SafePredict: A Machine Learning Meta-Algorithm That Uses Refusals to Guarantee Correctness

Mustafa A. Kocak, David Ramírez, Elza Erkip, and Dennis E. Shasha

SafePredict [1] is a novel meta-algorithm that works with any base prediction algorithm for online data to guarantee an arbitrarily chosen correctness rate, $1-\epsilon$, by allowing refusals. Allowing refusals means that the meta-algorithm may, on occasion, refuse to predict so that the error rate on non-refused predictions does not exceed ϵ. *SafePredict* does not rely on any assumptions on the data distribution or the base predictor. David Ramírez, the presenting author of Mexican ancestry, collaborated with completing the mathematical proofs which show that when the base predictor happens not to exceed the target error rate ϵ, SafePredict refuses only a finite number of times; additionally, when the error rate of the base predictor changes through time SafePredict makes use of a weight-shifting heuristic that adapts to these changes without knowing when the changes occur yet still maintains the correctness guarantee. Empirical results show that SafePredict compares favorably with state-of-the-art confidence-based refusal mechanisms, and combining SafePredict with such refusal mechanisms can further reduce the number of refusals. Details of the datasets, setup, and implementations of our algorithm are available online and in [1].

We assume access to a base predictor that produces a label prediction on an observed object. We denote the base predictor as P and a sequence of (object, label) pairs by (x_t, y_t) with $t \in \{1, \ldots, T\}$ where T is an arbitrary horizon. At each time t the base predictor observes the object x_t, predicts the corresponding label $\hat{y}_{P,t}$, observes the true label y_t, and suffers the loss $l_{P,t}$ with $0 \leq l_{P,t} \leq 1 \;\forall\; t$. We stay agnostic to the data sequence and inner workings of the base predictor P (which may be an ensemble of predictors).

Once the base predictor P is chosen, our meta-algorithm M decides either to follow the prediction made by P *or* to refuse to make a prediction for each data point. We characterize this meta-algorithm by the following:

- **Parameter**: Target error/loss rate, $\epsilon \in (0,1)$, is the maximum average loss over time the user has specified.
- **Input**: In full generality, the input of M at time t consists of $x_i, \hat{y}_{P,i} \;\forall\; i \in \{1,...,t\}$, and $y_j, l_{P,j} \;\forall\; j \in \{1,...,t-1\}$. Note these are all the observed quantities *before* the label y_t is revealed.
- **Output**: A randomized decision to predict (or refuse) at time t. Define $\varnothing$ as a refusal. We represent the output of M as a probability value $w_{P,t} \in [0,1]$ that is used to decide the final prediction $\hat{y}_t$ as follows:

$$\hat{y}_t = \begin{cases} \hat{y}_{P,t} & \text{with prob. } w_{P,t} \\ \varnothing & \text{with prob. } 1 - w_{P,t}. \end{cases}$$

Note that the number of (non-refused) predictions by M is a random variable and we compute the expected value T^* of this quantity as $T^* = \sum_{t=1}^{T} w_{P,t}$. We ascribe no loss from refusing to predict and define the expected cumulative loss of M as $L^*_{P,T} = \sum_{t=1}^{T} w_{P,t} l_{P,t}$. Finally, we define the error rate for this randomized meta-algorithm by normalizing the cumulative expected loss via the expected number of non-refused predictions, i.e., $L^*_{P,T}/T^*$.

We seek to guarantee that the error rate of non-refused predictions made by M does not exceed the target error rate ϵ as the number of predictions increases. Following nomenclature in [2], [3], we call this property the *validity* of the algorithm. Given a target error ϵ, M is called *valid* if $T^* = O(1)$ or $\limsup_{T^* \to \infty} \frac{L^*_{P,T}}{T^*} \leq \epsilon$. The *efficiency* of a *valid* meta-algorithm M is denoted by $\rho_T = T^*/T$ and M is called *efficient* if $T^* = o(T)$. Furthermore, M is said to have the *finite refusal property* if $T - T^* = O(1)$.

We compute the variance V^* of the number of non-refused predictions with respect to the randomness of the meta-algorithm (i.e., $w_{P,t}$) as $V^* = \sum_{t=1}^{T} w_{P,t}\,(1 - w_{P,t})$. Denote the cumulative loss for any predictor P (typically for the base predictor) as $L_{P,T} = \sum_{t=1}^{T} l_{P,t}$. We introduce a third sub-index t_0 for any cumulative quantity to represent that the corresponding sum starts from $t_0 + 1$, e.g., for the cumulative loss of P, $L_{P,T,t_0} = \sum_{t=t_0+1}^{T} l_{P,t}$.

Note the journal version of this abstract is available in arXiv. M. A. Kocak is with the Broad Institute of Harvard and MIT. D. Ramirez is with the Department of Electrical Engineering at Princeton University, e-mail dard@princeton.edu. E. Erkip is with the Department of Electrical and Computer Engineering at NYU Tandon School of Engineering. D. E. Shasha is with Courant Institute of Mathematical Sciences at New York University. Work supported in part by NYU Seed Grant, NYU WIRELESS, and the National Science Foundation grants CNS-1302336, MCB-1158273, IOS-1339362, and MCB-1412232. Refernces at https://scholar.princeton.edu/sites/default/files/dramirez/files/lxai_references.pdf.

To achieve validity, we introduce a trivial predictor which can meet the target error rate by refusing to predict all the time. We refer to this particular predictor as the "dummy predictor" and denote it by D, i.e., $\hat{y}_{D,t} = \varnothing \ \forall \ t$. We also assume that the predictor D suffers a constant loss ϵ for each time t, i.e., $l_{D,t} = \epsilon$ and $L_{D,t} = \epsilon t \ \forall \ t$. At each instance, *SafePredict* follows the base predictor with probability $w_{P,t}$ and computes the prediction probability for the next round after observing the corresponding loss of P as $w_{P,t+1} = \frac{w_{P,t} e^{-\eta l_{P,t}}}{w_{P,t} e^{-\eta l_{P,t}} + w_{D,t} e^{-\eta \epsilon}}$, where $w_{D,t} = 1 - w_{P,t}$ is the dummy's weight. The following theorems (proofs in [1]) show the validity and efficiency of *SafePredict*.

Theorem .1. *For any P, $\eta > 0$, $\epsilon < 1/2$, and $0 < w_{P,1} < 1$, SafePredict satisfies $\frac{L^*_{P,T}}{T^*} \leq \epsilon - \frac{\log(w_{D,1})}{\eta T^*} + \frac{(1-\epsilon)^2 \eta V^*}{T^*}$. Consequently, by choosing the learning rate η to minimize the RHS of this bound, we get$\frac{L^*_{P,T}}{T^*} \leq \epsilon + (1-\epsilon) \frac{2\sqrt{\log(1/w_{D,1}) V^*}}{T^*}$ for $\eta^* = \frac{\sqrt{\log(1/w_{D,1})/V^*}}{1-\epsilon}$.*

Theorem .2. *If $L_{P,T}/T < \epsilon$ and $\eta T \to \infty$ in the limit $T \to \infty$, then the expected number of refusals made by SafePredict is finite, i.e., $\lim_{T \to \infty} T - T^* = \sum_{t=1}^{\infty} w_{D,t} \ < \ \infty$.*

We evaluate SafePredict on MNIST digit recognition [4] (vision application), IMDb sentiment analysis [5] (natural language problem), and Reuters topic recognition [6] (text classification). We compare SafePredict with a natural confidence-based refusal (CBR) mechanism, a method widely used in practice [7]–[11]. We investigate a heuristically promising method of combining SafePredict with the confidence-based mechanism to improve efficiency (SP+CBR), and include an "amnesic adaptive" version of the combined meta-algorithm and base predictor that considers excessive refusals of SafePredict as a sign that adaptation is needed.

For each dataset, we randomly permute the data and use the first 10000 data points. Artificial change points are introduced every 2000 data points by applying a random label permutation to all the data points after the change point, i.e., we effectively change the data distribution at each change point. The target error rate is fixed as $\epsilon = 0.05$.

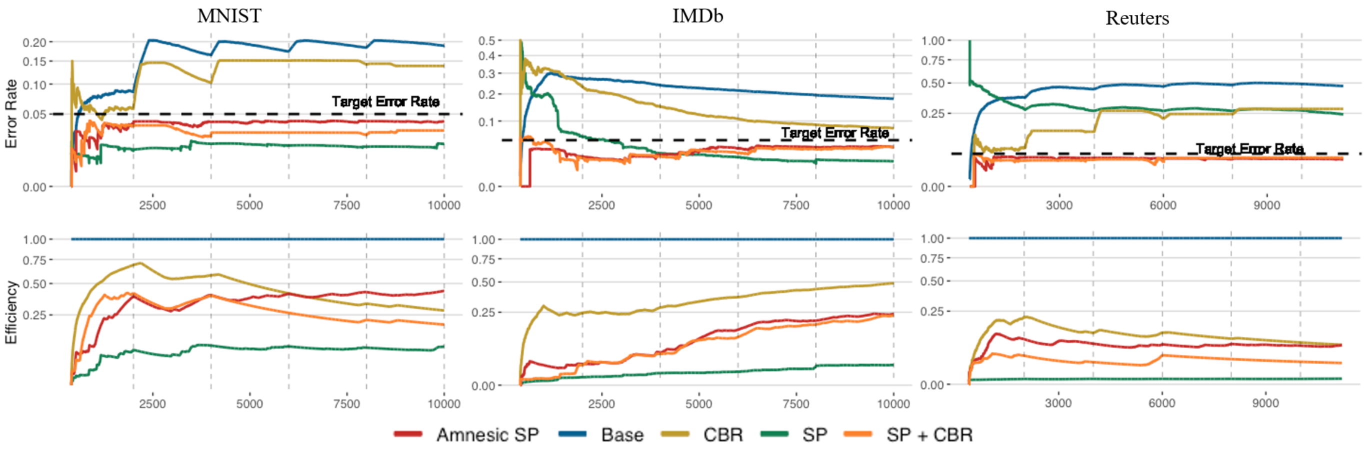

Fig. 1: Top three plots show error rates and bottom three show efficiency of evaluated algorithms. Each column of plots corresponds to the dataset labeled above. Efficiency for the base predictor is always 1.0, since it always predicts. All SafePredict variants rapidly approach an error rate value below the target error rate 0.05 as they make *good* predictions. The confidence-based meta-algorithm does not achieve validity due to changes in the underlying distribution (marked by vertical dashed lines). Two forms of adaptivity help reduce the number of refusals: weight-shifting especially with a high α value and amnesic adaptivity. Combining both leads to the highest efficiency while preserving validity.

The confidence-based refusal mechanism fails to be valid because it requires data points to be (at least approximately) exchangeable to achieve the required error guarantee. This assumption fails after the change point. SafePredict establishes validity by refusing when the base predictor cannot achieve the error rate without making any assumptions about the data points. For the Reuters dataset, the base predictor never reaches the target error rate. Thus, SafePredict refuses to make a prediction almost all the time.

While the confidence-based refusal method has a generally higher efficiency than SafePredict, the discrepancy in the efficiency is mitigated by the SP+CBR heuristic which performs almost as efficiently as confidence-based refusals before the change point and achieving validity throughout. Note, in SP+CBR, SafePredict will seldom refuse as long as the CBR algorithm has a validity rate as high as the desired correctness rate. Link to references.

References

[1] M. A. Kocak, D. Ramirez, E. Erkip, and D. E. Shasha, “Safepredict: A Meta-Algorithm for Machine Learning That Uses Refusals to Guarantee Correctness,” *arXiv preprint arXiv:1708.06425*, 2017.

[2] J. W. Tukey, “Sunset Salvo,” *The Amer. Statistician*, vol. 40, no. 1, pp. 72–76, 1986.

[3] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer Sci. & Bus. Media, 2005.

[4] Y. LeCun and C. Cortes, “MNIST Handwritten Digit Database,” 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[5] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning Word Vectors for Sentiment Analysis,” in *49th Annu. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015

[6] empty, “Reuters-21578 dataset,” empty. [Online]. Available: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[7] B. Zhang, Y. Zhou, and H. Pan, “Vehicle Classification with Confidence by Classified Vector Quantization,” *IEEE Intell. Transp. Syst. Mag.*, vol. 5, no. 3, pp. 8–20, 2013.

[8] B. Hanczar and E. R. Dougherty, “Classification with Reject Option in Gene Expression Data,” *Bioinformatics*, vol. 24, no. 17, pp. 1889–1895, 2008.

[9] C. De Stefano, C. Sansone, and M. Vento, “To Reject or Not to Reject: That is the Question-an Answer in Case of Neural Classifiers,” *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 30, no. 1, pp. 84–94, 2000.

[10] C. Cortes, G. DeSalvo, M. Mohri, and S. Yang, “On-line Learning with Abstention,” *arXiv preprint arXiv:1703.03478*, 2017.

[11] P. L. Bartlett and M. H. Wegkamp, “Classification with a Reject Option Using a Hinge Loss,” *J. of Mach. Learning Research*, vol. 9, pp. 1823–1840, Aug, 2008.