Criterion of calibration for Transductive Confidence Machine with limited feedback

Ilia Nouretdinov and Vladimir Vovk

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 #6

April 14, 2003

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

Abstract

This paper is concerned with the problem of on-line prediction in the situation where some data is unlabelled and can never be used for prediction, and even when data is labelled, the labels may arrive with a delay. We construct a modification of randomised Transductive Confidence Machine for this case and prove a necessary and sufficient condition for its predictions being calibrated, in the sense that in the long run they are wrong with a prespecified probability under the assumption that data is generated independently by same distribution. The condition for calibration turns out to be very weak: feedback should be given on more than a logarithmic fraction of steps.

Contents

1 Introduction

In this paper we consider the problem of prediction: given some training data and a new object x_n we would like to predict its label y_n. We use the randomised on-line version of Transductive Confidence Machine as basic method of prediction; first we explain why we are interested in this method and then formulate the main question of this paper.

Transductive Confidence Machine (TCM) [3, 4] is a prediction method giving "p-values" p_y for any possible value y of the unknown label y_n; the p-values satisfy the following property (proven in, e.g., [1]): if the data satisfies the i.i.d. assumption, which means that the data is generated independently by same mechanism, the probability that $p_{y_n} < \delta$ does not exceed δ for any threshold $\delta \in (0,1)$ (the *validity* property).

There are different ways of presenting the p-values. The one used in [3] only works in the case of pattern recognition: the prediction algorithm outputs a "most likely" label (y with the largest p_y) together with *confidence* (one minus the second largest p_y) and *credibility* (the largest p_y). Alternatively, the prediction algorithm can be given a threshold δ as an input and its answer will be that the label y_n should lie in the set of such y that $p_y > \delta$; this scenario of *set* (or *region*) *prediction* was used in [5, 2] and will be used in this paper. The validity property says that the set prediction will be wrong with probability at most δ. Therefore, we can guarantee some maximal probability of error; the downside is that the set prediction can consist of more than one element.

Randomised TCM (rTCM), which is described below, is valid in a stronger sense than pure TCM: the error probability is *equal* to δ.

In *on-line TCM* [5] it is supposed that machine learning is performed step-by-step: on the nth step TCM predicts the new label y_n using knowledge of the new object x_n and all the previous objects with their labels; after that the true information about y_n becomes available and TCM can use it on the next step $n+1$. In the paper [5] it was proven that the probability of error on each step is again δ; moreover, errors on different steps are independent of each other, so the mean percentage of errors asymptotically tends to δ (the *calibration* property).

In principle, it is easy to be calibrated in set prediction; what makes TCMs interesting is that they output few *uncertain* predictions (predictions containing more than one label). This can be demonstrated both empirically on standard benchmark data sets (see, e.g., [5]) and theoretically: a simple Nearest Neighbours rTCM produces asymptotically no more uncertain predictions than any other calibrated algorithm for set prediction.

The interest of this paper is a more general case of on-line TCM prediction, where only some subsequence of labels is available, possibly with a delay; a necessary and sufficient condition for calibration in probability is given in Theorem 1 below. Originally, we stated this result assuming that true labels were given without delay, but then we noticed that Daniil Ryabko's [2] device of "ghost rTCM" (in our terminology) makes it possible to add delays without any extra work.

2 On-line randomised TCM

Now we describe (mainly following [5]) how on-line rTCM works.

Suppose we observe a sequence $z_1, z_2, \dots, z_n, \dots$ of *examples*, where $z_i = (x_i, y_i) \in \mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, $x_i \in \mathbf{X}$ are *objects* to be labelled and $y_i \in \mathbf{Y}$ are the *labels*; $\mathbf{X}$ and $\mathbf{Y}$ are arbitrary measurable spaces.

"On-line" means that for any n we try to predict y_n using

$$z_1 = (x_1, y_1), \dots, z_{n-1} = (x_{n-1}, y_{n-1}), x_n.$$

The method is as follows. We need a symmetric function

$$f(z_1, \dots, z_n) = (\alpha_1, \dots, \alpha_n).$$

"Symmetric" means that if we change order of $z_1, \dots, z_n$, the order of $\alpha_1, \dots, \alpha_n$ will change in the same way. In other words, there must exist a function F such that

$$\alpha_i = F(\wr z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n \int, z_i),$$

where $\wr \cdots \int$ means a multiset. The output of on-line rTCM is a set Y_n of predictions for y_n; a label y is included in Y_n if and only if

$$\#\{i : \alpha_i > \alpha_n\} + \theta_n \#\{i : \alpha_i = \alpha_n\} > n\delta,$$

where

$$(\alpha_1, \dots, \alpha_n) = f(z_1, \dots, z_{n-1}, (x_n, y)),$$

$\theta_n \in [0, 1]$ are random numbers distributed uniformly and independently of each other and everything else, and $\delta > 0$ is a given threshold (called *significance level*). We will be concerned with the error sequence $e_1, \dots, e_n, \dots$, where $e_n = 0$ if the true value y_n is in Y_n, and $e_n = 1$ otherwise.

In the paper [5] it is proven that for *any* probability distribution P in the set $\mathbf{Z}$ of pairs $z_i = (x_i, y_i)$, the corresponding $(e_1, e_2, \dots)$ is a Bernoulli sequence: for each i, $e_i \in \{0, 1\}$, $e_i = 1$ with probability δ, and all e_i are independent.

3 Restricted TCM

In practice we are likely to have the true labels y_n only for a subset of steps n; moreover, even for this subset y_n may be given with a delay. In this paper we consider the following scheme. We are given a function $\mathcal{L} : N \to \mathbb{N}$ defined on an infinite set $N \subseteq \mathbb{N}$ and required to satisfy

$$\mathcal{L}(n) \leq n$$

for all $n \in N$ and

$$m \neq n \Longrightarrow \mathcal{L}(m) \neq \mathcal{L}(n)$$

for all $m \in N$ and $n \in N$; a function satisfying these properties will be called the *teaching schedule*. The teaching schedule $\mathcal{L}$ describes the way the data is disclosed to us: at the end of step n we are given the label $y_{\mathcal{L}(n)}$ for the object $x_{\mathcal{L}(n)}$. The elements of $\mathcal{L}$'s domain N in the increasing order will be denoted n_i: $N = \{n_1, n_2, \dots\}$ and $n_1 < n_2 < \cdots$.

We transform the on-line randomised TCM algorithm to what we call the *$\mathcal{L}$-restricted rTCM*. We again use a symmetric function $f(\zeta_1, \dots, \zeta_k) = (\alpha_1, \dots, \alpha_k)$ and for any $n = n_{k-1} + 1, \dots, n_k$ and any $y \in \mathbf{Y}$ we include y in Y_n if and only if

$$\#\{i = 1, \dots, k : \alpha_i > \alpha_k\} + \theta_n \#\{i = 1, \dots, k : \alpha_i = \alpha_k\} > k\delta,$$

where

$$(\alpha_1, \dots, \alpha_k) = f(z_{\mathcal{L}(n_1)}, \dots, z_{\mathcal{L}(n_{k-1})}, (x_n, y)),$$

θ_n are random numbers and δ is a given significance level. As before, the error sequence is: $e_n = 1$ if $y_n \notin Y_n$ and $e_n = 0$ otherwise.

Let U be the uniform distribution in $[0, 1]$. If a probability distribution P in $\mathbf{Z}$ generates the examples z_i, the distribution $(P \times U)^\infty$ generates z_i and the random numbers θ_i and therefore determines the distribution of all random variables, such as the errors e_i, considered in this paper.

We say that a restricted rTCM is *(well-)calibrated in probability* if the corresponding error sequence $e_1, e_2, \ldots$ has the property that

$$\frac{e_1 + \cdots + e_n}{n} \to \delta$$

in $(P \times U)^\infty$-probability for any significance level δ and distribution P in $\mathbf{Z}$. (Remember that, by definition, $\xi_1, \xi_2, \ldots$ converges to a constant c in Q-probability if

$$\lim_{n\to\infty} Q\left\{|\xi_n - c| > \varepsilon\right\} \to 0$$

for any ε.)

Our aim is to prove the following statement.

Theorem 1 *Let $\mathcal{L}$ be a teaching schedule with domain $N = \{n_1, n_2, \ldots\}$, where $n_1, n_2, \ldots$ is an increasing infinite sequence of positive integers.*

- *If $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$, any $\mathcal{L}$-restricted rTCM is calibrated in probability.*
- *If $\lim_{k\to\infty}(n_k/n_{k-1}) = 1$ does not hold, there exists an $\mathcal{L}$-restricted rTCM which is not calibrated in probability.*

In words, the theorem asserts that the restricted rTCM is guaranteed to be calibrated in probability if and only if the growth rate of n_k is sub-exponential.

4 Proof that $n_k/n_{k-1} \to 1$ is sufficient

We start from a simple general lemma about martingale differences.

Lemma 1 *If $\xi_1, \xi_2, \ldots$ is a martingale difference w.r. to σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ such that, for all $i \geq 1$,*

$$\mathbb{E}(\xi_i^2 \mid \mathcal{F}_{i-1}) \leq 1$$

and $w_1, w_2, \ldots$ is a sequence of positive numbers, then

$$\mathbb{E}\left(\left(\frac{w_1\xi_1 + \cdots + w_n\xi_n}{w_1 + \cdots + w_n}\right)^2\right) \leq \frac{w_1^2 + \cdots + w_n^2}{(w_1 + \cdots + w_n)^2}.$$

Proof Since elements of a martingale difference sequence are uncorrelated, we have

$$\begin{aligned}\mathbb{E}\left((w_1\xi_1+\cdots+w_n\xi_n)^2\right) &= \sum_{1\le i\le n} w_i^2\,\mathbb{E}(\xi_i^2) + 2\sum_{1\le i<j\le n} w_iw_j\,\mathbb{E}(\xi_i\xi_j)\\ &\le \sum_{1\le i\le n} w_i^2. \qquad\blacksquare\end{aligned}$$

Fix a probability distribution P in $\mathbf{Z}$ generating the examples z_i; let $\mathbb{P}$ stand for $(P\times U)^\infty$ (the probability distribution generating the examples z_i and the random numbers θ_i) and $\mathbb{E}$ stand for the expected value w.r. to $\mathbb{P}$.

Along with the original $\mathcal{L}$-restricted rTCM making errors $e_1, e_2, \ldots$ we also consider the *ghost rTCM* (introduced in [2]) which uses the same alpha function as the $\mathcal{L}$-restricted rTCM but is fed with the examples

$$z'_1 := z_{\mathcal{L}(n_1)}, z'_2 := z_{\mathcal{L}(n_2)}, \ldots$$

and random numbers $\theta'_1, \theta'_2, \ldots$ (independent from each other and anything else); the error sequence of the ghost rTCM is denoted $e'_1, e'_2, \ldots$ (remember that an error is encoded as 1 and the absence of error as 0). The ghost rTCM is given all labels and each label is given without delay. Notice that the input sequence $z_{\mathcal{L}(n_1)}, z_{\mathcal{L}(n_2)}, \ldots$ to the ghost rTCM is also distributed according to P^∞.

Set, for each $n = 1, 2, \ldots$,

$$d_n = \mathbb{P}\{e_n = 1 \mid z_1, \ldots, z_{n-1}\}$$

(it is clear that, for each k, d_n will be the same for all $n = n_{k-1}+1, \ldots, n_k$) and

$$d'_k = \mathbb{P}\left\{e'_k = 1 \mid z'_1, \ldots, z'_{k-1}\right\}.$$

Notice that, for all $k = 1, 2, \ldots$,

$$d_{n_k} = d'_k. \tag{1}$$

Corollary 1 *For each k,*

$$\mathbb{E}\left(\left(\frac{(e'_1-\delta)n_1 + (e'_2-\delta)(n_2-n_1) + \cdots + (e'_k-\delta)(n_k-n_{k-1})}{n_k}\right)^2\right) \le \frac{n_1^2 + (n_2-n_1)^2 + \cdots + (n_k-n_{k-1})^2}{n_k^2}.$$

Proof It is sufficient to apply Lemma 1 to $w_1 = n_1, w_2 = n_2 - n_1, \ldots, w_k = n_k - n_{k-1}$, the independent zero-mean (by the result of [5] described at the end of §2) random variables $\xi_k = e'_k - \delta$, and the trivial σ-algebras. ∎

Corollary 2 *For each k,*

$$\mathbb{E}\left(\left(\frac{(e'_1 - d'_1)n_1 + (e'_2 - d'_2)(n_2 - n_1) + \cdots + (e'_k - d'_k)(n_k - n_{k-1})}{n_k}\right)^2\right)$$
$$\leq \frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2}.$$

Proof Use Lemma 1 for $w_1 = n_1, w_2 = n_2 - n_1, \ldots, w_k = n_k - n_{k-1}$, $\xi_k = e'_k - d'_k$, and the σ-algebras $\mathcal{F}_k$ generated by $z'_1, \ldots, z'_{k-1}$. ∎

Corollary 3 *For each k,*

$$\mathbb{E}\left(\frac{(e_1 - d_1) + (e_2 - d_2) + \cdots + (e_{n_k} - d_{n_k})}{n_k}\right)^2 \leq \frac{1}{n_k}.$$

Proof Apply Lemma 1 to $w_i = 1$, $\xi_i = e_i - d_i$, and the σ-algebras $\mathcal{F}_i$ generated by $z_1, \ldots, z_i$. ∎

Lemma 2 *If $\lim_{k\to\infty} \frac{n_{k+1}}{n_k} = 1$ for some increasing sequence of positive integers $n_1, n_2, \ldots, n_k, \ldots,$ then*

$$\lim_{k\to\infty} \frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} = 0.$$

Proof For any $\varepsilon > 0$, there exists K such that $\frac{n_k - n_{k-1}}{n_{k-1}} < \varepsilon$ for any $k \geq K$. Therefore,

$$\frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2}$$
$$\leq \frac{n_K^2}{n_k^2} + \frac{(n_{K+1} - n_K)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2}$$
$$\leq \frac{n_K^2}{n_k^2} + \frac{n_{K+1} - n_K}{n_K}\frac{n_{K+1} - n_K}{n_k} + \frac{n_{K+2} - n_{K+1}}{n_{K+1}}\frac{n_{K+2} - n_{K+1}}{n_k} + \cdots$$
$$+ \frac{n_k - n_{k-1}}{n_{k-1}}\frac{n_k - n_{k-1}}{n_k} \leq \frac{n_K^2}{n_k^2} + \varepsilon\frac{(n_{K+1} - n_K) + \cdots + (n_k - n_{k-1})}{n_k} \leq 2\varepsilon$$

from some k on. ∎

Now it is easy to finish the proof of the first part of the theorem. In combination with Chebyshev's inequality and Lemma 2, Corollary 1 implies that

$$\frac{(e'_1 - \delta)n_1 + (e'_2 - \delta)(n_2 - n_1) + \cdots + (e'_k - \delta)(n_k - n_{k-1})}{n_k} \to 0$$

in probability; using the notation $k(n) := \min\{k : n_k \geq n\}$, we can rewrite this as

$$\frac{1}{n_k} \sum_{n=1}^{n_k} \left(e'_{k(n)} - \delta\right) \to 0. \tag{2}$$

Similarly, (1) and Corollary 2 imply

$$\frac{1}{n_k} \sum_{n=1}^{n_k} \left(e'_{k(n)} - d'_{k(n)}\right) = \frac{1}{n_k} \sum_{n=1}^{n_k} \left(e'_{k(n)} - d_n\right) \to 0 \tag{3}$$

and Corollary 3 implies

$$\frac{1}{n_k} \sum_{n=1}^{n_k} (e_n - d_n) \to 0 \tag{4}$$

(all convergences are in probability). Combining (2)–(4), we obtain

$$\frac{1}{n_k} \sum_{n=1}^{n_k} (e_n - \delta) \to 0; \tag{5}$$

the condition $n_{k+1}/n_k \to 1$ allows us to replace n_k with n in (5).

5 Proof that $n_k/n_{k-1} \to 1$ is necessary

As a first step, we construct the example space $\mathbf{Z}$, the probability distribution P in $\mathbf{Z}$ and an rTCM for which d'_k deviate consistently from δ. Let $\mathbf{X} = \{0\}$, $\mathbf{Y} = \{0, 1\}$, so z_i is, essentially, always 0 or 1. The probability P is defined by $P\{0\} = P\{1\} = \frac{1}{2}$. Define the alpha function $(\alpha_1, \ldots, \alpha_k) = f(\zeta_1, \ldots, \zeta_k)$ as follows:

$$(\alpha_1, \ldots, \alpha_k) = (\zeta_1, \ldots, \zeta_k)$$

if $\zeta_1 + \cdots + \zeta_k$ is even and

$$(\alpha_1, \ldots, \alpha_k) = (1 - \zeta_1, \ldots, 1 - \zeta_k)$$

if $\zeta_1 + \cdots + \zeta_k$ is odd.

It follows from the central limit theorem that

$$\frac{\#\{i = 1, \ldots, k : z'_i = 1\}}{k} \in (0.4, 0.6) \tag{6}$$

with probability more than 99% for k large enough. Let $\delta = 5\%$. Consider some $k \in \{1, 2, \ldots\}$; we will show that d'_k deviates significantly from δ with probability more than 99% for sufficiently large k; namely, that d'_k is significantly greater than δ if $z'_1 + \cdots + z'_{k-1}$ is odd (intuitively, in this case both potential labels are strange) and d'_k is significantly less than δ if $z'_1 + \cdots + z'_{k-1}$ is even (intuitively, both potential labels are typical). Formally:

- If $z'_1 + \cdots + z'_{k-1}$ is odd, then

  $$\begin{aligned} z'_k = 1 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is even} \Longrightarrow \alpha_k = z'_k = 1 \\ z'_k = 0 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is odd} \Longrightarrow \alpha_k = 1 - z'_k = 1; \end{aligned}$$

  in both cases we have $\alpha_k = 1$ and, therefore, with probability more than 99%,

  $$\begin{aligned} d'_k &= \mathbb{P}\left\{\theta'_k \#\{i = 1, \ldots, k : \alpha_i = 1\} \le k\delta\right\} \\ &= \frac{k\delta}{\#\{i = 1, \ldots, k : \alpha_i = 1\}} \ge \frac{k\delta}{0.7k} = \frac{10}{7}\delta. \end{aligned}$$

- If $z'_1 + \cdots + z'_{k-1}$ is even, then

  $$\begin{aligned} z'_k = 1 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is odd} \Longrightarrow \alpha_k = 1 - z'_k = 0 \\ z'_k = 0 &\Longrightarrow z'_1 + \cdots + z'_{k-1} + z'_k \text{ is even} \Longrightarrow \alpha_k = z'_k = 0; \end{aligned}$$

  in both cases $\alpha_k = 0$ and, therefore, with probability more than 99%,

  $$\begin{aligned} d'_k = \mathbb{P}\left\{\#\{i = 1, \ldots, k : \alpha_i = 1\} + \theta'_k \#\{i = 1, \ldots, k : \alpha_i = 0\} \le k\delta\right\} \\ \le \mathbb{P}\left\{0.3k \le k\delta\right\} = 0. \end{aligned}$$

To summarise, for large enough k,

$$|d'_k - \delta| = |d_{n_k} - \delta| > \delta/3 \tag{7}$$

with probability more than 99%.

Suppose that

$$\frac{1}{n}\sum_{i=1}^{n} e_i - \delta \to 0 \tag{8}$$

in probability; we will deduce that $n_k/n_{k-1} \to 1$. By (4) (remember that Corollary 3 and, therefore, (4) do not depend on the condition $n_k/n_{k-1} \to 1$) and (8) we have

$$\frac{1}{n}\sum_{i=1}^{n} d_i - \delta \to 0;$$

we can rewrite this in the form

$$\sum_{i=1}^{n} d_i = n(\delta + o(1))$$

(all $o(1)$ are in probability). This equality implies

$$\sum_{k=0}^{K} d_{n_k}(n_{k+1} - n_k) = n_{K+1}(\delta + o(1))$$

and

$$\sum_{k=0}^{K-1} d_{n_k}(n_{k+1} - n_k) = n_K(\delta + o(1));$$

subtracting the last equality from the penultimate one we obtain

$$d_{n_K}(n_{K+1} - n_K) = (n_{K+1} - n_K)\delta + o(n_{K+1}),$$

i.e.,

$$(d_{n_K} - \delta)(n_{K+1} - n_K) = o(n_{K+1}).$$

In combination with (7) and (1), this implies $n_{K+1} - n_K = o(n_{K+1})$, i.e., $n_{K+1}/n_K \to 1$ as $K \to \infty$.

References

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

[2] Daniil Ryabko, Vladimir Vovk, and Alex Gammerman. Online region prediction with real teachers. On-line Compression Modelling Project, Working Paper 7, `http://vovk.net/kp`, March 2003.

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

[4] Vladimir Vovk, Alex Gammerman, Craig Saunders. Machine-learning applications of algorithmic randomness. *Proceedings of the 16th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, pp. 444–453, 1999.

[5] Vladimir Vovk. On-line Confidence Machines are well-calibrated. *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, IEEE Computer Society, 2002. On-line Compression Modelling Project, Working Paper 1, `http://vovk.net/kp`, April 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.