Transductive Confidence Machine is universal

Ilia Nouretdinov, Vladimir Vovk,
Vladimir V'yugin, Alex Gammerman

Computer Learning Research Centre
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England

{ilia,vovk,vyugin,alex}@cs.rhul.ac.uk

Technical Report

CLRC-TR-02-05

December 2002

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

Abstract

Transductive Confidence Machine (TCM) is a practical prediction algorithm giving, in additions to its predictions, confidence information valid under the general iid assumption. The main result of this paper is that the prediction method used by TCM is universal under a natural definition of what "valid" means: any prediction algorithm providing valid confidence information can be replaced, without losing much of its predictive performance, by a TCM.

1 Introduction

In the last several decades new powerful machine-learning algorithms have appeared. A serious shortcoming of most of these algorithms, however, is that they do not directly provide any measures of confidence in the predictions they output. Two of the most important traditional ways to obtain such confidence information are provided by PAC theory (a typical result that can be used is Littlestone and Warmuth's theorem; see, e.g., [3]) and Bayesian theory. The former is discussed in detail in [8] and the latter is discussed in [7], but disadvantages of the traditional approaches can be summarized as follows: PAC bounds are valid under the general iid assumption but are too weak for typical problems encountered in practice to give meaningful results; Bayesian bounds give practically meaningful results, but are only valid under strong extra assumptions.

Papers [5, 15, 13, 9, 10] introduced a practical (as confirmed by numerous empirical studies reported in those papers) method of computing confidence information valid under the general iid assumption. There are several versions of this method: Transductive Confidence Machine (TCM), Inductive Confidence Machine, hashing, competitive transduction, etc. The basic method is TCM, and the rest are different ways of making TCM more computationally efficient. TCM is based on a specific formula,

$$p = \frac{\#\{i : \alpha_i \geq \alpha_{l+1}\}}{l+1}$$

(cf. (2) in Section 2). A natural question is whether there are better ways to produce valid confidence information. In this paper (Sections 2 and 3) we show that the first-order answer is "no": no way of producing valid confidence information is drastically better than TCM. In the last section we discuss questions raised by our results.

2 Specific randomness

Suppose we have two sets: the training set $(x_1, y_1), \ldots, (x_l, y_l)$ and the test set (x_{l+1}, y_{l+1}) containing only one example. The unlabelled examples x_i are drawn from a set $\mathbf{X}$ and the labels y_i are drawn from a finite set $\mathbf{Y}$; we assume that $|\mathbf{Y}|$ is small (i.e., we consider the problem of classification with a small number of classes). The examples (x_i, y_i) are assumed to be generated by some probability distribution P (same for all examples) independently of each other; we call this the *iid assumption*. Set $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$.

In this section we discuss four natural ways of predicting with confidence, which we call Randomness Predictors, Exchangeability Predictors, Invariant Exchangeability Predictors, and Transductive Confidence Machines. We start with the latter (following the papers mentioned above).

An *individual strangeness measure* is a family of functions A_n, $n = 1, 2, \ldots$, such that each A_n maps every pair (B, z), where B is a bag of $n-1$ elements of $\mathbf{Z}$ and z is an element of $\mathbf{Z}$, to a real (typically non-negative) number $A_n(B, z)$. (Intuitively, $A_n(B, z)$ measures how different z is from the elements of B.) The Transductive Confidence Machine associated with A_n works as follows: when given the data

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

(the training set and the known component x_{l+1} of the test example), every potential classification y of x_{l+1} is assigned the *p-value*

$$p(y) := \frac{\#\{i : \alpha_i \geq \alpha_{l+1}\}}{l+1}, \tag{2}$$

where

$$\alpha_i := A_{l+1}(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_{l+1}\}, z_i),$$

$z_j = (x_j, y_j)$ (except $z_{l+1} = (x_{l+1}, y)$), and $\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_{l+1}\}$ stands for a bag rather than a set. TCM's output $p : \mathbf{Y} \to [0, 1]$ can be further packaged in two different ways:

- ⋄ we can output $\arg\max_y p(y)$ as the *prediction* and say that $1 - p_2$, where p_2 is the second largest p-value, is the *confidence* and that the largest p-value p_1 is the *credibility*;
- ⋄ or we can fix some conventional threshold δ (such as 1% or 5%) and output as our prediction (*predictive region*) the set of all y such that $p(y) > \delta$.

For us, the essence of TCM is formula (2). For numerous examples of TCM, see the papers referred to above.

Next we define Randomness Predictors (RP). We say that a family of functions $t_n : \mathbf{Z}^n \to [0, 1]$, $n = 1, 2, \dots$, is a *randomness test* if, for every n, any probability distribution P in $\mathbf{Z}$, and any $\gamma \in [0, 1]$,

$$P^n\{(z_1 \dots z_n) \in \mathbf{Z}^n \,|\, t_n(z_1 \dots z_n) \leq \gamma\} \leq \gamma. \tag{3}$$

Instead of (2), we now define

$$p(y) := t_{l+1}((x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, y)). \tag{4}$$

Using this function, we can define the corresponding predictive region, or the prediction, confidence, and credibility, as above. Since we are making the iid assumption, this is the most general way of producing valid confidence information. This definition is, however, very non-constructive, and it is not clear how to use it directly.

A slightly more manageable definition is as follows. A family of functions $t_n : \mathbf{Z}^n \to [0, 1]$, $n = 1, 2, \dots$, is an *exchangeability test* if, for every n, any exchangeable probability distribution P in $\mathbf{Z}^n$, and any $\gamma \in [0, 1]$,

$$P\{(z_1 \dots z_n) \in \mathbf{Z}^n \,|\, t_n(z_1 \dots z_n) \leq \gamma\} \leq \gamma. \tag{5}$$

If we now define $p(y)$ by the same formula (4), we obtain the notion of an Exchangeability Predictor (EP). If we further require t_n to be *invariant*, in the sense that $t_n(z_1,\dots,z_n)$ does not change if any z_i and z_j, $i,j=1,\dots,n-1$, are swapped, then we arrive at the notion of an Invariant Exchangeability Predictor (IEP).

Our first proposition asserts that TCM and IEP are essentially the same notion. Formally, we identify TCM, RP, EP, and IEP with the functions mapping $(x_1,y_1),\dots,(x_l,y_l),x_{l+1}$ to the function $p=p_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}:\mathbf{Y}\to[0,1]$, according to (2) or (4), respectively. We say that a predictor (TCM, RP, EP, or IEP) $p^A_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)$ is (at least) *as good as* a predictor $p^B_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)$ if, for any training set $(x_1,y_1),\dots,(x_l,y_l)$, any unlabelled test example x_{l+1}, and any label y,

$$p^A_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)\le p^B_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y).$$

We say that a class $\mathcal{A}$ (such as TCM or RP) of predictors is *as good as* a class $\mathcal{B}$ of predictors if for any $B\in\mathcal{B}$ there exists $A\in\mathcal{A}$ such that A is as good as B (i.e., if every predictor in $\mathcal{B}$ can be replaced by an equally good predictor in $\mathcal{A}$).

Proposition 1 *Transductive Confidence Machines are as good as Invariant Exchangeability Predictors, and vice versa.*

Proof For simplicity we will assume that $\mathbf{X}$ is finite. First we show that Transductive Confidence Machines are Invariant Exchangeability Predictors; we only need to check the "validity property"

$$P\{(x_1,y_1),\dots,(x_l,y_l),(x_{l+1},y)\,|\,p_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)\le\gamma\}\le\gamma$$

for the values $p(y)=p_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)$ computed according to (2), where P is an exchangeable distribution which generates $(x_1,y_1),\dots,(x_l,y_l),(x_{l+1},y)$. (Invariance is obvious.) Inequality $p_{(x_1,y_1),\dots,(x_l,y_l),x_{l+1}}(y)\le\gamma$ means that α_{l+1} is among the top $100\gamma\%$ in the bag $\{\alpha_1,\dots,\alpha_{l+1}\}$; the validity follows from the fact that all permutations of α_i are equiprobable.

To show that Invariant Exchangeability Predictors can be replaced by Transductive Confidence Machines, we have to explicitly construct α's. Suppose we are given an IEP generated by an invariant exchangeability test t. If B is a bag in $\mathbf{Z}$ of size l and $z\in\mathbf{Z}$, define $A_n(B,z)$ to be $1/t_n(z_1,\dots,z_l,z)$, where $z_1\dots z_l$ is a list of all elements of B (in any order; because of invariance, the order does not matter). The corresponding TCM will be as good as the IEP. ∎

It is clear that EP are as good as IEP and that RP are as good as EP (see Figure 1). In the next section we will see that the opposite relations also hold, in some weaker sense.

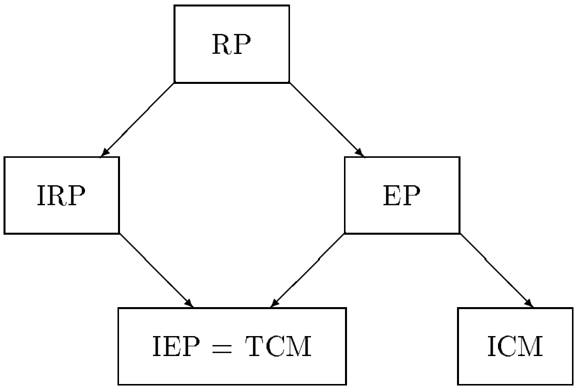

Figure 1 Relationship between different classes of predictors (a line connecting two classes means that the upper class is as good as the lower one). ICM (Inductive Confidence Machines, not discussed elsewhere in this paper) sacrifice invariance for the sake of computational efficiency. IRP (Invariant Randomness Predictors) are defined analogously to IEP.

3 From Randomness Predictors to Invariant Exchangeability Predictors

If the set $\mathbf{X}$ is "constructive" (we do not define this notion precisely here, but it is not restrictive), there exists an *optimal* randomness test δ, which means the following: δ is upper semicomputable and for any other upper semicomputable randomness test t there exists a constant $C > 0$ such that, for any n and $(z_1 \dots z_n) \in \mathbf{Z}^n$,

$$\delta(z_1 \dots z_n) \le Ct(z_1 \dots z_n).$$

(We omit the superfluous subscript n. Usually optimal tests are called "universal", but in this paper we use "universal" in a different sense.) We fix some optimal randomness test and denote it δ^{iid}; $\delta^{\mathrm{iid}}(z_1 \dots z_n)$ is called the *randomness level* of the sequence $z_1 \dots z_n$. Analogously, there exist an optimal exchangeability test δ^{exch} and an optimal invariant exchangeability test $\delta^{\mathrm{inv\ exch}}$. These optimal tests determine the Optimal Randomness Predictor, Optimal Exchangeability Predictor, and Optimal Invariant Exchangeability Predictor, respectively.

It is obvious that the Optimal Randomness Predictor is as good as the Optimal Exchangeability Predictor, and the latter is as good as the Optimal Invariant Exchangeability Predictor. The following proposition shows that the difference is not essential in the most interesting case where the training set and the unlabelled test example are random; this condition means that the randomness level

$$\delta^{\mathrm{iid}}((x_1, y_1), \dots, (x_l, y_l), x_{l+1}) := \sup_{y \in \mathbf{Y}} \delta^{\mathrm{iid}}((x_1, y_1), \dots, (x_l, y_l), (x_{l+1}, y))$$

is not small.

Proposition 2 *For any constant $\epsilon > 0$ there exists $C > 0$ such that, for all training sets $(x_1, y_1), \ldots, (x_l, y_l)$, unlabelled test examples x_{l+1}, and labels y,*

$$\frac{1}{C}\delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))$$
$$\leq \delta^{\mathrm{inv\ exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \leq$$
$$C\frac{\left(\delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))\right)^{1+\epsilon}|\mathbf{Y}|^{2+\epsilon}}{\left(\delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1})\right)^{2+\epsilon}}.$$

We will end this section with a proof sketch; now we briefly discuss Proposition 2. Intuitively, it asserts that if the data set $(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$ is random and the set $\mathbf{Y}$ is small,

$$\delta^{\mathrm{inv\ exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \approx \delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)). \tag{6}$$

This shows the universality of TCM: the optimal IEP (equivalently, TCM; see Proposition 1) is about as good as the optimal RP. The precise statement involves a multiplicative constant C; this is inevitable since randomness and exchangeability levels are only defined to within a constant factor.

Approximate equality (6) can be split into two:

$$\delta^{\mathrm{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \approx \delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \tag{7}$$

and

$$\delta^{\mathrm{inv\ exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \approx \delta^{\mathrm{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)). \tag{8}$$

The first of these, (7), is closely connected with de Finetti's theorem, which plays a prominent role in Bayesian statistics [1, 11]. In the case of infinite sequences, de Finetti's theorem implies that the randomness and exchangeability levels coincide. For finite sequences, the situation is much more complicated (see, e.g., [4] and [12]); in Proposition 2 we have to add the assumption that $(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}$ is random.

Actually, the following simple example shows that, if we do not ignore constant factors, Randomness and Exchangeability Predictors lead to different results for finite sequences. Suppose the training set is

$$(y_1,\ldots,y_l) = (0,\ldots,0)$$

(there are no xs, or they are trivial: $|\mathbf{X}| = 1$) and we are asked to predict y_{l+1}. The confidence in the natural prediction $y_{l+1} = 0$ will be $1 - \frac{1}{l}$ under the exchangeability model and will be close to $1 - \frac{1}{el}$ (for large l) under the iid model: indeed, the maximal probability of the event

$$(y_1,\ldots,y_l,y_{l+1}) = (0,\ldots,0,1)$$

under an iid distribution is

$$\frac{1}{l+1}\left(1-\frac{1}{l+1}\right)^l \sim \frac{1}{el}.$$

The second approximate equality, (8), corresponds to the invariance principle in statistics (see, e.g., [2], 2.3(vi)). Since we know that the elements of the training set are produced by an iid mechanism, the principle requires that we should only consider tests and predictors that are invariant w.r. to permutations of the training set.

Proof sketch of Proposition 2. The non-trivial assertion is

$$\delta^{\text{inv exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \le$$
$$C\frac{\left(\delta^{\text{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))\right)^{1+\epsilon}|\mathbf{Y}|^{2+\epsilon}}{\left(\delta^{\text{iid}}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1})\right)^{2+\epsilon}}.$$

This is equivalent to the analogous inequality for optimal "i-tests", where the validity conditions (3) and (5) are replaced by the stronger integral conditions $\int(1/t_n)dP^n \le 1$ and $\int(1/t_n)dP \le 1$, respectively. (The closeness of optimal i-tests to optimal tests is analogous to the closeness of prefix complexity K to Kolmogorov complexity C in the theory of Kolmogorov complexity [6].) In the rest of this sketch we will interpret the δs as optimal i-tests.

It can be shown (cf. [12]) that

$$\delta^{\text{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \stackrel{\times}{=} \delta^{\text{iid}}(\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\})$$
$$\times\delta^{\text{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\,|\,\delta^{\text{iid}}(\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\})), \tag{9}$$

where $\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\}$ is the bag of all elements of the data sequence $((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))$ and $\stackrel{\times}{=}$ stands for equality to within a constant factor. (Equation (9) is analogous to the well-known [6] equality $K(a,b) \stackrel{+}{=} K(a)+K(b\,|\,a,K(a))$ in the theory of Kolmogorov complexity.) Instead of the precise inequalities in this proof sketch we will content ourselves with approximate ones, such as

$$\delta^{\text{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \stackrel{*}{=} \delta^{\text{iid}}(\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\})$$
$$\times\delta^{\text{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \tag{10}$$

in place of (9), with the asterisk showing that the equality or inequality is approximate; it will be easy to check that the accuracy will be sufficient for our purpose. Combining (10) with

$$\delta^{\text{iid}}(\{(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)\}) \stackrel{*}{\ge} \delta^{\text{iid}}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1})/|\mathbf{Y}|$$

(this inequality follows from the following argument: if $((x_1,y_1),\ldots,(x_l,y_l),x_{l+1})$ is random when (x_i,y_i) are generated by some measure P in $\mathbf{Z}$, then

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

will be random when the measure P is corrupted by randomly changing all y_i and y with probability $\frac{1}{l+1}$), we obtain

$$\delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \overset{*}{\geq} \delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1}) \times \delta^{\mathrm{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))/|\mathbf{Y}|.$$

It remains to combine this with

$$\delta^{\mathrm{exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y)) \overset{*}{\geq} \delta^{\mathrm{iid}}((x_1,y_1),\ldots,(x_l,y_l),x_{l+1}) \times \delta^{\mathrm{inv\ exch}}((x_1,y_1),\ldots,(x_l,y_l),(x_{l+1},y))/|\mathbf{Y}|$$

(proof omitted). ∎

4 Future research

In conclusion, we list possible directions of future research:

- From the practical point of view, it is important to remove all multiplicative constants from Proposition 2, stating it in terms of specific rather than optimal tests. We only showed the optimality of TCM in an asymptotic sense; it is an open question whether TCM is universal in different non-asymptotic senses.
- The statistical invariance principle makes it interesting to study Invariant Randomness Predictors.

Acknowledgments

This work was partially supported by EPSRC, BBSRC and EU through grants: EPSRC grant GR/M16856 ("Comparison of Support Vector Machine and Minimum Message Length methods for induction and prediction"); 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"; EU grant IST-1999-10226 "EUREDIT: The Development and Evaluation of New Methods for Editing and Imputation". This paper is partly based on [14].

References

[1] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 2000.

[2] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Methods*. Cambridge, Cambridge University Press, 2000.

[4] P. Diaconis and D. A. Freedman. Finite exchangeable sequences. *Annals of Probability*, 8:745–764, 1980.

[5] Alex Gammerman, Vladimir Vapnik, and Vladimir Vovk. Learning by transduction. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 148–156, San Francisco, CA, 1998. Morgan Kaufmann.

[6] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 2nd edition, 1997.

[7] Tom Melluish, Craig Saunders, Ilia Nouretdinov, and Vladimir Vovk. Comparing the Bayes and typicalness frameworks. In L. De Raedt and P. Flash, editors, *Machine Learning: ECML 2001. Proceedings of the 12th European Conference on Machine Learning*, volume 2167 of *Lecture Notes in Artificial Intelligence*, pages 360–371, 2001. Full version published as Technical Report CLRC-TR-01-05; see `http://www.clrc.rhul.ac.uk`.

[8] 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`.

[9] 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.

[10] Craig Saunders, Alex Gammerman, and Vladimir Vovk. Computationally efficient transductive machines. In *Proceedings of ALT'00*, 2000.

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

[12] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986.

[13] Vladimir Vovk and Alex Gammerman. Statistical applications of algorithmic randomness. In *Bulletin of the International Statistical Institute. The 52nd Session, Contributed Papers*, volume LVIII, book 3, pages 469–470, 1999.

[14] Vladimir Vovk and Alex Gammerman. Algorithmic randomness for machine learning. Manuscript, 2001.

[15] 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.