Online Region Prediction with Real Teachers

Daniil Ryabko, Vladimir Vovk and Alex Gammerman

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project

Working Paper #7

March 22, 2003

Project web site:
http://vovk.net/kp

Abstract

In the online prediction scenario the predictor's task is to predict the label of an object given by Nature at each trial, based on the labels for the objects learned so far. At each trial Nature discloses the correct label for the current object, so the predictor is being taught. While this pure online scenario is convenient for theoretical studies, it is a poor model for many practical applications: the situation where the correct answers are given immediately after each prediction does not often occur in reality. In this work we suggest a more general scenario for online prediction, according to which correct answers may be given with some delay and not at every trial. We modify a particular class of region predictors, Transductive Confidence Machines, which have been proved to have several useful properties, to work in the new scenario, and find some sufficient conditions under which their rates of erroneous and uncertain predictions remain the same in the new online scenario as in the traditional one.

Contents

1 Introduction

Online region predictors work as follows. Nature outputs some examples, i.e., pairs consisting of an object and a classification (e.g., an object may be a handwritten digit and its classification may be any number between 0 and 9). We assume that examples are independent and identically distributed according to some probability distribution P. At every trial a predictor gives a region (a set) of possible labels for the current object. A region may consist of one label (the best case if the label is correct) but it also may consist of all possible labels (this practically useless prediction is always correct). A predictor may also output an empty prediction, which always leads to an error. A prediction which consists of more then one element is called *uncertain*. There are two main criteria for evaluating how good a region predictor is: the error rate and the rate of uncertain predictions.

While the online scenario is convenient for theoretical studies, in practice, however, rarely does one immediately get the true label for every object (otherwise the prediction is not needed). In this paper we suggest the following modified scenario for online prediction, which we call *online region prediction with real teachers*. At each trial Nature outputs an object, and at some trials it also outputs the correct label for one of the previous objects, so that the predictor can use this data afterwards. Thus, some labels may never be revealed while others may be revealed with some delay. As before, examples are assumed to be i.i.d. The trials at and for which true labels are given are chosen independently of data.

The suggested scenario is studied for a particular class of a region predictors, called Transductive Confidence Machines (TCMs). First introduced in [4], Transductive Confidence Machine is a way of constructing region predictors from machine-learning algorithms. One of the advantages of the TCM is that it is always *well-calibrated*: the number of errors it makes up to trial n divided by n tends to δ almost surely, where $\delta \in (0, 1)$ is any pre-specified "significance level" (see [5]). Moreover, the probability of error at each trial is δ and errors are made independently at different trials. (Strictly speaking, this statement is applicable to the version of TCM called "randomised TCM"; this is the version we consider in this paper. The reader can check easily that all results carry over to the case of deterministic TCM as well.)

The main result of this paper (Theorem 1) can be illustrated by the following simple example. Suppose only every kth label is revealed to a TCM, and even this is done with a delay of l, where k and l are positive

integer constants. Then the algorithm will remain well-calibrated, and its asymptotic rate of uncertain predictions will not suffer.

In this paper we find sufficient conditions on the amount of information revealed to the TCM up to the nth trial under which it remains well-calibrated and has the same asymptotic rate of uncertain predictions. A related paper is [2], which continues research on our model establishing some necessary conditions for a TCM to remain well-calibrated.

2 Preliminaries

Traditionally, online region prediction protocol is defined as follows. Nature outputs *examples* $(x_1, y_1), (x_2, y_2), \ldots$; each example (x_i, y_i) consists of an object $x_i \in \mathbf{X}$ and a label $y_i \in \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are measurable spaces called the *object space* and the *label space* correspondingly. The notation $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ is used for the measurable space of all examples. Examples are drawn according to some probability distribution P^∞ on $\mathbf{Z}^\infty$.

A *region predictor* is a measurable function

$$\Gamma_\gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n),$$

where $n \in \mathbb{N}$, the $(x_i, y_i) \in \mathbf{Z}$, $i = 1, \ldots, n-1$ are examples and $x_n \in \mathbf{X}$ is an object which satisfies

$$\Gamma_{\gamma_1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma_{\gamma_2}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

whenever $\gamma_1 \le \gamma_2$. We are interested in prediction with confidence, and so the predictor is given an extra input $\gamma = (1 - \delta) \in (0, 1)$ which is called the *confidence level*; the complementary value δ is called the *significance level*. An important modification of this definition is where the region predictor is allowed to depend on additional inputs, random numbers $\tau_i \in [0, 1]$ (τ_i are assumed to be independently distributed according to the uniform distribution in $[0, 1]$ and to be independent of the examples); however, this case reduces to the case of deterministic region predictors by extending the object space $\mathbf{X}$ to $\mathbf{X} \times [0, 1]$, so that τ_i becomes an element of the extended object x_i. Therefore, we need not mention the random numbers τ_i explicitly.

The number of errors $\mathrm{Err}_n(\Gamma_{1-\delta})$ which Γ makes at the confidence level $1 - \delta$ up to the trial $n \in \mathbb{N}$ is defined as

$$\#\{i = 1, \ldots, n \mid y_i \notin \Gamma_{1-\delta}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i)\}.$$

The indicator of an individual error err_n at trial n is defined to be 1 if $y_n \notin \Gamma_n$ and 0 otherwise (we often use the notation Γ_n instead of

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

when the values of other parameters are clear). Similarly, the number of uncertain predictions $\mathrm{Unc}_n(\Gamma_{1-\delta})$ that $\Gamma_{1-\delta}$ makes up to the trial $n \in N$ is defined to be

$$\#\{i = 1, \ldots, n \mid |\Gamma_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{i-1}, x_i)| > 1\},$$

and the indicator unc_n of uncertain prediction at individual trial n to be 1 if $|\Gamma_n| > 1$ and 0 otherwise.

A region predictor is called *well-calibrated* if, for any $\delta \in (0, 1)$,

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma_{1-\delta})}{n} \to \delta \quad \text{a.s.}$$

under any probability distribution P^∞ generating the examples.

Transductive Confidence Machine is a region predictor constructed from a machine-learning algorithm (more precisely, from the so-called individual strangeness measure). A complete description of how TCMs work can be found in [5, 6]. In this paper it is sufficient for us to observe that any TCM has the following properties. First, its errors err_i, $i \in \mathbb{N}$, are distributed as independent Bernoulli trials with probability of 1 equal to δ, where $1 - \delta$ is the chosen confidence level (in particular, any TCM is well calibrated). And second, its predictions do not depend on the order of the examples learnt so far: if Γ is a TCM then

$$\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) = \Gamma(x_{\pi(1)}, y_{\pi(1)}, \ldots, x_{\pi(n-1)}, y_{\pi(n-1)}, x_n)$$

for any permutation π. We call region predictors satisfying this property *invariant*.

3 Main Result

We suggest the following modified scenario for online region prediction.

We call a function $\mathcal{L} : N \to \mathbb{N}$ defined on an infinite set $N \subset \mathbb{N}$ a *learning function* if $\mathcal{L}(n) \in \{1, \ldots, n\}$ for all $n \in N$, and $m \neq n$ implies

$\mathcal{L}(m) \neq \mathcal{L}(n)$ for all $m, n \in \mathbb{N}$. The domain of the learning function $\mathcal{L}$ in this paper is always denoted by $N = \{n_1, n_2, \dots\}$ where $n_1 < n_2 < \cdots$. We define the total amount of information available at the beginning of trial n to a prediction algorithm taught according to the learning function $\mathcal{L}$ as $s(n) = \#\{i \mid i \in N, i < n\}$.

Suppose that $\Gamma_{1-\delta}$ is an online region predictor and $\mathcal{L}$ is a learning function. Then we define the *$\mathcal{L}$-taught version of* $\Gamma_{1-\delta}$ as follows:

$$\Gamma^{\mathcal{L}}_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) = \Gamma_{1-\delta}(x_{\mathcal{L}(n_1)}, y_{\mathcal{L}(n_1)}, \dots, x_{\mathcal{L}(n_{s(n)})}, y_{\mathcal{L}(n_{s(n)})}, x_n).$$

So at the end of each trial $n \in N$ the predictor $\Gamma^{\mathcal{L}}_{1-\delta}$ "learns" the label $y_{\mathcal{L}(n)}$ if $n \in N$ and "learns" nothing otherwise.

Consider several examples.

Ideal teacher. If $N = \mathbb{N}$ and $\mathcal{L}(n) = n$ for each $n \in N$, then $\Gamma^{\mathcal{L}}_{1-\delta}$ is equal to $\Gamma_{1-\delta}$.

Slow teacher with a fixed lag. If $N = \{l+1, l+2, \dots\}$ for some $l \in \mathbb{N}$ and $\mathcal{L}(n) = n - l$ for $n \in N$, then $\Gamma^{\mathcal{L}}_{1-\delta}$ is a predictor which learns true labels with the delay l.

Slow teacher. The previous example can be generalised as follows. Let $l(n) = n + \mathrm{lag}(n)$ where $\mathrm{lag} : \mathbb{N} \to \mathbb{N}$ is an integer function. We define $N = l(\mathbb{N})$ and $\mathcal{L}(n) = l^{-1}(n)$, $n \in N$. Then $\Gamma^{\mathcal{L}}_{1-\delta}$ models a predictor which learns the true label for each example x_n with the delay $\mathrm{lag}(n)$. This is what we call a *region predictor with slow teacher with delay* lag.

Lazy teacher. Suppose that $N \neq \mathbb{N}$ and $\mathcal{L}(n) = n$, $n \in N$; then $\Gamma^{\mathcal{L}}_{1-\delta}$ is a *region predictor with lazy teachers*: it is given true labels immediately but not on every step.

Prior to stating the main theorem about $\mathcal{L}$-taught TCMs we need to give one more definition. If Γ is a region predictor, set

$$U(\Gamma) = \left[\liminf_{n\to\infty} \frac{\mathrm{Unc}_n(\Gamma)}{n}, \limsup_{n\to\infty} \frac{\mathrm{Unc}_n(\Gamma)}{n}\right].$$

The interval $U(\Gamma)$ characterises the asymptotical uncertainty of Γ; of course, this is a random interval, since it depends on the actual examples output by Nature. It turns out, however, that in the most important case (covering TCM and $\mathcal{L}$-taught TCM) this interval is close to being deterministic.

Lemma 1 *For each invariant region predictor Γ and probability distribution P in $\mathbf{Z}$ there exists an interval $[a,b] \subseteq \mathbb{R}$ such that $U(\Gamma) = [a,b]$ P^∞-almost surely.*

Proof The statement of this lemma is an immediate consequence of the Hewitt-Savage zero-one law (see, e.g., [3]). ∎

We will use the notation $U(\Gamma, P)$ for the interval whose existence is asserted in the lemma; we will call it the *asymptotical uncertainty* of Γ with examples distributed according to P.

Theorem 1 *Let $\Gamma_{1-\delta}$ be a TCM, let $\mathcal{L}$ be a learning function, and let $\Gamma^{\mathcal{L}}_{1-\delta}$ be the $\mathcal{L}$-taught version of $\Gamma_{1-\delta}$. The following statements hold for any probability distribution P^∞ generating the examples.*

- *If*
$$\sum_{i=2}^{\infty} \frac{(n_i - n_{i-1})^2}{n_i^2} < \infty \tag{1}$$
then $\Gamma^{\mathcal{L}}_{1-\delta}$ is well calibrated.
- *If, for some $k \in \mathbb{N}$, $n_{i+1} - n_i = k$ from some n on, then $U(\Gamma^{\mathcal{L}}_{1-\delta}, P) = U(\Gamma_{1-\delta}, P)$.*

A discussion of cases in which the first statement of the theorem is satisfied is given in the next section.

Proof Suppose that B is a region predictor (such as $\Gamma_{1-\delta}$ or $\Gamma^{\mathcal{L}}_{1-\delta}$). We introduce the following "predictable" versions of $\mathrm{err}_n(B)$ and $\mathrm{unc}_n(B)$:

$$\overline{\mathrm{err}}_n(B) = P\Big\{(x,y) \in \mathbf{Z} \mid y \notin B_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x)\Big\},$$
$$\overline{\mathrm{unc}}_n(B) = P\Big\{(x,y) \in \mathbf{X} \mid \big|B_{1-\delta}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x)\big| > 1\Big\},$$
$$\overline{\mathrm{Err}}_n(B) = \sum_{i=1}^{n} \overline{\mathrm{err}}_i(B),$$
$$\overline{\mathrm{Unc}}_n(B) = \sum_{i=1}^{n} \overline{\mathrm{unc}}_i(B).$$

Since $\mathrm{Err}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{Err}}_n(\Gamma^{\mathcal{L}}_{1-\delta})$ and $\mathrm{Unc}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{Unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta})$ are martingales, and

$$|\,\mathrm{err}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{err}}_n(\Gamma^{\mathcal{L}}_{1-\delta})| \le 1,$$
$$|\,\mathrm{unc}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta})| \le 1,$$

the martingale strong law of large numbers (see, e.g., [3]) implies that

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{Err}}_n(\Gamma^{\mathcal{L}}_{1-\delta})}{n} = 0 \text{ a.s.}$$

and

$$\lim_{n\to\infty} \frac{\mathrm{Unc}_n(\Gamma^{\mathcal{L}}_{1-\delta}) - \overline{\mathrm{Unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta})}{n} = 0 \text{ a.s.;}$$

this actually means that we can study $\overline{\mathrm{Err}}_n(\Gamma^{\mathcal{L}}_{1-\delta})$ and $\overline{\mathrm{Unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta})$ instead of $\mathrm{Err}_n(\Gamma^{\mathcal{L}}_{1-\delta})$ and $\mathrm{Unc}_n(\Gamma^{\mathcal{L}}_{1-\delta})$.

In this proof, where the arguments of functions $\Gamma_{1-\delta}$ and $\Gamma^{\mathcal{L}}_{1-\delta}$ are not given explicitly, we assume that the predictor $\Gamma^{\mathcal{L}}_{1-\delta}$ receives the sequence $\big(z_i : i \in \mathbb{N}\big)$ while the predictor $\Gamma_{1-\delta}$ receives the sequence $\big(z_{\mathcal{L}(i)} : i \in N\big)$. The latter sequence is distributed according to P^∞, since the choice of $\mathcal{L}$ (and its domain N), by definition, does not depend on the examples z_i, $i \in \mathbb{N}$.

We first prove the second statement of the theorem. For any $n \in \mathbb{N}$, by definition of $\Gamma^{\mathcal{L}}_{1-\delta}$ and using the property of a TCM $\Gamma_{1-\delta}$ that it does not depend on the order of the learnt examples, we have

$$\begin{aligned}
\overline{\mathrm{unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta}) &= P\Big\{(x,y) \in \mathbf{Z} \mid y \notin \Gamma^{\mathcal{L}}_{1-\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x)\Big\} \\
&= P\Big\{(x,y) \in \mathbf{Z} \mid y \notin \Gamma_{1-\delta}(x_{\mathcal{L}(n_1)}, y_{\mathcal{L}(n_1)}, \dots, x_{\mathcal{L}(n_s(n))}, y_{\mathcal{L}(n_{s(n)})}, x)\Big\} \\
&= \overline{\mathrm{unc}}_{s(n)+1}(\Gamma_{1-\delta}).
\end{aligned}$$

Thus, since $s(n) = n/k + O(1)$, we have

$$\sum_{i=1}^{n} \overline{\mathrm{unc}}_i(\Gamma^{\mathcal{L}}_{1-\delta}) = k \sum_{i=1}^{\lfloor n/k \rfloor} \overline{\mathrm{unc}}_i(\Gamma_{1-\delta}) + O(1),$$

and so $\overline{\mathrm{Unc}}_n(\Gamma^{\mathcal{L}}_{1-\delta}) = k\overline{\mathrm{Unc}}_{\lfloor n/k \rfloor}(\Gamma_{1-\delta}) + o(n)$. The desired statement immediately follows.

Now we proceed with the first statement of the theorem. Clearly,

$$\overline{\mathrm{Err}}_{n_k}(\Gamma^{\mathcal{L}}_{1-\delta}) = n_1\overline{\mathrm{err}}_1(\Gamma_{1-\delta})+(n_2-n_1)\overline{\mathrm{err}}_2(\Gamma_{1-\delta})+\cdots+(n_k-n_{k-1})\overline{\mathrm{err}}_{k-1}(\Gamma_{1-\delta}) \tag{2}$$

for any $k \in \mathbb{N}$. Denote

$$\overline{e}_1 = n_1\overline{\mathrm{err}}_1(\Gamma_{1-\delta}),\ \overline{e}_2 = (n_2 - n_1)\overline{\mathrm{err}}_2(\Gamma_{1-\delta}),\ \ldots$$

and

$$e_1 = n_1\,\mathrm{err}_1(\Gamma_{1-\delta}),\ e_2 = (n_2 - n_1)\,\mathrm{err}_2(\Gamma_{1-\delta}),\ \ldots\ .$$

It is easy to see that $\overline{e}_i - e_i$, $i \in \mathbb{N}$, is a martingale difference sequence with respect to $\Gamma_{1-\delta}$'s input sequence, $z_{\mathcal{L}(n_i)}$, $i \in \mathbb{N}$. Moreover,

$$\mathbb{E}\big((\overline{e}_i - e_i)^2 \mid z_{\mathcal{L}(n_1)}, \ldots, z_{\mathcal{L}(n_{i-1})}\big) \le (n_i - n_{i-1})^2,$$

for $i \in \mathbb{N}$, assuming that $n_0 = 0$. Thus,

$$\sum_{i=1}^{\infty} \frac{1}{n_i^2}\,\mathbb{E}\big((\overline{e}_i - e_i)^2 \mid z_{\mathcal{L}(n_1)}, \ldots, z_{\mathcal{L}(n_{i-1})}\big) \le \sum_{i=1}^{\infty} \frac{(n_i - n_{i-1})^2}{n_i^2} < \infty.$$

We can use (2) and the martingale strong law of large numbers to conclude that, as $k \to \infty$,

$$\frac{1}{n_k}\Big(\overline{\mathrm{Err}}_{n_k}(\Gamma^{\mathcal{L}}_{1-\delta}) - \sum_{i=1}^{k-2} e_i\Big) = \frac{1}{n_k}\sum_{i=1}^{k-2}(\overline{e}_i - e_i) \to 0 \text{ a.s.}$$

Analogously,

$$\frac{1}{n_k}\left(\sum_{i=1}^{k-2} e_i\right) - \delta = \frac{1}{n_k}\sum_{i=1}^{k-2}\big(e_i - \delta(n_i - n_{i-1})\big) \to 0 \text{ a.s.}$$

And so

$$\frac{1}{n}\overline{\mathrm{Err}}_n(\Gamma^{\mathcal{L}}_{1-\delta}) \le \frac{1}{n}\left(\overline{\mathrm{Err}}_{n_{s(n)}}(\Gamma^{\mathcal{L}}_{1-\delta}) + (n_{s(n)+1} - n_{s(n)})\right) \to \delta \text{ a.s.},$$

which implies the first statement of the theorem: $\frac{1}{n}\,\mathrm{Err}_n(\Gamma^{\mathcal{L}}_{1-\delta}) \to \delta$ a.s. ∎

4 Remarks on the conditions of the theorem

In this short section we will see that $\Gamma^{\mathcal{L}}_{1-\delta}$ is well calibrated when

$$\frac{n_{k+1}}{n_k} = 1 + O\left(\frac{1}{\sqrt{k}\ln k}\right). \tag{3}$$

For example, it is well calibrated when n_k grows as $\exp(\sqrt{k}/\ln k)$; on the other hand, our result cannot guarantee that it is well calibrated if n_k grows as $\exp(\sqrt{k})$. Indeed, condition (1) can be rewritten as

$$\sum_{i=2}^{\infty}\left(\ln\frac{n_{i+1}}{n_i}\right)^2 < \infty;$$

therefore, it is satisfied when $\ln(n_{k+1}/n_k) = O(1/(\sqrt{k}\ln k))$; this is equivalent to (3).

References

[1] Luc Devroye, László Györfi, Gábor Lugosi, *A probabilistic theory of pattern recognition*. New York: Springer, 1996.

[2] Ilia Nouretdinov and Vladimir Vovk. Criterion of calibration for Transductive Confidence Machine with limited feedback. On-line Compression Modelling Project, Working Paper 6, `http://vovk.net/kp`, April 2003.

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

[4] Vladimir Vovk, Alex Gammerman and Craig Saunders. Machine-learning applications of algorithmic randomness. Proceedings of the Sixteenth International Conference on Machine Learning, pp. 444–453. San Francisco, CA: Morgan Kaufmann, 1999.

[5] Vladimir Vovk. On-line Confidence Machines are well-calibrated. On-line Compression Modelling Project, Working Paper 1, `http://vovk.net/kp`, April 2002. Conference version published in: Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science, pp. 187–196, IEEE Computer Society, 2002.

[6] Vladimir Vovk. Asymptotic optimality of Transductive Confidence Machine. On-line Compression Modelling Project, Working Paper 2, `http://vovk.net/kp`, May 2002. Conference version published in: Proceedings of the Thirteenth International Conference on Algorithmic Learning Theory, 2002.

On-line Compression Modelling Project
Working Papers

1. *On-line confidence machines are well-calibrated*, by Vladimir Vovk, April 2002.

2. *Asymptotic optimality of Transductive Confidence Machine*, by Vladimir Vovk, May 2002.

3. *Universal well-calibrated algorithm for on-line classification*, by Vladimir Vovk, November 2002.

4. *Mondrian Confidence Machine*, by Vladimir Vovk, David Lindsay, Ilia Nouretdinov and Alex Gammerman, March 2003.

5. *Testing exchangeability on-line*, by Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman, February 2003.

6. *Criterion of calibration for Transductive Confidence Machine with limited feedback*, by Ilia Nouretdinov and Vladimir Vovk, April 2003.

7. *Online region prediction with real teachers*, by Daniil Ryabko, Vladimir Vovk and Alex Gammerman, March 2003.

8. *Well-calibrated predictions from on-line compression models*, by Vladimir Vovk, April 2003.