POINTERWISE ADAPTATION VIA STAGEWISE AGGREGATION OF LOCAL ESTIMATES FOR MULTICLASS CLASSIFICATION

NIKITA PUCHKIN
National Research University Higher School of Economics, Moscow, Russia
Institute for Information Transmission Problems, Moscow, Russia
Moscow Institute of Science and Technology, Dolgoprudny, Russia
e-mail: nikita.puchkin@phystech.edu

In a problem of multiclass classification, one observes a training sample \( S_n = \{ (X_i, Y_i) \}_{i=1}^n \), where \( X_i \in \mathcal{X} \subseteq \mathbb{R}^d \) and \( Y_i \in \mathcal{Y} = \{ 1, \ldots, M \} \), \( M > 2 \). The goal of statistician is to construct a rule \( \hat{f}(X) \) predicting a label at the point \( X \). We consider a very general probabilistic model, where features \( X_i \) are generated from some distribution \( P_X \) over \( \mathcal{X} \) and \( Y_i \) are generated from the conditional distribution \( P( Y = m | X = x ) = \theta_m(x), 1 \leq m \leq M \), and \( \theta_1(x), \ldots, \theta_M(x) \) are unknown Lipschitz functions. We propose an algorithm of multiclass classification, which is robust against class imbalance and outliers and also computationally efficient. We extend the idea of pointwise adaptation proposed in [1] to the case of many classes. We study a plug-in classifier \( \hat{f}(X) = \arg\max_{1 \leq m \leq M} \hat{\theta}_m(X) \), where \( \hat{\theta}_m(X) \) stands for an estimate of \( \theta_m(X) \) at the point \( X \). Such problem of nonparametric estimation usually uses a localization technique, i.e. the weight \( w_i \in [0, 1] \) of observation \( (X_i, Y_i) \) depends on the distance between \( X_i \) and \( X \). One can consider a local likelihood estimate
\[
\tilde{\theta}_m(X) = \arg\max_{\theta} L_m(W, \theta) = \arg\max_{\theta} \sum_{i=1}^n w_i \left[ 1_{\{ Y_i = m \}} \log \frac{\theta}{1 - \theta} + \log(1 - \theta) \right].
\]
However, the estimate \( \tilde{\theta}_m(X) \) strongly depends on the localizing scheme \( W \) and its choice determines the performance of the classifier \( \hat{f} \). Moreover, in multiclass learning there is a common problem of class imbalance, i.e. some classes may be not presented in a small vicinity of a distinct point. Obviously, one localizing scheme is not enough for such situation. In our algorithm, we fix a set of \( K \) schemes \( W^{(1)}, \ldots, W^{(K)} \) and apply an aggregation procedure to corresponding likelihood estimates. Our theoretical study shows that under mild assumptions the algorithm has an optimal accuracy of classification with only a logarithmic payment for the number of weak estimates \( K \) and classes \( M \). We also study the behavior of procedure under small noise assumptions and show that it can achieve fast learning rates.

References