Competitive On-Line Statistics

Vladimir Vovk
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
vovk@dcs.rhbnc.ac.uk

A radically new approach to statistical modellingΓ which combines mathematical techniques of Bayesian statistics with the philosophy of the theory of competitive on-line algorithmsΓ has arisen over the last decade in computer science (to a large degreeΓ under the influence of Dawid's prequential statistics). In this approachΓ which we call "competitive on-line statistics"Γ it is not assumed that data are generated by some stochastic mechanism; the bounds derived for the performance of competitive on-line statistical procedures are *guaranteed* to hold (and not just hold with high probability or on the average).

Problem

Making rational decisions is a central problem in science and everyday life. (Polynomials of which degree should I use to fit my data sets? Should I take my umbrella todayΓ tomorrowΓ etc.? Which stocks should I buy and sell this year?) Only rarely we can readily choose the best course of action; more often we will have a more or less extensive (maybe infinite) family of potentially successful decision strategies. (Whether a decision strategy is successful will depend not only on the merits of this strategy but also on the future events which we do not know yet.) HoweverΓ at the end of the day we must choose one specific decision strategyΓ so we naturally arrive at this problem: **given a family of decision strategies, find a new decision strategy which will perform, under any circumstances, almost as well as the best (under those circumstances) decision strategy in the family**.

At firstΓ this task might appear hopeless for even moderately interesting families of decision strategies. (Recall that we want the constructed decision strategy to perform almost as well as the best strategy in the family *always*; we do not make any stochastic assumptions about the generation of the future events.) HoweverΓ for one specific loss function a good merging algorithm has been known for a long time (the Bayesian mixture) and for many more loss functions good merging algorithms have been found in recent years.

The usual *statistical model* is a family of probability distributions reflecting our knowledge or our assumptions about some piece of the world. The Bayesian framework involves another elementΓ the prior distribution on the parameter setΓ which allows the Bayesian to replace the statistical model by a single probability distribution: the statistical model $\{Q_\theta \mid \theta \in \Theta\}$ with the prior distribution $P(d\theta)$ in Θ is replaced by the Bayesian mixture $Q = \int_\Theta Q_\theta P(d\theta)$. This formulaΓ reinterpreted and generalizedΓ lies at the heart of competitive on-line statistics.

In competitive on-line statistics the statistical model is replaced by the *decision pool*Γ which is a family of decision strategies; the statistician's goal is to replace this family with a single decision strategy. Recall that the competitive on-line statistician is agnostic: she does not assume anything about the stochastic mechanism generating the data; she does not even assume the *existence* of such a mechanism.

Aggregating Algorithm

Let Ω be some *sample space* Γ be a *decision space* and Θ be a *parameter space* (the decision strategies in our decision pool will be indexed by $\theta \in \Theta$). We consider the following perfect-information game between three players Statistician Decision Pool and Nature: at each trial $t = 1, 2, \ldots$

- Decision Pool makes a prediction $\xi_t : \Theta \to \Gamma$; $\xi_t(\theta)$ is interpreted as the decision recommended by the decision strategy $\theta \in \Theta$;
- Statistician makes her own decision $\gamma_t \in \Gamma$;
- Nature chooses some outcome $\omega_t \in \Omega$.

There is some fixed *loss function* $\lambda : \Omega \times \Gamma \to [0, \infty]$; Statistician's goal is to ensure that her cumulative loss $L_T = \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t)$ is almost as good as the loss $L_T(\theta) = \sum_{t=1}^{T} \lambda(\omega_t, \xi_t(\theta))$ of all or most of the decision strategies $\theta \in \Theta$. Assuming that the number $n = |\Theta|$ of decision strategies in the pool is finite it is possible to prove for a wide class of *games* $(\Omega, \Gamma, \lambda)$ that Statistician can ensure that for all$T$ and θ

$$L_T \le c L_T(\theta) + a \ln n, \tag{1}$$

where c and a are some constants. For a wide class of games the *Aggregating Algorithm* (described in eg Vovk 1998a) ensures that (1) holds with optimal c and a. Since it is possible to improve c at the expense of deteriorating a and vice versa the algorithm involves a *learning rate* $\eta \in (0, \infty)$ and theoptimal constants $c = c(\eta)$ and $a = a(\eta)$ in (1) depend on η.

The constants $c(\eta)$ and $a(\eta)$ have been found for many games (the constants mentioned below were found by DeSantis Haussler Kivinen Littlestone Markowsky Vovk Warmuth and Wegman); especially important are the *perfectly mixable* games for which $c(\eta) = 1$ for some η. The most important for statistics games are perhaps the *log-loss* games; assuming for simplicity that Ω is finite in the log-loss game with the sample space Ω the decision space Γ is the set of all probability distribution in Ω and the loss function is $\lambda(\omega, \gamma) = -\ln \gamma\{\omega\}$. For this game the Aggregating Algorithm with learning rate $\eta = 1$ coincides with the Bayesian mixture (assuming the uniform prior); the constants are $c(1) = a(1) = 1$. Also important especially in the problems of regression is the following *square-loss game*: $\Omega = \Gamma = [-1, 1]$ and $\lambda(\omega, \gamma) = (\omega - \gamma)^2$. (We assume that the outcomes never exceed some known bound C; without loss of generality we take $C = 1$.) In the case of the square-loss game $c(\eta) = 1$ and $a(\eta) = 2$ for some η. In general a game is perfectly mixable if its loss function is "strictly convex" in some sense.

Some important games are not perfectly mixable such as the *simple prediction* game $\Omega = \Gamma = \{0, 1\}$ $\lambda(\omega, \gamma) = |\omega - \gamma|$ where $c(\eta) = \eta / \ln \frac{2}{1+\exp(-\eta)}$ and $a(\eta) = 1 / \ln \frac{2}{1+\exp(-\eta)}$. When we take $\Gamma = [0, 1]$ instead (the *absolute loss* game) $c(\eta)$ and $a(\eta)$ are halved and the game becomes "almost perfectly mixable" in the sense that $c(\eta) \to 1$ as $\eta \to 0$ (this is also true if $\Omega = [0, 1]$).

Notice that the philosophy of competitive on-line statistics only "works" when Nature is oblivious to Statistician's decisions (ω_T does not depend on $\gamma_1, \ldots, \gamma_{T-1}$): if Nature is not oblivious it is not longer possible to interpret inequality (1) as saying that Statistician performs not much worse than the best of the decision strategies: if we had followed strategy θ we would perhaps have observed a different sequence of outcomes. The assumption that Nature is oblivious is always justified when γ_t are predictions (say the atmosphere does not care about our predicting rain) but it can also be justified for decisions different from predictions such as portfolio selection by a small investor: see the work on universal portfolio selection originated by Cover (relevant references can be found in Vovk and Watkins 1998).

Linear regression

Even if the decision pool is infinite in a surprisingly wide classof problems it is possible to derive good bounds for competitive on-line procedures; in this note however we will only consider the problem of linear regression with the square loss. There are several competitive on-line algorithms for this problem: see eg the beautiful results in (Kivinen and Warmuth 1997) about their Exponentiated Gradient algorithm. We will consider just one of those algorithms (Vovk 1998b).

We have to extend slightly the protocol of the previous section: we will assume that at the beginning of every trial t Nature outputs a "signal" x_t to be used by Decision Pool and Statistician in making their decisions. We assume that the signals are taken from the ball $\{x \in \mathbb{R}^n \mid \|x\| \leq X\}$ of radius X; the decision pool is indexed by the ball $\Theta = \{\theta \in \mathbb{R}^n \mid \|\theta\| \leq C\}$ of radius C; the decision strategy θ recommends prediction $\theta \cdot x_t$ at trial t. Applying the Aggregating Algorithm to this decision pool and a Gaussian prior the standard bounds for that algorithm imply

$$L_T \leq L_T(\theta) + C^2X^2 + nC^2X^2 \ln(T+1), \qquad (2)$$

for all T and θ. (For the proof of this inequality and its elaborations see (Vovk 1998b); the assumptions that the signal and parameter spaces should be bounded were made only for simplicity; the essential assumption is that the responses ω_t should be bounded by a known constant. A similar inequality was proved earlier by Foster.) To see that bound (2) is tight assume that $n = 1$ $\theta \in [-1, 1]$ $x_t = 1$ for all t andω_t are generated by the iid process with the probability $\frac{1+\theta}{2}$ of $\omega_t = 1$ and the probability $\frac{1-\theta}{2}$ of $\omega_t = -1$. It is easy to check (for details see Vovk 1998b) that when Statistician uses the Maximum Likelihood estimator for computing γ_t and $\theta = 0$ the**expected value** of the difference between the right-hand and left-hand sides of (2) does not exceed the minute quantity of $\frac{1}{T}$.

Discussion and further research

It is not completely clear yet how far competitive on-line statistics can be developed; some of its limitations are well understood and others still wait to be disclosed. One serious limitation is that for some interesting games (especially those with non-compact Ω and Γ) the constants $c(\eta)$ and $a(\eta)$ are infinite; eg they become infinite if we remove the assumption that the response variable is bounded in the square-loss game. Another limitation is that Nature is implicitly assumed to be oblivious. However the advantages of competitive on-line statistics turned out to be clear enough to generate a lot of interesting research in computational learning community: see eg the work on tracking the best expert (Auer Herbster Littlestone Vovk Warmuth) applications to financial theory (Blum Cover Helmbold Kalai Ordentlich Schapire Singer Warmuth) pruning decision trees (Helmbold Hirai Maruoka Pereira Schapire Singer Takimoto Vovk) boosting (Freund and Schapire) predictors that specialize (Freund Schapire Singer Warmuth) etc.

An interesting application of the Aggregating Algorithm is to generalize the notion of

Kolmogorov complexity (seeΓegΓLi and VitanyiΓ1997). The idea is to apply the Aggregating Algorithm to the "universal decision pool" containing every computable decision strategy (such a pool can be constructed from a universal Turing machine). The loss of the resulting "decision strategy" (actually it will be a decision strategy only in a generalized sense) on a data sequence x is called the *predictive complexity* of x. When applied to the log-loss gameΓthis leads to a variant of Kolmogorov complexity. As well as being a fundamental concept *per se*Γ the notion of predictive complexity allows us to define the notion of randomness for prediction games different from the log-loss game; for detailsΓ see (VovkΓ 1999a).(Though even the standard notion of log-loss randomness seems to be grossly under-used presently: see Vovk and GammermanΓ 1999b.) Another application is to generalizing the MDL principle to games different from the log-loss one: see (Vovk and GammermanΓ 1999b).

REFERENCES

KivinenΓ J and WarmuthΓ M K (1997).Exponential Gradient versus Gradient Descent for linear predictors. *Inform Computation* **132**Γ1–63.

LiΓ M and VitányiΓ P (1997).*An Introduction to Kolmogorov Complexity and Its Applications.* Second edition. New York: Springer.

VovkΓV (1998a). A game of prediction with expert advice. *J Comput Syst Sci* **56**Γ153–173.

VovkΓV (1998b). Competitive on-line linear regression. In *Advances in Neural Information Processing Systems* (eds M I JordanΓM J Kearns and S A Solla)Γ364–370. CambridgeΓMA: MIT Press. Full version: Technical Report CSD-TR-97-13ΓDepartment of Computer ScienceΓ Royal HollowayΓ University of London.

VovkΓ V (1999a). Probability theory for the Brier game. *Theoretical Computer Science*Γto appear.

VovkΓ V and GammermanΓ A (1999b). Statistical applications of algorithmic randomness. *The 52nd Session of the International Statistical Institute, Conference Volume of Contributed Papers.*

VovkΓ V and GammermanΓ A (1999)Complexity Estimation Principle. *The Computer Journal*Γ to appear.

VovkΓ V and WatkinsΓ C (1998). Universal portfolio selection. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*Γ12–23.

RÉSUMÉ

Cet article décrit une approche nouvelle à modelage statistique combinant les techniques mathematiques de statistique Bayesienne avec la philosophie de la theorie de algorithmes compétitives en ligne. Dans cette approche, qui émergeait durant le décennie dernière dans l'informatique, on ne suppose pas que les données sont produites par une mécanisme stochastique ; au lieu de cela, il est prouvé que les procédures statistiques compétitives en ligne atteignent toujours (et non, par exemple, avec haute probabilité) quelque but desirable (explicitant la bonne performance sur les données réeles).