Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220413

A Confidence Interval-Based Method for Classifier Re-Calibration

Andrea CAMPAGNER [a,1] and Lorenzo FAMIGLINI [a] and Federico CABITZA [a,b]
[a] *University of Milano-Bicocca, Milano, Italy*
[b] *IRCCS Istituto Ortopedico Galeazzi, Milan, Italy*

Abstract. We propose a re-calibration method for Machine Learning models, based on computing confidence intervals for the predicted confidence scores. We show the effectiveness of the proposed method on a COVID-19 diagnosis benchmark.

Keywords. Calibration, confidence interval, medical ML, trustable AI

1. Introduction

Machine Learning (ML) has shown promising accuracy in the clinical domain, but there are important limitations in terms of other quality dimensions [1], including *calibration* [4]: this is the extent the confidence scores associated with each prediction are close to the observed frequency of events. A possible solution to this problem is to apply a re-calibration method, so to adjust confidence scores [3]. Several such techniques have been proposed in the literature; however, existing methods do not provide any guarantee and require additional data for re-calibration.

2. Method

In this article we propose a re-calibration method based on the computation of confidence intervals for the confidence scores provided by any ML model. Let $S = \{(x_i, y_i)\}_{i=1}^{n}$ be the training set, and h a classifier, where $h(x)$ is the confidence score associated the positive class. We first partition S into k bins $S_k^1, \ldots, S_k^k$, by sorting the instances' confidence scores. Then, this partition is used to compute confidence intervals around $h(x)$: given a confidence score $h(x)$ falling in a bin S_k^i, we compute a confidence interval (given $\alpha \in (0,1)$) as: $\left[\max\{0, h(x) - \frac{\sqrt{2\hat{\sigma}_i} \cdot erf^{-1}(\alpha)}{|S_k^i|+1}\}, \min\{1, h(x) + \frac{\sqrt{2\hat{\sigma}_i} \cdot erf^{-1}(\alpha)}{|S_k^i|+1}\}\right]$, where $h(x) \in S_k^i$, $\hat{\sigma}_i$ is the average confidence score in bin S_k^i, and erf is the error function.

We evaluated the proposed method, in comparison with other re-calibration methods, on a public dataset for the task of COVID-19 diagnosis from routine blood exams [1]. The training set consists of 1736 samples collected from February to May 2020 at the IRCCS Ospedale San Raffaele (OSR), in Milan, Italy. The test set consists of 224 sam-

[1] Corresponding Author: Andrea Campagner, a.campagner@campus.unimib.it

ples collected in November 2020 at IRCCS OSR. See [1] for further details. In regard to ML models, we used a state-of-the-art SVM-based model [1] as baseline and compared 4 re-calibration methods: Sigmoid regression (SR); Isotonic regression [3] (IR); Venn prediction [2] (VP); our proposed method ($\alpha = 0.90$). Models were compared in terms of the Brier score $\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$ and graphical analysis of the reliability diagrams.

3. Results and Discussion

The results of the experiment are shown in Figure 1. The proposed method reported a consistent improvement in calibration, as shown by the lower Brier score and the fact that the bisector line lies within the interval bounds. By contrast, the other methods did not provide any improvement compared to the baseline model, in terms of Brier score.

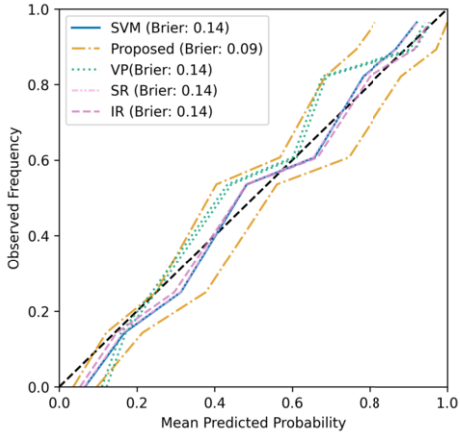

Figure 1. Reliability diagram for the considered models. The dashed line denotes perfect calibration.

4. Conclusion

We proposed a novel re-calibration method, based on computing confidence intervals for the confidence scores provided by a ML model. Through an illustrative experiment, we showed that the proposed technique provides better calibration than existing methods. However, further and more extensive experimental validation should be conducted.

References

[1] F. Cabitza, A. Campagner, et al. The importance of being external. Methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed*, 208:106288, 2021.

[2] A. Lambrou, I. Nouretdinov, et al. Inductive venn prediction. *Ann Math Artif Intell*, 74(1):181–201, 2015.

[3] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd ICML*, pages 625–632, 2005.

[4] B. Van Calster, D. J. McLernon, et al. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7, 2019.