The Quality of Prediction and Optimal Predictions Based on Two Experts

An. A. Muchnik, A. L. Semenov

Division of Cybernetics at Computing Center of Russian Academy of Sciences

We consider sequences of events, each of the events belongs to a given finite set (alphabet) A. We assume that several experts make their (may be probabilistic) predictions on the next upcoming event. Each expert knows all previous events.

So, formally, expert is any function that maps finite sequences of elements of A to distributions of probability over A: conditional probability of event a equals to $P(xa|x)$ provided previous events constitute the sequence x.

Let a set of experts is given. We consider the task of prediction of the upcoming event based on all previous events, all previous predictions of experts, and the predictions for the upcoming event by the experts. The rule of such prediction is called *predictor*. The predictor is *integral* if at any step it chooses one from the given experts and follows its (as we remember, probabilistic) prediction.

The general property of a good predictor is:

Let the events under consideration satisfy a distribution R. If one of the expert distributions P_1, P_2, P_3, $\ldots$ is close to the true distribution R, then the distribution Q given by the predictor is close to R as well.

Let me start with examples of well-known and well-studied predictors:

Bayes-mixture predictor was proposed by Ray Solomonoff in 1950-s. It is based on the concept of degree of trust to expert. Let us have experts: P_1, P_2, P_3, $\ldots$ We start with some a priori assignment of a weight (of trust) w_j to each expert P_j. The weights constitute a probability distribution over P_1, P_2, P_3, $\ldots$ (or over their numbers). After a sequence of events $z = y_1 \ldots y_n$ have happened we recalculate the weights using the classical formula $w_j P_j(z)/\sum_i w_i P_i(z) = W_j(z)$. The Bayes-mixture predictor Q is defined by formula $Q(za|z) = \sum_j W_j(z) P_j(za|z)$.

MDL (minimal description length, or maximum a posteriori, or Occam razor) predictor is an example of an integral predictor. It advises us to use P_j for which the weight $W_j(z)$, or the degree of trust, is maximal.

Our main result is a general method to construct predictors that are better then Bayes-mixture predictors, as well as MDL predictors. Our method gives integral predictor but it works in the case of two experts only.

Before giving formal results we shell consider one specific example that illustrates the difference between our and known approaches.

Example. We have two events only: a and b.
The predictions, we consider, are of two types:
Deterministic: a with probability 1.
Stochastic: a with probability 0.99, b with probability 0.01.
The actual sequence is: $a, a, a, \ldots$
We have two experts: I (First), and II (Second).
Experts behaviour:
For the first 10.000 steps:

- I deterministic,
- II stochastic.

Then, for the next 10.000 steps:

- I stochastic,
- II deterministic.

In this situation MDL predictor advises us to follow I expert all through the 20.000 steps (because II expert "lost our trust" fast in the beginning). The Bayes-mixture will advise us to predict the distribution very close to the given by I expert almost every time.

Our predictor in this situation advises us: to follow deterministic expert (I) for the first 10.000 steps, then, on step 10.001 follow stochastic expert (I), then follow deterministic expert (II). This behaviour is wiser. We emphasize that, of course, we propose not a solution to this particular problem, but a general method of constructing predictors.

Let us repeat that we want our predictions to be not much less precise than prediction of any of the experts.

For the "precision" or "closeness" measure of prediction we use the Hellinger distance $\rho(p,q) = \sum\limits_{a\in A} \left(\sqrt{p(a)} - \sqrt{q(a)}\right)^2$ as it was done in works of Vovk and Hutter–Poland.

Let events $\langle y_1, \ldots, y_N\rangle = Y$ happened and two predictions, $R(Y_{1:n}a|Y_{1:n})$ and $Q(Y_{1:n}a|Y_{1:n})$, on all initial segments of Y are given. We define distance between two sequences of predictions as $\sum\limits_{n=0}^{N} \rho(R(Y_{1:n}a|Y_{1:n}), Q(Y_{1:n}a|Y_{1:n}))$.

Let us assume now that in reality sequence Y is appearing in accordance with a distribution R. For s big enough our goal is to make as small as possible the probability in R of $\left\{Y \,\middle|\, \sum\limits_{n=0}^{N} \rho(R(Y_{1:n}a|Y_{1:n}), Q(Y_{1:n}a|Y_{1:n})) > s\right\}$. Let us denote this probability $\mathcal{F}_{R,Q}(s)$. Our term "precise" corresponds to the rate of decreasing of $\mathcal{F}_{R,Q}(s)$ and $\mathcal{F}_{R,P_j}(s)$, as functions of s. Here the distribution P_j describes predictions of the expert j, Q describes our predictions.

Quality of a set k of predictors is the supremum of all α such that $\exists\, \beta\ \forall s > 0$ there exists a predictor from k such as for any given experts P_j it produces predictions Q with the property $\forall j\ \mathcal{F}_{P_j,Q}(s) < e^{\beta - \alpha s}$.

Our main results correspond to the case of two experts.

I. Quality of set of all predictors is 2. The predictors to achieve this quality can be constructed explicitly (for any given s).

II. Quality of set of all integral predictors is 1. An integral predictor achieving this quality is described.

III. Quality that can be obtained by using Bayes-mixture is $\frac{1}{2}$, by MDL — not greater than $\frac{1}{2}$.

The natural open problem is to extend our results for the greater numbers of experts.