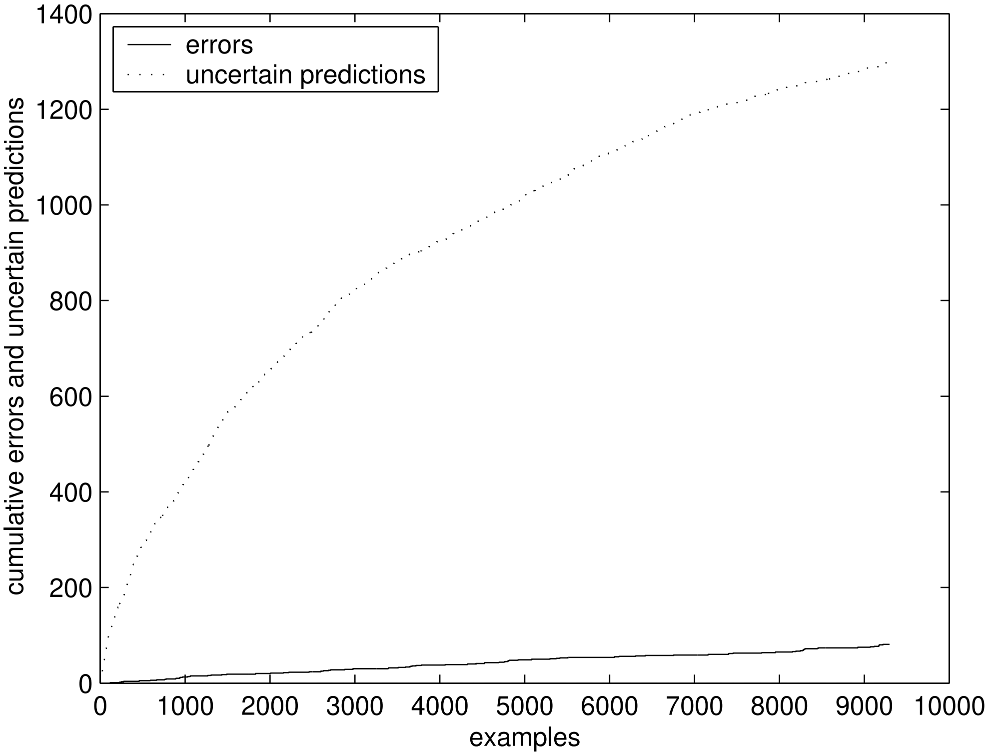

Figure 5 On-line performance of the Nearest Neighbour TCM on another random permutation of the USPS data set for the confidence level 99%. There are no empty predictions.

Summary statistics:

- the number of uncertain predictions at the confidence level 99% in the full data set is 1298, which gives the success rate

$$\frac{9298 - 1298}{9298} = 86.04\%;$$

- the number of uncertain predictions at the confidence level 99% in the second half of the data set is $1298 - 985 = 313$, which gives the success rate

$$\frac{4649 - 313}{4649} = 93.27\%.$$

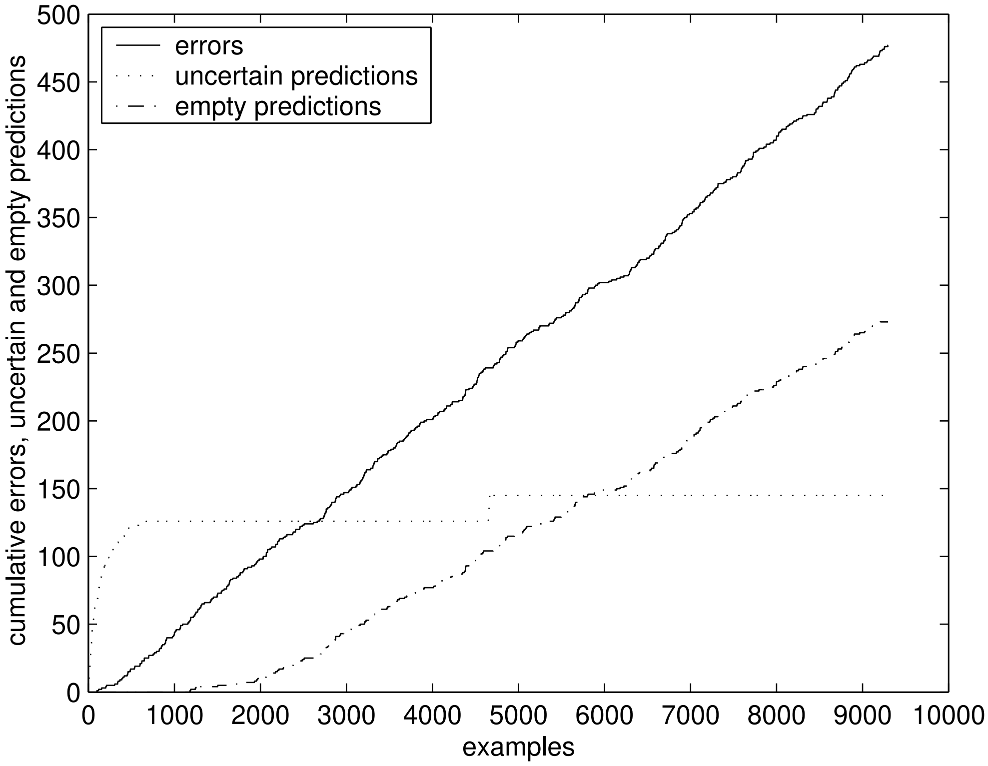

Figure 6 On-line performance of the Nearest Neighbour ICM with the training trial 4649 (= 9298/2) on the USPS data set for the confidence level 95%. In accordance with Theorem 2, starting from scratch at trial 4670 does not affect the error rate (solid line).

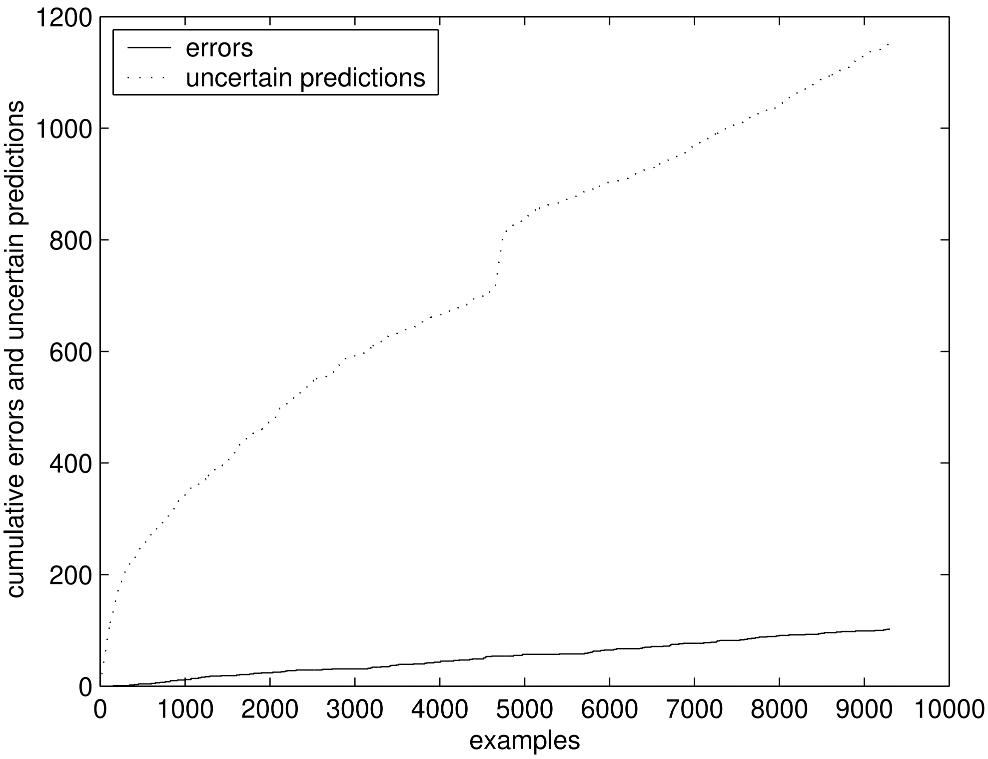

Figure 7 On-line performance of the Nearest Neighbour ICM with the training trial 4649 on the USPS data set for the confidence level 99%.

References

[1] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms.* MIT Press, Cambridge, MA, 1990.

[2] A. Philip Dawid. Statistical theory: the prequential approach. *Journal of the Royal Statistical Society* A, 147:278–292, 1984.

[3] A. Philip Dawid. Prequential analysis. In S. Kotz, N. L. Johnson, and C. B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 1 (update), pages 464–470. Wiley, New York, 1997.

[4] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

[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] Tom Melluish, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. Comparing the Bayes and typicalness frameworks. In *Proceedings of ECML'2001*, 2001. Full version published as a CLRC technical report TR-01-05; see `http://www.clrc.rhul.ac.uk`.

[8] Ilia Nouretdinov, Tom Melluish, and Vladimir Vovk. Ridge Regression Confidence Machine. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[9] Ilia Nouretdinov, Vladimir Vovk, Michael Vyugin, and Alex Gammerman. Pattern recognition and density estimation under the general i.i.d. assumption. In David Helmbold and Bob Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, volume 2111 of *Lecture Notes in Artificial Intelligence*, pages 337–353, 2001. Full version published as a CLRC technical report; see `http://www.clrc.rhul.ac.uk`.

[10] Ilia Nouretdinov, Vladimir Vovk, Vladimir V'yugin, and Alex Gammerman. Transductive Confidence Machine is universal. Manuscript, 2002.

[11] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the 2002 International Conference on Machine Learning and Applications*. CSREA Press, 2002. To appear.

[12] Craig Saunders, Alex Gammerman, and Vladimir Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 722–726, 1999.

[13] Mark J. Schervish. *Theory of Statistics.* Springer, New York, 1995.

[14] Albert N. Shiryaev. *Probability.* Springer, New York, second edition, 1996.

[15] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[16] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[17] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the 16th International Conference on Machine Learning*, pages 444–453, 1999.

Appendix: Proofs

Proof of Theorem 2

First we explain the basic idea of the proof. To show that $(\mathrm{err}_1, \ldots, \mathrm{err}_N)$ is distributed as B_δ^N (it will be easy to get rid of the assumption of a fixed horizon N), we use the standard idea of reversing the time (see, e.g., the proof of De Finetti's theorem in [13]). We can imagine that the sample $(z_1, \ldots, z_N)$ is generated in two steps: first, the bag $\langle z_1, \ldots, z_N \rangle$ is generated from some probability distribution (namely, the image of P under the mapping $(z_1, z_2, \ldots) \mapsto \langle z_1, \ldots, z_N \rangle$), and then the actual sample $(z_1, \ldots, z_N)$ is chosen randomly from the set of all orderings of the bag $\langle z_1, \ldots, z_N \rangle$. Already the second step ensures that, conditionally on knowing $\langle z_1, \ldots, z_N \rangle$ (and, therefore, unconditionally), the sequence $(\mathrm{err}_N, \ldots, \mathrm{err}_1)$ is distributed as B_δ^N. Indeed, roughly speaking (i.e., ignoring ties and borderline effects), err_N will be 1 if α_N is among the $N\delta$ largest α_i, and the probability of this is δ since all permutations are equiprobable; when z_N is disclosed, the value err_N will be settled; conditionally on knowing $\langle z_1, \ldots, z_N \rangle$ and z_N (and, therefore, knowing $\langle z_1, \ldots, z_{N-1} \rangle$), err_{N-1} will also be 1 with probability δ, and so on.

We start the proof by giving some preliminary definitions. The σ-algebra $\mathcal{G}_n$, $n = 0, 1, 2, \ldots$, is the collection of all measurable sets $E \subseteq \mathbf{Z}^\infty$ which satisfy

$$(z_1, z_2, \ldots) \in E \Longrightarrow (z_{\pi(1)}, \ldots, z_{\pi(n)}, z_{n+1}, z_{n+2}, \ldots) \in E$$

for any permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. In particular, $\mathcal{G}_0$ (the most informative σ-algebra) coincides with the original σ-algebra on $\mathbf{Z}^\infty$; $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$.

Fix a TCM Γ and a confidence level $1 - \delta$; these elements will always be left implicit in our notation. The proof will be based on the following lemma.

Lemma 1 *For any trial n,*

$$\mathbb{E}_{P \times U^\infty} (\mathrm{err}_n \mid \mathcal{G}_n) = \delta. \tag{34}$$

Proof Define

$$\mathrm{error}(z_1, \ldots, z_n) := \mathrm{err}_n(\Gamma, 1 - \delta, \omega),$$

where $\omega \in \mathbf{Z}^\infty$ is any continuation of the sequence $(z_1, \ldots, z_n)$; we do not reflect in the notation the dependence on the random numbers $(\tau_1, \tau_2, \ldots) \in [0, 1]^\infty$. First we prove that, for any $(z_1, \ldots, z_n)$,

$$\frac{1}{n!} \sum_\pi \mathbb{E}\, \mathrm{error}(z_{\pi(1)}, \ldots, z_{\pi(n)}) = \delta, \tag{35}$$

where π ranges over all permutations of $\{1, \ldots, n\}$ and $\mathbb{E}$ stands for the expected value w.r.t. to the distribution U^∞; it is intuitively clear, and will be easy to show formally, that (35) implies (34).

Let us fix a sequence $(z_1,\dots,z_n)\in\mathbf{Z}^n$. As usual we denote by α_i the result of applying the individual strangeness measure underlying Γ to the given examples: $(\alpha_1,\dots,\alpha_n)=A_n(z_1,\dots,z_n)$. For every $i=1,\dots,n$ define

$$p_i^+ := \frac{\#\{j=1,\dots,n \mid \alpha_j \ge \alpha_i\}}{n}, \quad p_i^- := \frac{\#\{j=1,\dots,n \mid \alpha_j > \alpha_i\}}{n}.$$

It is clear that $p_i^- < p_i^+$ and

$$p_i^+ - p_i^- = \frac{\#\{j \mid \alpha_j = \alpha_i\}}{n}.$$

Notice that the semi-closed intervals $[p_i^-, p_i^+)$, $i=1,\dots,n$, either coincide or are disjoint; it is also easy to see that they "lie next to each other", in the sense that their union is also a semi-closed interval (namely, $[0,1)$).

Let us say that an example z_i (more accurately, its index i) is

- *strange* if $p_i^+ \le \delta$
- *ordinary* if $p_i^- > \delta$
- *borderline* if $p_i^- \le \delta < p_i^+$.

We will use the notation $p^- := p_i^-$ and $p^+ := p_i^+$ where i is the index of any borderline example. Notice that the fraction of strange examples (i.e., the number of strange examples divided by n) is p^-, the fraction of ordinary examples is $1-p^+$, and the fraction of borderline examples is $p^+ - p^-$.

By the definition of rTCM, $\mathrm{error}(z_{\pi(1)},\dots,z_{\pi(n)})$ is 1 if the last example $z_{\pi(n)}$ is strange, is 0 if the last example is ordinary, and is 0 with probability

$$\frac{p^+ - \delta}{p^+ - p^-}$$

(cf. (16)) if the last example is borderline. Therefore, the expected value of $\mathrm{error}(z_{\pi(1)},\dots,z_{\pi(n)})$ over all equiprobable permutations π and random numbers $(\tau_1\tau_2\dots)\in[0,1]^\infty$ (in other words, the left-hand side of (35)) is

$$p^- + (p^+ - p^-)\frac{\delta - p^-}{p^+ - p^-} = \delta.$$

This proves (35).

It remains to prove that (35) implies (34). Let us assume (35). We say that a set $E\subseteq\mathbf{Z}^n$ is *symmetric* if

$$(z_1,\dots,z_n)\in E \Longrightarrow (z_{\pi(1)},\dots,z_{\pi(n)})\in E$$

for any permutation $\pi:\{1,\dots,n\}\to\{1,\dots,n\}$. We are required to prove that

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1,\dots,z_n)-\delta)\,dP = 0 \tag{36}$$

for any symmetric measurable set $E \subseteq \mathbf{Z}^n$.

First we notice that, if $G : \Omega \to \Omega$ is a bijection defined on a measurable space Ω and measurable in both directions, then for every measurable function $f : \Omega \to \mathbb{R}$, measurable set $E \subseteq \Omega$, and measure P on Ω,

$$\int_E f dP = \int_{E'} f' dP', \tag{37}$$

where the set E', function f', and measure P' are defined by

$$E' := G^{-1}(E), \;\; f'(\omega) := f(G(\omega)), \;\; P'(A) := P(G(A)).$$

Applying this to $\Omega := \mathbf{Z}^n$, $G(z_1, \ldots, z_n) := (z_{\pi(1)}, \ldots, z_{\pi(n)})$, and

$$f(z_1, \ldots, z_n) := \mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta, \tag{38}$$

where π is a permutation, we obtain

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP = \int_E \left(\mathbb{E}\,\mathrm{error}(z_{\pi(1)}, \ldots, z_{\pi(n)}) - \delta\right) dP$$

(remember that $E' = E$ and $P' = P$) and so, from (35),

$$\begin{gathered}\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP = \\ \frac{1}{n!} \sum_\pi \int_E \left(\mathbb{E}\,\mathrm{error}(z_{\pi(1)}, \ldots, z_{\pi(n)}) - \delta\right) dP = 0.\end{gathered}$$

∎

The other basic result that we will need is the following simple lemma.

Lemma 2 *For any trial $n \geq 1$, err_n is $\mathcal{G}_{n-1}$-measurable.*

Proof Fix a trial n. We are required to prove that the event $\{\mathrm{err}_n = 1\}$ is $\mathcal{G}_{n-1}$-measurable, i.e., invariant w.r.t. to permutations of the first $n-1$ examples. By the definition, (6), this follows from the invariance of the TCM Γ w.r.t. to permutations of the first $n-1$ examples, which, in its turn, follows (see (10) and (11)) from the invariance of the underlying individual strangeness measure (see (7)). ∎

The proof of Theorem 2 will use the following properties of conditional expectations:

A. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, ξ and η are bounded $\mathcal{F}$-measurable random variables, and η is $\mathcal{G}$-measurable,

$$\mathbb{E}(\xi\eta \,|\, \mathcal{G}) = \eta\, \mathbb{E}(\xi \,|\, \mathcal{G})$$

almost surely (see, e.g., [14], property K* in §II.7.4).

B. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, and ξ is a random variable,

$$\mathbb{E}(\mathbb{E}(\xi \,|\, \mathcal{F}) \,|\, \mathcal{G}) = \mathbb{E}(\xi \,|\, \mathcal{G})$$

almost surely (this is property G* [14], §II.7.4); in particular,

$$\mathbb{E}(\mathbb{E}(\xi \,|\, \mathcal{F})) = \mathbb{E}(\xi).$$

Fix temporarily positive integer N. To simplify the formulas, we will often omit curly braces. First we prove that, for any $n = 1, \ldots, N$,

$$\mathbb{P}\left((\mathrm{err}_n, \ldots, \mathrm{err}_1) = (\omega_n, \ldots, \omega_1) \,|\, \mathcal{G}_n\right) = \delta^k (1-\delta)^{n-k}, \tag{39}$$

where $\mathbb{P}$ refers to the probability distribution $P \times U^\infty$ and k is the number of 1s in $(\omega_n, \ldots, \omega_1)$. The proof is by induction in n. For $n = 1$, (39) immediately follows from Lemma 1. For $n > 1$ we obtain, making use of Lemmas 1 and 2, properties A and B of conditional expectations, and the inductive assumption:

$$\begin{aligned} &\mathbb{P}\left((\mathrm{err}_n, \ldots, \mathrm{err}_1) = (\omega_n, \ldots, \omega_1) \,|\, \mathcal{G}_n\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\mathbb{I}_{\mathrm{err}_n = \omega_n} \mathbb{I}_{(\mathrm{err}_{n-1}, \ldots, \mathrm{err}_1) = (\omega_{n-1}, \ldots, \omega_1)} \,|\, \mathcal{G}_{n-1}\right) \,|\, \mathcal{G}_n\right) \\ &= \mathbb{E}\left(\mathbb{I}_{\mathrm{err}_n = \omega_n} \mathbb{E}\left(\mathbb{I}_{(\mathrm{err}_{n-1}, \ldots, \mathrm{err}_1) = (\omega_{n-1}, \ldots, \omega_1)} \,|\, \mathcal{G}_{n-1}\right) \,|\, \mathcal{G}_n\right) \\ &= \mathbb{E}\left(\mathbb{I}_{\mathrm{err}_n = \omega_n} \delta^{k^\dagger} (1-\delta)^{(n-1)-k^\dagger} \,|\, \mathcal{G}_n\right) = \delta^k (1-\delta)^{n-k} \end{aligned}$$

almost surely, where $\mathbb{I}_E$ means the indicator of E, k and $k^\dagger$ are the number of 1s in $(\omega_n, \ldots, \omega_1)$ and $(\omega_{n-1}, \ldots, \omega_1)$, respectively, and the expected value $\mathbb{E}$ is taken over $\mathbb{P} = P \times U^\infty$.

By property B, (39) immediately implies

$$\mathbb{P}\left((\mathrm{err}_N, \ldots, \mathrm{err}_1) = (\omega_N, \ldots, \omega_1)\right) = \delta^k (1-\delta)^{N-k},$$

where k is the number of 1s in $(\omega_N \ldots \omega_1)$. Therefore, we have proved that the distribution of the random sequence $\mathrm{err} \in \{0,1\}^\infty$ coincides with B_δ^∞ on the σ-algebra $\mathcal{F}_N$ generated by the events $\{(\omega_1, \omega_2, \ldots) \in \{0,1\}^\infty \,|\, \omega_i = 1\}$, $i = 1, \ldots, N$. It is well known (see, e.g., [14], Theorem II.3.3) that this implies that the distribution of err coincides with B_δ^∞ on all measurable sets in $\{0,1\}^\infty$.

Precise statement of Theorem 3

First we define rICM formally. The rICM based on D, Δ, and the parametric sequence $m_1, m_2, \ldots$ (as before, m_i are required to satisfy $m_1 < m_2 < \cdots$) is defined to be the randomised region predictor Γ such that $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta, \tau_n)$ is computed as follows:

- if $n \le m_1$, $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta, \tau)$ is found using a fixed rTCM;

- ⋄ otherwise, find the k such that $m_k < n \le m_{k+1}$, and include $y \in \mathbf{Y}$ in $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta, \tau_n)$
  - with probability 1 if
  $$\frac{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j > \alpha_n\}}{n - m_k} > \delta$$
  - with probability 0 if
  $$\frac{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j \ge \alpha_n\}}{n - m_k} \le \delta$$
  - with probability
  $$\frac{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j \ge \alpha_n\} - (n - m_k)\delta}{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j = \alpha_n\}} \tag{40}$$
  (i.e., if
  $$\tau_n < \frac{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j \ge \alpha_n\} - (n - m_k)\delta}{\#\{j = m_k + 1, \ldots, n \,|\, \alpha_j = \alpha_n\}}) \tag{41}$$
  otherwise,

  where $\alpha_{m_k+1}, \ldots, \alpha_n$ are defined by (22).

The main assertion made in Theorem 3 is that the image of the probability distribution $P \times U^\infty$ under the mapping

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \mapsto \mathrm{err}(\Gamma, 1-\delta, \omega, \tau)$$

is B_δ^∞ for any rICM Γ.

Remark In our definitions of rTCM and rICM we assumed that the same random number τ_n is used for every potential label y of x_n. In fact, we can also use a separate random number τ_n^y for each $y \in \mathbf{Y}$, with the random numbers τ_n^y, $n = 1, 2, \ldots$, $y \in \mathbf{Y}$, independent. On the other hand, an arbitrary correlation between τ_n^y, $y \in \mathbf{Y}$, can be allowed; Theorems 2 and 3 will continue to hold as long as the random numbers $\tau_n^{y_n}$, $n = 1, 2, \ldots$, are independent.

Proof of Theorem 3

Let us set $m_0 := 0$. In this proof, the σ-algebra $\mathcal{G}_n$, $n = 1, 2, \ldots$, is the collection of all measurable sets $E \subseteq \mathbf{Z}^\infty$ which satisfy the following:

- ⋄ if k is the largest non-negative integer such that $m_k < n$,
$$(z_1, z_2, \ldots) \in E \Longrightarrow$$
$$(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}, z_{n+1}, z_{n+2}, \ldots) \in E$$
  for any permutation $\pi : \{m_k + 1, \ldots, n\} \to \{m_k + 1, \ldots, n\}$;

- if k is any positive integer such that $m_k < n$,

$$(z_1, z_2, \ldots) \in E \Longrightarrow$$
$$(z_1, \ldots, z_{m_{k-1}}, z_{\pi(m_{k-1}+1)}, \ldots, z_{\pi(m_k)}, z_{m_k+1}, z_{m_k+2}, \ldots) \in E$$

for any permutation $\pi : \{m_{k-1}+1, \ldots, m_k\} \to \{m_{k-1}+1, \ldots, m_k\}$.

As before, $\mathcal{G}_0$ is the original σ-algebra on $\mathbf{Z}^\infty$. It is obvious that $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$.

The proof of Theorem 2 obviously works for ICMs as well, with the possible exception of Lemma 1. Therefore, in this subsection we will only show how to prove the following analogue of Lemma 1.

Lemma 3 *Any rICM satisfies (34), for any confidence level $1-\delta$ and any trial n.*

Proof If $n \le m_1$, an rTCM is used and so we have nothing to prove; we will assume $n > m_1$. The proof is parallel to the proof of Lemma 1.

Analogously to the reduction of the proof of Lemma 1 to proving (35) for any $(z_1, \ldots, z_n)$, we first prove that, for any $(z_1, \ldots, z_n)$,

$$\frac{1}{(n-m_k)!} \sum_{\pi} \mathbb{E}\,\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) = \delta, \tag{42}$$

where k is the largest positive integer for which $m_k < n$ and $\pi : \{m_k + 1, \ldots, n\} \to \{m_k + 1, \ldots, n\}$ ranges over all permutations. Define α_j, $j = m_k + 1, \ldots, n$, by the formula

$$\alpha_j := \Delta\left(y_j, D_{\lbag(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\rbag}(x_j)\right)$$

(cf. (22)). For every $i = m_k + 1, \ldots, n$ define

$$p_i^+ := \frac{\#\{j = m_k + 1, \ldots, n \mid \alpha_j \ge \alpha_i\}}{n - m_k},$$
$$p_i^- := \frac{\#\{j = m_k + 1, \ldots, n \mid \alpha_j > \alpha_i\}}{n - m_k}.$$

It is clear that $p_i^- < p_i^+$ and

$$p_i^+ - p_i^- = \frac{\#\{j = m_k + 1, \ldots, n \mid \alpha_j = \alpha_i\}}{n - m_k}.$$

Again the semi-closed intervals $[p_i^-, p_i^+)$, $i = m_k + 1, \ldots, n$, either coincide or are disjoint; they also "lie next to each other".

Let us say that an example z_i, $i = m_k + 1, \ldots, n$, is

- *strange* if $p_i^+ \le \delta$
- *ordinary* if $p_i^- > \delta$

- $\diamond$ *borderline* if $p_i^- \le \delta < p_i^+$.

We will use the notation $p^- := p_i^-$ and $p^+ := p_i^+$, where i is the index of any borderline example. The fraction of strange examples (i.e., the number of strange examples divided by $n - m_k$) is p^-, the fraction of ordinary examples is $1 - p^+$, and the fraction of borderline examples is $p^+ - p^-$.

By the definition of rICM, $\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)})$ is 1 if the last example $z_{\pi(n)}$ is strange, is 0 if the last example is ordinary, and is 0 with probability

$$\frac{p^+ - \delta}{p^+ - p^-}$$

(cf. (40)) if the last example is borderline. Therefore, the expected value of $\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)})$ over all equiprobable permutations π of $\{m_k + 1, \ldots, n\}$ and random numbers $(\tau_1 \tau_2 \ldots) \in [0,1]^\infty$ (i.e., the left-hand side of (42)) is

$$p^- + (p^+ - p^-)\frac{\delta - p^-}{p^+ - p^-} = \delta.$$

This proves (42).

To finish the proof it remains to establish (34). We say that a set $E \subseteq \mathbf{Z}^n$ is *calibration symmetric* if

$$(z_1, \ldots, z_n) \in E \Longrightarrow (z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) \in E$$

for any permutation $\pi : \{m_k + 1, \ldots, n\} \to \{m_k + 1, \ldots, n\}$. It is sufficient to prove that

$$\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP = 0$$

for any calibration symmetric measurable set $E \subseteq \mathbf{Z}^n$.

Applying (37) to $\Omega := \mathbf{Z}^n$,

$$G(z_1, \ldots, z_n) := (z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}),$$

and (38), where π is a permutation of $\{m_k + 1, \ldots, n\}$, we obtain

$$\begin{aligned} &\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP \\ = &\int_E \left(\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) - \delta\right) dP \end{aligned}$$

(again $E' = E$ and $P' = P$) and so, from (42),

$$\begin{aligned} &\int_E (\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_n) - \delta)\, dP = \\ &\frac{1}{(n - m_k)!} \sum_\pi \int_E \left(\mathbb{E}\,\mathrm{error}(z_1, \ldots, z_{m_k}, z_{\pi(m_k+1)}, \ldots, z_{\pi(n)}) - \delta\right) dP = 0, \end{aligned}$$

π ranging over the permutations of $\{m_k + 1, \ldots, n\}$. ∎

Proof and discussion of Corollary 3

When discussing p-values corresponding to region predictors, it is convenient to deal with the latter's' "normalised" versions. Formally, the *normal form* Γ^{norm} of a region predictor Γ is defined by

$$\Gamma^{\mathrm{norm}}(x_1, y_1, \ldots, x_n, 1-\delta) := \bigcup_{(1-\delta')<(1-\delta)} \Gamma(x_1, y_1, \ldots, x_n, 1-\delta').$$

We say that Γ is *normalised* if $\Gamma^{\mathrm{norm}} = \Gamma$; These definitions are modified in the obvious way for randomised region predictors. The following lemma lists some basic properties of the operation $^{\mathrm{norm}}$.

Lemma 4 *For any region predictor Γ (randomised or deterministic),*

1. Γ^{norm} dominates Γ, *in the sense that*
$$\Gamma^{\mathrm{norm}}(x_1, y_1, \ldots, x_n, 1-\delta) \subseteq \Gamma(x_1, y_1, \ldots, x_n, 1-\delta)$$
for all $(x_1, y_1, \ldots, x_n)$ and $1-\delta$;
2. $(\Gamma^{\mathrm{norm}})^{\mathrm{norm}} = \Gamma^{\mathrm{norm}}$*;*
3. *Γ is normalised if and only if the set*
$$\{\delta \mid y_n \in \Gamma(x_1, y_1, \ldots, x_n, 1-\delta)\}$$
is always open in $(0,1)$;
4. *if Γ is normalised,*
$$\mathrm{err}_n(\Gamma, 1-\delta) = \mathbb{I}_{p_n(\Gamma)\le\delta};$$
5. *all TCMs, rTCMs, ICMs, and rICMs are normalised;*

Proof The first two properties are obvious. Properties 3 and 4 follow from the following restatement of the definition of Γ^{norm}:

$$y_n \in \Gamma^{\mathrm{norm}}(x_1, y_1, \ldots, x_n, 1-\delta)$$

if and only if

$$\exists (1-\delta') < (1-\delta) : y_n \in \Gamma(x_1, y_1, \ldots, x_n, 1-\delta').$$

Property 5 follows from the fact that all inequalities involving δ in (10), (17), (21), and (41) are strict. ∎

Lemma 4 shows that $\mathrm{err}_n(\Gamma, 1-\delta) = \mathbb{I}_{p_n(\Gamma)\le\delta}$ for CM and rCM. This shows that the random variables p_n have the uniform distribution U in $[0,1]$. Corollary 3 now follows from the following simple fact: if a sequence $p_1, p_2, \ldots$ of random variables distributed as U is such that, for any sequence $(\delta_1, \delta_2, \ldots) \in (0,1)^\infty$, the random variables $\mathbb{I}_{p_n\le\delta_n}$, $n = 1, 2, \ldots$, are independent, then the random variables p_n themselves are independent (see, e.g., [14], Theorem II.5).

Remark The following example (due to Ilia Nouretdinov) shows that we cannot deduce Corollary 3 from Theorems 2 and 3 themselves. Suppose the joint distribution of random variables p_1 and p_2 is uniform in any square $\left[\frac{i-1}{4}, \frac{i}{4}\right] \times \left[\frac{j-1}{4}, \frac{j}{4}\right]$, $i = 1, \ldots, 4$, $j = 1, \ldots, 4$, and the probabilities of these squares are

$$\mathbb{P}\left\{\frac{i-1}{4} < p_1 < \frac{i}{4} \ \& \ \frac{j-1}{4} < p_2 < \frac{j}{4}\right\} = \frac{A_{ij}}{16},$$

where A_{ij} is the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 2 & 0 & 1 & 1 \end{pmatrix}.$$

Then p_1 and p_2 are distributed uniformly in $[0, 1]$, the random variables $\mathbb{I}_{p_1 \le \delta}$ and $\mathbb{I}_{p_2 \le \delta}$ are independent for any $\delta \in (0, 1)$, but the variables p_1 and p_2 themselves are not independent.

Proof of Theorem 5

All work is done in [6]; we will only show how their construction fits our framework.

Fix a class F of binary functions $f : \mathbf{X} \to \{0, 1\}$ of finite VC-dimension V. For any finite sequence of objects $(x_1, \ldots, x_n) \in \mathbf{X}^n$, consider the following 1-*inclusion graph* $G(x_1, \ldots, x_n)$:

- ◇ the nodes of $G(x_1, \ldots, x_n)$ are the set

$$\{(f(x_1), \ldots, f(x_n)) \mid f \in F\} \subseteq \{0, 1\}^n;$$

- ◇ there is an edge between two nodes of $G(x_1, \ldots, x_n)$ if and only if the Hamming distance between them is 1 (i.e., they differ in just one place).

The following is the result which will imply Theorem 5.

Theorem 6 ([6], Theorem 2.2) *For any $(x_1, \ldots, x_n) \in \mathbf{X}^*$, it is possible to direct all edges of $G(x_1, \ldots, x_n)$ in such a way that the out-degree of each node is at most V.*

For any sequence $(x_1, \ldots, x_n)$ fix a directed graph $\overrightarrow{G}(x_1, \ldots, x_n)$ made from $G(x_1, \ldots, x_n)$ with out-degree at most V.

Now we can construct the required TCM. Define an individual strangeness measure (8) as follows:

- ◇ if $(z_1, \ldots, z_n)$ is not consistent with F, set, e.g., $\alpha_i := 0$, $i = 1, \ldots, n$;

- ⋄ otherwise, $(y_1, \dots, y_n)$ is a node of $\overrightarrow{G}(x_1, \dots, x_n)$ (where $(x_i, y_i) := z_i \in \mathbf{X} \times \mathbf{Y}$, $i = 1, \dots, n$); for each $i = 1, \dots, n$, set $\alpha_i := 1$ if

  $$(y_1, \dots, y_{i-1}, 1 - y_i, y_{i+1}, \dots, y_n)$$

  is a node of $\overrightarrow{G}(x_1, \dots, x_n)$ and

  $$((y_1, \dots, y_{i-1}, y_i, y_{i+1}, \dots, y_n), (y_1, \dots, y_{i-1}, 1 - y_i, y_{i+1}, \dots, y_n))$$

  is an edge of $\overrightarrow{G}(x_1, \dots, x_n)$; otherwise, set $\alpha_i := 0$.

To see that (33) holds, we will argue indirectly: the negation of (33) implies that the pair

$$\{(y_1, \dots, y_{n-1}, 0), (y_1, \dots, y_{n-1}, 1)\}$$

is an edge of $G(x_1, \dots, x_n)$; assuming, for concreteness, that this edge is directed from $(y_1, \dots, y_{n-1}, 0)$ to $(y_1, \dots, y_{n-1}, 1)$ in $\overrightarrow{G}(x_1, \dots, x_n)$, we can see that 0 will not be included in $\Gamma(z_1, \dots, z_{n-1}, x_n, 1 - V/n)$.