Influence of hyperparameters on aggregating predictions of infinite number of experts

S. M. Kunin-Bogoiavlenskii, A. V. Zukhba
kunin-bogoiavlenskii.sm@phystech.edu; a__l@mail.ru
Moscow Insitute of Physics and Technology

Using aggregation of expert forecasts is one of the important methods to improve prediction accuracy. This paper focuses on examining the influence of hyperparameters on the accuracy of the aggregation algorithm with an infinite number of experts. We implemented a time series generator with specified properties and an aggregating forecasting model. We conducted a series of experiments with various hyperparameters of the algorithm. Finding optimal hyperparameters leads to high performance and wider adoption of this prediction method.
Keywords: *online learning; aggregating algorithm; prediction with experts' advice; Fixed Share, Mixing Past Posteriors (MPP)*

1 Introduction

This work is inspired by the algorithm, developed in the article [4], considering online game of prediction with experts' advice. The data is presented as a time series, consisting of outcome pairs — «signal» and «response». In contrast to the classical statistical theory of sequential prediction, we made no assumptions about the nature of the data (which could be deterministic, stochastic, etc.). We used machine learning methods to build forecasters within a game-theoretic approach. The online learning master model considers a series of reference forecasters, referred to as experts, to build its opinion by combining their predictions.

The general prediction algorithm with expert advice follows this structure: Learning progresses in trials at discrete points in time. During each step , expert models, based on past observational data subsamples, provide their predictions. The master model then makes a decision using the chosen aggregating algorithm. At the end of the trial, the generator presents the true outcome, and both the master and expert models are scored using a loss function. The difference between the master's cumulative losses and the expert's cumulative losses is defined as regret. The traditional goal of the aggregating algorithm is to keep it as small as possible.

We use special assumptions about the data generation structure when building forecasting strategies. It is assumed that there are multiple generators, whose structure is unknown to the predictors. These generators switch, producing a time series that is subdivided into a sequence of segments - areas of stationarity, which can be studied using machine learning methods. Each corresponding local predictive model will be constructed based on data from the area of stationarity and can be then successfully applied in other areas of stationarity generated by the same generator.

In this formulation of the forecasting problem, the series of prediction steps is divided into segments that frame arbitrary sequences of expert strategies. The sequence of segments and its associated sequence of experts is called a partition. The modified goal of the aggregating algorithm is to perform well relative to the best partition. Accordingly, the new concept of algorithm regret is the difference between the algorithm's losses and the cumulative losses of the sequence of experts. This change allows for a more accurate modeling of real-life conditions, where the nature of outcomes may change over time and different experts may predict with varying degrees of success depending on the current trend.

The corresponding algorithm is called Fixed Share [5]. A further proposed generalization of it is the Mixing Past Posteriors (MPP) method [6]. The cumulative losses of the aggregating algorithm are related to convex combinations of expert losses. The concept of regret also changes. Now the algorithm's cumulative losses are compared to the cumulative losses of convex combinations of expert strategies.

A characteristic feature of the problem considered in [4] is the absence of a predefined set of competing expert strategies, as was the case in the works cited above. Instead, new expert strategies are constructed at each step of the online learning process. The master must aggregate the forecasts of all expert strategies constructed up to that time in real-time at each step. Algorithm GMPP, proposed in [4], is the foundation of our experiments.

2 Problem statement

2.1 Definitions

Let $x_1, x_2, \ldots$ be the sequence of signals belonging to the sample space X. The forecaster's goal is to predict the sequence of the corresponding responces $y_1, y_2, \ldots$ belonging to the outcome space Y. We assume that there is an infinite number of experts $i \in \mathcal{N}$, where $\mathcal{N}$ is the natural numbers set. The predictions of the master and experts belong to the decision space D. Let $\lambda : \mathsf{D} \times \mathsf{Y} \to \mathbb{R}_+$ be the nonnegative loss function.

At each step t every expert $i \in \mathcal{N}$ provides his prediction $f_t^i = f_t^i(x_t) \in \mathsf{D}$. After obtaining them the master gives his prediction $p_t = p_t(x_t) \in \mathsf{D}$. Next, the generator reveals the true outcome y_t, and the losses are computed. Let $l_t^i = \lambda(f_t^i, y_t)$ be the loss of expert i, and $h_t = \lambda(p_t, y_t)$ be the master's loss

Using designations $L_T^i = \sum_{t=1}^T l_t^i$ — the cumulative loss of expert i during the first T steps, and $H_T = \sum_{t=1}^T h_t$ — the master's cumulative loss during the first T steps, we can define the master's regret relative to the expert i as $R_t^i = H_T - L_T^i$

2.2 Data generator

2.3 Algorithm GMPP

Initialization: Initialize the weights w_1^i so that $\sum_{i \in \mathcal{N}} w_1^i = 1$
FOR $t = 1, 2, \ldots$:

1. Expert f^t initialization
2. Signal x_t received from the generator
3. Experts' predictions $f_t^i = f_t^i(x_t), 1 \leqslant i \leqslant t$
4. Compute normalized weights of experts $1 \leqslant i \leqslant t$:

$$\widehat{w}_t^i = \frac{w_t^i}{\sum_{j=1}^t w_t^j}$$

5. Master's prediction evaluation $\gamma_t = \mathsf{Subst}(\mathbf{f_t}, \widehat{\mathbf{w}}_\mathbf{t})$,
   where $\widehat{\mathbf{w}}_\mathbf{t} = (\widehat{w}_t^1, \widehat{w}_t^2, \ldots \widehat{w}_t^t)$, $\mathbf{f_t} = (f_t^1, f_t^2, \ldots f_t^t)$
6. Generator reveals true outcome y_t
7. Computation of master's loss $h_t = \lambda(p_t, y_t)$ and experts' losses: $l_t^i = \begin{cases} \lambda(f_t^i, y_t), & \text{if } i \leqslant t \\ h_t, & \text{if } i > t \end{cases}$

8. **Loss Update** weights modification

$$\widetilde{w}_t^i = \frac{w_t^i e^{\eta l_t^i}}{\sum\limits_{j=1}^{t} w_t^j e^{-\eta l_t^i} + e^{-\eta h_t}(1 - \sum\limits_{j=1}^{t} w_t^j)}$$

9. **Mixing Update** weights modification

$$\widetilde{w}_{t+1}^i = \alpha_t \widetilde{w}_1^i + (1 - \alpha_t \widetilde{w}_t^i)$$

ENDFOR

It was proved in [4] that for $\alpha = \frac{1}{t+1}$ and any partition of $k+1$ experts, the following is true (where $L_T(E)$ is the cumulative loss of partition E) :

$$\limsup_{T \to \infty} \frac{H_T - L_T(E)}{T} = 0$$

This allows formulating the main assumption underlying the application of the GMPP algorithm: In cases where it is possible to «attach» a valid predictive (expert) strategy to each local sub-sample from the generation area, carrying small losses on each local sub-sample generated by the generator, i.e., «learn» this generator, the GMPP algorithm will also predict with sufficiently small losses on the entire sample.

References

[1] N. Cesa-Bianchi, G. Lugosi. 2006. *Prediction, Learning, and Games*. Available at: `https://ii.uni.wroc.pl/~lukstafi/pmwiki/uploads/AGT/Prediction_Learning_and_Games.pdf`

[2] Hyndman, R. J. & Athanasopoulos, G., 2nd edition. 2018. *Forecasting: Principles and Practice*. *OTexts: Melbourne, Australia* . Available at: `https://otexts.com/fpp2/`

[3] V. V'yugin. 2022. Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya [Mathematical Foundations of Machine Learning and Forecasting]. Available at: `http://iitp.ru/upload/publications/6256/vyugin1.pdf`

[4] V. V'yugin, V. Trunov. 2023. Prognozirovanie lokal'no statsionarnykh dannykh s ispol'zovaniem predskazanii ekspertnykh strategiy [Prediction of Locally Stationary Data Using Prediction with Expert Advice]. Available at: `http://www.jip.ru/2023/470-487-2023.pdf`

[5] M. Herbster, M. Warmuth. 1998. *Tracking the best expert*. Available at: `https://link.springer.com/content/pdf/10.1023/A:1007424614876.pdf`

[6] O. Bousquet, M. Warmuth. 2002. *Tracking a small set of experts by mixing past posteriors*. Available at: `https://www.jmlr.org/papers/volume3/bousquet02b/bousquet02b.pdf`